MINITAB™ Hints and Tricks for
BASIC STATISTICS AND DATA ANALYSIS

Alan T. Arnholt
Appalachian State University

Suzanne R. Arnholt
Appalachian State University

May 4, 2002†

†© 2003 All rights reserved. No part of this document may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without written consent of the authors.
Preface

The authors thank Minitab’s Author Assistance Program for supplying the authors with the most recent version of MINITAB™. Portions of MINITAB™ Statistical software input and output contained in this book are printed with permission of Minitab Inc. MINITAB™ is a trademark of Minitab Inc. in the United States and other countries and is used herein with the owner’s permission.

As of the publication of MINITAB™ Hints and Tricks for Basic Statistics and Data Analysis, Minitab was offering a free 30 day MINITAB™ 13 Demo from www.minitab.com. Students may also rent MINITAB™ from e-academy (www.e-academy.com\minitab).
# Contents

## 0 — Using Hints and Tricks for BSDA

- 0.1 Using this Document ............................................... 1
- 0.2 Using Adobe .................................................. 1
- 0.3 Computing Hints .................................................. 2
  - 0.3.1 Installing the Data Sets from the CD ......................... 3
  - 0.3.2 Reading External Files ..................................... 3
  - 0.3.3 Navigating MINITAB™ Windows ............................... 3
- Lab 0.1 Introductory Lab ........................................... 6

## 1 — Organizing and Summarizing Univariate Data

- 1.0 Introduction .................................................... 12
- 1.1 Essential Elements of Statistics ................................ 12
  - 1.1.1 Categorizing Variables by Learning Their Properties ....... 12
  - 1.1.2 Understanding the Distribution of a Variable ............... 14
- 1.2 Displaying Data with Tables and Graphs ....................... 14
  - 1.2.1 Creating Frequency Tables .................................. 15
  - 1.2.2 Graphing Time Series Data ................................ 21
- 1.3 Displaying Numerical Data ..................................... 26
  - 1.3.1 Displaying Data with a Dotplot .............................. 26
  - 1.3.2 Displaying Data with Stem-and-Leaf Plots .................. 28
  - 1.3.3 Displaying Data with Histograms & LOWESS Lines .......... 29
- 1.4 Summarizing Data with Statistics ............................... 32
  - 1.4.1 Calculating and Interpreting Numerical Characteristics of Data .... 32
  - 1.4.2 The Empirical Rule ........................................ 34
1.4.3 Z-scores ........................................... 34

1.5 Describing a Distribution ..................................... 36
1.5.1 Examining Tails and Outliers ............................... 36
1.5.2 Describing the Center of a Uni-Modal Distribution ................... 37
1.5.3 Describing the Spread of a Uni-Modal Distribution ................... 38
1.5.4 Interpreting Boxplots .................................... 38
1.5.5 Knowing When Boxplots Are Appropriate ...................... 38

1.6 Summary and Review Labs .................................... 40
Lab 1.1 Graphing Practice ................................... 40
Lab 1.2 More Graphing Practice ................................ 43
Lab 1.3 Stem-and-Leaf and Histograms Lab ....................... 47
Lab 1.4 Stem-and-Leaf and Histograms Lab # 2 ....................... 49
Lab 1.5 Distribution Shape Detection Using the Boxplot ........... 51
Lab 1.6 Summarizing Data With Statistics #1 ....................... 54
Lab 1.7 Summarizing Data With Statistics #2 ....................... 55

2 — Summarizing Relationships Between Variables 56
2.0 Introduction ............................................ 56
2.1 Scatterplots ............................................. 56
2.1.1 Identifying Outliers ..................................... 59
2.2 Correlation ............................................. 61
2.2.1 Properties of the Correlation Coefficient ......................... 61
2.3 Least Squares Regression ..................................... 64
2.4 Assessing the Fit of a Line .................................... 68
2.4.1 Understanding the Effects of Outliers and Other Influential Observations ....... 71
2.4.2 Interpreting the Coefficient of Determination, $r^2$. .................... 71
2.5 Relationships among Categorical Variables ....................... 72
2.6 Summary and Review Labs .................................... 73
Lab 2.1 Scatterplots ....................................... 73
Lab 2.2 Correlation Coefficients .................................. 74
Lab 2.3 Least Squares Regression #1 ................................ 77
Lab 2.4 Least Squares Regression #2 ................................ 78
Lab 2.5 A Simple Analysis ....................................... 79
3 — Probability and Probability Distributions

3.0 Introduction ............................................ 81
3.1 Basic Probability Concepts .................................. 81
  3.1.1 Simulating Data ....................................... 82
3.2 The Binomial Distribution ..................................... 85
  3.2.1 Bernoulli and Binomial Random Variables ................. 85
  3.2.2 How to calculate binomial probabilities with MINITAB™ .................................. 86
  3.2.3 Hints & Tricks of solving Binomial problems ............... 89
3.3 The Normal Distribution ..................................... 91
  3.3.1 Finding Probabilities and Percentiles in a Normal Distribution . 91
  3.3.2 Normal Probability Plots .................................. 93
3.4 Summary and Review Labs .................................... 95
  Lab 3.1 Basic Probability Concepts .............................. 95
  Lab 3.2 The Binomial Distribution ............................... 97
  Lab 3.3 The Normal Distribution #1 ............................. 98
  Lab 3.4 The Normal Distribution #2 .............................. 103

4 — Sampling Distributions

4.0 Introduction ............................................ 105
4.1 Principals of Sampling ...................................... 105
  4.1.1 Generating a Simple Random Sample from a finite population using MINITAB™ 105
  4.1.2 Random Samples ...................................... 106
4.2 The Sampling Distribution of \( \bar{x} \) ................................. 106
4.3 The Sampling Distribution of the Sample Proportion \( p \) .................. 112
4.4 Summary and Review Labs .................................... 116
  Lab 4.1 The Sampling Distribution of \( \bar{x} \) ......................... 116
  Lab 4.2 The Sampling Distribution of the Sample Proportion .......... 119
5 — Introduction to Inference

5.0 Introduction ............................................ 121
5.1 Describing the Parent Distribution ............................... 121
  5.1.1 Examining Shape and Choosing Estimators ................. 121
5.2 Estimating Certain Population Parameters ....................... 122
  5.2.1 Properties of Estimators .................................. 123
  5.2.2 Estimating \( \pi \) ........................................ 123
  5.2.3 Estimating the Center of a Distribution ...................... 123
5.3 Confidence Interval for a Bernoulli Proportion, \( \pi \) ............. 124
  5.3.1 Calculating and Interpreting Confidence Intervals for a Proportion .......... 124
  5.3.2 Constructing a confidence interval for a Bernoulli proportion ............. 127
  5.3.3 Calculating Sample Size for a Given Margin of Error ............... 128
5.4 Confidence Interval for a Population Mean, \( \mu \) .................. 129
  5.4.1 The \( Z \)-Interval ....................................... 129
  5.4.2 Sample Size Determination for Estimating \( \mu \) .................. 130
  5.4.3 The \( t \) Distribution ..................................... 132
  5.4.4 Critical Values Based on the \( t \) Distribution ................. 132
  5.4.5 The \( t \) Interval ........................................ 134
5.5 Confidence Interval for a Population Median, \( \theta \) ............... 135
  5.5.1 Small Sample Confidence Interval For The Population Median ............ 136
  5.5.2 Large Sample Confidence Interval For The Population Median ............. 138
5.6 Summary and Review Labs ....................................... 139
  Lab 5.1 Estimating Certain Population Parameters #1 ............... 139
  Lab 5.2 Estimating Certain Population Parameters #2 ............... 142
  Lab 5.3 Confidence Interval for a Bernoulli Proportion, \( \pi \) ............. 144
  Lab 5.4 Confidence Interval for a Population Mean, \( \mu \) ............... 146
  Lab 5.5 Confidence Interval for a Population Median, \( \theta \) ............... 150
6 — Hypothesis Testing

6.1 Introduction to Hypothesis Testing ................................................................. 152

6.1.1 Choosing Test Statistics .............................................................................. 152

6.1.2 Working with $p$-Values ............................................................................ 153

6.1.3 5 Step Procedure For Testing Hypotheses ............................................... 153

6.1.4 Type I and II Errors ................................................................................... 153

6.1.5 The Power of a Test .................................................................................... 154

6.1.6 Testing $\mu$ When $\sigma$ Is Known: The $Z$-Test .................................... 154

6.1.7 Using MINITAB$^{TM}$ to do a $Z$-test ......................................................... 156

6.1.8 Using MINITAB$^{TM}$ to calculate Power .................................................. 158

6.2 Testing a Population Proportion ..................................................................... 159

6.2.1 Equivalence of Confidence Intervals and Two-tailed Tests .................... 161

6.2.2 Determining Power with a Bernoulli Population ..................................... 161

6.3 Testing a Population Mean When $\sigma$ Is Unknown: The $t$-Test .............. 164

6.3.1 Determining the Power of a One Sample $t$-Test ..................................... 167

6.4 Testing a Population Median ......................................................................... 169

6.4.1 Small-Sample Test Of The Population Median ....................................... 169

6.4.2 Large-Sample Test Of The Population Median ....................................... 171

6.5 Summary and Review Labs ......................................................................... 174

Lab 6.1 Introduction to Hypothesis Testing ......................................................... 174

Lab 6.2 Testing a Population Proportion #1 ......................................................... 177

Lab 6.3 Testing a Population Proportion #2 ......................................................... 178

Lab 6.4 Testing a Population Mean When $\sigma$ is Unknown ............................ 179
7 — Inference About the Difference Between Two Parameters

7.1 Overview of Selecting Procedures ........................................... 180

7.2 Inference about the Difference Between Two Population Proportions ........ ... 181

7.3 Comparing Two Population Means ........................................... 183

7.3.1 Two Sample $Z$-procedures ............................................... 183

7.3.2 Two Sample $t$ Procedures ............................................... 183

7.4 Comparing Two Population Centers: Variances Equal ..................... 186

7.4.1 Pooled $t$ Procedures .................................................. 186

7.4.2 The Mann-Whitney-Wilcoxon Test ..................................... 190

7.4.3 Student’s $t$ Approximation To The Null Distribution Of $R(X_i)$ ............... 190

7.4.4 MINITAB™’s Mann-Whitney Test ....................................... 193

7.5 Comparing Two Population Centers Using Matched Samples ............... 194

7.5.1 The Matched Pairs $t$-Test ............................................... 194

7.5.2 Consequences of Using the Matched Pairs $t$-Test When It Is Inappropriate .... 196

7.5.3 Wilcoxon Signed Rank Test for the Matched Pairs Experiment ................ 197

7.5.4 Student’s $t$ Approximation to the Null Distribution of $T_+$ ................ 198

7.5.5 MINITAB™’s Wilcoxon Signed Rank Test ................................ 200

7.6 Choosing the Right Procedure ............................................... 202

7.7 Summary and Review Labs .................................................. 203

Lab 7.1 Inference About the Difference Between Two Population Proportions .......... 203

Lab 7.2 Comparing Two Population Means ..................................... 204

Lab 7.3 Comparing Two Population Centers: Variances Equal ................... 206
8 — Analysis of Categorical Data  

8.0 Introduction ............................................ 209  
8.1 The Chi-Square Goodness-of-Fit Test  
8.1.1 Goodness-of-Fit Global Macro ............................... 211  
8.2 The Chi-Square Test of Independence  
8.2.1 Chi-Square Tests With MINITAB™ ........................... 213  
8.3 The Chi-Square Test of Homogeneity  
8.3.1 Chi-Square Test of Homogeneity with MINITAB™ ................. 216  
8.4 Summary and Review Labs .................................... 218  
Lab 8.1 Introduction to the Chi-Square Distribution ..................... 218  
Lab 8.2 Chi-Square Goodness-of-Fit Test  
Lab 8.3 Chi-Square Tests of Independence and Homogeneity ................ 220

9 — Regression Analysis  

9.0 Introduction ............................................ 221  
9.1 The Linear Regression Model  
9.1.1 Model Assumptions ..................................... 222  
9.1.2 Estimating $\sigma^2$ ........................................ 222  
9.2 Inference about the Regression Model  
9.2.1 Hypothesis Test For $\beta_1$ ................................... 223  
9.2.2 Hypothesis Test For $\beta_0$ ................................... 224  
9.2.3 Confidence Intervals For $\beta_1$ And $\beta_0$ .................... 225  
9.2.4 Confidence Interval For $\mu_y$ ................................. 226  
9.2.5 Prediction Interval For A New Observation $y'(\text{new})$ Given $x'$ .......... 227  
9.2.6 Confidence And Prediction Bands ............................. 229  
9.3 Checking Model Adequacy .................................... 230  
9.4 Summary and Review Labs .................................... 236  
Lab 9.1 The Linear Regression Model ................................ 236  
Lab 9.2 Inference About the Regression Model ......................... 237  
Lab 9.3 Checking Model Adequacy .................................. 238
10 — Analysis of Variance

10.1 Introduction ............................................ 239
10.2 The One-Way ANOVA ...................................... 239
  10.2.1 Multiple Comparisons ................................... 245
10.3 The Kruskal-Wallis Test ...................................... 248
  10.3.1 Kruskal-Wallis Approximation ............................... 251
10.4 Summary and Review Labs .................................... 253
  Lab 10.1 The One-Way ANOVA ................................ 253
  Lab 10.2 The Kruskal-Wallis Test ............................... 254

A — Glossary ............................................. 255

B — Statistics, Distributions, and Conditions ......................... 256

C — Confidence Intervals for Common Parameters ..................... 257

D — Key Words and Problem Hints .................................... 258
  D.1 Basic Probability .......................................... 258
  D.2 Sampling Distributions and Probability ............................. 258
  D.3 Probabilities with Normal Distributions ............................. 259

E — Macros ............................................. 260
  E.1 NPIE: Macro for finding sample size given error bound for $\pi$ CI ............... 260
  E.2 NMU: Macro for finding sample size given error bound and confidence level .................. 261
  E.3 GOF: Goodness of Fit Macro .................................. 262
  E.4 SIMP: Simulate Power for Pooled versus Un-pooled $t$ Tests ................................. 263
  E.5 SIMH: Simulate Power for Pooled $t$ Test versus Wilcoxon Signed Rank Test ........... 265
List of Figures

0.1 Adobe Acrobat Toolbar .......................................................... 2
0.2 Screen View when MINITAB™ Opens ....................................... 2
0.3 Open Worksheet Dialog Window .............................................. 3
0.4 MINITAB™ Tutorials Window .................................................. 6
0.5 Tutorial Menu Window ............................................................ 7
0.6 Example of Split Screen ........................................................ 8
0.7 Project Manager with Entries in the Session Folder Highlighted .......... 8

1.1 Help Window for Furnace.MTW description ............................... 13
1.2 Coding Window to change 1’s and 2’s to EVD and TVD ................... 13
1.3 Worksheet after coding .......................................................... 14
1.4 Graph>Dotplot Menu ............................................................ 15
1.5 Graph>Dotplot Window for Weight of Bears by Sex ....................... 15
1.6 Bears’ Weight by Sex Dotplot ............................................... 16
1.7 Tally Dialog Box for Frequency Tables ....................................... 16
1.8 Chart Dialog Box for Furnace Data .......................................... 17
1.9 Bar Chart of Counts for Furnace Types ...................................... 17
1.10 Title Dialog Box for Furnace Chart ......................................... 18
1.11 Axis Dialog Box for Furnace Chart .......................................... 18
1.12 Worksheet Example of Tally Output ........................................ 18
1.13 Mean Energy Usage by Furnace Type and Damper Category ............ 19
1.14 Chart Dialog Box for Furnace Data with a Clustered Chart ............ 19
1.15 Chart Options Box for Furnace Data with a Clustered Chart .......... 20
1.16 Pie Chart Options Box for Furnace Data ................................... 20
LIST OF FIGURES

1.17 Pie Chart of Frequencies of Furnace Types .......................................................... 21
1.18 Time Series Plot for Problem 1.37 ........................................................................... 21
1.19 Time Series Dialog Box ......................................................................................... 22
1.20 Axis Dialog Box for Time Series ............................................................................. 22
1.21 Text Dialog Box for Time Series ............................................................................. 23
1.22 Time Series Plot for Problem 1.37 ........................................................................... 23
1.23 Time Series Plot done with Plot command .............................................................. 23
1.24 Time Series Plot Worksheet Rearrangement ............................................................ 24
1.25 Plot Dialog Box for Worksheet Rearrangement to Make Time Series Plot .............. 25
1.26 Time Series Plot for Problem 1.37 with a Rearranged Worksheet ............................ 25
1.27 Text Values Dialog Box to name Male and Female .................................................. 26
1.28 Dotplot Dialog Window for Framingham Heart Study ......................................... 27
1.29 Dotplot For Framingham Heart Study With No Grouping ....................................... 27
1.30 Dotplot For Framingham Heart Study With Grouping ............................................. 27
1.31 Unstack Columns from Framingham Heart Study .................................................... 28
1.32 Stem-and-Leaf Dialog Window ............................................................................... 29
1.33 Histogram Dialog Window ...................................................................................... 30
1.34 Selecting Graph under Data Display ..................................................................... 31
1.35 A Density Histogram with a LOWESS Curve ......................................................... 31
1.36 Variable gpa Sorted and Stored in gpas .................................................................. 33
1.37 Calculation of Standard Deviation with MINITAB™’s Calculator ............................ 33
1.38 Histogram with LOWESS Curve for gpa ................................................................. 34
1.39 Standardizing Variable gpa ..................................................................................... 35
1.40 Boxplot of Female Bears with Fences Labeled ....................................................... 36
1.41 Boxplot Dialog Window for ALFALFA.MTW Comparative Boxplot ....................... 38
LIST OF FIGURES

1.42 Boxplot of Yield by Variety for ALFALFA.MTW ........................................ 39
1.43 Time Series Plot Window for TIAA CREF stock ............................................. 41
1.44 Time Series Plot for TIAA CREF data .............................................................. 42
1.45 Rearrangement of OWNCAR.MTW Worksheet ................................................. 43
1.46 Chart Window configuration for Data OWNCAR1.MTW for Comparison Bar Chart .............................................................. 44
1.47 Chart Window Options for Data OWNCAR1.MTW for Comparison Bar Chart .............................................................. 44
1.48 Final Comparative Bar Chart for Data OWNCAR1.MTW ....................................... 45
1.49 Comparative Bar Chart for Data HEAT1.MTW ...................................................... 45
1.50 Time Series Graph of Men’s 1500 Olympic Times ............................................. 46
1.51 Code Numeric to Text Dialog Window for FRAMINGH.MTW Lab ..................... 47
1.52 Generation of Standard Normal Random Data .................................................... 51
1.53 Stack Columns for Normal Distribution Shape Boxplot Lab ................................. 52
1.54 Side-By-Side Boxplots of Random Normal Samples ........................................... 52
2.1 Searching for Help on the BEARS.MTW Dataset .............................................. 57
2.2 MINITAB™’s Help for the BEARS.MTW Dataset ............................................... 57
2.3 Matrix Plot of BEARS.MTW Numerical Variables .............................................. 58
2.4 Draftsman Plot Dialog Window ........................................................................... 58
2.5 Draftsman Plot of BEARS.MTW Numerical Variables ........................................... 59
2.6 Scatterplot Dialog Window for Weight versus Chest.G ...................................... 59
2.7 Scatterplot of Weight versus Chest.G ................................................................. 60
2.8 Scatterplot Dialog Box of Weight versus Chest.G with Groups ............................ 60
2.9 Scatterplot of Weight versus Chest.G with Groups ............................................. 61
2.10 Five BEARS Scatterplots with Brushed Value .................................................... 62
2.11 Stat>Basic Statistics>Correlation Menu ............................................................. 62
LIST OF FIGURES

2.12 Correlation Calculation with MINITAB™ .................................. 63

2.13 Fitted Line Plot Window for Weight versus Chest Girth .................. 64

2.14 Fitted Line Plot Graph for Weight versus Chest Girth .................... 65

2.15 Fitted Line Plot Graph for Weight versus Chest Girth with (0,0) included 65

2.16 Calculator Window for \( \hat{y}_i \) Calculation .................................. 66

2.17 Fitted Line Plot – Storage Window of Residuals and Fits .................. 66

2.18 Calculator Window for Residuals Calculation ................................ 67

2.19 Regression–Options Window for Prediction Calculation .................... 67

2.20 Regression–Storage Window ...................................................... 68

2.21 Residual Graphs Window ............................................................ 69

2.22 Prototypical Residual Plot ............................................................. 69

2.23 Curvilinear Relationship Between the Residuals and the Fitted Values .......... 70

2.24 Increasing Variability in Residuals .................................................. 70

2.25 Increasing Variability in Residuals and a Curvilinear Relationship ............ 70

2.26 Influential observations from Value regressed on Revenue ................... 72

2.27 Scatterplot by Quadrants ............................................................... 74

2.28 Standardize Window for Weight ..................................................... 75

2.29 Calculator Example for Sum-Product ........................................... 75

2.30 Resized, Split Screen Example ....................................................... 79

2.31 Project Manager ................................................................. 80

2.32 Report Title ................................................................. 80

3.1 Generation of Random Integer Data ........................................... 82

3.2 Code Numeric to Text for Head/Tail Simulation ................................ 83

3.3 Cross Tabulation Dialog Window ................................................ 83
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>Binomial Distribution Dialog Window</td>
</tr>
<tr>
<td>3.5</td>
<td>Calculator Dialog Window</td>
</tr>
<tr>
<td>3.6</td>
<td>Row Statistics Dialog Box for Basketball Shots Made</td>
</tr>
<tr>
<td>3.7</td>
<td>Tally Dialog Window for Basketball Shots Made</td>
</tr>
<tr>
<td>3.8</td>
<td>Normal Distribution Probability Calculation for Symmetric “Between”</td>
</tr>
<tr>
<td>3.9</td>
<td>Normal Distribution Probability Calculation for Asymmetric “Between”</td>
</tr>
<tr>
<td>3.10</td>
<td>Normal Distribution Probability Calculation for “Greater Than”</td>
</tr>
<tr>
<td>3.11</td>
<td>Normal Distribution Inverse CDF Calculation Given a Probability Less Than</td>
</tr>
<tr>
<td>3.12</td>
<td>Normal Distribution Inverse CDF Calculation Given a Probability Greater Than</td>
</tr>
<tr>
<td>3.13</td>
<td>Normal Distribution Inverse CDF Calculation Given a “Between” Area and One End Point</td>
</tr>
<tr>
<td>3.14</td>
<td>Shots Made Chart</td>
</tr>
<tr>
<td>3.15</td>
<td>Binomial Distribution Dialog Window</td>
</tr>
<tr>
<td>3.16</td>
<td>Example of Normal Overlayed with Binomial</td>
</tr>
<tr>
<td>3.17</td>
<td>Normal Overlayed with Binomial #2</td>
</tr>
<tr>
<td>3.18</td>
<td>Simple Set of Numbers Dialog Window</td>
</tr>
<tr>
<td>3.19</td>
<td>Normal Distribution Dialog Window</td>
</tr>
<tr>
<td>3.20</td>
<td>Plot Dialog Window</td>
</tr>
<tr>
<td>3.21</td>
<td>Three Overlayed Normal Graphs</td>
</tr>
<tr>
<td>3.22</td>
<td>Probability for Values 40 through 60 from a $Bin(100, 0.5)$</td>
</tr>
<tr>
<td>3.23</td>
<td>Commands for Example Binomial Calculations</td>
</tr>
<tr>
<td>3.24</td>
<td>Plot Dialog Window for Normal and Binomial Overlay</td>
</tr>
<tr>
<td>4.1</td>
<td>Simple Set of Numbers Dialog Window</td>
</tr>
<tr>
<td>4.2</td>
<td>Sample / / Rows From Column(s) Box</td>
</tr>
<tr>
<td>4.3</td>
<td>Generate A Random Sample Of Size 50 From A Finite Population Of Size 1,000</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

4.4 Uniform(0,10) ................................................. 107
4.5 Sampling Distribution of Size 2 Samples from a Uniform(0,10) ...................... 108
4.6 Normal Distribution with $\mu = 5$ and $\sigma = 2.0412$ .................................. 108
4.7 Superimposed Normal and Uniform Graphs ...................................................... 109
4.8 Histogram of Calculated Means ................................................................. 109
4.9 Code – Numeric to Numeric Dialog Window .................................................. 110
4.10 Tally Dialog Window Example ...................................................................... 111
4.11 Binomial Distribution Dialog Window ............................................................ 113
4.12 Calculator Dialog Window for Binomial Distribution ...................................... 113
4.13 Histogram of the Simulated Distribution of $p$ .............................................. 113
4.14 Calculation of Lower and Upper Confidence Values ..................................... 114
4.15 Coding UpperLimit for Confidence Interval Calculation ................................ 115
4.16 Cross Tabulation Dialog Window for Confidence Interval .......................... 115

5.1 Normal Distribution Dialog Window ............................................................. 125
5.2 Normal Distribution Dialog Window for Cumulative Probability .................... 126
5.3 CRANKSH.MTW Description ....................................................................... 130
5.4 Boxplot of data in AtoBDist .......................................................................... 130
5.5 1-Sample Z Dialog Window for Example 5.4.1 .............................................. 131
5.6 Cross Tabulation Dialog Window for Confidence Interval .......................... 131
5.7 Comparison of $Z$ and $t$ Distributions ........................................................... 132
5.8 $t$ Distribution Dialog Window .......................................................... 133
5.9 $t$ Distribution Dialog Window for area to the Left of $-1.96$ ......................... 134
5.10 1-Sample $t$ (Test and Confidence Interval) Window ................................... 135
5.11 1-Sample Sign Dialog Window for Median CI .............................................. 137
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.12</td>
<td>Example of How to Calculate ( \sum x_i )</td>
<td>140</td>
</tr>
<tr>
<td>5.13</td>
<td>Example Histogram for Simulated Chi-Square Distribution</td>
<td>140</td>
</tr>
<tr>
<td>5.14</td>
<td>Integer Distribution Dialog Window for OSP Simulation</td>
<td>142</td>
</tr>
<tr>
<td>5.15</td>
<td>Row Statistics Dialog Window for OSP Simulation</td>
<td>143</td>
</tr>
<tr>
<td>5.16</td>
<td>Simulation of 10,000 Binomial Random Variables with ( n = 500 ) and ( \pi = 0.4 )</td>
<td>144</td>
</tr>
<tr>
<td>5.17</td>
<td>Calculation of ( p )</td>
<td>145</td>
</tr>
<tr>
<td>5.18</td>
<td>Generation of Random ( \mathcal{N}(100, 30) )</td>
<td>147</td>
</tr>
<tr>
<td>5.19</td>
<td>Calculation of Lower Confidence Values Using Equation (5.4)</td>
<td>147</td>
</tr>
<tr>
<td>5.20</td>
<td>Coding Lower Limit for Confidence Interval Calculation</td>
<td>147</td>
</tr>
<tr>
<td>5.21</td>
<td>Cross Tabulation Dialog Window for Confidence Interval</td>
<td>148</td>
</tr>
<tr>
<td>5.22</td>
<td>Font Dialog Window</td>
<td>149</td>
</tr>
<tr>
<td>6.1</td>
<td>( \alpha = .05 ), ( \beta ), and Power for ( H_0 : \mu = 100 ) vs. ( H_A : \mu &gt; 100 ) when ( \mu = 105 )</td>
<td>154</td>
</tr>
<tr>
<td>6.2</td>
<td>Functional Relationship of Values of ( \mu ) and Power</td>
<td>155</td>
</tr>
<tr>
<td>6.3</td>
<td>1-Sample Z (Test and Confidence Interval) Dialog Window</td>
<td>156</td>
</tr>
<tr>
<td>6.4</td>
<td>Two Distributions for Example 6.1.2</td>
<td>157</td>
</tr>
<tr>
<td>6.5</td>
<td>Graphical Representation of Power for Example 6.1.2</td>
<td>158</td>
</tr>
<tr>
<td>6.6</td>
<td>Probability Distributions of ( X \sim \text{Bin}(500, 0.2) ) and ( Y \sim \text{Bin}(500, 0.22) ) With Decision Criterion</td>
<td>162</td>
</tr>
<tr>
<td>6.7</td>
<td>Cumulative Probability Distributions of ( X \sim \text{Bin}(500, 0.2) ) and ( Y \sim \text{Bin}(500, 0.22) ) With Decision Criterion</td>
<td>162</td>
</tr>
<tr>
<td>6.8</td>
<td>( \text{Stat} \rangle \text{Basic Statistics}\rangle \text{Normality Test} ) Graph</td>
<td>165</td>
</tr>
<tr>
<td>6.9</td>
<td>( \text{Graph} \rangle \text{Probability Plot} ) Graph</td>
<td>165</td>
</tr>
<tr>
<td>6.10</td>
<td>1-Sample t Dialog Window</td>
<td>166</td>
</tr>
<tr>
<td>6.11</td>
<td>Power and Sample Size for 1-Sample t</td>
<td>167</td>
</tr>
<tr>
<td>6.12</td>
<td>Three Power Curves for a Two Tailed Alternative Hypothesis</td>
<td>168</td>
</tr>
<tr>
<td>6.13</td>
<td>Normal Probability Plot of the Length of Long Distance Phone Calls</td>
<td>170</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>6.14</td>
<td>1-Sample Sign Dialog Window for Example 6.4.1 .................................................. 171</td>
<td></td>
</tr>
<tr>
<td>6.15</td>
<td>Stem-and-Leaf Plot of Income Distribution for North Carolina Social Workers ............ 172</td>
<td></td>
</tr>
<tr>
<td>6.16</td>
<td>Connect Dialog Window ......................................................................................... 175</td>
<td></td>
</tr>
<tr>
<td>6.17</td>
<td>Line Dialog Window .............................................................................................. 175</td>
<td></td>
</tr>
<tr>
<td>6.18</td>
<td>Final Graph for Step 10 of Lab 6.1 ...................................................................... 175</td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>Flow Chart for Selecting Procedures Covered in Chapter 7 ................................... 180</td>
<td></td>
</tr>
<tr>
<td>7.2</td>
<td>2 Proportions (Test and Confidence Interval) Dialog Window .................................. 182</td>
<td></td>
</tr>
<tr>
<td>7.3</td>
<td>2-Sample (Test and Confidence Interval) Dialog Window Example ............................ 184</td>
<td></td>
</tr>
<tr>
<td>7.4</td>
<td>2-Sample t - Graphs Dialog Window Example .......................................................... 184</td>
<td></td>
</tr>
<tr>
<td>7.5</td>
<td>Side-by-Side Boxplots from Example 7.3.1 ............................................................. 185</td>
<td></td>
</tr>
<tr>
<td>7.6</td>
<td>Side-by-Side Boxplots from Example 7.4.1 ............................................................. 188</td>
<td></td>
</tr>
<tr>
<td>7.7</td>
<td>Normal Probability Plots from Example 7.4.1 ........................................................ 188</td>
<td></td>
</tr>
<tr>
<td>7.8</td>
<td>Side-by-Side Boxplots from Example 7.4.2 ............................................................. 191</td>
<td></td>
</tr>
<tr>
<td>7.9</td>
<td>Mann-Whitney Test with MINITAB™ ......................................................................... 194</td>
<td></td>
</tr>
<tr>
<td>7.10</td>
<td>Boxplot of Differences from Example 7.5.1 ........................................................... 195</td>
<td></td>
</tr>
<tr>
<td>7.11</td>
<td>Normal Probability Plot of Differences from Example 7.5.1 .................................. 196</td>
<td></td>
</tr>
<tr>
<td>7.12</td>
<td>Calculator Dialog Window Commands to Create Signed Ranked Differences ................ 198</td>
<td></td>
</tr>
<tr>
<td>7.13</td>
<td>Boxplot for Example 7.5.2 ...................................................................................... 199</td>
<td></td>
</tr>
<tr>
<td>7.14</td>
<td>Normal Probability Plot for Example 7.5.2 ............................................................. 199</td>
<td></td>
</tr>
<tr>
<td>7.15</td>
<td>1-Sample Wilcoxon Dialog Window ........................................................................... 201</td>
<td></td>
</tr>
<tr>
<td>8.1</td>
<td>Three Different Chi-Square Distributions ............................................................... 210</td>
<td></td>
</tr>
<tr>
<td>8.2</td>
<td>Chi-Square Test of Independence Dialog Window for Example 8.2.1 ....................... 214</td>
<td></td>
</tr>
<tr>
<td>8.3</td>
<td>Chi-Square Test of Homogeneity Dialog Window for Example 8.3.1 ....................... 217</td>
<td></td>
</tr>
<tr>
<td>9.1</td>
<td>Regression–Storage Window ..................................................................................... 223</td>
<td></td>
</tr>
</tbody>
</table>
9.2 MSE Value From Regressing eggs on feed ........................................ 223
9.3 Regression - Options Dialog Window .................................................. 227
9.4 Fitted Line Plot Dialog Window .............................................................. 229
9.5 95% Confidence and Prediction Bands for Regressing Weight on Chest.G ........................................ 230
9.6 Regression - Graphs Dialog Window ....................................................... 231
9.7 Histogram of Residuals from Example 9.2.2 ........................................ 232
9.8 Normal Probability Plot of Residuals from Example 9.2.2 ...................... 232
9.9 Residuals Plotted Against Fitted Values from Example 9.2.2 ............... 233
9.10 Residuals Plotted Against Observation from Example 9.2.2 ............... 233
9.11 Checked Circles in Fitted Line Plot Dialog Window ......................... 233
9.12 Checked Boxes in Fitted Line Plot - Storage Dialog Window ............... 234
9.13 Fitted Quadratic Model for Example 9.2.2 ....................................... 234
9.14 Fitted Quadratic Model Diagnostics for Example 9.2.2 ...................... 235
10.1 Two Different $F$ Distributions ............................................................ 240
10.2 Side-By-Side Boxplots for Example 10.2.1 ....................................... 242
10.3 Normal Probability Plots of the Four Groups for Example 10.2.1 .......... 243
10.4 One-way Multiple Comparisons Dialog Window ................................ 246
10.5 Side-by-Side Boxplots from Example 10.3.1 on page 249 .................... 249
# List of Tables

0.1 Color Coding for this Document .................................. 1
0.2 Font Coding for this Document ................................... 1
0.3 Basic Toolbar Buttons ........................................... 4
0.4 *Data Window* Toolbar Buttons .................................. 4
0.5 Additional *Session Window* Toolbar Buttons .................. 5
0.6 Additional *Graph Window* Toolbar Buttons .................... 5
0.7 *Project Manager* Toolbar Buttons .............................. 5

1.1 Rejecting Normality with Outliers ................................ 37
1.2 TIAA Cref Stock Worksheet ....................................... 41

3.1 Theoretical Binomial Values ...................................... 96

4.1 Theoretical Normal Values versus Simulated Values for \( n = 2 \) from a Uniform(0, 10) .......... 111
4.2 Output for Sample Proportion Lab ................................ 119

5.1 Order Statistic 92\% Confidence Levels ........................... 151
5.2 Order Statistic 96\% Confidence Levels ........................... 151

6.1 Relationship Between Type I and Type II Errors ................ 153
6.2 Applicability of the *t* Distribution ............................. 164

7.1 Study Habits from Example 7.5.2 ................................ 199
7.2 Comparison of *Satterthwaite* and *Pooled t-Test* ............. 205
7.3 Comparison of *Wilcoxon Signed Rank* and *Pooled t-Test* .... 207

8.1 Results of a Study About Snoring Reported in the *British Medical Journal* ................. 213
8.2 Sports Preferences from Example 8.3.1 .......................... 216

10.1 Format for an ANOVA Table ...................................... 241
10.2 Test Results from Example 10.2.1 ............................... 242
10.3 ANOVA Table for Example 10.2.1 ................................ 243
10.4 Number of Successful Free Throws ................................ 249
# List of Output

1.1  **Tally Output of a Frequency Table**  ................................................. 16  
1.2  **Stem-and-Leaf Graph of cholest from Framingh.MTW**  .................................. 29  
1.3  **Basic Statistics Output for gpa from Simpson.MTW**  ..................................... 32  

2.1  **Output from Correlation Calculation with MINITAB™**  .................................. 63  
2.2  **Output from Multiple Correlation Calculations with MINITAB™**  ................. 63  
2.3  **Fitted Line Plot’s Numerical Output for Weight versus Chest Girth**  ................... 65  
2.4  **Output from Regression Including FIT**  .................................................. 68  
2.5  **Value regressed on Revenue**  ................................................................. 71  

3.1  **Output Head/Tail Tally**  ........................................................................... 83  
3.2  **Cross Tabulation of Two Coin Flipping**  ................................................... 84  
3.3  **CDF and PDF for X ∼ Bin(10,.33)**  ......................................................... 86  
3.4  **Additional Functionality of MINITAB™ with a Binomial**  ............................ 87  
3.5  **Total Shots Made**  ..................................................................................... 89  
3.6  **Various ways to solve a Binomial Problem**  ............................................. 90  
3.7  **MINITAB™ commands and storage locations so that all answers agree in Binomial calculation**  ........................................... 90  
3.8  **Example when X ∼ Bin(100,0.5)**  ............................................................... 104  

4.1  **Calculations of Various Statistics for Simulation**  ...................................... 110  
4.2  **Cross Tabulation Results for Confidence Interval**  .................................... 115  

5.1  **Session Window Commands to Calculate Z_{0.975}**  ................................... 125  
5.2  **Session Window Commands to Calculate Z_{0.975}**  ................................... 126  
5.3  **Calculating P(IQ < 132) and P(X ≤ x) = 0.90**  ............................................. 127  
5.4  **Session Window Commands to Calculate Z_{0.965} and Confidence Interval Limits for \pi**  ................................................................. 127
LIST OF OUTPUT

5.5 Session Window for NPIE Macro ............................................................. 129
5.6 Session Window Output from NMU Macro ............................................... 131
5.7 Session Window Commands to find \(t_{0.975;3}\) and \(t_{0.025;3}\) .................. 133
5.8 \(t\) Interval Limits .................................................................................. 136
5.9 Output for Median CI Calculation .......................................................... 137
5.10 Output of Binomial for Median CI Calculation ........................................ 137
5.11 Median CI Exact Calculation with MINITAB™ ......................................... 138
5.12 Calculation of Median CI ...................................................................... 150

6.1 Session Window Commands to Calculate \(Z_{0.975}\) ................................. 156
6.2 Area Calculations for Example 6.1.2 ....................................................... 157
6.3 MINITAB™ Results for Example 6.1.2 ................................................... 158
6.4 \(\pi\) Test Using Both Methods ................................................................. 160
6.5 Session Window Commands for Example 6.2.1 ...................................... 162
6.6 Session Window Commands to Calculate Power with Continuity Correction ......................................................................................... 163
6.7 Session Window Commands to Calculate Power without a Continuity Correction ......................................................................................... 163
6.8 Using MINITAB™'s Built In Power Function ............................................ 164
6.9 \(t\) Test of \(H_0 : \mu = 98.6\) versus \(H_A : \mu < 98.6\) ....................................... 166
6.10 Power for One Sample \(t\) Test .................................................................. 168
6.11 Descriptive Statistics of the Variable Time ................................................ 170
6.12 Session Window Output from 1-Sample Sign .......................................... 171
6.13 MINITAB™'s 1-Sample Sign Test of \(H_0 : \theta = 27,500\) versus \(H_A : \theta < 27,500\) ................................................................................................. 173

7.1 MINITAB™ Output for Example 7.3.1 ....................................................... 186
7.2 MINITAB™ Output for Example 7.4.1 ....................................................... 189
7.3 Session Window Commands and Output for Example 7.4.2 ..................... 192
List of Videos

1.1 For Problem 1.15 — (Duration: 4 minutes 21 seconds) ........................................ 14
1.2 For Problem 1.33 — (Duration: 5 minutes 44 seconds) ........................................ 19
1.3 For Problem 1.126 — (Duration: 3 minutes 47 seconds) ...................................... 29
1.4 For Problem 1.56 — (Duration: 4 minutes 27 seconds) ........................................ 30
1.5 For Problem 1.75 — (Duration: 4 minutes 27 seconds) ........................................ 35
1.6 For Problem 1.91 — (Duration: 5 minutes 7 seconds) .......................................... 37

2.1 For Problem 2.15 — (Duration: 5 minutes 37 seconds) ........................................ 61
2.2 For Problem 2.30 — (Duration: 4 minutes 53 seconds) ........................................ 64
2.3 For Problem 2.46 — (Duration: 4 minutes 13 seconds) ........................................ 67
2.4 For Problem 2.59 — (Duration: 4 minutes 27 seconds) ........................................ 71
2.5 For Problem 2.97 — (Duration: 5 minutes 2 seconds) .......................................... 71
2.6 For Problem 2.79 — (Duration: 6 minutes 59 seconds) ........................................ 72

3.1 For Problem 3.26 — (Duration: 3 minutes 24 seconds) ........................................ 84
3.2 For Problem 3.51 — (Duration: 3 minutes 4 seconds) .......................................... 84
3.3 For Problem 3.108 — (Duration: 5 minutes 55 seconds) ....................................... 84
3.4 For Problem 3.66 — (Duration: 4 minutes 31 seconds) ........................................ 93
3.5 For Problem 3.112 — (Duration: 2 minutes 58 seconds) ....................................... 94

4.1 For Problem 4.26 — (Duration: 7 minutes 1 second) ........................................... 111
4.2 For Problem 4.83 — (Duration: 5 minutes 8 seconds) .......................................... 112
4.3 For Problem 4.84 — (Duration: 5 minutes 53 seconds) ........................................ 112
4.4 For Problem 4.46 — (Duration: 7 minutes 55 seconds) ........................................ 112

5.1 For Problem 5.14 — (Duration: 4 minutes 21 seconds) ....................................... 122

xxii
5.2 For Problem 5.25 — (Duration: 4 minutes 8 seconds) ... 122
5.3 For Problem 5.44 — (Duration: 10 minutes 26 seconds) ... 128
5.4 For Problem 5.66 — (Duration: 4 minutes 53 seconds) ... 135
5.5 For Problem 5.83 — (Duration: 5 minutes 11 seconds) ... 138
5.6 For Problem 5.119 — (Duration: 4 minutes 43 seconds) ... 138

6.1 For Problem 6.45 — (Duration: 4 minutes 43 seconds) ... 167
6.2 For Problem 6.99 — (Duration: 5 minutes 24 seconds) ... 171
6.3 For Problem 6.65 — (Duration: 6 minutes 43 seconds) ... 173

7.1 For Example 7.7 from Basic Statistics and Data Analysis — (Duration: 2 minutes 23 seconds) ... 182
7.2 For Problem 7.35 — (Duration: 4 minutes 42 seconds) ... 186
7.3 For Problem 7.49 — (Duration: 7 minutes 23 seconds) ... 189
7.4 For Problem 7.87 — (Duration: 9 minutes 38 seconds) ... 189
7.5 For Problem 7.62 — (Duration: 10 minutes 41 seconds) ... 202

8.1 For Problem 8.35 — (Duration: 6 minutes 6 seconds) ... 217

9.1 For Problem 9.20 — (Duration: 7 minutes 1 second) ... 230
9.2 For Problem 9.34 — (Duration: 6 minutes 6 seconds) ... 231
9.3 For Problem 9.62 — (Duration: 4 minutes 14 seconds) ... 235

10.1 For Problem 10.19 — (Duration: 7 minutes 30 seconds) ... 248
10.2 For Problem 10.35 — (Duration: 11 minutes 43 seconds) ... 252
Chapter 0

Using Hints and Tricks for Basic Statistics and Data Analysis

0.1 Using this Document

This manual has color coded boxes to draw your attention to important things in the text. It also has “font coded” text to indicate different types of commands, menus, windows, and the like.

One of the great benefits of this document is an interweaving of videos that demonstrate how to do problems. Any time you see the icon, you can click the icon to see a movie that will show you how to solve problems or use techniques that are applicable for the section you are reading.

0.2 Using Adobe

Your Hints and Tricks for Basic Statistics and Data Analysis document was designed to be read on screen. When you first open the Hints and Tricks for Basic Statistics and Data Analysis, the text should take up the whole window. If the text does not take up the whole window choose View > Fit Width from the Adobe Acrobat toolbar. Provided you have a rectangular screen, you should only have to use the scroll bar on a few pages of the entire document.

<table>
<thead>
<tr>
<th>Steps you will take to solve problems by hand</th>
<th>Steps you will follow to solve problems in MINITAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Important equations</td>
<td>Cautions you should heed</td>
</tr>
<tr>
<td>Notes you do not want to ignore</td>
<td></td>
</tr>
</tbody>
</table>

Table 0.1: Color Coding for this Document

Table 0.2: Font Coding for this Document

- Clicked commands
- Non-clicked commands
- Definitions
- Things we can see in output
- Things you should type
- Window, box, or menu names
- Macro names
- Things you should absolutely remember
- Data set names
- Variable and column names
- Tests you can run on your data
- Notes in the flow of the text
- Problem numbers in the flow of the text
- Text for the introductions of labs


0.3 Computing Hints

When you start MINITAB™ you are faced with a daunting array of choices for how you want to begin your session. The opening window will resemble Figure 0.2. The toolbars at the top of the screen provide you with many options depending on whether you want to view the Session Window, the Data Window, or the Project Manager.
Before you get started working with the windows, your book, *Basic Statistics and Data Analysis*, has data sets you will want to install.

### 0.3.1 Installing the Data Sets from the CD

For your book, *Basic Statistics and Data Analysis*, all data sets can be installed from a CD.

To install the *Basic Statistics and Data Analysis* data sets from the CD follow the directions that appear when you insert the CD. If your computer does not have autorun enabled, click **start** > **run** and type `NAME:\install.exe` where `NAME` is the directory letter of your CD drive. Your other option is to copy the files from the BSDA directory to a directory under your MINITAB directory.

If your book does not come with a CD, you can download the files from: http://www.duxbury.com/datasets.htm. You will need to scroll down to BSDA and choose the files beside your book. Then, click **MiniTab** and choose **save file**. Choose a directory to put the mini.exe file. If you have MINITAB™ loaded on your own machine, this should be under the MTBWIN directory and called “BSDA.” If you do not have MINITAB™ loaded on your own machine, you should be aware that all the data sets require approximately 5 megabytes of disk space and cannot be loaded onto a standard 3.5 inch disk.

### 0.3.2 Reading External Files

MINITAB™ can read data as well as store data in numerous formats in addition to its MINITAB™ formats of *.MTW, *.MTP, and *.MPJ. These formats include Excel, Quatro Pro, Lotus 1-2-3, dBase, and plain Text. To read in data from an external file choose **File** > **Open worksheet** and select the appropriate file type from the **Drop Down Menu Files of type** in the **Open Worksheet Dialog Window**. Navigate to the appropriate directory and double click on the external file of interest. See Figure 0.3 for a picture of the **Open Worksheet Dialog Window**. To store a MINITAB™ worksheet in another format, select **File** > **Save Worksheet As** and select your desired format by using the **Drop Down Menu Save as type** then click on the **Save** button in the **Save Worksheet As Dialog Window**.

### 0.3.3 Navigating MINITAB™ Windows

One of the easiest ways to navigate the windows in MINITAB™ is with the toolbars. The following tables outline which buttons on the toolbars are used for various purposes.
### Table 0.3: Basic Toolbar Buttons

<table>
<thead>
<tr>
<th>Button</th>
<th>Image</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Project</td>
<td><img src="Open_Project.png" alt="Image" /></td>
<td>Opens a MINITAB™ Project (MPJ) file</td>
</tr>
<tr>
<td>Save Project</td>
<td><img src="Save_Project.png" alt="Image" /></td>
<td>Saves the current project in a MINITAB™ Project (MPJ) file</td>
</tr>
<tr>
<td>Print</td>
<td><img src="Print.png" alt="Image" /></td>
<td>Prints the current window’s contents</td>
</tr>
<tr>
<td>Cut</td>
<td><img src="Cut.png" alt="Image" /></td>
<td>Cuts the selection to the clipboard</td>
</tr>
<tr>
<td>Copy</td>
<td><img src="Copy.png" alt="Image" /></td>
<td>Copies the selection to the clipboard</td>
</tr>
<tr>
<td>Paste</td>
<td><img src="Paste.png" alt="Image" /></td>
<td>Inserts the contents of the clipboard</td>
</tr>
<tr>
<td>Undo</td>
<td><img src="Undo.png" alt="Image" /></td>
<td>Undoes the last action</td>
</tr>
<tr>
<td>Edit Last Command</td>
<td><img src="Edit_Last_Command.png" alt="Image" /></td>
<td>Opens the most recently used dialog box, with the same</td>
</tr>
<tr>
<td>Dialog Box</td>
<td><img src="Dialog_Box.png" alt="Image" /></td>
<td>settings from the last time you used it</td>
</tr>
<tr>
<td>Go to the Session Window</td>
<td><img src="Go_to_Session_Windows.png" alt="Image" /></td>
<td>Makes the Session Window active</td>
</tr>
<tr>
<td>Go to the Data Window</td>
<td><img src="Go_to_Data_Windows.png" alt="Image" /></td>
<td>Makes the Data Window active</td>
</tr>
<tr>
<td>Go to the Project Manager</td>
<td><img src="Go_to_Project_Manager.png" alt="Image" /></td>
<td>Makes the Project Manager active</td>
</tr>
<tr>
<td>Close all graphs</td>
<td><img src="Close_all_graphs.png" alt="Image" /></td>
<td>Closes all active graphs</td>
</tr>
<tr>
<td>Find</td>
<td><img src="Find.png" alt="Image" /></td>
<td>Finds the requested text or number</td>
</tr>
<tr>
<td>Find Next</td>
<td><img src="Find_Next.png" alt="Image" /></td>
<td>Finds the requested text or number again</td>
</tr>
</tbody>
</table>

### Table 0.4: Data Window Toolbar Buttons

<table>
<thead>
<tr>
<th>Button</th>
<th>Image</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insert Cells</td>
<td><img src="Insert_Cells.png" alt="Image" /></td>
<td>Inserts one or more empty cells above the active cell of the Data window and moves the remaining cells in the column down. The number of cells inserted is equal to the number of cells selected before you choose the command.</td>
</tr>
<tr>
<td>Insert Rows</td>
<td><img src="Insert_Rows.png" alt="Image" /></td>
<td>Inserts one or more empty rows above the active row of the Data window and moves the remaining rows down. The number of rows inserted is equal to the number of rows selected before you choose the command.</td>
</tr>
<tr>
<td>Insert Columns</td>
<td><img src="Insert_Columns.png" alt="Image" /></td>
<td>Inserts one or more empty columns to the left of the active column of the Data window and moves the remaining columns to the right. The number of columns inserted is equal to the number of columns selected before you choose the command.</td>
</tr>
<tr>
<td>Move Columns</td>
<td><img src="Move_Columns.png" alt="Image" /></td>
<td>Moves columns to location you specify in the dialog box</td>
</tr>
<tr>
<td>Clear Cells</td>
<td><img src="Clear_Cells.png" alt="Image" /></td>
<td>Erases the contents of the selected cells, without moving other cells</td>
</tr>
<tr>
<td>Go to the Previous Brushed Row</td>
<td><img src="Go_to_Prev_Brush.png" alt="Image" /></td>
<td>Moves the active cell to the previous row (up) that contains a brushed data point</td>
</tr>
<tr>
<td>Go to the Next Brushed Row</td>
<td><img src="Go_to_Next_Brush.png" alt="Image" /></td>
<td>Moves the active cell to the next row (down) that contains a brushed data point</td>
</tr>
<tr>
<td>Cancel</td>
<td><img src="Cancel.png" alt="Image" /></td>
<td>Cancels the current operation</td>
</tr>
<tr>
<td>Help</td>
<td><img src="Help.png" alt="Image" /></td>
<td>Opens the MINITAB™ Help Window</td>
</tr>
</tbody>
</table>
### 0.3. Computing Hints

#### Table 0.5: Additional Session Window Toolbar Buttons

<table>
<thead>
<tr>
<th>Button</th>
<th>Image</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Go to the Previous</td>
<td><img src="image" alt="up-arrow" /></td>
<td>Allows you to scroll backward from command to command in the Session Window</td>
</tr>
<tr>
<td>Command</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Go to the Next</td>
<td><img src="image" alt="down-arrow" /></td>
<td>Allows you to scroll forward from command to command in the Session Window</td>
</tr>
<tr>
<td>Command</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stat Guide</td>
<td><img src="image" alt="Stat Guide" /></td>
<td>Displays help from the Stat Guide for your most recently completed procedure</td>
</tr>
</tbody>
</table>

#### Table 0.6: Additional Graph Window Toolbar Buttons

<table>
<thead>
<tr>
<th>Button</th>
<th>Image</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Put the Active Graph in</td>
<td><img src="image" alt="view-only" /></td>
<td>Makes the current graph uneditable</td>
</tr>
<tr>
<td>View Only Mode</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Put the Active Graph in</td>
<td><img src="image" alt="edit-mode" /></td>
<td>Makes the current graph editable</td>
</tr>
<tr>
<td>Edit Mode</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Put the Active Graph in</td>
<td><img src="image" alt="brush-mode" /></td>
<td>In some graphs, brush mode will give you the line on which a clicked data point is</td>
</tr>
<tr>
<td>Brush Mode</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Table 0.7: Project Manager Toolbar Buttons

<table>
<thead>
<tr>
<th>Button</th>
<th>Image</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show Session Folder</td>
<td><img src="image" alt="folder" /></td>
<td>Shows the contents of the Session Folder</td>
</tr>
<tr>
<td>Show Worksheets Folder</td>
<td><img src="image" alt="folder" /></td>
<td>The Worksheet Folders and their subfolders (Columns, Constants, and Matrices) contain an automatically updated summary of the current worksheet: columns, stored constants, and matrices</td>
</tr>
<tr>
<td>Show Graphs Folder</td>
<td><img src="image" alt="folder" /></td>
<td>You can use the Graphs Folder in the Project Manager to manage multiple Graph windows</td>
</tr>
<tr>
<td>Show Info</td>
<td><img src="image" alt="info" /></td>
<td>Shows the Worksheet Folders and their subfolders (Columns, Constants, and Matrices) contain an automatically updated summary of the current worksheet: columns, stored constants, and matrices</td>
</tr>
<tr>
<td>Show History</td>
<td><img src="image" alt="history" /></td>
<td>The History Folder displays all session commands, without the output</td>
</tr>
<tr>
<td>Show the Report Pad</td>
<td><img src="image" alt="report-pad" /></td>
<td>Displays the contents of the Report Pad</td>
</tr>
<tr>
<td>Show Related Documents</td>
<td><img src="image" alt="related-documents" /></td>
<td>Shows all related information for your current project</td>
</tr>
<tr>
<td>Show Design</td>
<td><img src="image" alt="design-subfolder" /></td>
<td>The Design Subfolder displays Factors, Runs, Blocks, Base Design, Replicates, and Center points, as well as the display order and display units, and the factors and their uncoded levels</td>
</tr>
</tbody>
</table>
Lab 0.1 — Introductory Lab

Objectives:

I. To become acquainted with the MINITAB™ environment
II. To produce MINITAB™ output
III. To work with the Report Pad to create a professional report

Introduction:

MINITAB™ tutorials are a quick way to become acquainted with the MINITAB™ environment and some of MINITAB™’s most important features. In this first lab, we will use a slightly modified version of the MINITAB™ tutorial Session One: MINITAB Basics. The overview shown in Figure 0.4 indicates the specific ideas you will learn during the tutorial.

Figure 0.4: MINITAB™ Tutorials Window

In addition to the ideas the tutorial lists, you will also learn how to append selected Session Window output and any graphs you make to the report pad. The report pad is a basic word processing tool that comes with MINITAB™. The report you will create will start with a brief description of why you are studying this particular problem and explain from where and how the data were collected.
Directions:

1. Start MINITAB™ and begin the tutorial. To launch the MINITAB™ tutorial, click Help>Tutorials from the MINITAB™ menu bar and a Help Window similar to Figure 0.5 will appear.

   Resize the tutorial window to occupy \( \frac{1}{3} \) of the screen on the right side.

   If you again click on MINITAB™ (on the left side of your screen), the tutorial window disappears. So that you can see both MINITAB™ and the tutorial, click on the Restore button (\( \square \)). Resize the MINITAB™ window so that it no longer covers the tutorial screen and covers the left \( \frac{2}{3} \) of the screen. See Figure 0.6 on the next page for an example. Start the tutorial Session One: MINITAB Basics and follow the directions. Work through the tutorial through step 9. Move through the tutorial directions by clicking on the green arrow pointing to the right \( \Rightarrow \).

   Ignore Step 1 since MINITAB™ is already launched.

   In Step 5 of the tutorial, you are directed to save your project. If you are working on a network computer, you may not be able to save the project in MINITAB™’s default saving folder (Data). It will be necessary to switch to your own floppy drive (often A:\). If you are working on your own computer, this does not present a problem.

   When you finish step 9 of the tutorial, click on the project manager icon (\( \square \)). Click once on the Session Folder. Highlight the appropriate output as seen in Figure 0.7 on the following page.

   Right click the highlighted output and choose Append to Report Pad. This may take your computer several minutes to complete, so be patient.

   As you become more proficient in working with MINITAB™, you may choose to Append to the Report Pad as you are working. Most things in the Session Window or Graphics Windows can be appended to the Report Pad by right clicking them.

   Use the back green arrow (\( \Rightarrow \)) to go back to the Overview of Session One and do the following:

   a. Highlight the first two paragraphs of The story provided in the Overview of Session One with your mouse.
   b. Right click and Copy the highlighted portion.
   c. Click on the project manager icon (\( \square \))
   d. Then, click on the ReportPad folder.
e. At this point, position your cursor after the default title MINITAB PROJECT REPORT and before the first output and click right to paste these paragraphs. An example of what your lab should resemble starts on page 9.

2. Advance to Step 6 of the tutorial.

   **Step 6: Compute Descriptive Statistics**

   Copy the information at the bottom of Step 6 that analyzes the boxplots (Judging from the boxplots ...). Paste this paragraph after the descriptive statistics and before the first boxplot in your report pad.

3. Advance to Step 8.
Step 8: Create a Scatter Plot

Copy the information at the bottom of Step 8 that analyzes the scatterplot (Looking at the scatterplot . . .), and paste it after the third boxplot and before the scatterplot in your report pad.

4. Advance to Step 9.

Step 9: Compute a Correlation Coefficient

Click on Verify Session Window Output. A new window will open. Right click in the new window and choose Copy. Paste what you have copied into your report pad after the correlation output and delete the duplicate material.

5. Title the report:

Change the title from MINITAB™ Project Report to Session One: Minitab Basics on the first line, Your Class (STAT 2810-101) on the second line, and Your Name (Alan T. Arnholt) on the third line. See the sample lab on page 9 as an example.

6. Save what you have been doing by clicking File>Save Project As. Choose and appropriate location and click Save.

When you save the project, MINITAB™ saves everything you have done during your MINITAB™ session. This file can be very large. Many times, all you will need to save is the work you created in the Report Pad. To save the Report Pad document, right click directly on the Report Pad folder in the Project Manage and select Save Report As. Give your work a descriptive name (for example SessOne). Then, click the Save button.

Report Pad saves in *.rtf format which is fairly large, yet readable by most word processors. It will save you space if you open the *.rtf file in your favorite word processor and then save the file in whatever format is native to your program. (For example Microsoft Word uses a *.doc format and Word Perfect uses *.wp.) This typically reduces the file 75% in size.

Group Work Suggestions: If your instructor asks you to work in a group, it is beneficial to use footers to add group members’ names to every page automatically. In Microsoft Word: This can be done by choosing View>Header and Footer, clicking on the Insert Auto Text button, picking Author, Page #, Date, and changing the Author to the names of those in the group.

Example Lab:

Session One: Minitab Basics
STAT 2810-101
Alan T. Arnholt

Clones are genetically identical cells descended from the same individual. Researchers have identified a single poplar clone that yields fast-growing, hardy trees. These trees may one day be an alternative energy resource to conventional fuel.

Researchers at The Pennsylvania State University planted Poplar Clone 252 on two different sites: one site was by a creek with rich, well-drained soil, and the other site was on a ridge with dry, sandy soil. They measured the diameter in centimeters, height in meters, and dry weight of the wood in kilograms of a sample of three-year-old trees. These researchers want to see if they can predict how much a tree weighs from its diameter and height measurements.
0.3. Computing Hints

Descriptive Statistics: Diameter, Height, Weight by Site

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter</td>
<td>1</td>
<td>10</td>
<td>2.598</td>
<td>2.320</td>
<td>2.604</td>
<td>0.916</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10</td>
<td>3.028</td>
<td>3.250</td>
<td>3.041</td>
<td>1.284</td>
</tr>
<tr>
<td>Height</td>
<td>1</td>
<td>10</td>
<td>4.098</td>
<td>4.120</td>
<td>4.175</td>
<td>1.103</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10</td>
<td>4.257</td>
<td>4.865</td>
<td>4.351</td>
<td>1.250</td>
</tr>
<tr>
<td>Weight</td>
<td>1</td>
<td>10</td>
<td>0.3090</td>
<td>0.2050</td>
<td>0.2863</td>
<td>0.2528</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10</td>
<td>0.399</td>
<td>0.380</td>
<td>0.356</td>
<td>0.366</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site</th>
<th>SE Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter</td>
<td>1</td>
<td>0.290</td>
<td>1.060</td>
<td>4.090</td>
<td>2.120</td>
<td>3.245</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.406</td>
<td>1.180</td>
<td>4.770</td>
<td>1.488</td>
<td>4.053</td>
</tr>
<tr>
<td>Height</td>
<td>1</td>
<td>0.349</td>
<td>1.850</td>
<td>5.730</td>
<td>3.518</td>
<td>4.853</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.395</td>
<td>2.220</td>
<td>5.540</td>
<td>2.775</td>
<td>5.143</td>
</tr>
<tr>
<td>Weight</td>
<td>1</td>
<td>0.0800</td>
<td>0.0200</td>
<td>0.7800</td>
<td>0.1575</td>
<td>0.4600</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.116</td>
<td>0.030</td>
<td>1.110</td>
<td>0.063</td>
<td>0.648</td>
</tr>
</tbody>
</table>

Judging from the boxplots, poplars grown at Site 2 are larger than those grown at Site 1. The Session Window data confirm that the median values for Diameter, Height, and Weight of the poplars are larger for Site 2 than for Site 1. Also, the variable Weight has a very large standard deviation relative to its mean. At Site 2, the minimum weight is only 0.03 kg while the maximum is 1.11 kg. It appears that some of our poplars are doing very well, while others are barely alive.
Looking at the scatter plot, you see a positive linear relationship between Weight and D2H. That is, as D2H increases, so does Weight. You also notice an unusual data point—a tree that has a very low weight for a relatively high D2H value. For now, you decide to ignore it, but it is something you may want to check on later. Next, you will compute the correlation between these two variables to quantify the relationship.

Correlations: Weight, D2H

Pearson correlation of Weight and D2H = 0.913
P-Value = 0.000

The correlation coefficient measures the linear relationship between two variables and assumes a value between −1 and +1. The high positive correlation coefficient of 0.913 is close to 1, thus quantifying the relationship that you already saw in the scatter plot—there is a strong linear association between Weight and D2H (diameter squared times height) for the trees in our sample.
Chapter 1
Organizing and Summarizing Univariate Data

1.0 Introduction

Welcome to the wonderful world of statistics! This introductory chapter will introduce you to statistical vocabulary, variables and their properties, some ideas about distributions, ways to display your data, and numerical measures that can describe your data. Be diligent in your studies, and the material in this manual will serve you well as you become a more capable analyzer of data.

1.1 Essential Elements of Statistics

One of the most effective ways to learn your statistical vocabulary is to write note cards with the word you want to learn on the front and its definition on the back. Color coding the words by chapter can also prove helpful when you go to review for your final. There is nothing like constant review and using the correct term as you do your assignments to help you learn statistical vocabulary.

1.1.1 Categorizing Variables by Learning Their Properties

Your book divides variables into two main types: categorical and numerical. A categorical variable is a variable whose values are classifications, categories, or groups. Many times numbers are used for the various classifications or categories a categorical variable may take on. This often leads to confusion since a numerical variable is a variable whose values are numbers obtained from a count or a measurement. Assigning a numerical code to a categorical variable does not make it numerical. One should get in the habit of using what MINITAB calls text columns to represent categorical data. When categorical data is stored in a text column, we can no longer calculate statistics on the categorical data. You cannot perform meaningful arithmetic operations on coded categorical variables. This is a good thing since statistics performed on categorical data DO NOT have meaning. Statistics ONLY have meaning when calculated on numerical variables. A numerical variable is a number that actually measures something. Numerical variables may be either discrete, countable, or continuous, usually a measurement with no smallest division. Most of the time, MINITAB will not require you to distinguish what type of data you have, but you should be able to do so to make appropriate charts and graphs.

Think: Categorical—group, Numerical—measurement.

In the example that follows, we examine a data set that contains both categorical and numerical variables that uses only numbers in all of the variables. Open the data set Furnace.MTW stored in the MTBWIN\DATA directory by choosing File>Open worksheet and selecting the Furnace.MTW worksheet from the Open Worksheet Dialog Box. To gain an understanding of the data and what the various variables and categories represent, we should read the description of the data given in the help menu. To read the description choose Help>Search Help then type Furnace.MTW and click on Furnace.MTW once it is highlighted. The Furnace.MTW Help Window will open resembling the window in Figure 1.1 on the next page. Notice from the Help Window that Type, CH.Shape, CH.Liner, House, and Damper are all categorical variables that use numbers to represent their various categories. We want to change the numerical entries in each of the categorical variables to reflect the category the
1.1. Essential Elements of Statistics

1.1. Essential Elements of Statistics

numbers are representing. In other words, for the variable Damper, we want to replace all 1’s with EVD and all 2’s with TVD. Although it is possible to edit the worksheet one cell at a time, there are more efficient and less time consuming methods to make the desired changes.

An easy way to change all of the numerical entries for a categorical variable to text is to choose Manip > Code > Numeric to Text. When the Code - Numeric to Text Dialog Box appears, fill in the information exactly as shown in Figure 1.2. When you are finished, click OK and view the changes in your worksheet. Use the same procedure to change the numerical quantities in the categorical variables Type, CH.Shape, CH.Liner, and House to reflect their respective categorical quantities. Your finished worksheet should resemble Figure 1.3 on the next page. Note in the example from Figure 1.2, that we created a new column named DamperC while leaving the old column intact. We could have stored the new information in the same column; however, if a mistake where made in coding, we would no longer have access to the old information.
If you want to eliminate the categorical variables with numerical entries, simply left click on the column number (C1, C3, C5, C6, and C10) to highlight the entire column then right click and choose delete cells. You will have to do this one column at a time.

### 1.1.2 Understanding the Distribution of a Variable

The distribution of a variable consists of the general pattern of the data, the variation in the data, and any unusual observations in the data. It is simply the values a variable assumes and how often these values occur. All of the features of a distribution of a continuous variable can be observed in a graph called a dotplot. To make a dotplot with MINITAB\textsuperscript{TM}, click Graph>Dotplot as on Figure 1.4 on the next page. If you have the data set Bears.MTW (found in MTBWIN\DATA directory) open and you want to make a graph of the weights by sex, you would fill in the resulting window as Figure 1.5 on the following page. The final graph would resemble Figure 1.6 on page 16. Note that Sex 2 has more tightly clustered weights than Sex 1 and that Sex 1 has a clearly skew right distribution.

More will be said about dotplots in conjunction with the rest of the continuous data’s graphs. Some additional features of MINITAB\textsuperscript{TM}’s Dotplot command are demonstrated in Video 1.1.

### 1.2 Displaying Data with Tables and Graphs

For categorical variables, the distribution is often represented with frequency tables or graphs such as bar graphs and piecharts.
1.2. Displaying Data with Tables and Graphs

Figure 1.4: Graph>Dotplot Menu

Figure 1.5: Graph>Dotplot Window for Weight of Bears by Sex

1.2.1 Creating Frequency Tables

To produce a frequency table for a single variable choose Stat>Tables>Tally and then select the variable of interest when the Tally Dialog Box appears. An example of the Tally Dialog Box appears in Figure 1.7 on the next page. After selecting the variables of interest, click OK and the results will appear in the Session Window such as Output 1.1 on the following page for the variable TypeC. Note that Count, Percents, Cumulative Counts, and Cumulative Percents were selected in the Tally Dialog Box and subsequently information on Count, Cumulative Counts (CumCnt), Percent, and Cumulative Percents (CumPct) are displayed in the Session Window output of Output 1.1 on the next page.

The information in frequency tables may also be shown in graphical form as a chart. Suppose we want to produce a graph depicting the number of houses with various furnace types. There are two ways we can produce a chart depicting the number of houses with various furnace types.

The first way is to choose Graph>Chart. When your Chart Dialog Window opens, click on the Function drop down menu and select Count. Next, select any numerical variable as your measurement variable (we choose BTU.In). Finally, select TypeC as your categorical variable then click OK. The Chart Dialog Box is
1.2. Displaying Data with Tables and Graphs

Figure 1.6: Bears’ Weight by Sex Dotplot

Bear Weights by Sex

Figure 1.7: Tally Dialog Box for Frequency Tables

Output 1.1: Tally Output of a Frequency Table

Tally for Discrete Variables: TypeC

<table>
<thead>
<tr>
<th>TypeC</th>
<th>Count</th>
<th>CumCnt</th>
<th>Percent</th>
<th>CumPct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forced Air</td>
<td>76</td>
<td>76</td>
<td>84.44</td>
<td>84.44</td>
</tr>
<tr>
<td>Forced Water</td>
<td>7</td>
<td>83</td>
<td>7.78</td>
<td>92.22</td>
</tr>
<tr>
<td>Gravity</td>
<td>7</td>
<td>90</td>
<td>7.78</td>
<td>100.00</td>
</tr>
<tr>
<td>N=</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

shown in Figure 1.8 on the following page. To have your chart resemble the chart in Figure 1.9 on the next page, you will need to select a few more options that are described next. In Figure 1.9, the title was created by clicking on the drop down Annotation menu in the Chart Dialog Box and choosing Title. When the Title Dialog Box opens, you can choose a title and options as shown in Figure 1.10 on page 18. The footnote appearing in the
1.2. Displaying Data with Tables and Graphs

Figure 1.8: Chart Dialog Box for Furnace Data

Figure 1.9: Bar Chart of Counts for Furnace Types

bottom left was also created using the drop down Annotation menu. Experiment using different text fonts and different text colors on your own. Finally, Figure 1.9 does not have the default variables on the X and Y axes. We choose not to have any variable display on the X axis and to have the Y axis say Count. To open the Axis Dialog Box and make changes to the axes, choose Frame>Axis from within the Chart Dialog Box as in Figure 1.8. The changes we made to the axes are shown in Figure 1.11 on the next page.

The second way to create this chart is to first create a frequency distribution for the furnace types with the tally command. (Stat>Tables>Tally) Then, either copy or directly type the output from the Session Window into a worksheet. An example of the Worksheet Window is given in Figure 1.12 on the following page.

Once the information is entered in the worksheet, select Graph>Chart. When the Chart Dialog Box opens, select Sum or Mean from the drop down menu of functions; select Frequency for the measurement variable, and select Type1 for the Category. Provided you select the same title and axes options you selected previously, the two methods will produce identical graphs. Video 1.2 on page 19 demonstrates coding numerical data into five classes and displaying the coded categorical data with both a frequency table and a bar graph.
1.2. Displaying Data with Tables and Graphs

Figure 1.10: Title Dialog Box for Furnace Chart

Figure 1.11: Axis Dialog Box for Furnace Chart

Figure 1.12: Worksheet Example of Tally Output
Video 1.2: For **Problem 1.33** — (Duration: 5 minutes 44 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select **Start > Settings > Control Panel > Display > Settings.**

To create charts with subgroups within the main categorical variables with appropriate legends, we use both the **group** and **cluster** options simultaneously. For example, suppose we want to create a chart which will depict the mean energy usage by damper for the various furnace types (See Figure 1.13). To create such a graph,

Figure 1.13: Mean Energy Usage by Furnace Type and Damper Category

![Clustered Chart](image)

choose **Graph > Chart**. When the **Chart Dialog Box** opens, select **Mean** as your function from the drop down menu. Choose **BTU.In** as your measurement variable and **TypeC** as your categorical variable. At this point, your **Chart Dialog Window** should resemble Figure 1.14. Click on the drop down menu in the **For each Box** and select **group**. Next, select **DamperC** for the **Group variables Box.** Finally, click on the **options** button and select **Cluster** with **DamperC** in the **Cluster Box** as shown in Figure 1.15 on the following page. You should create an appropriate
1.2. Displaying Data with Tables and Graphs

Figure 1.15: Chart Options Box for Furnace Data with a Clustered Chart

Pie charts may be created by using a single column containing the categorical variables of interest, or from a frequency table showing the categories of interest and the frequency of occurrence for each category. We use the information on furnace types stored in C11 (TypeC) first and then the information we stored after creating a frequency table using the Tally command in columns C16-T (Type1) and C17 (Frequency) to illustrate the creation of identical pie charts. To create a pie chart, choose Graph > Pie Chart, and the Pie Chart Dialog Window will appear as shown in Figure 1.16. When your categorical information is stored in a column as the information in TypeC, click in the Chart data in circle (to) and select TypeC and provide an appropriate title in the Title Box. If you have both the categories and frequency of the categories stored in separate columns like Type1 and Frequency, click in the Chart table circle (to) then select TypeC for the Categories Box and Frequency for the Frequencies in Box. The Pie Chart Dialog Box is shown in Figure 1.16. When your Pie Chart Dialog Box resembles Figure 1.16, click OK and the result should resemble Figure 1.17 on the next page.
1.2. Displaying Data with Tables and Graphs

1.2.2 Graphing Time Series Data

Data that are observed at regular intervals are often depicted using time series plots. Time series plots typically display measurement data on the vertical (Y) axis and time information on the horizontal (X) axis. There are numerous options available for displaying the date information using a time series plot available on line from the Time series plot help menu. We will illustrate a few of the options to produce a time series plot for Problem 1.37 which uses the worksheet Life.MTW. One of our final time series plots for Problem 1.37 is shown in Figure 1.18.

To produce a time series plot using the worksheet Life.MTW, choose Graph>Time Series Plot then under Graph Variables:, select Men for the Graph 1 Y, and Women for the Graph 2 Y as shown in Figure 1.19 on the next page. Click OK and notice that separate time series plots are created for the variables Men and Women. To get both time series plots on the same graph, click Edit>Edit Last Dialog or the Edit Last Command Dialog Button . When the Time series plot Dialog Box opens, click on the Frame drop down menu and select Multiple graphs. After the Multiple graphs Dialog Box opens, click in the circle to the left of Overlay graphs on the same page . When you are finished, click OK. At this point, there is simply an index counter on the X axis and the Y axis shows the first graphing variable selected, Men. Ideally, we would like the tick marks on the X axis to display a year and have the Y axis label say Life Expectancy. To change the index values to dates, click the Edit Last Command Dialog Button, then select Options from the Time series plot Dialog Box. Beneath the
1.2. Displaying Data with Tables and Graphs

Figure 1.19: Time Series Dialog Box

words *Cycle through values type 1920:1990/10*. This will tell MINITAB™ to place the years 1920 through 1990 by 10 year increments where there were previously index values. Click the OK button twice and your graph will appear. Notice that we still need to fix the label on the Y axis. To change a label, first click on the Edit Last Command Dialog Button, then select Frame>Axis. When the Axis Dialog Box opens as shown in Figure 1.20, change the value in the Y axis label Box from Auto to *Life Expectancy*. When you are finished click OK twice.

Figure 1.20: Axis Dialog Box for Time Series

Titles and appropriate footnotes may be added by selecting the drop down annotation button inside the Time series plot Dialog Box in the same fashion titles and footnotes were created for charts. One additional feature that was used in the creation of Figure 1.18 on the page before not previously discussed was the use of text in a graph other than titles and footnotes. To place text in a time series plot, first click on the annotation drop down menu inside the Time series plot Dialog Box. Next, select text and the Text Dialog Box will open. You should type in values to mirror those shown in Figure 1.21 on the following page. The commands tell MINITAB™ to place the text *Women* in the graph at position 5, 75, and the text *Men* at position 5, 68.

Time series plots can also be produced using the Plot command. The graph shown in Figure 1.22 on the next page, was produced using the Minitab plot command.

To produce a time series plot with the plot command, select Graph>Plot. Fill in the Plot Dialog Box as shown in Figure 1.23 on the following page.
The title and footnote are produced in the same fashion used for the time series plot. Just as with the time series plot, you will have to change the label on the Y axis if you want it to read Life Expectancy. One small difference in placing text in the graph between the Time series plot command and the plot command is the numbering for the X axis used for text placement. The Time series plot used the index values for placement along...
1.2. Displaying Data with Tables and Graphs

the X axis while the plot command uses the actual numbers we provided for the X axis (the actual dates). Consequently, the X and Y placements change from (5, 75) and (5, 68) in the Time series plot to (1960, 75) and (1960, 68) respectively for the plot command.

A third way to create a time series plot for Problem 1.37 is to stack the data in the worksheet and subsequently group the variables by gender. To use this third approach, first you must rearrange your worksheet data to appear as shown in Figure 1.24. Once the data are rearranged, select Graph>Plot And fill in the Plot Figure 1.24: Time Series Plot Worksheet Rearrangement

Dialog Box as shown in Figure 1.25 on the following page. Notice that Life Expectancy is selected as the Y variable, while year is chosen as the X variable. Item 1 beneath the For each Box was changed from the default value of graph to group, and Sex was selected as the grouping variable for Item 1. The resulting graph shown in Figure 1.26 on the next page automatically produces a legend depicting the various categories in the grouping variable. In our case, the grouping values for Sex are Male and Female.
1.2. Displaying Data with Tables and Graphs

Figure 1.25: *Plot Dialog Box* for Worksheet Rearrangement to Make Time Series Plot

![Plot Dialog Box](image1)

Figure 1.26: Time Series Plot for Problem 1.37 with a Rearranged Worksheet

![Time Series Plot](image2)
1.3 Displaying Numerical Data

When variables are continuous, the distribution is displayed with a graph illustrating the data such as a dotplot, stem-and-leaf plot, histogram, or a density curve. We examine each of these graphs using the data set Framingh.MTW. This data set is also used for Problem 1.62 in your book. The data in Framingh.MTW are the cholesterol levels for 62 subjects that participated in the Framingham Heart Study.

1.3.1 Displaying Data with a Dotplot

MINITAB™'s dotplot command displays a dot for each observation along a number line. When values are identical or close in value, MINITAB™ stacks the dots vertically. Several options exist when creating dotplots such as to group by a second (discrete) variable, or to view two continuous variables on the same scale. For purposes of illustration, suppose that the first 31 cholesterol levels in Framingh.MTW correspond to Males and the second 31 cholesterol levels correspond to Females. To create the variable SEX with appropriate entries type SEX in the gray box beneath C2, then type Male in the first row of the SEX column. To fill in the remaining 30 rows with the value Male, left click with your mouse the first row of column SEX where you have typed Male. Move your mouse to the lower right of the highlighted box until the solid black cross hair appears (+). Hold the left mouse button down and drag the cross hair down 30 rows. Repeat the same procedure using Female to create values for the variable SEX for rows 32-62. An alternative method to fill in the SEX values for C2 is to use the command Calc>Make Patterned Data>Text Values. When the Text Values Dialog Box appears as shown in Figure 1.27, type MALE FEMALE in the Text Values Box, and be sure to type 31 in the List Each Value Box. When your Text Values Dialog Box appears identical to Figure 1.27, click OK. At this point it is a good idea to save the new worksheet you have created. Save the worksheet as Framingh1.MTW.

To create a dotplot of all of the cholesterol values, choose Graph>Dotplot and title the graph Framingham Heart Study in the Dotplot Dialog Window of the Title Box as shown in Figure 1.28 on the next page. To split the dotplot according to a categorical variable, click in the circle to the left of By variable (C to O) in the Dotplot Dialog Window. When the By Variable Box changes background colors from gray to white, click once with your mouse in the By Variable Box to make it active. Next, select the categorical variable of interest (in our case SEX) by either double clicking on the categorical variable (SEX) or clicking once on the categorical variable (SEX) then clicking on the Select Button. Dotplots with no grouping and grouping by SEX are shown in Figure 1.29 on the following page and Figure 1.30 on the next page respectively.
1.3. Displaying Numerical Data

Figure 1.28: Dotplot Dialog Window for Framingham Heart Study

Figure 1.29: Dotplot For Framingham Heart Study With No Grouping

Figure 1.30: Dotplot For Framingham Heart Study With Grouping

Unstacking and Stacking Data  At times you may want to have several columns each representing a category of a numerical variable. Suppose you want to split the data in the Framingham Heart Study into two columns. One
1.3. Displaying Numerical Data

Figure 1.31: Unstack Columns from Framingham Heart Study

[Unstack Columns dialog box image]

To display numerical data, you can use the Unstack Columns feature. This is accomplished by choosing Manip > Unstack Columns and filling in the appropriate boxes as shown in Figure 1.31. Note that if you follow Figure 1.31, you will create a new worksheet named Unstacked that will contain the two columns of interest. To stack data, choose Manip > Stack > Stack Columns. As an exercise, see if you can reproduce the original data set based on the data in the worksheet Unstacked.

To compare two or more numerical variables using a dotplot, first select the variables of interest in the Variables Box of the Dotplot Dialog Window. Second, click in the circle to the left of Each column constitutes a group. Verify for yourself that the dotplot you create using two separate columns is identical to the dotplot created by using a categorical variable.

1.3.2 Displaying Data with Stem-and-Leaf Plots

A stem-and-leaf plot is similar to a dotplot in that it displays numerical data along a number line. It is possible to create a stem-and-leaf plot with MINITAB™ using any one of three procedures that lead to the Stem-and-Leaf Window. We will list the three procedures, and you may use your choice of the three. The first way is to select Graph > Stem-and-Leaf; the second procedure is to select Stat > EDA > Stem-and-Leaf; and the last method to arrive at the Stem-and-Leaf Window is to select Graph > Character Graphs > Stem-and-Leaf. We will routinely use the first method as it only requires two mouse clicks whereas the other two procedures require three mouse clicks each. Figure 1.32 on the next page depicts the Stem-and-Leaf Dialog Window and the options needed to create a stem-and-leaf plot of the variable cholest in the Framingh.MTW data set. The resulting stem-and-leaf graph is depicted in the Session Window of MINITAB™ after clicking OK and is shown in Output 1.2 on the following page. There are three parts to the stem-and-leaf plot. The first column displays the depth values, the second column displays the stems, and the remaining values are the leaves. For the most part, we will not use the depth values displayed in the first column. Notice in Output 1.2 on the next page that MINITAB™ used a stem increment of 20 units for this data set. If you want a different stem increment, type the desired stem increment in the Increment Box of Figure 1.32 on the following page prior to clicking OK. We should make a few observations about the stem-and-leaf plot before continuing. First, the leaf values (for Output 1.2) are in units of 10. This is shown in the third row of text in the stem-and-leaf display. The first leaf (6) represents the number 67. All values between 60 and 69 are represented with a 6. The stem of 1 to the left of the leaf (6) represents units of 100. Consequently, taken as a whole, the stem of 1 and the leaf of 6 represents the smallest cholesterol reading in the Framingh.MTW data set which is 167. In a similar fashion, the third leaf (9) represents the number 92 while the fourth leaf (9) represents the number 98. Video 1.3 on the next page uses stem-and-leaf plots as well as dotplots to analyze the inaugural ages of United States presidents.
1.3. Displaying Numerical Data

Figure 1.32: Stem-and-Leaf Dialog Window

Output 1.2: Stem-and-Leaf Graph of cholesst from Framingh.MTW

Stem-and-leaf of cholesst  N  =  62
Leaf Unit  =  10

1   1 6
4   1 899
13  2 00111111
30  2 222233333333333
(11)  2 4444455555
21  2 66666677777
10  2 88
8   3 000
5   3 233
2   3 5
1   3
1   3 9

Video 1.3: For Problem 1.126 — (Duration: 3 minutes 47 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

1.3.3 Displaying Data with Histograms & LOWESS Lines

A histogram is another graph used to depict numerical information. It is similar to the bar chart from Section 1.2. However, the categories for the bar chart from Section 1.2 corresponded to discrete categories while the categories for the histogram correspond to intervals of continuous data. Histograms may be easily constructed from stem-and-leaf plots. However, in doing so, information is obscured in the resulting histogram. To create a histogram with MINITAB select Graph>Histogram and then fill in the appropriate boxes in the Histogram Dialog Window. MINITAB selects intervals (class limits) based on what it believes will render the best representation of the data. Often times we will want to override the default interval width MINITAB chooses. To create a histogram using the same class limits as the stem-and-leaf plot from Output 1.2, select Graph>Histogram. Select cholesst as the
1.3. Displaying Numerical Data

Graph Variable in the Histogram Dialog Window. Next, click on the Options Button and fill in the Histogram Options Dialog Window as shown in Figure 1.33. Be sure to click in the circle to the left of CutPoint (to )

![Figure 1.33: Histogram Dialog Window](image)

and then click in the circle to the left of Midpoint/cutpoint positions ( to ) and fill in the box to the right of Midpoint/cutpoint positions with 160:400/20 before clicking OK twice. The numbers 160:400/20 tell MINITAB to create class limits every 20 units starting at 160 and going up to 400. The default histogram is a frequency histogram. If you want a percent histogram instead of a frequency histogram, click in the circle to the left of Percent ( to ) beneath Type of Histogram in the Histogram Options Dialog Window. Titles and footnotes can be added to a histogram with the Drop Down Annotation Menu available in the Histogram Dialog Window in the same fashion as titles and footnotes were added to bar charts in Section 1.2. Video 1.4 creates a histogram from a sample of 62 subjects from the Framingham Heart Study using class limits from MINITAB's stem-and-leaf plot default settings. Video 1.4 also superimposes a LOWESS line over the histogram similar to Figure 1.35 on the next page.

Video 1.4: For Problem 1.56 — (Duration: 4 minutes 27 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

Sometimes, a histogram obscures more data than it reveals, so you will want to examine the LOWESS line. LOWESS stands for “locally weighted scatterplot smoother” and is an option available for use with a MINITAB histogram. A LOWESS line superimposed over a histogram or shown by itself provides a smoothed line to approximate the distribution of interest. A LOWESS line is not the same as a density since the method used to construct LOWESS lines does not force the area underneath the line to equal 1. However, the LOWESS function is the best option currently available in MINITAB to approximate a density with a smoothed function. To create a LOWESS line in MINITAB, select Graph>Histogram. Use the Display Drop Down Menu to select LOWESS for the second item of the Data Display. Be sure to select Graph in the For Each Box of the second item of the Data Display as shown in Figure 1.34 on the following page. A density histogram with a title, footnote, and a superimposed LOWESS line is shown in Figure 1.35 on the next page.
1.3. Displaying Numerical Data

Figure 1.34: Selecting Graph under Data Display

Figure 1.35: A Density Histogram with a LOWESS Curve
1.4 Summarizing Data with Statistics

One of the principal goals of statistics is to make generalizations about a population based on data from an appropriately selected sample. The numerical summary measures we calculate from samples are called statistics. The hope is that we select “good” statistics to generalize the population. Numerical measures calculated for a population are called parameters. Parameters are usually unknown. If we know the parameter of interest, there is no need to use statistics! The algorithms in Minitab™ use formulas to calculate statistics. In other words, Minitab™ assumes you are working with a sample.

1.4.1 Calculating and Interpreting Numerical Characteristics of Data

One of the more useful commands to summarize data with Minitab™ is to select Stat > Basic Statistics > Display Descriptive Statistics. The data set Simpson.MTW is described in Section 1.5 of your book and will be used to illustrate the Stat > Basic Statistics > Display Descriptive Statistics command’s Session Window output for the variable gpa shown in Output 1.3. The number (100) beneath N provides the number of non-missing observations in the data. The number (2.5230) beneath Mean provides the sample mean and is calculated by adding all of the values in gpa and then dividing by the total number of non-missing values (100). The sample mean is denoted with the symbol \( \bar{x} = \frac{\sum x}{n} \). The number (2.4700) beneath Median is the sample median and is denoted with an uppercase m (M). Minitab™ uses the formula \( L(M) = (n + 1)/2 \) to determine the location of the median after the data have been sorted. The notation \( L(M) \) stands for the location of the median, not the value of median. In our example, \( L(M) = (100 + 1)/2 = 50.5 \). This tells us the median is located between the 50\(^{th} \) and 51\(^{st} \) ordered data points. To calculate the value of the median, we average the value at the 50\(^{th} \) and 51\(^{st} \) ordered data points. For this problem, the values are identical and the median is calculated as 2.47. Note: To sort the data, use the command Manip > Sort. When the Sort Dialog Window opens, you will need to specify variables for three boxes. Figure 1.36 on the next page provides an example where the sorted variable (gpa) is stored in a column named gpas. The value (2.5247) beneath TrMean is a 5% trimmed mean. To calculate a 5% trimmed mean, first sort the observations. Next delete 5% (5 = 100 \times 0.05) observations from each end of the sorted numbers. An easy way to do this is to highlight the 5 smallest values with your mouse by holding down your left mouse button as you move over the values, then right click and choose Delete Cells. (Make sure you are highlighting and deleting the sorted values!) Follow the same procedure for deleting the largest 5% of your values. Once you have deleted the smallest and largest 5% of observations, calculate the mean of the resulting observations. If you use the Descriptive Statistics command to calculate the mean of the remaining 90 observations, make sure you look under Mean and not TrMean for the 5% trimmed mean. If you look under TrMean after deleting 5% of your observations for this problem, the value you will see actually corresponds to a 9% trimmed mean of the original data. The value for the sample standard deviation is reported beneath StDev (0.335) and is calculated using the formula \( s = \sqrt{\sum (x - \bar{x})^2}/n - 1 \). We will verify the value (0.335) using Minitab™’s calculator. To calculate the standard deviation using Minitab™’s calculator, choose Calc > Calculator. There are numerous mathematical and statistical functions you can select by scrolling down the Functions Box. The needed commands to calculate the sample standard deviation are illustrated in Figure 1.37 on the following page. The number (0.0335) beneath SE Mean is the standard error of the mean. The standard error of the mean is calculated by dividing the sample standard deviation by the square root of the sample number.
1.4. Summarizing Data with Statistics

Figure 1.36: Variable gpa Sorted and Stored in gpas

Figure 1.37: Calculation of Standard Deviation with MINITAB’s Calculator

size (s_x = s/\sqrt{n}). The numbers 1.83 and 3.14 beneath Minimum and Maximum respectively are the smallest and largest values in the data set. The numbers 2.29 and 2.775 beneath Q1 (Q_I) and Q3 (Q_3) respectively are the first and third quartile. The first and third quartiles are often referred to as the 25th and 75th percentiles. MINITAB determines the locations of Q_1 and Q_3 by using the formulas L(Q_1) = (n + 1)/4, and L(Q_3) = 3 \times (n + 1)/4. The location of Q_1 and Q_3 for gpa respectively are (100 + 1)/4 = 25.25 and 3 \times (100 + 1)/4 = 75.75. Remember, the location values refer to the values in your data set once they have been sorted (order statistics). To determine the value of Q_1, we take the value at the 25th order statistic (2.29) and then add 0.25 of the distance between the 25th and 26th order statistics to the value of the 25th order statistic. In this case, the 25th and 26th order statistics are both 2.29, so 25% of their difference is 0 and we report the value of Q_1 as 2.29. To find Q_3, take 75% of the distance between the 75th and 76th order statistics (0.75 \times (2.78 - 2.77) = 0.0075) and add this value to the value of the 75th order statistic (2.77). Note that 2.77 + 0.0075 = 2.775 which is agreement with the value MINITAB reports for Q_3.

Two common measures of variability MINITAB does not directly report are the range and the interquartile range (IQR). However, since the range is defined as Maximum – Minimum, it is easily determined based on the output from the Display Descriptive Statistics command. IQR is defined as Q_3 – Q_1. Consequently, IQR is likewise easily determined from the output of the Display Descriptive Statistics command. The range for
the variable gpa is 1.31 (3.14 – 1.83 = 1.31), and the IQR is 0.4875 (2.7775 – 2.29 = 0.4875). The IQR is a measure of the range of the middle 50% of the observations. When your data is skewed or contains outliers, the IQR provides a more informative summary of the data’s variability than does the standard deviation since the standard deviation is easily influenced by outliers whereas the IQR is fairly robust to outliers.

### 1.4.2 The Empirical Rule

Recall that the Empirical Rule indicates we can expect approximately 68% of the values in a sample to fall within one standard deviation of the mean, provided the shape of the data is approximately normal (bell-shaped). It goes on to say we can expect approximately 95% of the values in a sample to fall within two standard deviations of the mean provided the shape of the data is approximately normal. Finally, the Empirical Rule tells us that we can expect virtually all of our data in a sample to fall within three standard deviations of the mean. To evaluate the shape of a distribution, we might use a stem-and-leaf plot, a dotplot, a density curve, or a histogram.

Consider the histogram for gpa with a superimposed LOWESS curve shown in Figure 1.38. Based on

![Figure 1.38: Histogram with LOWESS Curve for gpa](image)

Figure 1.38, we can argue that gpa values have a moderately normal shape. Since the sample mean ($\bar{x}$) and sample standard deviation ($s$) for the variable gpa are 2.523 and 0.335 respectively, we can expect approximately 68% of the values in gpa to fall in the interval $(2.523 - 0.335 = 2.188, 2.523 + 0.335 = 2.858)$; approximately 95% of the values in gpa to fall in the interval $(2.523 - 2 \times 0.335 = 1.853, 2.523 \times 0.335 = 3.193)$; and virtually all of the values in gpa to fall in the interval $(2.523 - 3 \times 0.335 = 1.518, 2.523 + 3 \times 0.335 = 3.528)$. Actually, 65% of the values for the variable gpa fall in the interval 2.188 to 2.858, while 99% of the values fall in the interval 1.853 to 3.193, and all of the values fall in the interval 1.518 to 3.528. Video 1.5 on the following page illustrates how to standardize numerical values and to assess the normality of the data using knowledge of the empirical rule.

### 1.4.3 Z-scores

To calculate a Z-score for either a single value of a column of values, click on Calc > Standardize. When the Standardize Dialog Window opens, select the variable(s) you want to standardize by either double clicking the variable(s) or clicking once on the variable(s) then choosing Select. Type a name(s) for the column(s) where you want the standardized variables to be stored. Be sure you have Subtract mean and divide by std. dev. option selected (the second to the right) in the Standardize Dialog Window. The Standardize Dialog Window for standardizing all of the gpa values and storing the resulting Z-scores in a column named Zgpa is shown in Figure 1.39 on the next page.

Since virtually all of the values in a bell shaped distribution fall within three standard deviations of the mean, a Z-score greater than 3 is considered unusual.
1.4. Summarizing Data with Statistics

Figure 1.39: Standardizing Variable gpa

Video 1.5: For Problem 1.75 — (Duration: 4 minutes 27 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels. To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.
1.5 Describing a Distribution

When describing a distribution we need to discuss shape, center, and spread. One should first decide whether the distribution is uni-modal (one mode), or multi-modal (multiple modes). Graphical tools such as the dotplot, the stem-and-leaf, and the histogram are useful in determining modality. If a distribution is uni-modal, we need to classify the distribution as either symmetric or skewed. When distributions are symmetric, we need to decide whether the tails of the distribution are longer than the tails we might expect to see in a normal distribution. To help us identify outliers and check the symmetry of a distribution we will use the boxplot. The boxplot is a graph depicting the middle 50% of a distribution with a box. Lines, often called whiskers, extend from each end of the box to the adjacent values. The adjacent values are the largest and smallest values in a data set that do not exceed the upper and lower fences respectively. The upper and lower fence are not actually depicted on the boxplot but are used to determine whether an observation is classified as an outlier. The box is drawn by creating three vertical lines at the first, second, and third quartiles ($Q_1$, $Q_2$, $Q_3$) and connecting the vertical lines at $Q_1$ and $Q_3$ with horizontal lines. The upper fence is determined by adding 1.5 times the IQR to $Q_3$. The lower fence is determined by subtracting 1.5 times the IQR from $Q_1$. Values greater and smaller than the upper and lower fences respectively are called outliers and are typically depicted with a star. Figure 1.40 actually shows the upper and lower fence for the weights of a group of female bears. Note that the fences are not normally shown with the boxplot.

![Boxplot of Female Bears with Fences Labeled](image)

1.5.1 Examining Tails and Outliers

The boxplot in Figure 1.40 shows a relatively symmetrical distribution with four outliers. However, just because there are outliers does not make a distribution long tailed! When we talk about a long tailed distribution, we are using the normal distribution as a reference to decide whether or not the specific distribution we are examining has long tails. That is, does the distribution we are examining have longer tails than a normal distribution. The normal distribution will have approximately 0.007 of its values classified as outliers with a boxplot. Consequently, we see that the number of outliers we might see in a distribution is intimately linked with the sample size. We can expect approximately 7 outliers if we take a random sample of size 1000 from a normal distribution. Try this on your own by using the commands Calc>Random Data>Normal generating 1000 rows of data. Once you have generated the random data, create a boxplot of your random numbers by selecting Graph>Boxplot. Remember you may not get exactly seven. However, if you keep performing this experiment and look at the distribution of outliers when taking samples of size 1,000 from a normal distribution, the center of the distribution will be close to seven. To help determine whether you are observing values that might warrant classifying a distribution as long tailed, we have created Table 1.1 on the next page. Table 1.1 assumes you have already decided the distribution is uni-modal and provides a column with the number of outliers and various sample sizes to help you decide whether or not your distribution is long tailed. To use Table 1.1, you should first count the number of outliers you have in your
distribution. Next, you want to see where your sample size falls in relation to those in the row with the number of outliers you have discovered. You can be at least Level confident that you have a long tailed distribution where Level is the number at the top of the column corresponding to the sample size given that is larger than the one you have. For example, if you have 5 outliers and a sample size of 250, you can be at least 95\% confident that you have a long tailed distribution because 284 is larger than 250. If you had 9 outliers and a sample size of 400, you would be at least 99\% confident that you had a long tailed distribution because 505 is larger than 400. For a sample size of 700 with 9 outliers, you would only be 90\% confident that you had a long tailed distribution.

<table>
<thead>
<tr>
<th>outliers</th>
<th>Level .90</th>
<th>Level .95</th>
<th>Level .99</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>77</td>
<td>52</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>159</td>
<td>118</td>
<td>64</td>
</tr>
<tr>
<td>4</td>
<td>251</td>
<td>197</td>
<td>120</td>
</tr>
<tr>
<td>5</td>
<td>350</td>
<td>284</td>
<td>185</td>
</tr>
<tr>
<td>6</td>
<td>453</td>
<td>376</td>
<td>257</td>
</tr>
<tr>
<td>7</td>
<td>560</td>
<td>473</td>
<td>336</td>
</tr>
<tr>
<td>8</td>
<td>669</td>
<td>573</td>
<td>419</td>
</tr>
<tr>
<td>9</td>
<td>780</td>
<td>675</td>
<td>505</td>
</tr>
<tr>
<td>10</td>
<td>894</td>
<td>780</td>
<td>595</td>
</tr>
<tr>
<td>11</td>
<td>1008</td>
<td>887</td>
<td>687</td>
</tr>
<tr>
<td>12</td>
<td>1124</td>
<td>995</td>
<td>781</td>
</tr>
<tr>
<td>13</td>
<td>1241</td>
<td>1105</td>
<td>878</td>
</tr>
<tr>
<td>14</td>
<td>1360</td>
<td>1216</td>
<td>975</td>
</tr>
<tr>
<td>15</td>
<td>1479</td>
<td>1328</td>
<td>1075</td>
</tr>
</tbody>
</table>

Video 1.6 uses a boxplot in conjunction with descriptive statistics and a histogram to explore the distribution of the average monthly payments for families participating in AFDC (Aid to Families with Dependent Children).

Video 1.6: For Problem 1.91 — (Duration: 5 minutes 7 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start > Settings > Control Panel > Display > Settings.

1.5.2 Describing the Center of a Uni-Modal Distribution

From here forward, we will restrict our discussions to uni-modal distributions. When a distribution is symmetrical, the center can accurately be described by using either the mean or the median. This is true since the mean and median coincide for symmetrical distributions. However, since many of the statistical procedures we use later are based on normal distributions, it turns out that the mean has nicer “smaller variance” statistical properties than does the median for normal distributions and hence is to be preferred. When the distribution is skewed, the median provides a more representative measure of center than does the mean since it is more robust to outliers than is the mean. When the distribution is symmetrical with long tails, the trimmed mean is the preferred method of describing the center of a distribution.
1.5.3 Describing the Spread of a Uni-Modal Distribution

The most common method of describing the spread of a distribution is the standard deviation. However, this measure should only be used when the distribution is relatively symmetric. When the distribution is skewed, the preferred measure of spread is the interquartile range. So, when the mean is an appropriate measure of center, the standard deviation will be an appropriate measure of spread. If the distribution is skewed, the appropriate measure of center is the median, and the appropriate measure of spread is the IQR.

1.5.4 Interpreting Boxplots

The box portion of the boxplot contains roughly 50% of the data with the median splitting the data itself into equal halves. When the median splits the IQR box into equal halves, the middle 50% is symmetric. If the median is closer to one end of the box, then the distribution is skewed in the opposite direction. The tails of the boxplot are also useful in determining symmetry and skew of the data. When the tails are similar in length, the median splits the IQR box into equal halves, and no outliers are present, the distribution is symmetric.

1.5.5 Knowing When Boxplots Are Appropriate

In addition to using boxplots to help identify outliers, boxplots are useful when comparing several groups of data whose measurements are all on the same variable. Consider the MINITAB worksheet ALFALFA.MTW located in the MTBWIN\DATA folder. The data in ALFALFA.MTW are from a University of Wisconsin researcher who tested the yields of six varieties of alfalfa on each of four separate fields. We will use side-by-side boxplots to help us decide which variety has the highest yield. To create side by side boxplots, you must have both a numerical and a categorical variable. In the case of the ALFALFA.MTW worksheet, Yield is the numerical (measurement) variable and Variety is the categorical variable. Select Graph > Boxplot, and then select Yield as the measurement variable and Variety as the categorical variable. An example of the Boxplot Dialog Window is shown in Figure 1.41. After assigning appropriate titles and footnotes using the Annotation drop down menu, the result is shown in Figure 1.42 on the following page. Note that varieties 1, 4, 5, and 6 are skewed to the right, variety 3 is slightly skewed to the left, and variety 2 is relatively symmetric.
Figure 1.42: Boxplot of Yield by Variety for *ALFALFA.MTW*
1.6 Summary and Review Labs

Lab 1.1 — Graphing Practice

Objectives:

To create and read a bar chart, a histogram, and a time series plot.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

The following question all deal with data that comes with your book.

Questions and Directions:

Problem 1: Fifty families were interviewed and the number of dependent children were recorded and stored in the worksheet DEPEND.MTW.

a. Create a table using (Stat>Table>Tally) showing the frequency for the various numbers of dependents. Append the results to the report pad.

b. Copy the numerical values beneath the variables number and count from the Session Window and paste the results into columns 2 and 3. (Do not copy the N = 50) Name column 2 (C2) Dependents, and column 3 (C3) Frequency. Produce a Chart depicting the Frequency for the possible outcomes in the data set. Change the Y axis to read Frequency and the X axis to read Number of Dependents. (Hint: Graph>Chart>Frame>Axis then change the Labels.) Place a left justified footnote reading “Data: Depend.mtw”; and a two line center justified title reading “Bar Chart Depicting Frequency / for Number of Dependents”. Make the bar fill type solid and the back color green. Append your graph to the report pad. If you do not have a color printer, attach the history for your chart directly beneath your appended chart so your instructor can verify that you changed the back type to green.

c. Is the bar chart symmetrical with respect to number of dependents? Is it possible for a family to have more than 7 dependents? Explain your answer.

d. Create a histogram for the data in DEPEND.MTW to mirror all of the labels, footnote text and placement you created in part 1b. Do however change the title to read “Histogram Depicting Number of Dependents”. Append your histogram to the report pad. Does your histogram give identical information to the bar chart and the frequency table? Which format (1a, 1b, or 1d) seems to convey the data the best?

Problem 2: The closing prices for three funds are stored in the MINITAB™ worksheet TIAA.MTW. This worksheet also has a date-time formatted column showing the dates for the respective closing prices. The first five rows of information are shown in Table 1.2 on the next page. If your worksheet does not display the date in exactly the same format as shown below, you will need to delete the current information in the date column and then type the information as shown below for the first two rows. Highlight the dates you entered (4/16/99 and 4/17/99) then move your mouse to the bottom right corner of the highlighted dates until the crosshair appears. When the crosshair appears, drag the crosshair by holding down and dragging with the left mouse button to row 365. (The reason your worksheet may not come properly formatted is because .mtp files do not support date-time formats.)

a. Create a time series graphs for both crefstk (CREF Stock) and crefgwt (CREF Growth) by using the command Graph>Time Series. Append both graphs to your report pad.
1.6. Summary and Review Labs

Table 1.2: TIAA Cref Stock Worksheet

<table>
<thead>
<tr>
<th>crefstk</th>
<th>crefgwt</th>
<th>tiaa</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>179.287</td>
<td>79.656</td>
<td>71.9043</td>
<td>4/16/99</td>
</tr>
<tr>
<td>179.287</td>
<td>79.656</td>
<td>71.9043</td>
<td>4/17/99</td>
</tr>
<tr>
<td>179.287</td>
<td>79.656</td>
<td>71.9043</td>
<td>4/18/99</td>
</tr>
<tr>
<td>176.454</td>
<td>77.444</td>
<td>70.3376</td>
<td>4/19/99</td>
</tr>
<tr>
<td>177.672</td>
<td>78.635</td>
<td>71.2485</td>
<td>4/20/99</td>
</tr>
</tbody>
</table>

Figure 1.43: Time Series Plot Window for TIAA CREF stock

b. If you used all of the default options in your last graph you should have index values of 100, 200, and 300 along the X axis. This is not overly helpful. Create a times series graph for both CREF Stock and CREF Growth using the date/time stamp. When your Time Series Plot Window resembles Figure 1.43 click OK. Append both of your graphs to the report pad.

c. Create one time series plot that will simultaneously show both funds. (Frame>Multiple Graphs. Then, click in the circle to the left of Overlay graphs on same page to OK) Make both the Symbol and Connect for crefstk green and the Symbol and Connect for crefgwt blue by editing the attributes. Change the default value along the Y axis to read “Closing Price”. (Frame>Axis then type Closing Price in the Y label Box). Title the graph “Closing Prices for CREF Growth and CREF Stock”. Add text showing the name of each time series. (Annotation>Text) Add a right justified footnote with the name of the worksheet. Your final product should resemble Figure 1.44 on the next page. When your graph resembles Figure 1.44, append your result to the report pad.

d. If you purchased 20 shares of crefstk and 40 shares of crefgwt both on 4/16/99, and can only sell both on a single day, when should you sell and what will you make in profit? Report your answer in the report pad. (Assume you pay no transaction or brokerage fees.) Since we have all of the data, answering the last question is a matter of seeing when the combined value of both stocks is highest. If you can develop a model indicating when to buy and sell stocks, you will have a rich future! To develop such models, you will need a solid background in statistics.
Figure 1.44: Time Series Plot for TIAA CREF data
Lab 1.2 — More Graphing Practice

Objectives:

I. To rearrange data to create bar charts
II. To learn how to format axes’ labels

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

The following question all deal with data that comes with your book.

Questions and Directions:

Problem 1: How has the cost of owning a car changed since 1984? The worksheet OWNCAR.MTW gives the percentages of total costs for owning a car in 1984 and 1993 according to an article in USA Today, September 27, 1994.

a. The information in the worksheet OWNCAR.MTW needs to be rearranged slightly to create a meaningful graph. Most statistical packages operate on variables stored in columns and MINITAB™ is not an exception. Rearrange the data using the editing capabilities of MINITAB™ to resemble Figure 1.45. The three variables of interest, Category, Percentage, and Year, now each have their own column. Save the newly formatted information in a worksheet called OWNCAR1.MTW.

b. Use the worksheet you created in part 1a (OWNCAR1.MTW) to compare the changes from 1984 to 1993 using a bar chart (Graph>Chart). When the Chart Window opens, you will want to select Percentage as your Y (measurement) variable and Category as your X (category) variable. You will also need to select the grouping variable Year and change the For each drop down menu from graph to group as shown in Figure 1.46 on the next page. Once your Chart Window resembles Figure 1.46, click on Options and make your Options Window resemble Figure 1.47 on the following page. When you are finished, click OK and when the Chart Window reappears, click Annotation>Title and title your graph “Percentages of Total Cost for Owning a Car in 1984 and 1993”. Add a right justified footnote using Annotation>Footnote specifying the source of your data. Finally, change the Y axis label to read “Percentage of Total Costs” by selecting Figure>Axis then typing Percentage of Total Costs in the Y label. Your final result should resemble Figure 1.48 on page 45.
c. Search the internet and see if you can find similar information for the current year. Once you have more recent information, enter your values and create a similar graph to Figure 1.48 with three reference years.

Problem 2: **Problem 1.35** in your book shows various heat sources and the percentages of three people groups that use the heat sources.

a. Rearrange the data into three columns as you did in problem 1 on the preceding page and store the results in a worksheet named **Heat1.MTW**.

b. Create a bar chart resembling Figure 1.49 on the next page. Note that the Tick size in Figure 1.49 is 0.75 and is shown at an angle of 15 degrees. To change the size and angle of your tick marks, select **Figure > Tick** from within the Chart Window.

c. Using the information in **Heat1.MTW**, create a table similar to the one shown in your book. To create a table with two or more categorical variables, you will want to choose
1.6. Summary and Review Labs

Figure 1.48: Final Comparative Bar Chart for Data Owncar1.MTW

![Comparative Bar Chart for Owncar1.MTW](image)

**Perceptions of Total Costs for Owning a Car in 1984 and 1993**

- 1984
- 1993

Figure 1.49: Comparative Bar Chart for Data Heat1.MTW

![Comparative Bar Chart for Heat1.MTW](image)

**Heating of American Indian Homes**

- All Us
- Not Reserv
- Reserv

Stat Tables Cross Tabulation. When the Cross Tabulation Window opens select the two
categorical variables from Heat1.MTW as the classification variables. Next, click on the
Summaries Box in the Cross Tabulation Window. Type the name of your numerical variable
from Heat1.MTW in the Associated Variables Box and then click inside the Display Means Box.
Click OK twice and the output will appear in the Session Window. Append your output from
the Session Window to the report pad.

d. Do you find the table from part 2c. or the bar chart from part 2b. more helpful when comparing
the various heat sources across people groups?

Problem 3: The worksheet Track15.MTW (Problem 1.43 in the book) lists the running times (in seconds) for
the men’s 1,500 meter run in the Olympics from 1896 to 1996.

a. Produce a time series plot with output to resemble Figure 1.50 on the following page
(Graph Time Series Plot). When the time series plot opens, select Time as the variable to
go in the Graph Variables Box, and click in the circle to the left of calendar ( ) to ( ) and
select Year from the drop down box in the Calendar Box. To display the proper year on the X
axis, click on the Options Box and type 1896:1996/4 in the box directly beneath the text Cycle
through values. The dates displayed in Figure 1.50 have a size of .5 and are shown at a 30 degree
angle. To change text angle and size click on Frame Tick in the Time Series Plot Window
then select the appropriate values for text size and text angle using the drop down menus in the
Tick Window. Note that you will also need to change the label for the Y axis to read “Time in
Seconds for the 1500". Be sure to include the center justified title and the right justified footnote as shown in Figure 1.50.

b. Use the internet to look up the winning times for all the Olympic men’s 1500 meter runs since 1992. Enter the years and winning times in seconds in the Track15.MTW worksheet. Then, save the worksheet as Track15a.MTW. Create a time series plot similar to Figure 1.50 using the data in Track15a.MTW.
1.6. Summary and Review Labs

Lab 1.3 — Stem-and-Leaf and Histograms Lab

Objectives:

I. To learn different ways to display data
II. To evaluate how well different graphs and charts display data

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

The following questions all deal with data that comes with your book.

Introduction:

The MINITAB\textsuperscript{TM} worksheet \texttt{Framingh.MTW} provides the cholesterol levels from a sample of 62 subjects from the Framingham Heart Study. (Problem 1.62 in your book.)

Questions and Directions:

1. Construct a stem-and-leaf plot using an increment value of 50. Based on the shape of your stem-and-leaf plot, how would you characterize the shape of the cholesterol values?

2. Use MINITAB\textsuperscript{TM} to create a frequency table displaying information identical to the information you see in the stem-and-leaf plot from part 1. In other words, we want a frequency table that has class limits of [150,200), [200,250), [250,300), [300,350), and [250,400). Hint: One method is to convert the numerical data that fall inside the class limits to different discrete values which can be either text or numeric values using the Manip$>$Code$>$ (Numeric to Numeric or Text to Text) command. See Figure 1.51 for an idea of how to fill in the Code Numeric to Text Dialog Window. Once you have converted your numeric values to either text or discrete numerical values, use the Stat$>$Tables$>$Tally command to create your frequency table. When creating the frequency table, click in the Display Counts, Percents, Cumulative Counts, and Cumulative Percents boxes (change \texttt{false} to \texttt{true} for each one) in the Tally Dialog Window.

Figure 1.51: Code Numeric to Text Dialog Window for `Framingh.MTW` Lab
3. Based on the frequency table from part 2, what percent of the subjects have cholesterol values that fall in the interval \([200, 250)\)? What percent of the subjects have cholesterol values below 350? How many of the subjects have cholesterol values that fall in the interval \([250, 300)\)?

4. Create a percent histogram based on the same class limits used in part 2. (Remember to select \textbf{Percent} and to specify the class limits as 150:400/50 in the \textit{Histogram Options Dialog Window}. Title your histogram “Percent Histogram” and place a right justified footnote with your name on the first line and your class and section on the second line. Experiment changing both the color and the size of your fonts.

5. Superimpose a LOWESS curve on the histogram you created in part 4. Change the title from “Percent Histogram” to “Percent Histogram with LOWESS Curve.” Change the thickness of the LOWESS line from the default value of 1 to a line size of 2. (Hint: Click anywhere in the box labeled \textit{LOWESS} of the \textit{Histogram Dialog Window} then click on the \textbf{Edit Attributes Button}. Change the \textit{Line Size} from 1 to 2 in the \textit{LOWESS Dialog Window}.) To change the attributes of the histogram, follow the same procedure. That is, click on \textbf{Bar} in the \textit{Histogram Dialog Window} then click on \textbf{Edit Attributes}. Change the fill type to solid and the back color to green.

6. Create a cumulative percent histogram based on the same limits used in part 2. Title your cumulative histogram “Cumulative Percent Histogram” and place a right justified footnote with your name on the first line and your class and section on the second line. Do not paste your result to the report just yet. Note that there are only three values displayed along the vertical axis. The limited values make reading the cumulative histogram challenging. Change the default tick values on the horizontal axis by selecting \textbf{Frame}\textgreater\textbf{Tick} in the \textit{Histogram Dialog Window}. When the \textit{Tick Dialog Window} opens, change the \textit{Number of Major} for the Y-axis from \textbf{Auto} to 5, and change the \textit{Number of Minor} from \textbf{Auto} to 4. Change the \textit{fill type} to \textbf{solid} and the \textit{back color} to \textbf{blue}. Now, append this histogram to the \textit{Report Window}.

7. Superimpose a LOWESS curve on the histogram you created in part 6. Change the title from “Cumulative Percent Histogram” to “Cumulative Histogram with LOWESS Curve.” Use a LOWESS \textit{line size} of 2.

8. If you were interested in showing the distribution of cholesterol values for a report, would you prefer to use the frequency table from part 2 or a combination of the graphs from parts 5 and 7. Explain your choice.
Lab 1.4 — Stem-and-Leaf and Histograms Lab # 2

Objectives:

I. To learn different ways to display data
II. To evaluate how well different graphs and charts display data

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

The following question all deal with data that comes with your book.

Introduction:

The MINITAB™ worksheet CENSUS.MTW contains data relating to the problem of undercount in the 1980 census. The data set contains measurements of percent minority, crime rate, percent living in poverty, and the estimated 1980 census undercount for 66 statistical areas. (Problem 1.63 in your book.)

Questions and Directions:

1. Construct a stem-and-leaf plot of minority using an increment value of 10. Based on the shape of your stem-and-leaf plot, how would you characterize the shape of the minority percents’ values?

2. Use MINITAB™ to create a frequency table displaying information identical to the information you see in the stem-and-leaf plot from part 1. In other words, we want a frequency table that has class limits of [0,10), [10,20), [20,30), [30,40), [40,50), [50,60), and [60,70). Hint: One method is to convert the numerical data that fall inside the class limits to different discrete values which can be either text or numeric values using the Manip>Code>(Numeric to Numeric or Text to Text) command. See Figure 1.51 on page 47 for an idea of how to fill in the Code Numeric to Text Dialog Window. Note: the values in Figure 1.51 are for Lab 1.3. Once you have converted your numeric values to either text or discrete numerical values, use the Stat>Tables>Tally command to create your frequency table. When creating the frequency table, click in the Display Counts, Percents, Cumulative Counts, and Cumulative Percents boxes (change to for each one) in the Tally Dialog Window.

3. Based on the frequency table from part 2, what percent of the statistical areas have a minority percentage in [20,30)? What percent of the statistical areas have a minority percentage below 30%? How many of the statistical areas have a minority percentage that falls in the interval [40,50)?

4. Create a percent histogram based on the same class limits used in part 2. (Remember to select Percent and to specify the class limits as 0:80/10 in the Histogram Options Dialog Window.) Title your histogram “Percent Histogram” and place a right justified footnote with your name on the first line and your class and section on the second line. Experiment changing both the color and the size of your fonts.

5. Superimpose a LOWESS curve on the histogram you created in part 4. Change the title from “Percent Histogram” to “Percent Histogram with LOWESS Curve.” Change the thickness of the LOWESS line from the default value of 1 to a line size of 2. (Hint: Click anywhere in the box labeled LOWESS of the Histogram Dialog Window then click on the Edit Attributes Button. Change the Line Size from 1 to 2 in the LOWESS Dialog Window.) To change the attributes of the histogram, follow the same procedure. That is, click on Bar in the Histogram Dialog Window then click on Edit Attributes. Change the fill type to solid and the back color to green.

6. Create a cumulative percent histogram based on the same limits used in part 2. Title your cumulative histogram “Cumulative Percent Histogram” and place a right justified footnote with your name on the first line and your class and section on the second line. Do not paste your result to the report just yet. Note that their are only three values displayed along the vertical axis. The limited values make reading the cumulative histogram challenging. Change the default tick values on the horizontal axis by selecting Frame>Tick in the Histogram Dialog Window. When the Tick Dialog Window opens, change the Number of Major for the Y-axis from Auto to 5, and change the Number of Minor from Auto to 4. Change the fill type to solid and the back color to blue. Now, append this histogram to the Report Window.
7. Superimpose a LOWESS curve on the histogram you created in part 6. Change the title from “Cumulative Percent Histogram” to “Cumulative Histogram with LOWESS Curve.” Use a LOWESS line size of 2.

8. If you were interested in showing the distribution of the percent of minorities for a report, would you prefer to use the frequency table from part 2 or a combination of the graphs from parts 5 and 7. Explain your choice.
Lab 1.5 — Distribution Shape Detection Using the Boxplot

Objective:

This lab is designed to help you detect different shape distributions using the boxplot.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Directions:

Generate 1,000 rows for eight columns (C1-C8) of standard normal data (mean = 0, standard deviation = 1). Click on Calc>Random Data>Normal. When the Normal Distribution Dialog Box appears, fill in all of the empty boxes to resemble Figure 1.52. When your Normal Distribution Dialog Box looks identical to Figure 1.52, click OK. At this point, you have created eight separate columns each filled with 1,000 values randomly selected from a standard normal distribution. To examine all eight columns of information with boxplots, we will first rearrange the information. Click Manip>Stack>Stack Columns. When your Stack Columns Dialog Window resembles Figure StackW2, click OK. All of your randomly generated standard normal data is now stacked in one column and stored in a new worksheet. Column 2 (C2) contains subscripts corresponding to the original values. In other words, all of the values that were originally in column 1, column 2, ..., column 8, (C1, C2, ..., C8) now have a C1, C2, ..., C8 in column 2 (C2) the subscript column.

Produce side-by-side boxplots of the eight randomly generated standard normal samples each of size 1,000. To produce side-by-side boxplots, click Graph>Boxplot and then select C2 as your measurement variable, and Subscripts as the categorical variable. Click on the Options... button and then click in the box to transpose X and Y (change X to Y). Make sure you title your graph (Annotation>Title). An example of what your graph should resemble appears in Figure 1.54 on the following page. For the remainder of the lab, you should follow the same steps you used to produce your first boxplot with the standard normal data. However, the data you will be using will not be normally distributed. After you create the side-by-side boxplots and before starting on the next set of boxplots, you should create a new worksheet. To create a new worksheet, click File>New>MINITAB Worksheet>OK.

Generate eight columns each containing 1,000 rows of random data from a t distribution with 6 degrees of freedom. (Calc>Random Data>t) Produce side-by-side boxplots for your data. Be sure to title the graph.
Generate eight columns each containing 1,000 rows of random data from a uniform (Lower endpoint = 0, Upper endpoint = 10) distribution. (Calc>Random Data>Uniform) Produce side-by-side boxplots for your data. Be sure to title the graph.

Generate eight columns each containing 1,000 rows of random data from an exponential distribution with a mean of 10. (Calc>Random Data>Exponential) Produce side-by-side boxplots for your data. Be sure to title the graph.

Questions:

1. How many values out of the 1,000 for the standard normal data did you expect to be outliers?
2. What was the average number of outliers for your randomly generated standard normal data?
3. Characterize the shape for each of the four distributions.
4. If you take a random sample of size 100 and the resulting distribution is symmetric with one outlier on each end, would you consider the distribution long tailed? How many outliers should you expect from a normal distribution with with n = 100?
Extra Credit: Use the Graph>Layout with a width of 8.5 and height of 11. Individually change the Figure location (Regions>Figure) when you create each of your four graphs to place them on the same page. Hint: Use min and max X values of .02, .48, .52, and .98. Use min and max Y values of .02, .48, .52, and .98.
Lab 1.6 — Summarizing Data With Statistics #1

Objective:

This lab is designed to give you practice in calculating values of various statistics.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

The cholesterol levels of 62 subjects in the Framingham Heart Study is stored in the worksheet Framingh.MTW. This data set is also used in problems 1.62 and 1.87 of your book.

Questions and Directions:

1. Use the MINITAB™ command Display Descriptive Statistics (Stat>Basic Statistics>Display Descriptive Statistics) to calculate and report the mean, the median, and the 5% trimmed mean for the cholesterol observations stored in Framingh.MTW.

2. Why is the mean larger than both the trimmed mean and the median?

3. Among the estimators mean, median, and trimmed mean, which one is most representative of the center of the cholesterol observations?

4. Report the range, IQR, standard deviation, and variance for the cholesterol observations.

5. If 50 is subtracted from every cholesterol observation, what will be the new mean, median, and trimmed mean?

6. If 50 is subtracted form every cholesterol observation, what will be the new range, IQR, standard deviation, and variance for the cholesterol observations?

7. In general, adding or subtracting a constant to every observation has what effect on measures of center? What effect does it have on measures of variability?

8. Convert each of the cholesterol observations to a z-score using the MINITAB™ command Standardize (Calc>Standardize) and store the results in a column named Znorm. What are the mean and standard deviation for the z-scores stored in Znorm?

9. How many z-scores have an absolute value greater than 3?

10. What percent of z-scores have an absolute value greater than 2? Is this what you might expect according to the empirical rule?

11. Use the MINITAB™ calculator to calculate the Sum of Squares for the cholesterol observations. ( \( \sum (x_i - \bar{x})^2 \) )

12. Why is your answer in number 11. equal to the variance times 61? Note: If your answer in number 11. is not equal to the variance times 61, you have made a mistake.

13. Use the MINITAB™ calculator to calculate the ceiling of the absolute value for all of the standardized values you created in number 8 and store the results in ZN. Hint: If you stored your standardized values from number 8 in a column named Znorm, you should type CEIL(ABSO('Znorm')) in the Expression Box of the calculator. Use the MINITAB™ command Tally (Stat>Tables>Tally) and display the counts, percents, cumulative counts, and cumulative percents for the variable ZN.

14. Why did we calculate the ceiling of the absolute value of the values in Znorm in number 13? Hint: Look under the CumPct for 1 and 2.

15. Explain in detail how to calculate a 5/62 ≈ 8% trimmed mean for the cholesterol observations. Follow your own directions and then report your answer.
Lab 1.7 — Summarizing Data With Statistics #2

Objective:

This lab is designed to give you practice in calculating values of various statistics.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

The measurements of four flower parts on 50 specimens of each of three species of irises by R. A. Fisher (one of the founding fathers of statistics) is given in the worksheet IRIS.MTW. The sepal lengths of the species Iris Setosa are given in the first column labeled sepalL1.

Questions and Directions:

1. Use the MINITAB™ command Display Descriptive Statistics (Stat>Basic Statistics>Display Descriptive Statistics) to calculate and report the mean, the median, and the 5% trimmed mean for the sepal lengths of the Iris Setosa observations stored in IRIS.MTW.

2. Why are the mean, trimmed mean and median all similar in value?

3. Among the estimators mean, median, and trimmed mean, which one is most representative of the center for the sepal lengths of the Iris Setosa observations?

4. Report the range, IQR, standard deviation, and variance for the sepal lengths of the Iris Setosa observations.

5. If 50 is subtracted from every sepal length observation, what will be the new mean, median, and trimmed mean?

6. If 50 is subtracted form every sepal length observation, what will be the new range, IQR, standard deviation, and variance?

7. In general, adding or subtracting a constant to every observation has what effect on measures of center? What effect does it have on measures of variability?

8. h. Convert each of the sepal length observations to a z-score using the MINITAB™ command Standardize (Calc>Standardize) and store the results in a column named Znorm. What are the mean and standard deviation for the z-scores stored in Znorm?

9. How many z-scores have an absolute value greater than 3?

10. What percent of z-scores have an absolute value greater than 2? Is this what you might expect according to the empirical rule?

11. Use the MINITAB™ calculator to calculate the Sum of Squares for the sepal length observations. \( (\sum(x_i - \bar{x})^2) \)

12. Why is the your answer in number 11. equal to the variance times 49? Note: If your answer in number 11. is not equal to the variance times 49, you have made a mistake.

13. Use the MINITAB™ calculator to calculate the ceiling of the absolute value for all of the standardized values you created in number 8 and store the results in ZN. Hint: If you stored your standardized values from number 8 in a column named Znorm, you should type `CEIL(ABSD('Znorm'))` in the Expression Box of the calculator. Use the MINITAB™ command Tally (Stat>Tables>Tally) and display the counts, percents, cumulative counts, and cumulative percents for the variable ZN.

14. Why did we calculate the ceiling of the absolute value of the values in Znorm in number 13? Hint: Look under the CumPct for 1 and 2.

15. Explain in detail how to calculate a 5/50 = 10% trimmed mean for the sepal length observations. Follow your own directions and then report your answer.
Chapter 2

Summarizing Relationships Between Variables

2.0 Introduction

With computerized data collection, we are seeing an overwhelming amount of data on a diversity of topics from public opinion polls to the typical time required to serve customers using the drive through window at fast food restaurants. Data is often collected on a wide range of variables because the researcher is either new to the area or does not understand how the variables interact with one another. When information is collected on a wide range of variables, the data analyst must look for relationships among the numerous variables. Additionally, the analyst will want to identify observations that do not fit the general pattern of the data and to describe the strength of the relationships within the data. Finally, it will be necessary to assess how well the proposed relationships fit the data at hand and to provide an interpretation of the final analysis.

2.1 Scatterplots

When exploring multivariate data, it is possible to use a Matrix Plot to examine all possible 2-way relationships where order matters. What we are really saying is that a Matrix Plot produces all permutations of your $n$ variables taken 2 at a time. Consequently, if you have 6 variables, the Matrix plot will produce $6P_2 = 30$ individual scatterplots. When we are simply looking for relationships between variables, we will only consider half of the scatterplots in a Matrix Plot. The upper right triangle of scatterplots in a Matrix Plot provides identical information to the lower left triangle of scatterplots when exploring the data for relationships. Why is the last sentence true? If you can not figure out why, ask your instructor to explain why the upper right triangle of scatterplots provides identical information to the lower left triangle of scatterplots.

Consider the MINITAB\textsuperscript{TM} worksheet Bears.MTW located in the MTBWIN\textbackslash DATA folder. The information in Bears.MTW was collected when wild bears were anesthetized, and their bodies were measured and weighed. One goal of the study was to make a table (or perhaps a set of tables) for hunters, so they could estimate the weight of a bear based on other measurements. This would be used because in the forest it is easier to measure the length of a bear, for example, than it is to weigh the bear. To read about the variables in Bears.MTW select Help\textgreater Search Help\textgreater Index, then either type Bears.MTW in the top empty box or use the scroll bar on the left to scroll down to and then click on Bears.MTW as shown in Figure 2.1 on the next page. The information you should see is shown in Figure 2.2 on the following page. Before we start our analysis, remember that it is always a good habit to change the numbers for any categorical variables to text values. Specifically, change the numerical values in Sex from 1 and 2 to male and female respectively. (Manip\textgreater Code\textgreater Numeric to Text)

Recall that we are looking for a way to estimate a bear’s weight. Since the age of a bear would not be known to a hunter, we will not consider age in our analysis. The month the measurement was taken will most likely influence the various body measurements since bears hibernate in the winter. It also seems reasonable to assume the gender of the bear might influence the various measurements in the data set. However, for starters, let us consider only the numerical variables (Head.L, Head.W, Neck.G, Length, and Chest.G) and how they relate to Weight.

To produce a Matrix Plot select Graph\textgreater Matrix Plot. When the Matrix Plot Dialog Box opens, type Weight - "Head.L" in the Graph Variables Box. This will select the variables Weight, Chest.G, Length, Neck.G, Head.W, and Head.L to graph. Click OK, and your graph should resemble Figure 2.3 on page 58. Since we are
2.1. Scatterplots

Figure 2.1: Searching for Help on the Bears.MTW Dataset

Figure 2.2: MINITAB™’s Help for the Bears.MTW Dataset

looking for relationships to help explain the weight of the bear, we can restrict our search to either the top row or the first column. If you use the top row, note that the first scatterplot shows the relationship between Weight and Chest.G, while the second scatterplot in the first row shows the relationship between Weight and Length. Since the only graphs of interest at the moment are in the top row, we will use another graphical tool called a Draftsman Plot which will allow us to display only the scatterplots of interest. To produce the first row of scatterplots from Figure 2.3 on the following page, choose Graph>Draftsman Plot, and fill in the boxes as shown in Figure 2.4 on the next page. The resulting Draftsman Plot from the previous commands is shown in Figure 2.5 on page 59.

Based on the scatterplots in Figure 2.5 on page 59, we conclude that the relationship between Chest.G and Weight appears to be the strongest among the variables Chest.G, Length, Neck.G, Head.W, and Head.L with Weight. Based
2.1 Scatterplots

Figure 2.3: Matrix Plot of BEARS.MTW Numerical Variables

Figure 2.4: Draftsman Plot Dialog Window

on the scatterplots in Figure 2.5, we could also say that the strength of relationship between Weight and the other variables is ordered as follows: Chest.G, Neck.G, Length, Head.L, and Head.W. That is, Chest.G is most related to Weight and Head.W is the least related to Weight among the five variables. At this point, we decide to focus on the Weight versus Chest.G scatterplot. To produce a single scatterplot, select Graph>Plot. When the Plot Dialog Box opens, enter Weight as the Y variable and Chest.G as the X variable as shown in Figure 2.6 on the following page. The resulting scatterplot with title and footnote is shown in Figure 2.7 on page 60. It seems reasonable to say that as Chest.G increases, so does the Weight of a bear. We would also like to examine the data for relationships between weight and bear gender, as well as relationships between chest girth and bear gender. An easy way to examine the relationships between weight and bear gender as well as the relationship between chest girth and bear gender is to label the points in Figure 2.9 on page 61 according to gender. To label the points in Figure 2.7 on page 60 according to gender, choose Graph>Plot. When the Plot Dialog Box is open, proceed as before by selecting Weight as the Y variable and Chest.G as the X variable. Before clicking OK, click on the drop down menu For each and select
2.1. Scatterplots

Figure 2.5: Draftsman Plot of Bears.MTW Numerical Variables

![Draftsman Plot of Bears.MTW Numerical Variables](image)

Figure 2.6: Scatterplot Dialog Window for Weight versus Chest.G

![Scatterplot Dialog Window](image)

**Group.** Next, click inside the **Group Variables Box** inside the data display and select **Sex.** When your **Plot Dialog Box** resembles Figure 2.8 on the following page click **OK.** Figure 2.9 on page 61 shows the relationships between weight and bear gender, chest girth and gender, as well as the relationship between weight and chest gender. Although female bears tend to have smaller chest girths and weigh less than male bears, there does not appear to be any clear indication that the underlying relationship between **Weight** and **Chest.G** varies with gender.

### 2.1.1 Identifying Outliers

A bivariate outlier is a point in a scatterplot that does not follow the same pattern as the majority of the points in the scatterplot. MINITAB™ will allow you to brush selected points and see how they relate to other points in your
2.1. Scatterplots

Figure 2.7: Scatterplot of Weight versus Chest.G

![Weight versus Chest Girth](image)

Data: bears.mtw

Figure 2.8: Scatterplot Dialog Box of Weight versus Chest.G with Groups

![Scatterplot Dialog Box](image)

data set. This feature can be particularly helpful in performing visual quality checks of multivariate data since the brushed values are highlighted in all of your graphs. To brush a single point or a group of points, click on the brush icon or select Editor > Brush. Next, we use the brush feature to explore bivariate outliers in Figure 2.5 on the preceding page. First we recreate Figure 2.5 and decide that we would also like to display gender in the scatterplots and consequently choose Sex as the grouping variable. In the first scatterplot of Weight versus Head.L, we decide to brush a bivariate outlier located at (18.5, 204) which appears in the far right of the plot and approximately 1/3 of the distance up the Y-axis. The brushed point is depicted in green in all five of the scatterplots (Observation 77 in the Bears.MTW data set). However, the brushed point appears as a bivariate outlier in only the first scatterplot. This immediately draws our attention. If an observation is not a mistake, we would expect the brushed value also to appear as an outlier in the other graphs. The brushed value (Observation 77) corresponds to the bear named Bill. Bill has been measured on two separate occasions, so we look at the head length measurement from the other occasion. The second head length measurement is 15 inches and was taken one month after the first measurement.
of 18.5 inches. Since it is not very likely Bill’s head would shrink 3.5 inches in the span of one month, we become very suspicious of the original 18.5 inch measurement for Bill’s head length. Most likely, the head length measurement for Bill of 18.5 inches is an error in either transcription or in the original measurement. Figure 2.10 on the next page depicts the five scatterplots and the selected (Observation 77) brushed value. Video 2.1 illustrates creating scatterplots for both original and log-transformed data.

Video 2.1: For Problem 2.15 — (Duration: 5 minutes 37 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 x 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

2.2 Correlation

The correlation coefficient, $r$, can be calculated with MINITAB with the command Stat>Basic Statistics>Correlation as in Figure 2.11 on the next page.

2.2.1 Properties of the Correlation Coefficient

Pearson’s product moment correlation coefficient, $r$, measures the direction and strength of the linear relationship between two numerical variables. The value for $r$ will always be between $-1$ and $+1$. The closer $r$ is to either $-1$ or $+1$ the stronger the linear relationship between the two variables. A positive $r$ value indicates a positive linear relationship. A negative $r$ value indicates a negative linear relationship. If you are calculating the correlation
2.2. Correlation

Figure 2.10: Five Bears Scatterplots with Brushed Value

Figure 2.11: Stat>Basic Statistics>Correlation Menu

In Section 2.1, we used a scatterplot to explore the relationship between a bear’s chest girth and the bear’s weight. We now revisit the Bears.MTW data set and calculate the sample correlation coefficient between the variables Weight and Chest.G. To calculate the sample correlation coefficient, \( r \), with MINITAB™, choose Stat>Basic Statistics>Correlation. Select the variables of interest as shown in Figure 2.12 on the following page then click OK. The results after clicking OK are shown in Output 2.1 on the next page. It is also possible to calculate the correlations among several variables. The results in Output 2.2 on the following page show the correlations among all of the numerical variables in the Bears.MTW data set (Weight, Chest.G, Length, Neck.G, Head.W, and Head.L). Notice that the variables that are most linearly related to Weight are Chest.G (\( r = 0.966 \)), Neck.G (\( r = 0.943 \)), Length (\( r = 0.875 \)), Head.L (\( r = 0.833 \)), and Head.W (\( r = 0.756 \)) respectively. Be careful not to

\[
 r = \frac{\sum \left( \frac{(x - \bar{x})}{s_x} \times \frac{(y - \bar{y})}{s_y} \right)}{n - 1}
\]

(2.1)
2.2. Correlation

Output 2.1: Output from Correlation Calculation with MINITAB™

**Correlations: Weight, Chest.G**

Pearson correlation of Weight and Chest.G = 0.966

P-Value = 0.000

Output 2.2: Output from Multiple Correlation Calculations with MINITAB™


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chest.G</td>
<td>0.966</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>0.875</td>
<td>0.889</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Neck.G</td>
<td>0.943</td>
<td>0.940</td>
<td>0.873</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Head.W</td>
<td>0.756</td>
<td>0.756</td>
<td>0.736</td>
<td>0.805</td>
<td>0.000</td>
</tr>
<tr>
<td>Head.L</td>
<td>0.833</td>
<td>0.854</td>
<td>0.895</td>
<td>0.862</td>
<td>0.744</td>
</tr>
</tbody>
</table>

**Cell Contents: Pearson correlation**

P-Value

Confuse the p-value with the correlation coefficient. The p-values for all of the correlations in Output 2.2 are 0.000s. The p-value that is reported with the correlation is the p-value for testing the null hypothesis that the population correlation coefficient ($\rho$, the Greek letter “rho”) is zero versus the alternative hypothesis which is that $\rho$ is not zero. Video 2.2 on the next page examines a crime data set using scatterplots and numerically assesses the linear
2.3. Least Squares Regression

Simple linear regression is often used to approximate an unknown model where the relationship appears linear. In a more advanced class, specific models and underlying assumptions for the regression model are covered. For now, we will simply be concerned with fitting a straight line to data that appears to follow a linear relationship.

MINITAB\textsuperscript{TM} refers to the dependent and independent variables in a linear relationship as the response variable (\(Y\)) and the predictor variable (\(X\)) respectively. The regression equation is an algebraic representation of the regression line and is used to describe the relationship between the response and predictor variables. The response from a regression equation is what the fitted model (using least squares) predicts for a given value of the predictor variable. The difference between the actual observed response (\(Y\)) and the predicted response (called a fitted value in MINITAB\textsuperscript{TM} and often denoted as \(\hat{y}\) in text books) is called a residual and is denoted \(e_i\). The definition of the \(i^{th}\) residual is written mathematically as 
\[ e_i = y_i - \hat{y}_i. \]

In Section 2.1 we produced a scatterplot of the bears' weights versus their chest girths (Figure 2.7 on page 60), and we also measured the amount of linear relationship between those variables with Pearson’s correlation coefficient in Section 2.2 (\(r = 0.966\)). At this point, we will now decide to use simple linear regression to fit a straight line to the data in Figure 2.7. The most useful command to choose when working with simple linear regression in MINITAB\textsuperscript{TM} is Stat > Regression > Fitted Line Plot. Remember, we are trying to predict the weight of a bear based on the bear’s chest girth. Consequently, the response variable is Weight and the predictor variable is chest girth (Chest.G). Your Fitted Line Plot Window should resemble Figure 2.13. Note that the type of regression model automatically defaults to linear, which is what we want to use in this situation. Click on the OK and a scatterplot with the least squares regression line will appear in your Fitted Line Plot Window (Figure 2.14 on the following page) and the numerical answers associated with the problem will appear in the Session Window (Output 2.3 on the next page). When viewing the graph in Figure 2.14, you cannot visually judge whether or not the line actually has a slope of 12.968 and a y-intercept of \(-278.749\). However, the scatterplot in Figure 2.15 on the following page shows the same data as does Figure 2.14 and includes the point (0, 0) = (\(x, y\)) so that you can more easily estimate whether the computer has given you a reasonable answer. The sum of squares due to error is defined to be 
\[ \text{SSE} = \sum (y_i - \hat{y}_i)^2 \]
and can be found in the Session Window by looking at the intersection of the row labeled Residual Error and the column labeled SS. In Figure 2.15, the sum of squares due to error is reported as

---

2.3 Least Squares Regression

For optimal video viewing, set your computer’s display panel resolution to 1024 x 768 pixels. To verify or change your display panel resolution, select \textit{Start} > \textit{Settings} > \textit{Control Panel} > \textit{Display} > \textit{Settings}.
2.3. Least Squares Regression

Figure 2.14: Fitted Line Plot Graph for Weight versus Chest Girth

![Graph for Weight versus Chest Girth](image)

Output 2.3: Fitted Line Plot’s Numerical Output for Weight versus Chest Girth

Regression Analysis: Weight versus Chest.G

The regression equation is

\[
\text{Weight} = -279 + 13.0 \times \text{Chest.G}
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-278.75</td>
<td>10.88</td>
<td>-25.61</td>
<td>0.000</td>
</tr>
<tr>
<td>Chest.G</td>
<td>12.9680</td>
<td>0.2923</td>
<td>44.36</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[
S = 28.68 \quad R-Sq = 93.3\% \quad R-Sq(adj) = 93.3\%
\]

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1619241</td>
<td>1619241</td>
<td>1968.00</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>141</td>
<td>116012</td>
<td>823</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>142</td>
<td>1735253</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.15: Fitted Line Plot Graph for Weight versus Chest Girth with (0,0) included

![Graph for Weight versus Chest Girth with (0,0) included](image)
2.3. Least Squares Regression

To verify and ensure that we understand the formula and the definition of a residual, we calculate the SSE. First we need to calculate the predicted values ($\hat{y}_i$). Recall that the definition for $\hat{y}_i$ is $\hat{y}_i = b_0 + b_1 \times x_i$. In our problem with the bears data set, we determined $b_0$ to be $-278.749$ and $b_1$ to be $12.968$. Certainly, we do not want to calculate the one hundred forty-three $\hat{y}_i$ values by hand! We will calculate the $\hat{y}_i$ values two different ways. Choose Calc>Calculator and fill in the Calculator Window as shown in Figure 2.16. When you are finished, click OK; and the results will be stored in the next available column in your current worksheet under the name predicted. The second way to calculate the predicted and fitted values is to ask for them directly by clicking on the Storage Button in the Fitted Line Plot Dialog Window and then clicking in the square to the left of Fits (change $\square$ to $\blacksquare$) as shown in Figure 2.17. Notice that one can also calculate the residuals directly from the Fitted Line Plot Dialog Window by clicking in the square to the left of Residuals (change $\square$ to $\blacksquare$). At this point we should have three new columns. The first new column should be labeled predicted, the second new column should be labeled RESI1, and the third new column should be labeled FITS1. Note that RESI1 and FITS1 were created directly from the options we selected in the Fitted Line Plot Dialog Window. To create a column of residuals using the calculator, again we select Calc>Calculator and fill in the Calculator Dialog Window as shown in Figure 2.18 on the next page. Note that we are simply specifying the formula for creating the residuals. In this particular problem Weight contains the observed values and predicted contains the predicted(fitted) values based on the least squares equation. The columns residual and RESI1 should contain identical numbers. Finally, to calculate the sum of squares due to error, we need to square all of the values in either column residual or column RESI1 and subsequently add the squared values. This last step is easily done with the Calculator Window in MINITAB\textsuperscript{TM}. Attempt this last step on your own and verify that the value you get is the same as the value shown for SSE in your Session Window (116,012).

Another way to calculate the equation for the least squares line is to use the command Stat>Regression>Regression. However, this method will not automatically show a graph of the data with the least
2.3. LEAST SQUARES REGRESSION

Figure 2.18: Calculator Window for Residuals Calculation

squares line superimposed over the data. It does, on the other hand, have a few options not available with the Fitted Line Plot. Suppose a hunter bags a bear with a chest girth measurement of 46 inches. What weight does the least squares regression equation predict for the bear? Clearly the answer is weight = \(-278.749 + 12.968 \times 46 = 317.779\) pounds. MINITAB\textsuperscript{TM} will calculate this answer directly when using the Stat>Regression>Regression command if you click on the Options Button in the Regression Dialog Window and type 46 in the box beneath Prediction intervals for new observations as shown in Figure 2.19. Note that the answer is 317.78 and is printed beneath the word Fit in the Session Window Output as shown in Output 2.4 on the next page.

Figure 2.19: Regression–Options Window for Prediction Calculation

Video 2.3 uses the “Prediction intervals for new observations” option to predict the number of icebergs in Newfoundland based on the number in the Grand Banks.

Video 2.3: For Problem 2.46 — (Duration: 4 minutes 13 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 \times 768 pixels.

To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.
2.4. Assessing the Fit of a Line

Output 2.4: Output from Regression Including FIT

**Regression Analysis: Weight versus Chest.G**

The regression equation is

\[
\text{Weight} = -279 + 13.6 \times \text{Chest.G}
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-279</td>
<td>10.80</td>
<td>-25.61</td>
<td>0.000</td>
</tr>
<tr>
<td>Chest.G</td>
<td>13.6</td>
<td>0.2923</td>
<td>46.36</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[S = 28.68 \quad R^2 - \text{Rq} = 93.3\% \quad R^2 - \text{Rq(adj)} = 93.3\%\]

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1619241</td>
<td>1619241</td>
<td>1968.00</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>142</td>
<td>1735253</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>143</td>
<td>1753164</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Predicted Values for New Observations**

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95.0% CI</th>
<th>95.0% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>317.78</td>
<td>3.71</td>
<td>(310.44, 325.11)</td>
<td>(260.60, 374.96)</td>
</tr>
</tbody>
</table>

**Values of Predictors for New Observations**

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Chest.G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46.0</td>
</tr>
</tbody>
</table>

Finding the least squares regression line using Stat > Regression > Regression will also allow you to store the residuals and fits. Actually, the command Stat > Regression > Regression followed by clicking on the storage button will allow the user to store many more items than will the Storage Window in the Fitted Line Plot – Storage Dialog Window. Figure 2.20 shows the Regression – Storage Dialog Window with residuals, coefficients, and fits selected. Notice that there are many more elements one can select in Figure 2.20 than there are in Figure 2.17 on page 66. The additional features in the Regression – Storage Dialog Window are extremely useful for more advanced regression-type problems than those covered in our book.

**Figure 2.20: Regression–Storage Window**

2.4 Assessing the Fit of a Line

Residual plots are used to assess the adequacy of a given model. For the time being, the given model will be a simple linear model. MINITAB™ is capable of finding and storing residuals as well as fitted values when using either the Stat > Regression > Fitted Line Plot or the Stat > Regression > Regression command. Once we have stored the residuals and fitted values, we have several options from which to choose to produce residual plots. One method is to use the Stat > Regression > Residuals Plots command. This command produces four separate graphs. We defer discussion of the four graphs to chapter nine. Another option is to produce a scatterplot (using Graph > Plot) of the stored values placing the residuals on the y-axis and the fitted values on the x-axis. The last technique to produce a residual plot we will cover is to click on the Graphs Box in the Regression Dialog Box (the Regression
2.4. Assessing the Fit of a Line

Dialog Box appears when using the command Stat>Regression>Regression) and clicking in the Residual Versus Fits Box as shown in Figure 2.21. When examining a residual plot, we should see a band around the horizontal line if the model is adequate. Any patterns such as increasing or decreasing variability, or curvilinear relationships in the residual plots indicate some problem with the selected model. The residual plot in Figure 2.22 depicts a prototypical residual plot. The residual plots in Figures 2.23-2.25 display typical problems often seen in regression. Figure 2.23 on the next page shows a curvilinear relationship between the residuals and the fitted values indicating a need for another term (quadratic). Increasing variability often comes hand in hand with increasing size for living organisms. An example of increasing variability is seen in Figure 2.24 on the following page. At times, a residual plot may appear to signal both a need for a variance stabilizing transformation as well as the need for an additional term as illustrated in Figure 2.25 on the next page. Remedies to the various problems shown in Figures 2.22-2.25 are discussed more in depth in chapter nine. For now, we should simply be cognizant that if the residual plots do not follow a prototypical band like the one shown in Figure 2.22 we may have problems with our chosen model.
Figure 2.23: Curvilinear Relationship Between the Residuals and the Fitted Values

Figure 2.24: Increasing Variability in Residuals

Figure 2.25: Increasing Variability in Residuals and a Curvilinear Relationship
2.4. Assessing the Fit of a Line

Video 2.4 and Video 2.5 each illustrate several features of MINITAB™’s regression diagnostic capabilities.

Video 2.4: For Problem 2.59 — (Duration: 4 minutes 27 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

Video 2.5: For Problem 2.97 — (Duration: 5 minutes 2 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

2.4.1 Understanding the Effects of Outliers and Other Influential Observations

Values MINITAB™ considers as outliers or influential observations are marked in the Session Window output from the Stat>Regression>Regression command. Consider the data set NAME.MTP which provides the market value of several brand names and their corresponding revenue. The Session Window output from regressing value on revenue is shown in Output 2.5. Note that MINITAB™ highlights two values (2 and 11) it considers to have large residuals (outliers) and one value it considers to have large influence (1). Two of these values (1 and 2) are enclosed in a square in Figure 2.26 on the following page. The brushing feature was used to create the rectangular box around the points in Figure 2.26. Create a graph similar to Figure 2.26 on your own and use the brushing features of MINITAB™ to point to the various values in the graph until you identify the Nestle brand which is in row 11.

2.4.2 Interpreting the Coefficient of Determination, $r^2$.

The coefficient of determination should be interpreted as the percent of variation in the dependent variable, $y$, that is explained by changes in the independent variable, $x$. An $r^2$ value above 65% (0.65) should be interpreted as a strong relationship between $x$ and $y$. A value at least 25% (0.25) but less than 65% (0.65) should be interpreted as a moderate relationship, while a value of $r^2$ less than 25% (0.25) should be interpreted as a weak relationship between the variables. When using the fitted line plot command with MINITAB™, the coefficient of variation is reported in the Fitted Line Plot Window, as well as in the Session Window as shown in Figure 2.26 on the next page and Output 2.5 respectively.
2.5 Relationships among Categorical Variables

Relationships among categorical variables and how to graph categorical data were covered in section 1.2. Video 2.6 shows how to create comparison bar graphs with crime index information from 1992 and 1999.

Video 2.6: For Problem 2.79 — (Duration: 6 minutes 59 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels. To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.
2.6 Summary and Review Labs

Lab 2.1 — Scatterplots

Objective:

This lab is designed to help you explore relationships among several variables.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad. Use the data set Salary.MTW located in the MTBWIN\Student9 folder for the questions in this lab.

Introduction:

A small private college does a yearly study of its faculty’s salaries. Information on gender, department, years at the school, beginning salary, current salary, and an “experience” variable are recorded. Many other variables are combined to yield an “experience” score. The higher the score, the more experience the person had when they first started. The faculty member’s current rank is also measured. To read about the variables in Salary.MTW data set click Help>Search Help>Index. Then, either type Salary.MTW in the empty box beneath the text “Type the first few letters of the word you’re looking for” or scroll down the list using your mouse until you reach Salary.MTW. Select Salary.MTW then choose Display.

Questions and Directions:

1. Produce a scatterplot showing the current salary of a faculty member (\$1991) versus the starting year of the faculty member (StartYr). Title your scatterplot “Current Salary versus Start Year” and put your name in the bottom right of the graph as a footnote.

2. Provide an explanation of the relationship between current salary and starting year.

3. Use MINITAB™’s brushing capability to help identify any outliers in your plot from problem 1. Specifically, do you think the point (1978, 49200) is an outlier? Is this point possibly a mistake due to transcription or can you come up with a reasonable explanation why this particular faculty started working for this particular small private college making more money than the other faculty members who started in 1978?

4. One of your classmates claims that faculty who started at very low salaries are now making the most money. Do you agree with your classmate’s claim? Use appropriate scatterplots to provide support for your answer.

5. Produce a scatterplot of starting year (StartYr) versus starting salary (Begin $). Use MINITAB™’s brushing feature to help identify and report any outliers in your scatterplot. How would you describe the relationship between starting year and starting salary?
2.6. Summary and Review Labs

Lab 2.2 — Correlation Coefficients

Objective:

This lab is designed to help you calculate the sample correlation coefficient according to the formula and then confirm your results using the MINITAB\textsuperscript{TM} correlation command.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad. The questions in this lab will refer to the Bears.MTW worksheet located in the MTBWIN\DATA folder.

Introduction:

Recall that the definition for Pearson’s product moment correlation coefficient is given as:

\[ r = \frac{\sum \left( \frac{x - \bar{x}}{s_x} \times \frac{y - \bar{y}}{s_y} \right)}{n - 1} \]

One can think of Pearson’s correlation as the summed product of two Z-scores divided by the sample size minus one. One can consider the quantity \( \frac{x - \bar{x}}{s_x} \) as a standardized X-score (\( Z_x \)) and the quantity \( \frac{y - \bar{y}}{s_y} \) as a standardized Y-score (\( Z_y \)). In other words, we will have \( n \) standardized X-scores and \( n \) standardized Y-scores. In this lab, you will use MINITAB\textsuperscript{TM}’s standardize function to create the \( n \) standardized X-scores and \( n \) standardized Y-scores. Next, we will use MINITAB\textsuperscript{TM}’s calculator to sum the product of the \( n \) standardized X and Y values. Finally, we will take the sum of the product of the \( n \) standardized values and divide the result by \( n - 1 \) using MINITAB\textsuperscript{TM}’s calculator.

The scatterplot in Figure 2.27 depicts the standardized weight values versus the standardized chest girth values for the bears in worksheet Bears.MTW. Figure 2.27 has split the output into four quadrants labeled Q1, Q2, Q3, and Q4. The products of all \( Z_x \) and \( Z_y \) values located in Q1 and Q3 are positive. Note that the products of all \( Z_x \) and \( Z_y \) values located in Q2 and Q4 are negative.

Figure 2.27: Scatterplot by Quadrants
Questions and Directions:

Complete the following steps to create the data used in creating Figure 2.27.

Step 1: Standardize the Weight values by using MINITAB\textsuperscript{TM}'s standardize function. (Calc>Standardize) Be sure the default value, which is to subtract the mean and divide by the standard deviation, is selected and store your result in Zy. The Standardize Dialog Window with the options and appropriate variables per the instructions is shown in Figure 2.28.

Step 2: Repeat step 1 for the variable Chest.G and store the results in Zx.

Step 3: Sum the product of Zy times Zx using MINITAB\textsuperscript{TM}'s calculator. See Figure CalcW1 for an example.
Step 4: Divide the result from step 3 by \( n - 1 \). Perform this operation using MINITAB’s calculator and store the result in a column named \( r \). Report the answer in column \( r \) in your report pad.

Step 5: Use the MINITAB command Correlation (Stat>Basic Statistics>Correlation) to calculate the correlation between Chest.G and Weight. Append the result to your report pad.

Step 6: Are your results from steps 4 and 5 identical. Explain why your results either should or should not be identical for steps 4 and 5.

Extra Credit — Create a scatterplot identical to the scatterplot depicted in Figure 2.27.
Lab 2.3 — Least Squares Regression #1

Objective:

This lab is designed to reinforce the least squares regression concepts covered in section 2.3.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Use the data set Submarin.MTW that comes with your book to answer the questions in this lab (Submarin.MTW is also used for problems 2.19 and 2.51 in Basic Statistics and Data Analysis). The data set Submarin.MTW depicts the number of German submarines sunk each month by the U.S. Navy during World War II. The number sunk by the U.S. Navy does not agree with the number the U.S. Navy actually reported.

Questions and Directions:

1. Produce a scatterplot showing the actual (actual) number of submarines sunk versus the number reported (reported) by the U.S. Navy. Title your scatterplot “Actual versus Reported sunk Submarines” and put your name at the bottom right as a footnote.

2. Explain the type of relationship you see in the scatterplot created in part 1.

3. Determine the amount of linear relationship between the variables actual and reported and append your results to the report pad. Be sure to include an interpretation of the numerical answer you report.

4. Create a fitted line plot relating the actual count to the reported count. Title your fitted line plot “Actual versus Reported sunk Submarines” by clicking on the Options Box inside of the Fitted Line Plot Dialog Window and typing the appropriate verbiage in the Title Box.

5. Does a straight line seem to describe the relationship between the actual and reported number of submarines sunk adequately?

6. Have MINITAB™ store the residuals created from regressing actual on reported number of submarines. Use the calculator in MINITAB™ (Calc>Calculator) to sum the squared residuals. Report your answer and explain what the number you calculated represents.

7. How many submarines does your regression line predict were actually destroyed if 10 were reported destroyed? Do you think your answer is reasonable?
2.6. Summary and Review Labs

Lab 2.4 — Least Squares Regression #2

Objective:

This lab is designed to reinforce the least squares regression concepts covered in section 2.3.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Use the data set SOPHOMOR.MTW that comes with your book to answer the questions in this lab. SOPHOMOR.MTW is used in Problem 2.48 of Basic Statistics and Data Analysis. The data set SOPHOMOR.MTW lists the grade point averages (GPA), SAT math scores (SAT), and the final exam grades in college algebra (Final Exam) for a group of 20 sophomores.

Questions and Directions:

1. Produce a matrix plot showing the grade point averages (GPA), SAT math scores (SAT), and the final exam grades in college algebra (Final Exam). Title your matrix plot “Matrix plot for Lab Least Squares Regression #2” and put your name at the bottom right as a footnote.

2. Explain the types of relationships you see in the matrix plot created in part 1.

3. Determine the amount of linear relationship between the variables GPA, SAT, and Final Exam. Append your results to the report pad. Be sure to include an interpretation of the numerical answer you report.

4. Create a fitted line plot using the variable that has the greatest amount of linear relationship with Final Exam. Provide an appropriate title for your fitted line plot by clicking on the Options Box inside of the Fitted Line Plot Dialog Window and typing the appropriate verbiage in the Title Box.

5. Have MINITAB™ store the residuals created from the least squares regression used in part 4. Use the calculator in MINITAB™ (Calc>Calculator) to sum the squared residuals. Report your answer and explain what the number you calculated represents.

6. What Final Exam score does your regression line predict for a student with a GPA of 3.5? Do you think your answer is reasonable?
Lab 2.5 — A Simple Analysis

Objective:

To complete a simple analysis of a dataset.

Directions:

Complete the MINITAB™ tutorial, Session Two: Doing a Simple Analysis, per the directions in the tutorial with exception of exiting the tutorial after step 10. To access the tutorial click on Help>Tutorials>Session Two: Doing a Simple Analysis. It will be helpful to resize your screen so you can read the tutorial in one side of the screen and have MINITAB™ available on the other side. Use the restore button on both the tutorial and MINITAB™ and subsequently resize both the tutorial and MINITAB™ so neither overlaps the other. An example is given in Figure 2.30.

Figure 2.30: Resized, Split Screen Example

When you reach step 10 of the tutorial, save your MINITAB™ project as per the directions. However, do not exit MINITAB™ at this point. Click on the project manager icon, and subsequently click on the Session Window folder. Your screen should resemble Figure 2.31 on the next page.

Highlight everything from the first Regression Analysis: Weight versus D2H to the last Regression Analysis: Weight versus D2H. This can be accomplished by clicking once on the first Regression Analysis: Weight versus D2H then holding down on the Shift key and clicking on the last Regression Analysis: Weight versus D2H. Once everything is highlighted, right click with your mouse and append the highlighted portions to the report pad. Change the automatically generated title of MINITAB™ Project Report in the report pad to Chapter 2 Section 2.4 Lab Report. Include beneath the revised title your name, class, and date. An example is given in Figure 2.32 on the following page of what your title should resemble.

Answer the following questions beneath the fitted line plot (the last thing you have appended in your report pad). When you finish answering all of the questions, save your project and print out a copy of the contents of your report pad to turn in to your instructor.
Questions:

1. Did your coefficient of determination change when you corrected the Weight value in row 15 from 0.07 to 0.7? Be sure to provide numerical answers along with complete sentences to answer the question.

2. Do the residuals exhibit any patterns you might want to investigate further?

3. Are there any outliers or influential observations in the final model?

4. Predict the Weight of a tree with a D2H (diameter squared times height) value of 60. Be sure to include the appropriate units for weight in your answer.

5. Would the same regression equation you used in question 4 yield a reliable Weight value for a tree with a D2H value of 200?
Chapter 3

Probability and Probability Distributions

3.0 Introduction

While statistics deals with data and discovering the underlying patterns of that data, probability asks questions about what occurs when we assume an underlying pattern is true. Knowing a few basic probability concepts and how to use simulation will allow you to understand where much of the statistics you learn is based. This chapter will also examine two distributions in detail that are found often in real world problems: the binomial distribution and the normal distribution. You will learn how to calculate probabilities from these distributions which will help you in discovering p-values in the later chapters on hypothesis testing.

3.1 Basic Probability Concepts

A good way to learn terms from probability is to make note cards with the word you want to learn on the front and the definition and any relevant formulas on the back. Be sure to use a different color from that you used in chapter 1.

There are three main probability laws that you must know to work problems in introductory statistics. They are the law of complement, the addition law, and the law of multiplication for independent events.

The law of complement says that the probability of the complement of an event is equal to the difference of the probability of the original event and one:

\[ P(A^c) = 1 - P(A). \]

The addition law says that the probability of the union of two events is equal to the sum of the probabilities of the individual events minus the probability of the intersection of those events:

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B). \]

The law of multiplication for independent events says that for independent events ONLY the probability of the intersection of the events is equal to the product of the probabilities of the events:

\[ P(A \cap B) = P(A) \times P(B). \]
3.1.1 Simulating Data

This section provides an alternative approach to solving probability problems using simulation. While simulation is not a replacement for the methods of classical probability, it is extremely useful as a teaching and learning tool. There will also be times when it is much quicker to get a “crude” answer using simulation than to get the exact answer using the tools of mathematical probability. MINITAB will automatically generate random data from nineteen named distributions in addition to allowing the user to resample from in any given column (data set). For now, we will concentrate on the discrete uniform distribution. When simulating or generating data from a discrete uniform distribution, we are generating integer values such as 1, 2, 3, etc. in such a way that all the values in the range we specify have an equal chance of occurring.

Consider flipping a fair coin 50 times. How many of the flips do you expect to come up heads? Next consider flipping two fair coins 100 times. How many times do you expect the outcome \{Head, Head\} to occur? If you are thinking the answer to both of the last questions should be around 25 you are doing great. In the first example, there are only two possible outcomes in the sample space: \{Head\}, and \{Tail\}. Provided we have a fair coin, we would expect half of the flips (25) to appear as Heads and the other half to appear as Tails. In the second scenario, we have four possible outcomes in the sample space: \{Head, Head\}, \{Head, Tail\}, \{Tail, Head\}, and \{Tail, Tail\}. Each of the four possible outcomes is equally likely. Consequently, we expect the outcome \{Head, Head\} to appear about 25 times.

To simulate the last two scenarios, we will use MINITAB’s capability of generating random numbers from a discrete Uniform distribution. Let 0 and 1 represent a Tail and a Head respectively in our simulation. To generate 50 values where 0 is as equally likely to appear as a 1, choose Calc > Random Data > Integer. When the Integer Distribution Dialog Box appears, complete the screen exactly as shown in Figure 3.1. Note that we have typed a 50 in the Generate Rows of Data Box and Trials in the Store in Column, with a Minimum value of 0 and a Maximum value of 1. Once your screen looks identical to Figure 3.1, click the OK. Before counting the number of 0s and 1s, Tails and Heads, we will create a new text column actually containing the words Head or Tail. To convert the column Trial into a text column, select Manip > Code > Numeric to Text and fill in the Code - Numeric to Text Dialog Box as shown in Figure 3.2 on the following page. When you are finished, exit the window by clicking the OK. At this point, we would like to use the tally command (Stat > Tables > Tally) to count the number of heads and tails. Output 3.1 on the next page provides the results of this particular simulation. Keep in mind that you will most likely get a different number of Heads and Tails each time the experiment is performed. To simulate flipping two fair coins 100 times, generate 100 random values (0s and 1s) in two columns using MINITAB’s random Integer distribution. Store the generated values in two columns named coin1 and coin2. Once you have generated the two columns of 0s and 1s, code the 0s and 1s to Tails and Heads using Manip > Code > Numeric to Text. The four possible outcomes from our experiment are easily displayed with a contingency table. To create a contingency table select Stat > Tables > Cross Tabulation. The Cross Tabulation
3.1. Basic Probability Concepts

Figure 3.2: Code Numeric to Text for Head/Tail Simulation

Output 3.1: Output Head/Tail Tally

<table>
<thead>
<tr>
<th>Text</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>20</td>
<td>40.00%</td>
</tr>
<tr>
<td>Tail</td>
<td>30</td>
<td>60.00%</td>
</tr>
<tr>
<td>N</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.3: Cross Tabulation Dialog Window

_Dialog Window_ is shown in Figure 3.3 with the appropriate options selected for this example. The summarized data in the form of a contingency table from the two coin flipping experiment is displayed in Output 3.2 on the next page. Note that _coin1_ resulted in a _Head_ and _coin2_ resulted in a _Head_ 30 times, _coin1_ resulted in a _Head_ while _coin2_ resulted in a _Tail_ 27 times. What percent of the time did _coin1_ result in a _Tail_ and _coin2_ result in a _Head_? Hopefully you arrive at the correct answer of 25% by looking at the bottom left corner of Figure 3.2 on the following page where _coin1 Tail_ and _coin2 Head_ intersect. The probabilities associated with the outcomes of many games and scenarios can be simulated. Video 3.1 on the next page uses simulation to calculate probabilities for the game of roulette while Video 3.2 on the following page simulates the number of requests for assistance a poison control center receives in 20 days. Finally, Video 3.3 on the next page simulates rolling a pair of fair dice.
Output 3.2: *Cross Tabulation* of Two Coin Flipping

<table>
<thead>
<tr>
<th></th>
<th>coin1</th>
<th></th>
<th>coin2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Head</td>
<td>Tail</td>
<td>All</td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>30</td>
<td>27</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30.00</td>
<td>27.00</td>
<td>57.00</td>
<td></td>
</tr>
<tr>
<td>Tail</td>
<td>25</td>
<td>12</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25.00</td>
<td>12.00</td>
<td>37.00</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>55</td>
<td>39</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>55.00</td>
<td>45.00</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

*Cell Contents*:
- Count
- % of Tbl

---

Video 3.1: For **Problem 3.26** — (Duration: 3 minutes 24 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.

To verify or change your display panel resolution, select *Start > Settings > Control Panel > Display > Settings*.

---

Video 3.2: For **Problem 3.51** — (Duration: 3 minutes 4 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.

To verify or change your display panel resolution, select *Start > Settings > Control Panel > Display > Settings*.

---

Video 3.3: For **Problem 3.108** — (Duration: 5 minutes 55 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.

To verify or change your display panel resolution, select *Start > Settings > Control Panel > Display > Settings*.  

---
3.2 The Binomial Distribution

Properties of a Binomial Experiment A binomial experiment possesses the following properties:

1. The experiment consists of $n$ identical trials.
2. The experiment consists of two outcomes. We will call one outcome a success ($S$) and the other outcome a failure ($F$).
3. The probability of success on a single trial is equal to $\pi$ and remains constant from trial to trial. The probability of failure is $1 - \pi$.
4. The trials are independent.
5. The random variable of interest is $X$, the number of observed successes during the $n$ trials.

The binomial probability formula for the number of successes, $x$, is

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

for $x = 0, 1, 2, 3, \ldots n$ and $0 \leq \pi \leq 1$.

3.2.1 Bernoulli and Binomial Random Variables

Example 3.2.1: Consider the problem of calculating the probability a basketball player will score 6 or more baskets in 10 shots from a spot 19.75 feet from the basket. Assume that the basketball player has a 33% success rate on shots from that particular location.

Solution: For practical purposes, we can assume the five properties of the binomial experiment are satisfied. Let $X$ denote the number of shots made. The probability of making the shots is $\pi = 0.33$, while the number of trials is $n = 10$. To find the probability of making 6 or more baskets, we need to find the sum of the individual probabilities of making 6, 7, 8, 9, and 10 baskets. Mathematically this is written

$$P(X \geq 6) = P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$$

$$P(X = 6) = \frac{10!}{6!(10 - 6)!} \times 0.33^6 \times (1 - 0.33)^{(10 - 6)} = 0.0546515$$

$$P(X = 7) = \frac{10!}{7!(10 - 7)!} \times 0.33^7 \times (1 - 0.33)^{(10 - 7)} = 0.0153817$$

$$\vdots$$

$$P(X = 10) = \frac{10!}{10!(10 - 10)!} \times 0.33^{10} \times (1 - 0.33)^{(10 - 10)} = 0.0000153$$

$$P(X \geq 6) = 0.0546515 + 0.0153817 + 0.0028410 + 0.0003110 + 0.0000153 = 0.0732005$$
3.2.2 How to calculate binomial probabilities with MINITAB™

MINITAB™ can calculate the probability distribution, the cumulative distribution and the inverse cumulative distribution of a random variable for several named distributions including the binomial. First, we will consider how to produce the cumulative binomial distribution from the Session Window. Before we can type commands in the Session Window, we need to ensure the MINITAB™ prompt is displayed. To display the MINITAB™ prompt (MTB>) first ensure the active window is the Session Window by clicking anywhere in the Session Window. Second, click on Editor then make sure the Enable Commands is checked. To display a cumulative distribution for a random variable regardless of the distribution the initial command is always:

MTB> cdf;

Once you push enter, a subcommand prompt(SUBC>) appears in your MINITAB™ Session Window. At the subcommand prompt, we need to specify the distribution we would like to work with and all of its parameters. Since the binomial distribution has two parameters (n and π), we type `binomial 10, 0.33.` at the subcommand prompt making sure to end the subcommand with a period. Try this command for yourself. Note that all commands in MINITAB™ can be abbreviated to the first four letters of a command. Consequently, if you desire you can shorten the word `binomial` to simply `bino` in the subcommand. Our Session Window should now contain:

```
MTB> cdf;
SUBC> bino 10, 0.33.
```

Output 3.3 shows the Cumulative Distribution and the Probability Distribution for a binomial random variable with \( n = 10 \) and \( \pi = 0.33 \). When working in the Session Window, you may comment your commands provided you use a # after the command and before your comments. The commands displayed in Output 3.4 on the next page show some of the additional functionality of MINITAB™ as well as providing comments for each of the lines. We can naturally arrive at the same answer by using MINITAB™’s extensive drop down menus. To calculate the probability that \( P(X \geq 6) \) when the random variable \( X \) represents the number of baskets made and
Output 3.4: Additional Functionality of MINITAB™ with a Binomial

can be assumed to follow a binomial distribution with \( n = 10 \) and \( \pi = 0.33 \), choose **Calc>Probability Distributions>Binomial**. When the Binomial Distribution Dialog Window appears, click in the Cumulative probability circle (\( \int \) to \( \exists \)), type 10 for number of trials, 0.33 for the probability of success, click in the input constant circle (\( \int \) to \( \exists \)), and type 5 in the Input Constant Box. If you select **OK** at this point, the answer for \( P(X \leq 5) \) will be displayed in the Session Window. Recall, however, that we are really looking for \( P(X \geq 6) = 1 - P(X \leq 5) \). One option is to store the result for \( P(X \leq 5) \) in an optional storage location such as \( k1 \), then to use the calculator (**Calc>Calculator**) to find the final answer. Figure 3.4 shows the Binomial Distribution Dialog Window while Figure 3.5 on the next page shows the Calculator Dialog Window.

The final result from the calculator is not displayed in the Session Window. Instead, the result is found in the current worksheet in the column labelled Answer. If you want to display the value in the Session Window, at the MINITAB™ prompt type

`MTB>print 'Answer'`

**Note:** MINITAB™ is not case sensitive, so you may type *Answer, answer, AnSwEr* or any combination of upper and lower case letters as long as they spell “answer”.
3.2. The Binomial Distribution

The next problem we will consider is modelled quite often with the binomial distribution. However, we will now consider how to model the problem using simulation. What is the probability that a basketball player will score 6 or more baskets out of 10 shots from a spot 19.75 feet from the basket, if he averages a 33% success rate on shots from that particular location. To model the problem, consider a fair die. Let the outcomes 1 or 2 stand for a successful shot, and the outcomes 3, 4, 5, and 6 for unsuccessful shots. Repeat the process 10 times to model the 10 shots. Now, what we need to do is repeat the above process a large number of times. For now, we will simulate shooting 10 shots from a point 19.75 feet from the basket 1000 times and count the number of times 6 or more shots are made. The simulation will be done by generating 1000 Integer values (1 through 6) in 10 columns. When this is done, we can consider each of the 1000 rows of 10 values to be one of the 1000 experiments where an individual attempted 10 shots from a point 19.75 feet from the basket. Of course what we need to determine for each of the 1000 experiments is the number of the 10 attempted shots that were successful. One way to do this is to recode the values indicating a basket was made (1 and 2) to 1s and the missed shots (3, 4, 5, and 6) to 0s, and then add the coded values (0s and 1s) in each row, storing the result in a new column (ShotsMade) which will indicate the number of the successful shots. To sum values in a row use Calc > Row Statistics. This new column (ShotsMade) will simply provide the number of successful shots in each of our 1000 experiments. An example of the Row Statistics Dialog Box is given in Figure 3.6. To answer the original question, we need to find out how many times 6 or more shots were made in the 1000 experiments. This can be answered by selecting all four of the display options in the Tally Dialog Window created from using the Stat > Tables > Tally command as shown in Figure 3.7 on the next page. Observe Output 3.5 on the following page. Note that we see that of the 1000 experiments 47 times exactly 6 shots were made, 20 times exactly 7 shots were made, 8 times exactly 8 shots were made, and 1 time exactly 9 shots were made.
3.2. The Binomial Distribution

3.2.3 Hints & Tricks of solving Binomial problems

Always specify the random variable by writing a sentence explaining what the random variable represents. This sentence should take the form: The random variable \( X \) represents the number of ______. Next, ensure the random variable \( X \) actually satisfies the conditions for a binomial distribution. Once you have a verbal description of the random variable, write down the probability of success for the random variable (\( \pi \)), as well as the number of trials (\( n \)). Finally, write a mathematical statement for the quantity you are trying to find in terms of your random variable. This mathematical statement usually takes one of the forms \( P(X = x) \), \( P(X \leq x) \), or \( P(X \geq x) \). If the problem is of the form \( P(X \geq x) \), it is usually easier to rewrite the problem using the complement rule \( \{P(X \geq x) = 1 - P(X \leq (x - 1))\} \) and find the answer to the original problem with MINITAB™’s cdf command.

Consider the following problem which is a slight variation from a similar problem in your book. A box contains six marbles, two of which are white. Three are drawn with replacement. What is the probability at least two of the three are white?

First we will let the random variable \( X \) be the number of white marbles drawn. Since we are sampling with replacement, we have \( n = 3 \) identical trials where the probability of getting a white marble is \( \pi = \frac{2}{6} = .3 \) for each
3.2. The Binomial Distribution

The mathematical statement for the probability for which we are looking is $P(X \geq 2)$ which can we found using any of the following statements which are all equivalent

$P(X \geq 2) = P(X=2) + P(X=3) = 1 - [P(X=1) + P(X=0)]$. Commands to calculate the answer all three ways are given in Output 3.6. The three different methods $P(X=2) + P(X=3)$, $1 - P(X \leq 1)$,

Output 3.6: Various ways to solve a Binomial Problem

and $1 - [P(X=1) + P(X=0)]$ to calculate $P(X \geq 2)$ displayed in Output 3.6 display slightly different answers. This difference is not significant and is produced because the probabilities displayed in the *Session Window* are rounded to four decimal places, while the numbers stored in a MINITAB™ location are stored with double precision. This means they may have up to 15 or 16 digits without round-off error. Clearly, if one is concerned with round-off error, the best approach is to store results in MINITAB™ locations. Output 3.7 uses MINITAB™ commands and storage locations so that all answers agree.

Output 3.7: MINITAB™ commands and storage locations so that all answers agree in Binomial calculation
3.3 The Normal Distribution

The empirical rule for a normal distribution says that approximately 68% of the data will fall within one standard deviation of the mean while 95% of the data will fall within two standard deviations of the mean.

3.3.1 Finding Probabilities and Percentiles in a Normal Distribution

In this section we will use Minitab’s cdf (cumulative distribution function) command to solve the typical scenarios encountered when working with normal distributions. These scenarios include finding the area between two quantities that are symmetric about the mean (Figure 3.8), finding the area between two known quantities that are not symmetric about the mean (Figure 3.9 on the following page), finding the area to the right of a known quantity (Figure 3.10 on the following page), finding the value (b) when given an area to the left of b (Figure 3.11 on page 93), finding a value b when given an area to the right of b (Figure 3.12 on page 93), and a problem to tie all of the concepts together (Figure 3.13 on page 94).

Consider the problems based on scores on the Stanford-Binet IQ test which are assumed to be normally distributed with a standardized mean of 100 and a standard deviation of 16. Many authors will simply write IQ ∼ N(100, 16) to convey the information in the previous sentence in a much more compact fashion. The statement IQ ∼ N(100, 16) tells us that the random variable in this case IQ has a normal distribution (N) with a mean of the first entry, in this case 100, and a standard deviation of the second entry, in this example, 16.

The MINITAB commands to calculate the shaded areas or to find the unknown values (b) are given in the upper left hand corners of Figures 3.8–3.13. Probability statements are given in the upper right hand corner of each window indicating the quantity for which we are looking in concise mathematical notation. NOTE: The cdf command will only give areas to the left of a given number.

To find the area between two values we use the fact that MINITAB™ can find the area to the left of a given value. First, we find the area to the left of the greater value in the inequality. Second, we find the area to the left of the smaller value in the inequality. When we take the difference between the area to the left of the larger value and the area to the left of the smaller value, we are left with the area between the two values. Consider the concrete example in Figure 3.8 where we want to find the probability a randomly selected individual scores between 84 and 116 (one standard deviation from the mean in each direction). The same procedure is used for a “between”
3.3. The Normal Distribution

calculation where the values are not symmetric around the mean. For example, to find the probability a randomly selected individual scores between 110 and 130, (Figure 3.9), we find the area to the left of the larger number (130), and from that value, we subtract the area to the left of the smaller number (110). To find the area to the right of a given quantity, (Figure 3.10), we first find the area to the left of the given quantity and then use the Complement Law to find the area to the right of a original value by subtracting the area to the left of the given quantity from 1. To find a particular percentile \( p \) for a distribution, we need to find the value \( b \) such that \( p \) percent of the distribution is to the left of the value \( b \). When we used the \( cdf \) command, we had a value \( b \) and were looking for the area to the left of \( b \). Now, we have the area to the left of \( b \) but need to find the area. In essence, we are doing an inverse process. To this end, we will use the MINITAB\textsuperscript{TM} command \textit{invcdf} which returns a value that has a given amount of area to its left. For example, to find the value \( b \) such that 80\% of the area under the normal distribution is to its left, we use the \textit{invcdf} command as shown in Figure 3.11 on the next page. When a problem asks to find the value \( b \) such that one is in the top 10\% of the distribution, one needs to reword the original problem and realize that a person in the top 10\% of the distribution is identical to saying the person is in the 90\textsuperscript{th} percentile. Figure 3.12 on the following page uses this approach. Figure 3.13 on page 94 requires us to use all that we have learned about probabilities and the normal distribution as well as some of our algebra skills to reach the correct answer. We wish to discover the \( b \) such that the probability that an IQ score is between 120 and \( b \) is 8\%. First, we calculate the area less than 120. Next, we add that value to the area we already know (.08). Finally, we execute an \textit{invcdf} command to discover the value \( b \).
3.3. The Normal Distribution

3.3.2 Normal Probability Plots

Normal probability plots can be used as a tool to assess the shape of the original population when working with a sample. Normal probability plots can be created by choosing either Graph > Probability Plot or Stat > Basic Statistics > Normality Test. Video 3.4 examines the use of both commands and provides guidelines for assessing the normality of the original population. Video 3.5 on the next page simulates data from a normal distribution and

Video 3.4: For Problem 3.66 — (Duration: 4 minutes 31 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels. To verify or change your display panel resolution, select Start > Settings > Control Panel > Display > Settings.

then uses normal probability plots to assess the original population’s normality.
3.3. The Normal Distribution

Figure 3.13: Normal Distribution Inverse CDF Calculation Given a “Between” Area and One End Point

Video 3.5: For Problem 3.112 — (Duration: 2 minutes 58 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.
3.4 Summary and Review Labs

Lab 3.1 — Basic Probability Concepts

Objective:

To use simulation to approximate probabilities and solve problems

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Questions and Directions:

1. Use simulation to approximate the probability of making more than 6 out of 10 free throws if the individual shooting the free throws has a constant probability of \( \frac{2}{3} \) of making any given free throw. Specifically, simulate 10,000 experiments of shooting 10 free throws.

   a. Use the Tally command and append your output to the report pad.
   
   b. Based on the results from the Tally command, create a chart showing the probability of the various outcomes in your 10,000 experiments. An example of a chart for a different distribution is depicted in Figure 3.14.

   c. Based on your chart depicting the probability of making different numbers of free throws, how would you characterize the shape of the distribution for the number of shots made?
   
   d. What are the mean and the median number of shots made from the free throw line?
   
   e. The information in Table 3.1 on the next page provides theoretical values based on the binomial distribution for the probability of making X free throws. Duplicate the information in Table 3.1 in the report pad and fill in the column labelled Empirical values with the answers you get from your simulation.

2. A charity fund raiser is charging participants $3.50 to play a simple game where they either win $10 or they lose $3.50 (the price to play the game). The game consists of throwing two fair dice and taking the sum of the face showing dice. The participant is awarded $10 if the sum of the dice is 3, 5 or 7. Use simulation to approximate the probability of winning $10. The expected payoff is calculated by taking the probability of winning times the take home amount (10 – 3.50 = 6.50) added to the probability of losing times the cost of playing (−$3.50). If the expected pay off is positive, you will make money in the long run by playing this particular game.
Table 3.1: Theoretical Binomial Values

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
<th>Empirical Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000017</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.000339</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.003048</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.016258</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.056902</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.136565</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.227608</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.260123</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.195092</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.086708</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.017342</td>
<td></td>
</tr>
</tbody>
</table>

a. In the long run, how much can you expect to either win or lose each time you play this game?
b. Pretend you have $350 in your right front pocket. You are going to play this particular game 100 times. When you win money, you place your winnings in your left front pocket (which starts out empty). How much money should you expect to have in your left front pocket when the cash in your right front pocket is gone?
c. This particular game takes on average 10 seconds to play. That is, you are able to pay the game attendant $3.50, collect and throw two dice, and receive your winnings when you have any in an average of 10 seconds. How long (minutes) should you expect to play this particular game before you run out of money if you start the game with $350 and play continuously until you run out of money?

Hints: Simulate rolling two fair dice 10,000 times with the Calc > Random Data > Integer command. Use the Calc > Row Statistics command to add the two rolls. Finally, use the Stat > Tables > Tally command to help approximate the odds of winning and losing. (Calculate one of the probabilities and use the complement rule to find the other probability.)
3.4. Summary and Review Labs

Lab 3.2 — The Binomial Distribution

Objectives:

To create and analyze a binomial distribution graph

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Recall that the mean and standard deviation for any random variable $X$ are given by:

$$\mu_x = \sum (X \times P(X = x)) \quad \text{and} \quad \sigma_x = \sqrt{\sum ((X - \mu_x)^2 \times P(X = x))}.$$  

For a random variable $X$ that follows a binomial distribution, it can be shown that $\mu_x = n \times \pi$ and $\sigma_x = \sqrt{n \times \pi \times (1 - \pi)}$.

Questions and Directions:

Verify that the general formulas for finding the mean and standard deviation for any random variable give the same answers as the formulas $\mu_x = n \times \pi$ and $\sigma_x = \sqrt{n \times \pi \times (1 - \pi)}$ provide for a random variable $X \sim Bin(n = 100, \pi = 0.5)$. Start by creating a column of values 0 through 100 and storing the values in a column named $X$. Calculate the $P(X = x)$ where $x$ is the all of the values in the column $X$. Store these values in a column named $PX$.

Use the commands Calc>Probability Distributions>Binomial and be sure to change the default value in Binomial Distribution Dialog Window from Cumulative Probability to Probability. An example of the Binomial Distribution Dialog Window is given in Figure 3.15. If you verify the formulas by typing in the Session Window, copy all of your commands and append them to your report pad. If you use MINITAB™’s calculator, copy the relevant portions of your history window so your instructor can verify your work.

Graph the random variable $X \sim Bin(n = 100, \pi = 0.5)$. Since the binomial distribution is discrete, the graph should reflect the fact that the distribution is positive only at whole numbers less than or equal to $n$. One method to reflect this property is by using a line equal to the height of the probability at the discrete values. To produce a line at each of the discrete values whose height is equal to the probability at each discrete value use the Graph>Plot command in MINITAB™ and change the default Display in the drop down menu from Symbol to Project. Keep in mind that you want to display probability on the Y axis which you stored in a column named $PX$.

Determine $P(\mu_x - 2 \times \sigma_x \leq X \leq \mu_x + 2 \times \sigma_x)$. Are your answer and graph consistent with the Empirical rule?
Lab 3.3 — The Normal Distribution #1

Objectives:

I. To draw distributions with MINITAB™

II. To propose general guidelines about when a binomial distribution can be adequately approximated with a normal distribution

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Questions and Directions:

Create 9 binomial plots using \( n = 100 \) and \( \pi = 0.03, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.97 \). For each of the 9 plots, superimpose on the projected lines a normal distribution with the same mean and standard deviation as each binomial random variable. Title each graph appropriately and include your name and class in a footnote. Based on the graphs in the lab, propose general guidelines relating \( n \) and \( \pi \) about when you can approximate a binomial distribution with a normal distribution.

Hints and Examples:

When you superimpose a normal distribution over the binomial distribution, your normal curve should only cover values the binomial random variable can assume. Remember, the values the binomial random variable can assume in this particular problem are always between 0 and 100 inclusive. Refer to Figure 3.17 on the next page for an example where the density for the normal distribution stops at an \( X \) value of 0. In other cases, you will have the normal density stopping at an \( X \) value of 100. Two of the nine graphs (Figure 3.16, and Figure 3.17 on the next page) you are required to create to answer the questions in the lab are illustrated as examples. It is probably best to ensure you can duplicate Figure 3.16 and Figure 3.17 before attempting to create the other seven graphs in the lab.

Figure 3.16: Example of Normal Overlayed with Binomial

To produce a graph of the normal density function, we will calculate several values of the PDF and use a plot that connects the values of the PDF. The more values of the PDF that are calculated, the smoother the density curve will appear. When drawing normal distributions, we recommend calculating values of the PDF that are within plus and minus 4 standard deviations from the mean.
To produce a graph of a normal distribution with a mean of 50 and a standard deviation of 5, first determine \( \mu \pm 4 \times \sigma \). In this particular case, we will generate \( X \) values for the PDF between 30 and 70. To create 801 equally spaced values between 30 and 70, choose Calc > Make Patterned Data > Simple Set of Numbers. Fill in the Simple Set of Numbers Dialog Window as shown in Figure 3.18. Note that we are storing values from 30 to 70 in increments of 0.05 in a column named \( X \).

Next, determine the corresponding PDF value for each value in column \( X \). To calculate the PDF values choose Calc > Probability Distributions > Normal, and store the PDF values in a column named \( PX \) as illustrated in Figure 3.19 on the next page.

Be sure to change the default setting of Cumulative probability to Probability density. To draw the normal distribution, have MINITAB\textsuperscript{TM} plot the ordered pairs in columns \( X \) and \( PX \) using the Graph > Plot command. Further, change the default value Symbol to Connect in the Display drop down menu of the Plot Dialog Window as shown in Figure 3.20 on the following page.

Many graphs may be displayed in the same window. Suppose we want to draw three normal distributions all with mean 50 and standard deviations of 5, 2, and 1 respectively. First, we create a column with the PDF values for a normal distribution with mean 50 and standard deviation of 2. Store these values in a column named \( PX2 \). Repeat the process of finding the PDF values for the normal distribution with mean 50 and a standard deviation of 1 storing the results in \( PX1 \). To display all three plots at once, click on the Frame drop down menu and select...
Multiple graphs. In the Multiple Graphs Dialog Window, click in the circle to the left of Overlay graphs on the same page. Following this procedure, you should end up with a plot that resembles the graph in Figure 3.21.
Binomial distributions can be depicted in much the same fashion using the Plot command. However, since the binomial distribution is discrete and only has positive probability for nonnegative integers, we will depict the probability of the binomial using a straight line whose height is equal to the PDF value at all the values for which the binomial random variable is defined.

Recall the mean and standard deviation for the Binomial distribution are \( n \times \pi \), and \( \sqrt{n \times \pi \times (1 - \pi)} \) respectively. Suppose we have a random variable say \( X \) which follows a binomial distribution with parameters \( n \) and \( \pi \), which from now on we will write as \( X \sim \text{Bin}(n, \pi) \). Further, suppose the values for \( n \) and \( \pi \) are 100 and 0.5 respectively. That is, \( X \sim \text{Bin}(100, 0.5) \). The mean of the random variable \( X \) is \( \mu_x = n \times \pi = 100 \times 0.5 = 50 \), and the standard deviation of the random variable \( X \) is \( \sigma_x = \sqrt{100 \times 0.5 \times (1 - 0.5)} = 5 \). Suppose we are interested in calculating \( P(40 \leq X \leq 60) \). The answer is calculated by finding \( P(X \leq 60) - P(X \leq 39) \). However, when \( n \) becomes large (greater than 20), calculating the individual binomial probabilities presents a slight obstacle. Why? Consider the magnitude of 20!, which equals 2,432,902,008,176,640,000. The probabilities for values of \( X \) between 40 and 60 inclusive are illustrated with vertical lines that have a small solid circle at the top of the vertical line in Figure 3.22.

If one moves half a unit to the left and right of each solid circle, we can construct rectangles of width 1 and height equal to the probability that the binomial equals the \( X \) value of the vertical line. Note that these rectangles do a fair job of approximating the superimposed normal distribution. Consequently, if we were to approximate the probability \( P(40 \leq X \leq 60) \) using the normal distribution, we would want to find the area between 39.5 and 60.5 of a normal distribution with mean of 50 and standard deviation of 5. This alteration to the original values in order to use a normal approximation to solve a binomial problem is known as a continuity correction. The answer to the question \( P(40 \leq X \leq 60) = 0.9648 \) using the binomial distribution. An approximation to \( P(40 \leq X \leq 60) \) is calculated using the normal distribution by finding \( P(39.5 \leq Y \leq 60.5) = 0.9643 \) where \( Y \sim \mathcal{N}(50, 5) \). Note that the answers are quite close and differ by only 0.0005. If one simply uses the normal distribution to approximate the binomial without applying continuity corrections, we would calculate \( P(40 \leq Y \leq 60) = 0.9545 \) which is not nearly as close to the “correct” answer. See Figure 3.23 on the next page for commands used to calculate the previous answers.

To display both vertical lines for a binomial distribution and to superimpose a normal distribution with mean and standard deviation equal to the mean and standard deviation of the binomial random variable, first determine the values \( \mu_x \pm 4 \times \sigma_x \). Follow the directions in the fourth paragraph of the lab to create the normal curve. At this point we should have a column named \( X \) with 801 equally spaced values between 30 and 70 as well as the 801 corresponding PDF values for column \( X \) stored in a column named \( PX \). Although the binomial random variable can take on values between 0 and 100, the probability for values below and above 30 and 70 respectively is extremely small. Consequently, we will only graph values for the binomial between 30 and 70 inclusive. Create a column of integers from 30 to 70 inclusive and store the results in a column named \( XB \). This can be accomplished using the same MINITAB commands you used earlier to create the 801 values for \( X \) except be sure to make the increment value 1 instead of 0.05. Now, find the PDF values for the values stored in column \( XB \) based on the
3.4. Summary and Review Labs

Figure 3.23: Commands for Example Binomial Calculations

| MTH > cdf 60 k1; | # Finding V(X<=60) and storing result in k1 |
| SUBC> bino 100 .5. | # Using X-Binomial(100, 0.5) |
| MTH > cdf 39 k2; | # Finding V(X<=39) and storing result in k2 |
| SUBC> bino 100 .5. | # Using X-Binomial(100, 0.5) |
| MTH > let k3=k1-k2 | # Finding V(X<=<60) and storing result in k3 |
| MTH > print k3 |

Data Display

\[ k3 \approx 0.964800 \]

| MTH > cdf 60.5 k1; | # Finding V(Y<=60.5) and storing result in k1 |
| SUBC> norm 50 5. | # Using X-Normal(50, 5) |
| MTH > cdf 39.5 k2; | # Finding V(Y<=39.5) and storing result in k2 |
| SUBC> norm 50 5. | # Using X-Normal(50, 5) |
| MTH > let k3=k1-k2 | # Finding V(39.5<=Y<=60.5) and storing result in k3 |
| MTH > print k3 |

Data Display

\[ k3 \approx 0.964271 \]

binomial distribution and store your results in a column named PXB. Go to Graph>Plot and select the inputs for the two graphs you want to display. Be sure to select Connect for the Normal graph and Project for the Binomial graph from the Display drop down menu. See Figure 3.24 for an example.

Figure 3.24: Plot Dialog Window for Normal and Binomial Overlay

Do not hit the return yet! We need to request that both graphs be displayed on the same page. To produce both graphs on the same page, click on the Frame drop down menu and select Multiple Graphs. Click in the circle to the left of Overlay graphs on the same page. Click OK once and you should be back in the Plot Dialog Box. Now comes the slightly tricky part! Click once on the word Connect which should be in the first item’s Display Box. Once the word Connect is highlighted, click on the Edit Attributes Box. In the Connect Dialog Window, change the line type of the first graph to Solid and the line type of the second graph to None. Click OK and you should be back in the Plot Dialog Window. Click once on the word Project. Once the word Project is highlighted, click on the Edit Attributes Box. This time you will be in the Project Dialog Window as opposed to the Connect Dialog Window. Change the line type for the first graph to None and the line type of the second graph to Solid. Click OK once and you will be back in the Plot Dialog Window, click OK again and you should have a graph resembling Figure 3.16 on page 98.
Lab 3.4 — The Normal Distribution #2

Objectives:

I. To compute probabilities from binomial distributions
II. To propose general guidelines about when a binomial distribution can be approximated by a normal distribution

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

This is an extension of the concepts covered in the previous lab. You will compute several probabilities in the lab designed to reinforce the guidelines you developed in the previous lab. Finally, based on the computed probabilities, you will propose general guidelines relating \( n \) to \( \pi \) describing when you can approximate a binomial distribution with a normal distribution.

Questions and Directions:

For each binomial random variable \( X \), compute the indicated probabilities. Use a normal distribution with a mean equal to the mean of the binomial random variable and standard deviation equal to the standard deviation of the binomial random variable with appropriate continuity corrections to calculate the binomial probabilities. Keep in mind that the probabilities computed using the two different methods will only be similar under certain conditions. It is your job to decide based on the probabilities you calculate with the different methods what those general guidelines should be.

Reminders: For the binomial distribution, \( \mu_x = n \times \pi \), \( \sigma_x = \sqrt{n \times \pi \times (1 - \pi)} \). You may want to refer back to Figure 3.22 on page 101 and reread the material in the previous lab to refresh your understanding of continuity corrections.

<table>
<thead>
<tr>
<th>( X \sim \text{Bin}(100, 0.03) )</th>
<th>( X \sim \text{Bin}(100, 0.3) )</th>
<th>( X \sim \text{Bin}(100, 0.9) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a) P(1 \leq X \leq 2) = )</td>
<td>( a) P(22 \leq X \leq 26) = )</td>
<td>( a) P(86 \leq X \leq 88) = )</td>
</tr>
<tr>
<td>( b) P(4 \leq X \leq 5) = )</td>
<td>( b) P(34 \leq X \leq 38) = )</td>
<td>( b) P(92 \leq X \leq 94) = )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X \sim \text{Bin}(100, 0.05) )</th>
<th>( X \sim \text{Bin}(100, 0.5) )</th>
<th>( X \sim \text{Bin}(100, 0.95) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a) P(3 \leq X \leq 4) = )</td>
<td>( a) P(40 \leq X \leq 45) = )</td>
<td>( a) P(93 \leq X \leq 94) = )</td>
</tr>
<tr>
<td>( b) P(6 \leq X \leq 7) = )</td>
<td>( b) P(55 \leq X \leq 60) = )</td>
<td>( b) P(96 \leq X \leq 97) = )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X \sim \text{Bin}(100, 0.1) )</th>
<th>( X \sim \text{Bin}(100, 0.7) )</th>
<th>( X \sim \text{Bin}(100, 0.97) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a) P(6 \leq X \leq 8) = )</td>
<td>( a) P(62 \leq X \leq 66) = )</td>
<td>( a) P(95 \leq X \leq 97) = )</td>
</tr>
<tr>
<td>( b) P(12 \leq X \leq 14) = )</td>
<td>( b) P(74 \leq X \leq 78) = )</td>
<td>( b) P(98 \leq X \leq 99) = )</td>
</tr>
</tbody>
</table>

Output 3.8 on the next page shows output to answer question a) when \( X \sim \text{Bin}(100, 0.5) \).
Output 3.8: Example when $X \sim Bin(100, 0.5)$

```
MTB > cdf 45 k1;    # P(X<=45)
SUBC> binc 100 .5.  # X-Bin(100, 0.5)
MTB > cdf 39 k2;    # P(X<=39)
SUBC> binc 100 .5.  # X-Bin(100, 0.5)
MTB > let k3=k1-k2; # P(40<=X<=45)=P(X<=45)-P(X<=39)
MTB > print k3

Data Display

  k3  0.166501
MTB > cdf 45.5 k1;  # P(Y<=45.5)
SUBC> norm 50 5.     # Y-Norm(50, 5)
MTB > cdf 39.5 k2;  # P(Y<=39.5)
SUBC> norm 50 5.     # Y-Norm(50, 5)
MTB > let k3=k1-k2; # P(39.5<=Y<=45.5)=P(Y<=45.5)-P(Y<=39.5)
MTB > print k3

Data Display

  k3  0.166196
```
Chapter 4
Sampling Distributions

4.0 Introduction

Every statistic is calculated from a sample. The sampling distribution of a statistic is the set of all possible values the statistic could have along with the probability that each of those values occurs. This chapter introduces you to the principles of sampling and discusses the sampling distributions of $\bar{x}$, the sample mean, and $p$, the sample proportion of successes. As you complete the simulations, you will gain an understanding of the Central Limit Theorem and how it produces normal distributions for the sampling distributions of these two common statistics.

4.1 Principals of Sampling

What exactly is the difference between a random sample and a simple random sample? Let us answer the question by first reviewing the definition of a simple random sample. A simple random sample of size $n$ consists of $n$ elements chosen from a population in such a way that all samples of that size have the same chance of being selected. In other words, if we have a finite population of size $N$, each of the $\frac{N!}{n!(N-n)!}$ samples of size $n$ taken without replacement has the same probability of occurring. The key concept to keep in mind with a simple random sample is that we are sampling without replacement. In a random sample of $n$ objects from a finite population of size $N$, the observations of the random sample are independent and identically distributed. In essence, this means that we are sampling with replacement. If we remove the condition of a finite population and assume we are working with an infinite population, then the difference between a simple random sample and a random sample becomes moot.

4.1.1 Generating a Simple Random Sample from a finite population using MINITAB™

To generate a simple random sample of size $n$ from a finite population consisting of $N$ objects, first we create a column of patterned data of length $N$ by selecting Calc>Make Patterned Data>Simple Set of Numbers. Store your patterned data in a column whose name you choose, enter the number 1 in the From First Value Box, and enter $N$ in the To last value Box. Make sure you have the default value of 1 for the In steps of, List each value, and List the whole sequence boxes. Figure 4.1 on the following page shows the Simple Set of Numbers Dialog Window for $N = 1000$. Now that we have the possible numbers to represent our population, we want to take a sample of size $n$ without replacement. Suppose we want a sample of size 50 without replacement from the values stored in column Numbers. To select a sample without replacement choose Calc>Random Data>Sample from Columns. Fill in a 50 in the Sample / / Rows From Column(s) Box. Select the column from which you will sample and a column where you will store the results. Note that Figure 4.2 on the next page stored the results from the simple random sample of size 50 in a column named SRS.
4.2 The Sampling Distribution of $\bar{x}$

There are several ways to generate random samples with MINITAB\textsuperscript{TM}. If the population is finite, we might proceed in the same fashion explained for the simple random sample with the exception of clicking in the Sample With Replacement Box in the Sample From Columns Dialog Window. Another method to generate a random sample is to select Calc> Random Data> Integer. Once in the Integer Distribution Dialog Window, select the number $n$ for your random sample and enter the number in the Generate Rows Of Data Box. Store your random sample in an appropriately named column, and specify the Minimum Value as a 1 and the Maximum Value as $N$. Figure 4.3 on the following page shows the values required to generate a random sample of size 50 from a finite population of size 1,000.

4.2 The Sampling Distribution of $\bar{x}$

The sampling distribution of a statistic is the probability distribution associated with the values that the statistic can assume in repeated sampling. The standard deviation of a statistics’ sampling distribution is the standard error of the statistic once all unknown parameters have been estimated.
4.2. The Sampling Distribution of \( \bar{x} \)

Two important properties for the sampling distribution of the sample mean (\( \bar{x} \)) are:

1. \( \mu_{\bar{x}} = \mu \); where \( \mu_{\bar{x}} \) is the mean of the sampling distribution and \( \mu \) is the mean of the population.

2. \( \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \); where \( \sigma_{\bar{x}} \) is the standard deviation of the sampling distribution and \( \frac{\sigma}{\sqrt{n}} \) is the population standard deviation divided by the square root of the sample size.

The above statements are true regardless of the shape of the parent distribution. When the sample size, \( n \), is sufficiently large, the sampling distribution of \( \bar{x} \) is approximately normally distributed. This last statement is the Central Limit Theorem applied to \( \bar{x} \). It should be noted that the sampling distribution of \( \bar{x} \) when sampling from a normal distribution is normal regardless of sample size. Consequently, all sampling distributions for \( \bar{x} \) are either normal or approximately normal for sufficiently large sample sizes. Mathematically, the sampling distribution of \( \bar{x} \) is written \( \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \). Provided we are sampling from a normal distribution, \( \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \), and regardless of the parent distribution, \( \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \) provided \( n \) is sufficiently large.

The probability density function for the uniform distribution is written as \( f(x) = \frac{1}{b-a}; a < x < b \), with \( \mu = \frac{a+b}{2} \) and \( \sigma = \frac{b-a}{\sqrt{12}} \). Consider a uniform distribution with \( a = 0 \) and \( b = 10 \), \( f(x) = \frac{1}{10}; 0 < x < 10 \). A graph of this particular uniform distribution is shown in Figure 4.4. Note that the mean, \( \mu = \frac{0+10}{2} = 5 \) and that the

Figure 4.4: Uniform(0,10)
standard deviation, \( \sigma = \frac{10}{\sqrt{12}} = 2.8867 \) for this particular distribution. When all possible samples of size 2 are taken from a uniform distribution with \( a = 0 \) and \( b = 10 \), the resulting sampling distribution is triangular in shape, see Figure 4.5. We know the mean for the sampling distribution depicted in Figure 4.5 is 5 since \( \mu_\bar{x} = \mu = 5 \). We also know the standard deviation for the sampling distribution depicted in Figure 4.5 is 2.0412 since

\[
\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.8867}{\sqrt{2}} = 2.0412.
\]

Figure 4.5 clearly does not have a normal shape. Consider the Normal distribution with \( \mu = 5 \) and \( \sigma = 2.0412 \) depicted in Figure 4.6. The question is how large does \( n \) have to be before the sampling distribution of the sample mean becomes an adequate approximation of the normal distribution. Note that Figure 4.6 displays the percent of the distribution that is between negative infinity and two standard deviations to the left of the mean as 2.28%. The percent of the distribution that is between two and one standard deviations to the left of the mean is 13.59%. You should verify these percentages for all of the values shown in Figure 4.6. By using the formula to find the area of a triangle, we can also calculate the percent of the distribution that lies between 0 and the mean minus two standard deviations (1.684%), the mean minus two standard deviations and the mean minus one standard deviation (15.825%), and the mean minus one standard deviation and the mean (32.491%). Figure 4.7 on the following page displays a Normal distribution with \( \mu = 5 \) and \( \sigma = 2.0412 \) superimposed over the sampling distribution for the sample mean for samples of size 2 when sampling from a Uniform (0, 10) distribution. Note that the mean and the standard deviation for both distributions are identical although the shapes are different. It
4.2. The Sampling Distribution of \( \bar{x} \)

is important to note that to have a normal shape we need to have approximately 2.28% of the data be smaller than the mean minus two standard deviations. In the sampling distribution for the sample mean for samples of size 2, we have 1.684% of the data that is smaller than the mean minus two standard deviations. Next we show how to use simulation to approximate the sampling distribution of the sample mean for various sample sizes.

To simulate the sampling distribution for the sample mean when taking samples of size two from a Uniform (0, 10) distribution, we will generate 2 columns of K observations from a Uniform (0,10) distribution by using the commands Calc>Random Data>Uniform. The number K is some large number (we will usually use 10,000 or more for K). By generating two columns of K observations from a Uniform (0,10) distribution, we have simulated taking K samples of size 2. Next we use the Calc>Row Statistics to calculate the mean of the K samples of size 2 and store the result in a new column, perhaps named xbar2. A simulation generating 800,000 samples of size 2 from a Uniform (0,10) distribution and subsequently producing a histogram of the 800,000 calculated means produced Figure 4.8. Note how close the mean and standard deviation of the simulated sample mean distribution compare to the theoretical values for the mean and standard deviation for the sampling distribution of the sample mean when using samples of size 2 (4.9994 vs 5.0000, and 2.0415 vs 2.0412). The commands in Output 4.1 on the following page show how we have calculated the mean, standard deviation, and values like the mean minus two standard deviations.

Figure 4.7: Normal Distribution Superimposed over \( \bar{x} \)'s Distribution for Samples of Size 2 when Sampling from a Uniform (0, 10) Distribution

![Normal Distribution Superimposed over \( \bar{x} \)'s Distribution for Samples of Size 2 when Sampling from a Uniform (0, 10) Distribution](image)

Figure 4.8: Samples of Size 2 from a Uniform (0,10) Distribution and Subsequent Histogram of the 800,000 Calculated Means

![Simulated Sampling Distribution for the sample mean when n=2 and sampling from a Unif (0, 10)](image)
4.2. The Sampling Distribution of $\bar{x}$

Output 4.1: Calculations of Various Statistics for Simulation

4.99943 - 2 \times 2.0415 = 0.916441. Although you can follow the commands given in Output 4.1 to code your values, it is probably easier to use the Code Numeric to Numeric Dialog Window. Keep in mind that we have 800,000 values stored in xbar2 (C3). We want to know how many of those values are between 0 and 0.916441, 0.916441 and 2.95794, 2.95794 and 4.99943, 4.99943 and 7.04093, and 7.04093 and 10 respectively. One way to get an answer is to code all of the values in a given interval to an integer. This can be accomplished by selecting Manip > Code > Numeric to Numeric. Fill in the Code – Numeric to Numeric Dialog Window to look identical to the Code – Numeric to Numeric Dialog Window shown in Figure 4.9. When your window looks identical to Figure 4.9, click the OK. This procedure will code all values between 0 and 0.916441 to a 1, values between 0.916441 and 2.95794 to a 2, and so on. Now what we really want to do is to compare the percent of values in the various categories to the corresponding theoretical percentages of a normal distribution to see how well our simulated distribution of the sample mean approximates a normal distribution. To tally up the integer values we just created, choose Stat > Tables > Tally. Once in the Tally Dialog Window select cxbar2 (we used the name cxbar2 to stand for coded xbar2) as the variable.
4.2. **The Sampling Distribution of** $\bar{x}$

of interest and click in both the *Counts* and *Percents* Display Boxes before clicking the *OK*. See Figure 4.10 for an example. Table 4.1 shows the percent of values for the normal distribution falling between various categories

**Figure 4.10: Tally Dialog Window Example**

and the corresponding percent of values in the simulated sampling distribution of the sample mean.

<table>
<thead>
<tr>
<th>Normal Distribution</th>
<th>Simulated Sampling Distribution for the Sample Mean ($n = 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(-\infty, \bar{x} - 2s)$</td>
<td>2.28%</td>
</tr>
<tr>
<td>$(\bar{x} - 2s, \bar{x} - s)$</td>
<td>13.59%</td>
</tr>
<tr>
<td>$(\bar{x} - s, \bar{x})$</td>
<td>34.13%</td>
</tr>
<tr>
<td>$(\bar{x}, \bar{x} + s)$</td>
<td>34.13%</td>
</tr>
<tr>
<td>$(\bar{x} + s, \bar{x} + 2s)$</td>
<td>13.59%</td>
</tr>
<tr>
<td>$(\bar{x} + 2s, \infty)$</td>
<td>2.28%</td>
</tr>
</tbody>
</table>

When sampling from different parent populations the resulting sampling distribution for the sample mean will of course be different. However, the Central Limit Theorem tells us that resulting sampling distribution of the sample mean will have an approximate normal distribution provided the sample size is sufficiently large. So, just how large is this sufficiently large number? That depends on the parent distribution from which you are sampling. In Lab 4.1, you will have the opportunity to use simulation and determine for your self just how large the sample size needs to be with two different distributions before the sampling distribution of the sample mean adequately resembles a normal distribution. Note: Many authors of statistics texts have claimed that a sample size of 30 is sufficiently large for the Central Limit Theorem to work on the sampling distribution of the sample mean. Soon you will know whether or not this is a good rule of thumb. Video 4.1 illustrates the sampling distribution of $\bar{x}$ when sampling from a normal population while Video 4.2 on the next page and Video 4.3 on the following page illustrate the sampling distribution of $\bar{x}$ when sampling from an exponential distribution.

**Video 4.1: For Problem 4.26**  —  (Duration: 7 minutes 1 second)

For optimal video viewing, set your computer’s display panel resolution to 1024 x 768 pixels.
To verify or change your display panel resolution, select **Start>Settings>Control Panel>Display>Settings**.
4.3 The Sampling Distribution of the Sample Proportion $p$

If the sample size, $n$, is sufficiently large, then the sampling distribution of $p$ will have an approximately normal distribution centered at $\pi$ and a standard deviation of $\sqrt{\frac{\pi(1-\pi)}{n}}$. Mathematically, the sampling distribution of $p$ is written $p \sim (\mu_p, \sigma_p)$ where $\mu_p = \pi$ and $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$. When $n$ is sufficiently large, the sampling distribution of $p$ is written $p \sim N(\mu_p, \sigma_p)$. Notice the similarities between the mean and standard deviation for the sampling distribution of $p$ and the mean and standard deviation for a binomial distribution. Recall that the mean of the binomial distribution is $n\pi$ and the standard deviation for the binomial distribution is $\sqrt{n\pi(1-\pi)}$. Specifically, if $p$ represents the proportion of successes in a random sample of size $n$ from a Bernoulli distribution with parameter $\pi$, then $X = np$ is a Binomial random variable with parameters $n$ and $\pi$.

Most national polls take samples utilizing 1,000 to 1,500 respondents. Have you ever wondered how the polls can be so incredibly accurate? To gain an understanding of the sampling distribution of $p$, consider the following scenario. Suppose the true proportion of students in favor of a common final for all sections of STAT 101 is $\pi = 0.25$. Simulate the sampling distribution of $p$ by generating 50,000 samples of size $n = 1,000$ and $\pi = 0.25$. Since this procedure requires too much computer memory, we can generate 50,000 binomial random values with $n = 1,000$ and $\pi = 0.25$ to do a mathematically equivalent simulation. To complete the simulation, divide the binomial outcome by $n$ to get $p$.

To generate the 50,000 binomial random values from a distribution with $n = 1,000$ and $\pi = 0.25$, choose Calc $\rightarrow$ Random Data $\rightarrow$ Binomial and fill in the Binomial Distribution Dialog Window as shown if Figure 4.11 on the next page. Once the 10,000 values from the binomial distribution are created and stored in column a column named bino, we want to divide all of the values in bino by $n = 1,000$. One way to do this is to select Calc $\rightarrow$ Calculator. By selecting Calc $\rightarrow$ Calculator, you will be led to the Calculator Dialog Window shown in Figure 4.12 on the following page. Once in the Calculator Dialog Window, we store the result of our expression which is ‘bino’/1000 in a column named p. Next we use the command Graph $\rightarrow$ Histogram to produce the graph in Figure 4.13 on the next page which is a histogram of the simulated distribution of $p$. Video 4.4 uses these ideas to simulate the proportion of patients admitted to a hospital that have private health insurance. Based on the simulated sampling distribution of $p$, shown in Figure 4.13 on the next page, is it very likely one might get a sample where $p$ is either greater than 0.30 or less than 0.20? Quite often statements such as “we are 95% confident that...
the true proportion of students favoring a common final is $25\% \pm 2.68\%$” are made by people giving the results of their polls. Does this statement make sense after looking at the simulated sampling distribution? What this
4.3. The Sampling Distribution of the Sample Proportion $p$

The statement really means is that when repeated samples of size 1,000 are taken from a population where $\pi = 0.25$, about 95% of the samples will contain the true parameter $\pi$ between the 2.5th percentile and the 97.5th percentile of that sampling distribution.

The following simulation demonstrates the previous idea. First simulate 10,000 samples of size $n = 1,000$ with $\pi = 0.5$ from a binomial distribution. Next, determine the proportion of the simulated samples where $\pi = 0.5$ is between the 2.5th percentile and the 97.5th percentile of the sampling distribution. Recall that the standard deviation of the sampling distribution is $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$. Since $\pi = 0.5$, $\sigma_p = 0.015811$; therefore, the 2.5th and the 97.5th percentiles of the sampling distribution will equal $p - 1.96 \times 0.015811$ and $p + 1.96 \times 0.015811$ respectively. Why did we use 1.96? (Hint: What is the value in a standard normal distribution such that 2.5% of the area is to its right?) In other words, for each one of our simulated $p$ values we need to subtract and add the margin of error ($1.96 \times 0.015811 = 0.03098956$) and determine if the resulting interval contains the value $\pi = 0.5$. First we create two columns, one to contain the quantity $p - 1.96 \times 0.015811$ and the other to contain the quantity $p + 1.96 \times 0.015811$. This can be accomplished using the MINITAB$^{\text{TM}}$ Calculator (See Figure 4.14). Note that we store the lower value in a column named LowerLimit and when we create the upper value we store the result in a column named UpperLimit. Now, we would like MINITAB$^{\text{TM}}$ to count the number of times 0.5 is contained between the 10,000 lower and upper limits rather than manually looking through the entire 10,000 lower and upper limits to see if they contain 0.5. To accomplish this task, we will recode all values in the column named LowerLimit to a 0 if the value is less than or equal to 0.5, and to a 1 if the value is greater than 0.5 and store the result in a column named CLL for coded lower limit. We also code the values in UpperLimit except this time if the value in UpperLimit is less than or equal to 0.5 we code the value to a 1, and if the value in UpperLimit is greater than 0.5 we code the value to a 0 and store the results in a column named CUL for coded upper limit. In essence, anytime we see a 1 in either CLL or CUL it indicates 0.5 is not contained between the 2.5th and 97.5th percentiles for that sampling distribution. To code the values in UpperLimit and LowerLimit select Manip $\rightarrow$ Code $\rightarrow$ Numeric to Numeric and fill in the appropriate values. Figure 4.14 on the next page illustrates the coding of values from UpperLimit to CUL. Keep in mind that you will need to do this procedure for both columns UpperLimit and LowerLimit. Finally, we select Stat $\rightarrow$ Tables $\rightarrow$ Cross Tabulation to count the 0s and 1s in columns CUL and CLL. When the Cross Tabulation Dialog Window appears, select both CLL and CUL as the classification variables as well as clicking in both the Display Counts and Total Percents boxes (Figure 4.16 on the following page). The results from the Cross tabulation command appear in Output 4.2 on the next page. Note from Output 4.2 that 94.78% of the simulated sampling distributions contained the true parameter, $\pi = 0.5$, between the 2.5th percentile and the 97.5th percentile of that sampling distribution.
4.3. The Sampling Distribution of the Sample Proportion $p$

Figure 4.15: Coding UpperLimit for Confidence Interval Calculation

Figure 4.16: Cross Tabulation Dialog Window for Confidence Interval

Output 4.2: Cross Tabulation Results for Confidence Interval

<table>
<thead>
<tr>
<th></th>
<th>CLL</th>
<th>CUL</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9470</td>
<td>245</td>
<td>9723</td>
</tr>
<tr>
<td></td>
<td>94.78</td>
<td>2.45</td>
<td>97.23</td>
</tr>
<tr>
<td>1</td>
<td>277</td>
<td>0</td>
<td>277</td>
</tr>
<tr>
<td></td>
<td>2.77</td>
<td>--</td>
<td>2.77</td>
</tr>
<tr>
<td>All</td>
<td>9755</td>
<td>245</td>
<td>10000</td>
</tr>
<tr>
<td></td>
<td>97.55</td>
<td>2.45</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Cell Contents --

- Count
- % of Tbl
4.4 Summary and Review Labs

Lab 4.1 — The Sampling Distribution of $\bar{x}$

Objectives:

I. To determine how large the sample size must be before a distribution begins to look normal for different population distributions

II. To understand how the central limit theorem is applied to the sample mean

III. To display results in graphs and tables

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

In this lab, you will be generating random data from various parent populations and trying to determine how large a sample from the parent population you need before you start to see the Central Limit Theorem take effect. You will also discover how well simulated values approximate the true population distributions.

Commands and Reminders:

- Calc>Row Statistics>Mean will give you the sample mean of your random samples. Naming your storage column something meaningful will help you keep things straight.

- Calc>Random Data>Distribution Name (Uniform, Normal, etc.)
- The number of rows you simulate is the number of samples.
- The number of columns you simulate is the sample size.

Directions:

1. Create graphs of the parent populations from which you will sample with which you are not yet familiar. For extra credit, create a graph of a standard normal distribution. 
   Append the graphs from parts A. and B. to the report pad. Once you have created these graphs, say File>New Worksheet.

   A. Uniform(0,10)
     i. Calc>Make Patterned Data>Simple Set of numbers
        a) Store in: c1
        b) From: 0
        c) To: 10
        d) In Steps of: 0.01
   ii. Calc>Probability Distributions>Uniform
        a) Click Probability Density
        b) Lower: 0
        c) Upper: 10
        d) Input column: c1
4.4. Summary and Review Labs

iii. **Graph>Plot**
- a) X: c1
- b) Y: c2
- c) Display: connect
- d) **Annotation>title:** Uniform(0,10) Population

B. Exponential(10)

- i. **Calc>Make Patterned Data>Simple Set of numbers**
  - a) Store in: c3
  - b) From: 0
  - c) To: 40
  - d) In Steps of: 0.01
- ii. **Calc>Probability Distributions>Exponential**
  - a) Click **Probability Density**
  - b) Mean: 10
  - c) Input column: c3
  - d) Storage: c4
- iii. **Graph>Plot**
  - a) X: c3
  - b) Y: c4
  - c) Display: connect
  - d) **Annotation>title:** Exponential(10) Population

2. Use simulation to determine the approximate sample sizes required to have the sampling distribution of \( \bar{x} \) look approximately normal when sampling from a uniform distribution and an exponential distribution. To solve the problem for each distribution, you should consider the general shape of the sampling distribution and how close the sampling distribution resembles a normal distribution.

A. Sample from a uniform (0, 10).

- i. A uniform \((a, b)\) has \(\mu = \frac{b+a}{2}\) and \(\sigma = \frac{b-a}{\sqrt{12}}\).
- ii. Generate your random data (**Calc>Random Data>Uniform**). Use 10,000 samples [rows] and start with samples of size 5 (**Store in column(s) C1-C5**). Type a 0 in the **Lower Endpoint Box** and a 10 in the **Upper Endpoint Box**.
- iii. Calculate sample means (**calc>row statistics, store in xbars**) and create a histogram of the xbars column.
- iv. Give the percent of observations between 0 and 1, 1 and 2, 2 and \(\infty\), 0 and \(-1\), \(-1\) and \(-2\), and \(-2\) and \(-\infty\) standard deviations from the mean. An easy way to count the number of observations between two values is to code your numbers. See Output 4.1 on page 110 for an idea of how to do this.
- v. Record your results from step 2(A)iv in a table resembling the one that follows. Be sure to give the sample size used and the resulting \(\bar{x}\) and \(s\) values for each of the empirical percentages columns.

<table>
<thead>
<tr>
<th>Normal Percentages</th>
<th>UNIFORM</th>
<th>(n = )</th>
<th>(\bar{x} = )</th>
<th>(s = )</th>
<th>(\bar{x} = )</th>
<th>(s = )</th>
<th>(\bar{x} = )</th>
<th>(s = )</th>
</tr>
</thead>
<tbody>
<tr>
<td>((-\infty, \bar{x} - 2s))</td>
<td>(\infty)</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td></td>
</tr>
<tr>
<td>((\bar{x} - 2s, \bar{x} - s))</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((\bar{x} - s, \bar{x}))</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((\bar{x}, \bar{x} + s))</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((\bar{x} + s, \bar{x} + 2s))</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((\bar{x} + 2s, \infty))</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td>(\bar{x} = )</td>
<td>(s = )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

vi. If your histogram is not bell shaped or your empirical percentages do not match those for the normal distribution, delete the xbars column and add five new uniform columns with the same distribution as
those from 2(A)ii. Then, starting at 2(A)iii rework the calculations. You may stop when your histogram is bell shaped and your empirical percentages closely match those for the normal distribution. Include your final histogram in your lab report.

Before you start the exponential experimentation, it is important to get a file new project after you complete the uniform experimentation.

B. Sample from an exponential (10).
   i. An exponential(θ) has μ = θ and σ = θ.
   ii. Generate your random data (Calc>Random Data>Exponential). Use 10,000 samples [rows] and start with samples of size 5 (Store in column(s) C1-C5). Type a 10 in the Mean Box.
   iii. Calculate sample means (calc>row statistics, store in xbars) and create a histogram of the xbars column.
   iv. Give the percent of observations between 0 and 1, 1 and 2, 2 and ∞, 0 and −1, −1 and −2, and −2 and −∞ standard deviations from the mean. An easy way to count the number of observations between two values is to code your numbers. See Output 4.1 on page 110 for an idea of how to do this.
   v. Record your results from step 2(B)iv in a table resembling the one that follows. Be sure to give the sample size used and the resulting $\bar{x}$ and $s$ values for each of the empirical percentages columns.

<table>
<thead>
<tr>
<th>Normal Percentages</th>
<th>Range</th>
<th>Empirical Percentages</th>
<th>Empirical Percentages</th>
<th>Empirical Percentages</th>
<th>Empirical Percentages</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXPONENTIAL</td>
<td>$n = \bar{x}$</td>
<td>$n = s$</td>
<td>$n = \bar{x}$</td>
<td>$n = s$</td>
<td>$n = \bar{x}$</td>
</tr>
<tr>
<td></td>
<td>$x = s$</td>
<td>$x = \bar{x}$</td>
<td>$x = s$</td>
<td>$x = \bar{x}$</td>
<td>$x = s$</td>
</tr>
<tr>
<td></td>
<td>$x = s, \bar{x}$</td>
<td>$x = s, \bar{x}$</td>
<td>$x = s, \bar{x}$</td>
<td>$x = s, \bar{x}$</td>
<td>$x = s, \bar{x}$</td>
</tr>
<tr>
<td></td>
<td>$x + s$</td>
<td>$x + s$</td>
<td>$x + s$</td>
<td>$x + s$</td>
<td>$x + s$</td>
</tr>
<tr>
<td></td>
<td>$x + s, \bar{x} + 2s$</td>
<td>$x + s, \bar{x} + 2s$</td>
<td>$x + s, \bar{x} + 2s$</td>
<td>$x + s, \bar{x} + 2s$</td>
<td>$x + s, \bar{x} + 2s$</td>
</tr>
<tr>
<td></td>
<td>$x + 2s, \infty$</td>
<td>$x + 2s, \infty$</td>
<td>$x + 2s, \infty$</td>
<td>$x + 2s, \infty$</td>
<td>$x + 2s, \infty$</td>
</tr>
</tbody>
</table>

vi. If your histogram is not bell shaped or your empirical percentages do not match those for the normal distribution, delete the xbars column and add fifty new exponential columns with the same distribution as those from 2(B)ii. Then, starting at 2(B)iii rework the calculations. You may stop when your histogram is bell shaped and your empirical percentages closely match those for the normal distribution. Include your final histogram in your lab report.
Lab 4.2 — The Sampling Distribution of the Sample Proportion

Objectives:

I. To simulate the distribution of the sample proportion

II. To compare the simulation to the theoretical distribution of the sample proportion

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

In this lab, you will be working with the distribution of the sample proportion, \( p \). The directions will lead you to create various simulations and to use those to create graphs to answer the questions.

Questions and Directions:

1. Simulate 10,000 samples of size \( n = 1,000 \) with \( \pi = 0.5 \) from a binomial distribution. Store the results of the simulation in a column named \( \text{bino} \). Divide the values in \( \text{bino} \) by 1000 and store the results in a column named \( \text{p} \). Determine the proportion of the simulated samples (\( p \)) between .473 and .526. Hint: Use Manip>Code>Numeric to Numeric and then Stat>Tables>Tally.

2. Produce a histogram of the simulated sampling distribution of \( p \). Title your histogram “Simulated Sampling Distribution of \( p \)”, and provide a right justified footnote with your name.

3. Based on your simulated distribution, what are the 10\(^{th} \), 20\(^{th} \), 50\(^{th} \), 80\(^{th} \), and 90\(^{th} \) percentiles respectively for the sampling distribution of \( p \)? (Hint: Sort your simulated \( p \) values then answer the question.)

4. What are the theoretical values for the 10\(^{th} \), 20\(^{th} \), 50\(^{th} \), 80\(^{th} \), and 90\(^{th} \) percentiles respectively for the sampling distribution of \( p \) based on a normal distribution?

5. What are the theoretical values for the 10\(^{th} \), 20\(^{th} \), 50\(^{th} \), 80\(^{th} \), and 90\(^{th} \) percentiles respectively for the sampling distribution of \( p \) based on a binomial distribution? (Hint: Use the Inverse cumulative probability function (\( \text{invcdf} \)) with the Binomial distribution and divide your answer by \( n \). Keep in mind you will not get exact percentiles in many cases do to the discrete nature of the problem. Report the numbers closest to the desired percentiles.)

6. Report percentiles based on your simulation, the theoretical normal, and the theoretical binomial in a Table resembling Table 4.2.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Simulated</th>
<th>Based on Normal</th>
<th>Based on Binomial</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7. Are their any discrepancies among the three methods for their respective percentiles? Do the values you reported for the binomial distribution provide the exact percentile requested?

8. Produce a graph showing the theoretical distribution of \( p \) based on a normal distribution. Make the Item display in the Plot Dialog Window connect. (Hint: Think about using a random variable \( X \), where \( X \) = number of successes. Then divide \( X \) by \( n \) to get your \( p \) values.)
9. Produce a graph showing the probability associated with values of $p$ based on the binomial distribution. (Consider the hint from 8) Make the Item display in the Plot Dialog Window project.

10. Do the graphs from 8 and 9 have similar shapes?

11. Append the graphs from 8 and 9 with appropriate titles to the report pad.

12. If you take a random sample of size $n = 1,000$ from a Bernoulli population where $\pi = .5$, is it likely that your sample will yield a $p$ value greater than 0.54? Answer this question by providing a probability and by visually inspecting your histogram.
5.0 Introduction

In statistics, we are often concerned with forming conclusions based on the data we have collected. We will try to discover the shape of the underlying distribution from which we have collected our data based on characteristics of the resulting sample. After we have made shape determinations, we will want numerical values to describe various aspects of our parent population including center and spread and occasionally, the proportion of successes. We will use the statistics that we can calculate from our samples to create confidence intervals for appropriate parameters using appropriate underlying distributions based on the information about the parent population in our possession. Finally, the labs from this chapter will lead you in the process of determining estimators for several population parameters so that you will understand how the more well known estimators were discovered and evaluated.

5.1 Describing the Parent Distribution

Statistics texts often use the terminology underlying or parent distribution. When you read underlying or parent distribution in a book or hear a statistician say underlying or parent distribution, what the book or statistician is referring to is the distribution of measurements in the original population. Clearly the original distribution of measurements can take on many shapes. However, while working with the distributions from various statistics, the two most common distributions we encounter are symmetric and skewed distributions. In chapters 1 and 2 we explored tools such as the stem-and-leaf plot, the histogram, the dotplot, and the boxplot which can be used to help assess the shape of a distribution. The stem-and-leaf plot, dotplot, and histogram show us the general appearance of the collected data. Assuming the data you are analyzing is a representative sample of the population, the stem-and-leaf plot, dotplot, and histogram of the sample should resemble the shape of the parent distribution. The boxplot is a great tool to assess symmetry and identify outliers.

The four graphical techniques discussed to assess the shape of a distribution are created by choosing Graph > and then one of Stem-and-Leaf, Histogram, Dotplot, or Boxplot.

5.1.1 Examining Shape and Choosing Estimators

When describing a distribution, we are concerned with three characteristics: shape, center, and spread. Let us revisit both symmetric and skewed distributions. With symmetric distributions (shape), the mean and median are measures that return the same value (center). Although the mean is used most often as a measure of center in symmetric distributions, there will be times the median will be preferred because the distribution of the sample median($M$) will have smaller variance than the distribution of the sample mean($\bar{x}$). When using the mean as a measure of center in a symmetric distribution, the spread is described using the standard deviation. With skewed distributions (shape),
5.2. Estimating Certain Population Parameters

the mean and median are measures that do not return the same value (center). Recall that the mean is pulled in the direction of the skew. The preferred measure of center for skewed distributions is the median. When using the median as a measure of center in a skewed distribution, the Spread is described using the inter-quartile range (IQR). Although MINITAB\textsuperscript{TM} does not directly calculate the sample IQR, Q3 and Q1 are given in the output generated from using Descriptive Statistics (Stat>Basic Statistics>Display Descriptive Statistics). Video 5.1 uses several graphical procedures to provide insight about the distribution of dopamine for two groups of schizophrenic patients.

Video 5.1: For Problem 5.14 — (Duration: 4 minutes 21 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

5.2 Estimating Certain Population Parameters

One often attempts to explain the shape of the parent population by using sample data. When describing the parent population, it is helpful to have both an idea of where the parent population is centered as well as some measure of its variability. The descriptive measures we use to describe the parent population are called parameters. When using statistics to make inferences, one first needs to decide which parameters need to be estimated. We might want to estimate $\mu$, $\theta$, $\pi$, $\sigma$, or some other parameter. The parameter one chooses to estimate often depends on the shape of the underlying distribution. For symmetric distributions, measures of central location, including the mean, median, and trimmed mean are identical. The population median, $\theta$, is often recommended as the parameter to use in describing the center of either a long tailed symmetric distribution or a skewed distribution. Once the parameter of interest is determined, an estimator is selected that will be reasonably close to the true value of the parameter. Video 5.2 examines several graphs to assess the parent population for the number of hours college freshmen spend studying and makes recommendations about which parameters accurately describe the center and variability for this distribution.

Video 5.2: For Problem 5.25 — (Duration: 4 minutes 8 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

Statistics are numerical summaries calculated from Samples.
Parameters are numerical summaries calculated from Populations.

Statisticians work with several classes of estimators. One of the more common classes of estimators is the class of so called Uniform Minimum Variance Unbiased Estimators. A statistic that is unbiased is one where the mean of the statistic’s sampling distribution is centered at the parameter being estimated. Mathematically, we say that a statistic is unbiased when the expected value of the statistic is equal to the parameter, $E(\hat{\psi}) = \psi$. When the mean of the sampling distribution is not equal to the parameter, the statistic is said to be biased. For a statistic to be UMVU (Uniform Minimum Variance Unbiased), the statistic must be both unbiased and must have the smallest variance among all the unbiased statistics. In Lab 5.2 you can see how two estimators could be unbiased yet have different variability.
Many of the common statistics used to estimate parameters are unbiased. The following is a list of several unbiased statistics often used to estimate parameters.

\[
\begin{align*}
\bar{x} & \rightarrow \mu \\
M & \rightarrow \theta \\
p & \rightarrow \pi \\
s^2 & \rightarrow \sigma^2
\end{align*}
\]

### 5.2.1 Properties of Estimators

Recall that if a distribution is symmetric, both the sample mean and median will coincide. Consequently, with symmetric distributions we can use either \( \bar{x} \) or \( M \) to estimate \( \mu \). When the parent population is approximately normal, the sampling distribution for \( \bar{x} \) has smaller variance than does the sampling distribution for the sample median. Further, the asymptotic (very large sample) variance of the median can be shown to equal \( \frac{\pi^2}{4} \). This is the answer to part of the extra credit question in Lab 5.2. Recall that the variance for the sample mean, \( \bar{x} \), is \( \frac{\sigma^2}{n} \). However, when the parent distribution has extremely long tails, the sample median is a better estimator of \( \mu \) than is the sample mean. Note that the Laplace distribution is a distribution with long tails. As a consequence, the sampling distribution for the sample median has less variance than does the sampling distribution for the sample mean. It can be shown that the asymptotic standard deviation of the sample median for the Laplace distribution is \( \frac{\sqrt{2} \sigma}{\sqrt{n}} \). This last number is the answer to the other part of the extra credit question in Lab 5.2.

Parameters are estimated using statistics. However, the choice of what parameter to estimate is not always clear, and neither is it always clear what statistic to use to estimate the parameter of choice. Before one can make a decision on what parameters need to be estimated, an understanding of the parent population is paramount. From the collected data, we infer the underlying distribution’s shape, measures of central tendency, and measures of variability. Some typical distributional shapes include approximately normal, symmetric short tails, symmetric long tails, skew right, skew left, and multimodal.

### 5.2.2 Estimating \( \pi \)

For Bernoulli populations, we are concerned with the proportion of successes, \( \pi \). To estimate the proportion of successes in a Bernoulli population, we use \( p \), the sample proportion of successes. Recall that \( p \) is an unbiased estimator of \( \pi \). Also recall that under certain conditions, the distribution of \( p \) is distributed approximately normally with mean \( \pi \) and standard deviation \( \sqrt{\pi(1-\pi)/n} \). Do you recall the two conditions that must be satisfied to claim the distribution of \( p \) is approximately normal?

**Answer:** when both \( n \times \pi \) and \( n \times (1 - \pi) \) are greater than or equal to 5, the distribution of \( p \) is approximately normal.

### 5.2.3 Estimating the Center of a Distribution

There are numerous estimators for the center of a distribution. However, we will focus on just two estimators. First, the sample mean, \( \bar{x} \), is the minimum variance unbiased estimator for the population mean, \( \mu \), of a normal distribution. Further, \( \bar{x} \) is recommended as an estimator of \( \mu \) for near normal or symmetric distributions without excessively long tails. Second, the sample median, \( M \), is recommended to estimate \( \theta \) when the parent population is symmetric with long tails or skewed in shape.
5.3 Confidence Interval for a Bernoulli Proportion, \( \pi \)

A confidence interval for a population parameter is an interval of possible values for the unknown parameter. The interval is computed based on the data obtained from a random sample. The interval is constructed in such a way that we will have a high level of confidence that the true parameter is contained within the interval. The validity of a confidence interval refers to the confidence we can place on the interval containing the true parameter. In addition to validity, confidence intervals that are of relatively small width are desirable. The precision of a confidence interval refers to the length of the confidence interval. In practice, estimates of parameters are usually given in the form

\[
\text{estimate} \pm \text{margin of error}
\]

The margin of error is the maximum error one would expect to see at a specified confidence level.

The probability assigned to a confidence interval is relevant only prior to sampling. Once values for the random variables are inserted into the confidence interval formula, we no longer have a probability statement. Once the end points for the confidence interval are calculated, the true parameter is either contained within the limits, or it is not. Consequently, the probability the true parameter is contained within the calculated confidence interval limits can be only 0 or 1. What is meant by a 95% confidence interval is that if 95% confidence intervals are always calculated in the same manner, we expect that 95% of them will contain the true parameter.

Recall that the sampling distribution of \( p \) provided \( n \) is large is distributed normally with a mean of \( \pi \) and a standard deviation of \( \sqrt{\pi(1-\pi)/n} \). The standard error of \( p \) is written \( SE(p) \), or \( \hat{\sigma}_p \), where

\[
SE(p) = \sqrt{\frac{p(1-p)}{n}}
\]

Note the difference between the standard deviation of a statistic and the standard error of a statistic: the standard error of a statistic is the standard deviation of a statistic when all of the unknown parameters have been estimated. The approximate \( (1-\alpha) \times 100\% \) CI for \( \pi \) is written

\[
P \left( p - Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} < \pi < p + Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right) \equiv (1-\alpha) \times 100\% \tag{5.1}
\]

5.3.1 Calculating and Interpreting Confidence Intervals for a Proportion

Several notations are in use to denote critical values. Some authors denote their critical values with a *, others use slightly more descriptive notation. The notation used in this manuscript may be slightly confusing at first; however, it imparts maximum information to the reader. Many charts and software packages, including MINITAB\textsuperscript{TM}, will calculate and give the area in a specific distribution to the left of a given quantity. The notation we adopt in this manuscript is to use a subscript on the distribution symbol that gives the area less than the critical value. Often, the subscripts of critical values in a confidence interval (CI) are denoted with the expressions \( \alpha/2 \) and \( 1 - \alpha/2 \) where \( \alpha \) is the first letter of the Greek alphabet. Critical values for the standard normal distribution are denoted \( Z_{\alpha/2} \) and \( Z_{1-\alpha/2} \). When developing confidence intervals, the confidence interval is frequently called a \( (1-\alpha) \times 100\% \) confidence interval. When this occurs, the critical value, \( Z_{\alpha/2} \), is the number such that \( \alpha/2 \) of the area under the normal distribution is to the left of \( Z_{\alpha/2} \), while the critical value \( Z_{1-\alpha/2} \) is the number such that \( 1 - \alpha/2 \) of the area under the normal distribution is to the left of \( Z_{1-\alpha/2} \).

When one calculates a 95% CI, the a value is obtained by solving \( (1-\alpha) \times 100\% = 95\% \) for \( \alpha \). Note in this particular case that the \( \alpha \) value is 0.05. The critical value \( Z_{1-\alpha/2} \) or \( Z_{1-0.05/2} = Z_{0.975} \) is found with MINITAB\textsuperscript{TM} by requesting the number from the standard normal distribution such that 0.975 of the area is to the left of that value. See Output 5.1 on the next page for an example of MINITAB\textsuperscript{TM} Session Window commands to calculate \( Z_{0.975} \).
Output 5.1: Session Window Commands to Calculate $Z_{0.975}$

```
MTB > InvCDF .975;   # Requesting the value with 0.975 area to its left
SUBC> Normal 0.0 1.0. # Based on a standard normal distribution

Inverse Cumulative Distribution Function
Normal with mean = 0 and standard deviation = 1.00000
$P(X \leq x) = 0.9750$ $x = 1.9600$

MTB > CDF 1.96;       # Requesting the area to the left of the value 1.96
SUBC> Normal 0.0 1.0. # Based on a standard normal distribution

Cumulative Distribution Function
Normal with mean = 0 and standard deviation = 1.00000
$x = P(X \leq x) = 0.9756$
```

To calculate a critical value using the point and click approach, choose Calc>Probability Distributions>Normal. When the Normal Distribution Dialog Window opens,

1. Click in the Inverse cumulative probability circle.

2. Leave the default setting of Mean 0.0 and Standard deviation 1.0 as they are since these are the values for the standard normal.

3. Click in the Input constant circle then type the area to the left of the value you are seeking, in this case 0.975.

4. If you do not specify an optional storage location, the answer will display in the Session Window. This is what you will want to do in most cases. On occasion, you may want to use the Optional storage location. Be aware that if you do this the answer will not display in the Session Window unless you specifically request MINITAB™ to display the answer.

5. Click the OK button.

Figure 5.1 shows the Normal Distribution Dialog Window with the appropriate boxes and options filled in to find $Z_{0.975}$. The Normal Distribution Dialog Window will also allow the user to specify a value in order to find the area to the left of the specified value. Suppose one wants to find the area to the left of the value $-1.96$ in a standard normal distribution. From the Normal Distribution Dialog Window click in the circle to the left of
5.3. Confidence Interval for a Bernoulli Proportion, $\pi$

Figure 5.2: Normal Distribution Dialog Window for Cumulative Probability

Cumulative probability and fill in the remaining options as shown in Figure 5.2. The Session Window commands to accomplish the same objective are given in Output 5.2. Note the command CDF ends with a semicolon. Also note that the subcommand always ends with a period. The CDF and InvCDF commands work in the same manner with any distribution.

Output 5.2: Session Window Commands to Calculate $Z_{0.975}$

MTB > InvCDF .975; # Requesting the value with 0.975 area to its left
SUBC> Normal 0.0 1.0. # Based on a standard normal distribution

Inverse Cumulative Distribution Function

Normal with mean = 0 and standard deviation = 1.00000

$P( X \leq x )$

x

0.9750 1.9600

MTB > CDF 1.96; # Requesting the area to the left of the value 1.96
SUBC> Normal 0.0 1.0. # Based on a standard normal distribution

Cumulative Distribution Function

Normal with mean = 0 and standard deviation = 1.00000

$P( X \leq x )$

x

1.9600 0.9750

The distribution of scores for individuals taking the Stanford Binet is known to follow a normal distribution with mean of 100 and standard deviation of 16. What proportion of the population has a score less than 132? This question is often written mathematically as the probability the random variable IQ is less than 132, $P( IQ < 132 )$, where $IQ \sim N(100, 16)$. To answer the above question, we ask MINITAB™ to give us the area to the left of 132 in a normal distribution with mean 100 and standard deviation 16. The MINITAB™ Session Window commands are illustrated in Output 5.3 on the next page. Note that 97.72% of the individuals taking the Stanford Binet Test score less than 132. Another way to interpret the result is to say that only 2.28% of the individuals taking the Stanford Binet score higher than 132.

Example 5.3.1: A professor in the Biology Department who grades “on the curve” told his class that their last exam scores followed a normal distribution with a mean of 75 and a standard deviation of 15. If the top 10% will receive an A on this particular exam, what is the minimum score a student can make on the test a still get an A?

Solution: Mathematically we write this problem as $P( X \geq x ) = 0.10$, and solve for $x$. Note that $P( X \geq x ) = 0.10$ implies that $P( X < x ) = 1 - 0.10$. Because most software packages and tables give the area to the left of a specified
5.3. Confidence Interval for a Bernoulli Proportion, $\pi$

Output 5.3: Calculating $P(IQ < 132)$ and $P(X \leq x) = 0.90$

value, we wish to rearrange the inequality with an equivalent less than or less than or equal to statement. Consequently, instead of writing $P(X \geq x) = 0.10$, we switch the direction of the inequality and write $P(X < x) = 0.90$. The statement $P(X \leq x) = 0.90$ is asking the user to find the $x$ value such that 90% of the values in the $X$ distribution are less than the given $x$ value. Note that this is the same as finding the $x$ value such that only 10% of the values are greater than that particular $x$ value. The MINITAB\textsuperscript{TM} commands used to determine the minimum numerical score required to receive an A are illustrated in Output 5.3. Note that in this particular problem, a student will have to earn a 94.2233 or higher to be awarded an A.

5.3.2 Constructing a confidence interval for a Bernoulli proportion

Example 5.3.2: Consider constructing a 93% confidence interval for the true proportion of college students that exercise at least three times a week. A researcher randomly samples 500 college students from across the United States and finds 200 exercise at least three times a week.

Solution: To use formula (5.1), $P \left( p - Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} < \pi < p + Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right) \cong (1 - \alpha) \times 100\%$, we first find the critical value $Z_{1-\alpha/2}$. The sample proportion of successes, $p$, is calculated to be 0.40 (200/500).

The MINITAB\textsuperscript{TM} code in Output 5.4 determines the confidence interval limits by first calculating $p - Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$, the lower confidence interval limit, and then calculating $p + Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$, the upper confidence interval limit. Note from Output 5.4 that the value of $Z_{1-\alpha/2} = Z_{0.965}$ is 1.8119. We are 93% confident that the true proportion of college students who exercise at least three times a week is between 0.360303 and 0.439697. Video 5.3 on the following page illustrates the concept of a $(1 - \alpha) \times 100\%$ confidence interval for a population proportion using simulation. This video is a great primer for Lab 5.3 on page 144.
5.3. Confidence Interval for a Bernoulli Proportion, $\pi$

Video 5.3: For **Problem 5.44** — (Duration: 10 minutes 26 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select **Start>Settings>Control Panel>Display>Settings**.

Now that we have seen how to use MINITAB$^\text{TM}$ to calculate all of the values that go into formula (5.1), we will use MINITAB$^\text{TM}$ to get the CI directly.

MINITAB$^\text{TM}$ calculates exact and approximate CIs for $\pi$. To calculate a CI for $\pi$ choose:

1. **Stat>Basic Statistics>1 Proportion**
2. Based on your information, do one of the following:
   a. If you have summarized data, click in the circle to the left of **Summarized Data**. Enter the number of Bernoulli trials in the **Number of trials Box**, and the number of times your Bernoulli random variable was a success in the **Number of successes Box**.
   b. If you have data in a MINITAB$^\text{TM}$ worksheet, click in the circle to the left of **Samples in columns** and select the column that contains your data.
3. Click on the **Options Box**.
   a. Enter the desired confidence level $(1 - \alpha)$ in the **Confidence level Box** and click **OK** if you want an exact CI.
   b. Enter the desired confidence level $(1 - \alpha)$ in the **Confidence level Box** and click in the box **Use test and interval based on the normal distribution** (跟随)$\text{to} \sqrt{\text{}}$ if you want a CI based on formula (5.1). This is the formula **Basic Statistics and Data Analysis** uses.

5.3.3 Calculating Sample Size for a Given Margin of Error

Recall that most Confidence intervals take the form “estimate ± margin of error.” The margin of error for the population proportion confidence interval formula when $n$ is sufficiently large is $Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$. To estimate $\pi$ with a confidence interval so that the bound on the margin of error is $B$, we solve equation (5.2) for $n$.

$$B = Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

(5.2)

The solution is written as

$$n = \left(\frac{Z_{1-\alpha/2}}{B}\right)^2 \frac{p(1-p)}{n}$$

(5.3)

**Example 5.3.3:** Consider Example 5.3.2 on the preceding page. Note that the estimated margin of error for the 93% confidence interval for the proportion of college students who exercise at least three times per week was calculated to be 0.039697. Next we will find the sample size necessary to reduce that margin to 0.03.

**Solution:**

$$n = \left(\frac{1.81291}{0.039697}\sqrt{0.4(1-0.4)}\right)^2 = 875.47$$

To retain the same validity (93%), the researcher needs to take a random sample of 876 college students to have a margin of error no greater than 0.03. Note that in order to ensure a validity of at least 93%, the sample size $n$ must always be rounded up.

The global macro **NPIE**, in Appendix E.1 on page 260 and in the **macros** directory of the CD can be entered into your MINITAB$^\text{TM}$ macro subdirectory and invoked in the MINITAB$^\text{TM}$ **Session Window** by typing
5.4. Confidence Interval for a Population Mean, \( \mu \)

%NPIE. The macro computes the required sample size for a given confidence level when the bound is specified. The macro prompts the user to enter the bound, confidence level, and estimate of the parameter.

When macros are stored in the MINITAB™’s macro directory you do not need to provide a complete path name for MINITAB™ to know the location of the macro nor add the extension *.mac to the macro name. However, if you are running the macro off a CD, you will need to specify the complete path to where the macro is found after typing % such as %G:\HATforBSDA\MACROS\npie.mac.

The macro prompts and the resulting output generated in the Session Window by invoking the macro NPIE are displayed in Output 5.5. Note that the macro answer of 876 is in agreement with the answer calculated earlier.

Output 5.5: Session Window for NPIE Macro

MINITAB > npie
Executing from file: D:\Program Files\Minitab\MACROS\npie.MAC
Press Enter after typing your answer to each question.
What is your bound?
DATA>.03
What is your confidence level?
Enter your confidence level as a decimal.
DATA>.93
What is your point estimate of the parameter?
DATA>.4
Required Sample Size is n
Data Display

n= 876

5.4 Confidence Interval for a Population Mean, \( \mu \)

5.4.1 The Z-Interval

When the parent population is normally or near normally distributed with known standard deviation \( \sigma \), then the sampling distribution of \( \bar{x} \) follows a normal distribution with mean \( \mu \) and standard deviation \( \sigma/\sqrt{n} \). A \((1 - \alpha) \times 100\%\) confidence interval for \( \mu \) based on \( \bar{x} \) is given in equation (5.4). The one sample confidence interval formula based on the standard normal distribution will be referred to as the Z-interval.

\[
P \left( \bar{x} - Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = (1 - \alpha) \times 100\% \tag{5.4}
\]

To calculate a \((1 - \alpha) \times 100\%\) Z-interval using MINITAB™, choose

1. Stat>Basic Statistics>1-Sample Z.
2. Specify a variable for MINITAB™ to use in the Variables Box once you are in the 1-Sample Z (Test and Confidence Interval) Dialog Window.
3. Enter the value for \( \sigma \) in the Sigma Box of the 1-Sample Z (Test and Confidence Interval) Dialog Window.
4. Click on the Options Box in the 1-Sample Z (Test and Confidence Interval) Dialog Window. When the 1-Sample Z - Options Dialog Window opens, specify the confidence level in the Confidence Level Box.
5. Click OK twice, and the confidence interval results will display in the Session Window.
Example 5.4.1: Consider the MINITAB\textsuperscript{TM} worksheet Cranksh.MTW located in the MINITAB\textsuperscript{TM} data folder. Recall that descriptive information for numerous MINITAB\textsuperscript{TM} work sheets can be accessed by choosing Help>Search Help>Contents>Sample Data Sets>Data set descriptions then clicking on the desired data set. Figure 5.3 provides a description of the data stored in Cranksh.MTW. The manager wants you to calculate a 95% confidence interval for the true distance from the actual position of the measuring point on the crankshaft to the baseline of the engine. The manager tells you that the historical standard deviation for the measurement is 3.5 mm.

Solution: Since the sample size \( n = 125 \) is relatively large, the formula in equation (5.4) will produce accurate results unless there are indications that the parent population is extremely skewed. A boxplot of the data in column AtoBDist is given in Figure 5.4 which appears fairly symmetrical without long tails. Since the standard deviation is known and the distribution is fairly symmetrical without long tails, it is reasonable to use the Z-interval to place a confidence interval around \( \mu \). The 1-Sample Z (Test and Confidence Interval) Dialog Window with appropriate selections to calculate a 95% confidence interval for \( \mu \) is shown in Figure 5.5 on the next page. The results from the Session Window are displayed in Output 5.6 on the following page. Note the limits for the 95% confidence interval extend from \(-0.172\) mm to \(0.055\) mm. In addition to showing the limits for the 95% confidence interval, the Session Window output also shows the variable name (AtoBDist), the number of observations (125), the mean (0.422), the standard deviation (3.491), and the standard error of the mean (0.313).

5.4.2 Sample Size Determination for Estimating \( \mu \)

Example 5.4.2: The sample size required to estimate \( \mu \) with a \((1 - \alpha) \times 100\%\) confidence interval so that the bound on the margin of error is no more than \( B \) is \( n = [Z_{1 - \alpha/2} \times \sigma / B]^2 \). Suppose your boss wants to know how many A to B measurements are required to be within 0.5 of the true value for \( \mu \) with a \((1 - \alpha) \times 100\%\) confidence interval.
5.4. Confidence Interval for a Population Mean, \( \mu \)

Solution: \( n = \left[ 1.95996 \times 3.5/0.5 \right]^2 = 188.231 \) To retain the same validity (95%), we need to take a random sample of 189 A to B measurements. MINITAB\textsuperscript{TM} macro \textbf{NMU} in Appendix E.2 on page 261 can be entered into your macro directory and invoked in the MINITAB\textsuperscript{TM} Session Window by typing \texttt{%NMU}. The macro computes the required sample size for a given confidence level when the user specifies the bound.

When macros are stored in the MINITAB\textsuperscript{TM}'s macro directory you do not need to provide a complete path name for MINITAB\textsuperscript{TM} to know the location of the macro nor add the extension \texttt{*.mac} to the macro name. However, if you are running the macro off a CD, you will need to specify the complete path to where the macro is found after typing \texttt{%} such as \texttt{%G:\HATforBSDA\MACROS\nmu.mac}.

The macro prompts and the resulting output generated in the Session Window by invoking macro \textbf{NMU} are displayed in Output 5.6. Note that the macro answer of 189 is in agreement with the answer calculated earlier.
5.4.3 The t Distribution

The development of the Z-interval relies on the assumption that the population standard deviation \( \sigma \) is a known quantity. However, \( \sigma \) is not typically known by researchers and is usually estimated from the sample. When sampling from a normal distribution, with a finite but unknown \( \sigma \), the random variable \( (\bar{x} - \mu)/(s/\sqrt{n}) \) follows a \( t \) distribution with \( n - 1 \) degrees of freedom. The \( t \) distribution is similar to the \( Z \) distribution. However, the \( t \) distribution has a different shape for each different value of its degrees of freedom. As the degrees of freedom (dof) in the \( t \) distribution increase, the shape of the \( t \) distribution becomes closer and closer to the shape of the \( Z \) distribution. In general, the \( t \) distribution has fatter tails than the \( Z \) distribution. See Figure 5.7 for an illustration of how the \( Z \) and \( t \) distribution compare to one another. From Figure 5.7 note that in order for the \( t \) distribution with 3 degrees of freedom to contain the top 2.5% of the distribution we have to go past 3. Next we address how we can calculate that critical \( t \) value that contains 2.5% of the area of the \( t \) distribution in its tails.

5.4.4 Critical Values Based on the \( t \) Distribution

The notation we use to denote a particular value in a \( t \) distribution is very similar to the notation used earlier with the standard normal (\( Z \)) distribution. As with the standard normal distribution, the subscripts \( \alpha/2 \) and \( 1 - \alpha/2 \) are used with the \( t \) distribution to denote \( \alpha/2 \) and \( 1 - \alpha/2 \) area to the left respectively of the particular \( t \) critical values. Since the shape of the \( t \) distribution is dependent on the degrees of freedom, we use an additional subscript not used with the standard normal distribution to denote the shape of the particular \( t \) distribution in question. The degrees of freedom for the one sample \( t \) distribution are written \( n - 1 \). The full notation to denote a particular value in a \( t \) distribution is written \( t_{1-\alpha/2;n-1} \). When one calculates a 95% CI, the \( \alpha \) value is obtained by solving \( (1 - \alpha) \times 100\% = 95\% \) for \( \alpha \). Note in this particular case that the \( \alpha \) value is 0.05. The critical value is found with MINITAB\textsuperscript{TM} by requesting the number from a \( t \) distribution with degrees of freedom such that 0.975 of the area is to the left of that value. Suppose the sample size of the particular problem of interest is \( n = 4 \). The Session Window commands to find \( t_{0.975;3} \) as well as \( t_{0.025;3} \) are illustrated in Output 5.7 on the following page. The command, \texttt{InvCDF} requests a value such that the area to the left of the requested value is 0.975, and \texttt{T 3.} specifies a \( t \) distribution with 3 degrees of freedom. Recall that the command \texttt{InvCDF} can be used with any distribution. The default distribution for the \texttt{InvCDF} command is the standard normal distribution. In other words, if one simply types \texttt{InvCDF .975} and presses the enter key without specifying the distribution, MINITAB\textsuperscript{TM} uses the standard normal distribution. Consequently, one must always remember to specify the distribution of interest.
5.4. Confidence Interval for a Population Mean, \( \mu \)

Output 5.7: Session Window Commands to find \( t_{0.975,3} \) and \( t_{0.025,3} \)

MTB > InvCDF .975:  # Requesting the value that has .975 area to its left.
SUBC>  T 3.  # In a t-distribution with 3 degrees of freedom.

Inverse Cumulative Distribution Function
Student's t distribution with 3 DF
\[
P( X \leq x ) \quad x
\]
\[
0.9750 \quad 3.1824
\]

MTB > InvCDF .025:  # Requesting the value that has .025 area to its left.
SUBC>  T 3.  # In a t-distribution with 3 degrees of freedom.

Inverse Cumulative Distribution Function
Student's t distribution with 3 DF
\[
P( X \leq x ) \quad x
\]
\[
0.0250 \quad -3.1824
\]

To calculate a critical value using the point and click approach, choose Calc>Probability Distributions>t. The t Distribution Dialog Window will open resembling Figure 5.8.

1. Click in the circle to the left of Inverse cumulative probability. The default value when the window opens is Cumulative probability.

2. Specify the degrees of freedom in the Degrees of Freedom Box.

3. Click in the circle to the left of Input constant and enter the area that is to the left of the value you wish to find.

4. Click OK. The result is displayed in the Session Window.

Figure 5.8: t Distribution Dialog Window

The t Distribution Dialog Window will also allow the user to specify a value in order to find the area to the left of the specified value. Suppose one wants to find the area to the left of the value \(-1.96\) in a \( t \) distribution with 9 degrees of freedom. From the t Distribution Dialog Window click in the circle to the left of Cumulative probability and fill in the remaining options as shown in Figure 5.9 on the following page. The Session Window commands to accomplish the same objective are illustrated in Output 5.7. Note that the command CDF ends with a semicolon. Also note that the subcommand always ends with a period. The CDF and InvCDF commands work in the same manner with any distribution. It is important to keep in mind that the value one specifies either in the Session Window or in the Dialog Window when using the InvCDF command is always a value between 0 and 1 since the area under a density function is always between 0 and 1.
5.4. Confidence Interval for a Population Mean, \( \mu \)

![Figure 5.9: t Distribution Dialog Window for area to the Left of \( -1.96 \)](image)

5.4.5 The \( t \) Interval

When the parent population is normally or near normally distributed with mean \( \mu \) and standard deviation \( \sigma \), then the sampling distribution of \( \bar{x} \) follows a normal distribution with mean \( \mu \) and standard deviation \( \sigma/\sqrt{n} \). A \((1 - \alpha) \times 100\%\) confidence interval for \( \mu \) based on \( \bar{x} \) when \( \sigma \) is unknown is given in equation (5.5). The one sample confidence interval for \( \mu \) based on \( \bar{x} \) when \( \sigma \) is unknown given in equation (5.5) will be referred to as the \( t \)-interval.

\[
\mathbb{P} \left( \bar{x} - t_{1-\alpha/2; n-1} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{1-\alpha/2; n-1} \frac{s}{\sqrt{n}} \right) = (1 - \alpha) \times 100\% \tag{5.5}
\]

To calculate a \((1 - \alpha) \times 100\%\) \( t \)-interval using MINITAB™,

1. Select Basic Statistics > 1-Sample t. The 1-Sample t (Test and Confidence Interval) Window will open and resemble Figure 5.10 on the following page.

2. Specify a variable for MINITAB™ to use in the Variables Box by either double clicking the variable of interest or clicking once on the variable of interest and subsequently clicking the Select Box.

3. Click on the Options button and specify the confidence level in the Confidence Level Box of the 1-Sample t-Options Dialog Window. Click the OK button twice and the results will display in the Session Window.

4. Note that the default value for confidence level is 95%. Consequently, if you desire a 95% CI you can simply Click OK once after step 2 and your results will display in the Session Window.

Although the \( t \) distribution is based on sampling from a normal distribution, the \( t \) distribution has been found to work in a wide variety of situations when the sampling distributions are not exactly normal.

The following general guidelines help ensure the appropriate use of the \( t \)-interval:

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Shape Of Underlying Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n &lt; 15 )</td>
<td>Normal</td>
</tr>
<tr>
<td>( 15 \leq n &lt; 30 )</td>
<td>Reasonably symmetric, no outliers</td>
</tr>
<tr>
<td>( n \geq 30 )</td>
<td>Reasonably symmetric, no extreme outliers or strong skewness</td>
</tr>
</tbody>
</table>
One of the better ways to assess whether or not it is reasonable to assume the underlying distribution is Normal is with a Normal probability plot. Although we can use graphical tools such as histograms, boxplots, and stem-and-leaf plots, to help assess the shape of a sample, such graphical procedures can be very misleading with small sample sizes. Video 5.4 uses probability plots and the *t*-Sample command to create a \((1 - \alpha) \times 100\%\) confidence interval for the average lipoprotein (LDL) cholesterol found in a group of quail fed a special diet.

**Example 5.4.3:** Open the MINITAB\textsuperscript{TM} worksheet **Cranksh.MTW**, which is located in the MINITAB\textsuperscript{TM} DATA subdirectory by choosing **File>Open Worksheet**. If the DATA subdirectory is not the default directory on your machine, move to the DATA subdirectory. The manager wants you to calculate a 95% confidence interval for the true distance from the actual position of the measuring point on the crankshaft to the baseline of the engine.

**Solution:** Recall from example 5.4.1 on page 130 that the underlying distribution for \(\text{AtoBDist}\) appeared fairly symmetrical with no outliers. Since the sample size for the variable \(\text{AtoBDist}\) exceeded thirty and exhibited a reasonably symmetric distribution without extreme outliers, we will use the *t*-interval to develop a 95\% confidence interval for the true mean distance from A to B. The *t*-interval formula is given below as a reminder.

\[
P\left( \bar{x} - t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}} \right) = (1 - \alpha) \times 100\%
\]

Three quantities are needed to calculate a 95\% confidence interval for the true mean distance from A to B. The sample mean, \(\bar{x}\), the sample standard deviation, \(s\), and the *t* critical value, \(t_{1-\alpha/2,n-1}\). The *Session Window* commands to calculate each of the three quantities \(\bar{x}\), \(s\), and \(t_{1-\alpha/2,n-1}\) as well as the lower CI limit (LLCI) and upper CI limit (ULCI) are illustrated in Output 5.8 on the next page. Output 5.8 also shows the results from requesting a 95\% CI using the *t*-Interval with the point and click approach. Note the agreement with both approaches.

### 5.5 Confidence Interval for a Population Median, \(\theta\)

The *Z*-interval and *t*-interval, covered in section 5.4, are used to find confidence intervals for the center of a population that is relatively symmetric without long tails. When the parent population is skewed or is symmetric with long tails, the population median provides a better estimate of center than does the population mean. In situations where the median is the estimate of the distribution’s center, confidence intervals for the median should be used instead of confidence intervals for the mean. Recall that the median of a continuous population is the value that divides the area under the density curve into two equal parts.
5.5. Confidence Interval for a Population Median, $\theta$

Output 5.8: $t$ Interval Limits

```
MTB > Name C4 = 'ybar'
MTB > Let 'ybar' = MEAN('AtobDist')
MTB > Name C5 = 'stdev'
MTB > Let 'stdev' = STDDEV('AtobDist')
MTB > Name CI = 'CT'
MTB > TnCDF .975 'CT';
SUBC> T 124.
MTB > Name C6 = 'LLCI'
MTB > Let 'LLCI' = 'ybar'-'CT' * 'stdev'/SQRT(125)
MTB > Name C7 = 'ULCI'
MTB > Let 'ULCI' = 'ybar'+'CT' * 'stdev'/SQRT(125)
MTB > print ct ybar stdev llci ulci

Data Display
CT  1.97912

Row  ybar  stdev  LLCI  ULCI
1  0.441704  3.49136  -0.176379  1.05979
```

```
MTB > OneT 'AtobDist';
SUBC> Confidence 95.

One-Sample T: AtobDist

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>95.0% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtobDist</td>
<td>125</td>
<td>0.442</td>
<td>3.491</td>
<td>0.312</td>
<td>(-0.176, 1.060)</td>
</tr>
</tbody>
</table>
```

5.5.1 Small Sample Confidence Interval For The Population Median

The confidence interval we give for the population median is based on the binomial distribution. Assume that $X_1, X_2, \ldots, X_n$ are a random sample of $n$ observations drawn from a continuous population with unknown median $\theta$. The confidence interval endpoints are simply the $k$th and $(n-k+1)$st order statistics of the sample. Order statistics are the sample data rearranged in order of relative magnitude, denoted by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The specific confidence level is based on the binomial distribution where the confidence level is given by

$$1 - 2 \times P(X < k)$$

where

$$P(X = x) = \binom{n}{x} (\frac{1}{2})^x (\frac{1}{2})^{n-x}$$

(5.6)

Example 5.5.1: Consider calculating an approximate 95% confidence interval for the population median with the following sample data: $(3, 6, 3, 2, 5, 2, 4, 2, 5, 1, 3, 1, 1, 2, 2, 1, 4, 0)$.

Solution: The first thing we do is order the sample. Following are the ordered numbers: $0, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 5, 5, 6$. Next, we determine $k$ such that $1 - 2 \times P(X < k) \approx 0.95$. Instead of calculating $P(X = x) = \binom{n}{x} (\frac{1}{2})^x (\frac{1}{2})^{n-x}$ by hand, we use MINITAB™ to generate a table of values when $n = 18$ and $\pi = 0.5$. The MINITAB™ Session Window commands and the subsequent Session Window output are illustrated in Output 5.9 on the following page.

From Output 5.9 we see that $P(X < 5) = P(X \leq 4) = 0.0154$ and $P(X < 6) = P(X \leq 5) = 0.0481$. Note that $1 - 2 \times P(X < 5) = 0.9692$ and $1 - 2 \times P(X < 6) = 0.9038$. Consequently, a 96.92% confidence interval for $\theta$ would be $X_{(5)} < \theta < X_{(14)}$ or $1 < \theta < 4$, while a 90.38% confidence interval for $\theta$ would be $X_{(6)} < \theta < X_{(13)}$ or $2 < \theta < 3$.

To solve the same problem using MINITAB™ choose Stat>Nonparametrics> 1-Sample Sign. Select the appropriate options in the 1-Sample Sign Dialog Window as shown in Figure 5.11 on the next page. Then, click OK.

The output generated in the Session Window after clicking OK is displayed in Output 5.10 on the following page.

Note how the achieved confidence displayed in Output 5.10 and the corresponding confidence level obtained using equation (5.6) are in agreement out to 3 decimal places. The fourth decimal differs by 0.0001 in each calculation and is due to MINITAB™'s rounding of the values displayed in Figure 5.9 on the next page. If we do not round the values in Figure 5.9, the answers are in exact numerical agreement.
Output 5.9: Output for Median CI Calculation

MTB > cdf;
SUBC> bin 10 .5.

Cumulative Distribution Function

Binomial with n = 18 and p = 0.500000

<table>
<thead>
<tr>
<th>x</th>
<th>P(X ≤ x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>0.0001</td>
</tr>
<tr>
<td>2</td>
<td>0.0007</td>
</tr>
<tr>
<td>3</td>
<td>0.0038</td>
</tr>
<tr>
<td>4</td>
<td>0.0154</td>
</tr>
<tr>
<td>5</td>
<td>0.0481</td>
</tr>
<tr>
<td>6</td>
<td>0.1189</td>
</tr>
<tr>
<td>7</td>
<td>0.2403</td>
</tr>
<tr>
<td>8</td>
<td>0.4073</td>
</tr>
<tr>
<td>9</td>
<td>0.5527</td>
</tr>
<tr>
<td>10</td>
<td>0.7597</td>
</tr>
<tr>
<td>11</td>
<td>0.8811</td>
</tr>
<tr>
<td>12</td>
<td>0.9519</td>
</tr>
<tr>
<td>13</td>
<td>0.9846</td>
</tr>
<tr>
<td>14</td>
<td>0.9962</td>
</tr>
<tr>
<td>15</td>
<td>0.9993</td>
</tr>
<tr>
<td>16</td>
<td>0.9999</td>
</tr>
<tr>
<td>17</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Figure 5.11: 1-Sample Sign Dialog Window for Median CI

Output 5.10: Output of Binomial for Median CI Calculation

MTB > Sinterval 95 CI.

Sign CI: CI

Sign confidence interval for median

<table>
<thead>
<tr>
<th>CI</th>
<th>N</th>
<th>Median</th>
<th>Achieved Confidence</th>
<th>Confidence interval</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI</td>
<td>10</td>
<td>2.0000</td>
<td>0.9037 (2.000, 3.000)</td>
<td>6</td>
<td>0.9500 (1.518, 3.482)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9691 (1.000, 4.000)</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
5.5. Confidence Interval for a Population Median, θ

5.5.2 Large Sample Confidence Interval For The Population Median

For samples where \( n > 20 \), the value of \( k \), the number of the order statistic such that \( 1 - 2 \times P(X < k) \approx 1 - \alpha \), based on the normal approximation to the binomial distribution is given by:

\[
k = \left( \frac{n + 1 - Z_{1-\alpha/2} \times \sqrt{n}}{2} \right)
\]

(5.7)

Note that the 1 in the numerator is added as a continuity correction for the discrete distribution. The following sample of size 40 is used to calculate a 95% confidence interval for the median when the normal distribution is used to approximate the binomial distribution: \{0, 1, 0, 1, 1, 1, 2, 5, 1, 4, 0, 2, 1, 1, 1, 2, 3 3, 1, 2, 3, 0, 1, 0, 2, 0, 1, 0, 1, 0, 1, 3, 1 2, 3, 1, 1\}. The value of \( k \) is determined to be \( k = \left( \frac{40 + 1 - 1.96\sqrt{40}}{2} \right) = 14.3 \).

As we can see, the value of \( k \) is not generally an integer. To be conservative, the value of \( k \) is truncated to get the next smallest integer. Consequently, for \( n = 40 \), and 95% confidence, the endpoints are the observed values \( X_{(14)} \) and \( X_{(27)} \), where 27 is \( 40 - 14 + 1 = n - k + 1 \). Note that the values corresponding to \( X_{(14)} \) and \( X_{(27)} \) are 1 and 2 respectively. The output from using MINITAB to calculate exact confidence intervals with the same data set is given in Output 5.11. Note the close agreement between the endpoints for the normal approximation to the binomial and the endpoints based on the binomial distribution. Video 5.5 shows a manual procedure for calculating median confidence intervals when working with summarized data. The data is also stored in a form so that it is possible to use MINITAB’s 1-Sample Sign command to create a median confidence interval. Video 5.6 works with the rearranged data from Video 5.5 and calculates confidence intervals for both mean and median lengths of fish caught with both small mesh and large mesh codends.

Video 5.5: For Problem 5.83 — (Duration: 5 minutes 11 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

Video 5.6: For Problem 5.119 — (Duration: 4 minutes 43 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.
5.6 Summary and Review Labs

Lab 5.1 — Estimating Certain Population Parameters #1

Objectives:

I. To verify empirically that \( s^2 \) underestimates \( \sigma^2 \)

II. To determine the constant by which \( s^2 \) underestimates \( \sigma^2 \)

III. To verify that \( s \) is an unbiased estimator of \( \sigma^2 \)

Basic Directions:

Append all 9 graphs to your report pad. Answer the questions at the end of the lab in complete sentences in the report pad.

Introduction:

There are several possible statistics that could be used to measure the variation in a set of normally shaped data. You will examine two of those statistics in this lab, \( s^2 \) and \( s^2 \). This lab will direct you in an empirical verification that \( s^2 \) underestimates \( \sigma^2 \). This makes \( s^2 \) a biased estimator of \( \sigma^2 \). Further, at the completion of the lab, you should be able to determine correctly the constant by which \( s^2 \) underestimates \( \sigma^2 \). You will also empirically verify that \( s \) is an unbiased estimator of \( \sigma^2 \).

Definitions:

\[
s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n \sum_{i=1}^{n} x_i^2}{n} - \frac{\left( \sum_{i=1}^{n} x_i \right)^2}{n}
\]

\[
s = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{\left( \sum_{i=1}^{n} x_i \right)^2}{n-1}
\]

\[
\chi^2_{n-1} \sim \frac{(n-1)s^2}{\sigma^2}
\]

Experiment:

1. Generate 20,000 samples of size \( n = 4 \) from a normal distribution with \( \mu = 100 \) and \( \sigma = 10 \) (Calc>Random Data>Normal).

2. Calculate the quantities \( s^2 \) and \( s^2 \) and store the results in columns named \text{var} and \text{varstar} respectively. \textbf{Hint:} First calculate \( \sum x_i \) and \( \sum x_i^2 \) using using Calc>Row Statistics. See Figure 5.12 on the next page for an example of how to calculate \( \sum x_i \). Finally, use MINITAB’s calculator to calculate the quantities \( s^2 \) and \( s^2 \) making use of the values of \( \sum x_i \) and \( \sum x_i^2 \).

3. Calculate the mean of columns \text{var} and \text{varstar} and record your results.

4. Produce histograms of columns \text{var} and \text{varstar}. Title the histograms, “Histogram of VAR” and “Histogram of VARSTAR” respectively. In a right justified footnote, indicate the sample size used (4, 5, or 10) for each histogram as well as the simulated sampling distributions’ means. (The answer from step 3.)
Figure 5.12: Example of How to Calculate $\sum x_i$

5. Multiply the column that contains your 20,000 $s^2$ values (var) by $(n - 1)/\sigma^2$ and store the result in a column named SCS# (Simulated Chi-Square # Degrees of Freedom, where # = 3, 4, or 9). Recall that $\sigma = 10$ or $\sigma^2 = 100$ and $n$ is the sample size used for your simulations.

6. Determine the mean and variance of the column SCS#, where # indicates the appropriate degrees of freedom 3, 4, or 9.

7. Determine the 95th percentile of your Monte Carlo simulated Chi-Square distribution with $(n - 1)$ degrees of freedom and record your result. Hint: Use Manip>Sort to order the values in SSC# and record the 20,000 $\times$ .95 = 19,000th largest ordered value.

8. Produce a histogram of SSC#. Title your Histogram “Simulated Chi-Square Distribution”. One the second line of the title, indicate the degrees of freedom for your simulated Chi-Square Distribution. **Note: degrees of freedom = sample size−1.** In other words, for the simulated Chi-Square distribution created from samples of size 4, you would type “3 Degrees of Freedom”. Use a right justified footnote to display the mean and variance calculated from step 6 for that particular distribution “mean = ______, variance = ______”. On the second line of the right justified footnote record the value for the 95th percentile of that distribution. See Figure 5.13 for an example.

Figure 5.13: Example Histogram for Simulated Chi-Square Distribution

9. Repeat the experiment using samples of size 5 and 10 (Step 1). Run each sample size three times, and record all of your results.
5.6. Summary and Review Labs

**Note:** It will probably be easiest to run the experiment with samples of size 4 three times, then run the experiment with samples of size 5 three times, and finally run the experiment with samples of size 10 three times. In step 4 you should change the name of the column where you store your values to SCS4 and SCS9 when using samples of size 5 and 10 respectively.

**Questions:**

1) By what factor does $s^2$ underestimate $\sigma^2$? We are looking for a general expression for the underestimation factor.

2) What happens to the Monte Carlo simulated Chi-Square distribution as $n$ increases?

3) What is the relationship between the degrees of freedom for the Monte Carlo simulated Chi-Square distribution and the mean and variance of the Monte Carlo simulated Chi-Square distribution?

4) Is $s^2$ an unbiased estimator of $\sigma^2$ based on your histogram of the $s^2$ values?
Lab 5.2 — Estimating Certain Population Parameters #2

Objective:

To discover what estimator to use for the number of OSP's \( (N) \) that have been manufactured.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Selection of an appropriate estimator is not always clear. In many instances, the researcher will need to evaluate several potential estimators. Many researchers select estimators that are **UMVUE** (Uniform Minimum Variance Unbiased Estimators). For now, we will consider "good" estimators to be UMVUE. Consider the following scenario where the choice of the estimator is not immediately clear. The Walther OSP is a high end 0.22 caliber target pistol. At a recent international pistol match, a curious gun collector asks one of the competitors if he knows how many Walther OSP's are in existence. Unfortunately, the competitor only knows the OSP is expensive! The gun collector records the serial numbers for the Walther OSP's present at the pistol match and asks you to determine how many Walther OSP's have been manufactured. The serial numbers from the OSP's are as follows: 0122, 0476, 0636, 0862, 0218, 0412, 0915, 0802, 0987, 0170, 0036, 0531, 0403, 0980, 0505, 0786, 0868, 0314, and 0384. You start your mission by calling the manufacturing company. Unfortunately, they will not release how many OSP's have been manufactured. However, they do admit that the serial numbers for the OSP's started at 0001 and that the serial numbers are sequential.

Consider the following three estimators for \( N \): **MAXIMUM** observed value, **MAXIMUM**+**MINIMUM**−1 observed value, and **2×MEDIAN**−1 observed value. The first question that should be addressed is which, if any, of the estimators are unbiased. From the collection of unbiased estimators, we will select the estimator with smallest variance. To help in the assessment of each of the estimators, Monte Carlo Simulation will be used to generate 20,000 samples of size 20 from a discrete uniform distribution with lower endpoint 1 and upper endpoint 1000 \( \{ \text{discrete } U(1,1000) \} \). In other words we are assuming that 1,000 OSP's have been produced, and we will try to find estimators which return values close to 1,000. Of the estimators that are unbiased (mean \( \approx 1,000 \)), we will select the estimator with smallest variance as the best. To generate 20,000 samples of size 20 from a discrete \( U(1,1000) \), choose **Calc>Random Data>Integer** and fill in the boxes of the Integer Distribution Dialog Window as shown in Figure 5.14.

Figure 5.14: Integer Distribution Dialog Window for OSP Simulation

Note that we started with 1 instead of 0 since we know at least 1 OSP has been manufactured. To create the various estimates for the 20,000 samples of size 20, we will use the row statistics command. For example, to
create the estimator \( \text{MAXIMUM} \), select \text{Calc} \rightarrow \text{Row Statistics} \) and fill in the Row Statistics Dialog Window as shown in Figure 5.15.

To create the estimator \( (\text{MAXIMUM} + \text{MINIMUM} - 1) \), use the row statistics command to create a column containing the minimum value for each of the 20,000 samples of size 20. Next, use MINITAB’s calculator to create a column that is the result of adding columns \( \text{MAXIMUM} \) and \( \text{MINIMUM} \) and subtracting 1.

Questions and Directions:

1. Create columns with values for the three estimators \( \text{MAXIMUM} \), \( (2 \times \text{MEDIAN} - 1) \), and \( (\text{MAXIMUM} + \text{MINIMUM} - 1) \), from the 20,000 samples of size 20 from a discrete uniform distribution \((1,1000)\).

2. Create two additional estimators in addition to the three estimators given in step 1. Hint: Consider calculating the average distance between observations and then adding that value to the maximum observed value in the sample for one of your estimators.

3. Use \text{Stat} \rightarrow \text{Basic Statistics} \rightarrow \text{Display Descriptive Statistics} \) to compute the descriptive statistics for the five columns of estimates created in steps 1 and 2. Append the results of the descriptive statistics to the report pad.

4. Create histograms for the simulated sampling distribution for each of your five estimators. Set the minimum and maximum values on the \( x \)-axis to 500 and 1500 respectively for all graphs. Provide appropriate titles for all graphs, and place your full name in right justified footnotes. (To set the min and max values of a histogram, click on the Frame Drop Down Menu, select Min and Max and fill in the appropriate values in the Min and Max Dialog Window.)

5. Which of the five estimators are unbiased, and which are biased? Justify your answers with numbers.

6. From the unbiased estimators, select the estimator with the smallest variance. Use this estimator with the serial numbers: 0122, 0476, 0636, 0674, 0862, 0218, 0412, 0915, 0802, 0987, 0170, 0036, 0531, 0403, 0980, 0505, 0786, 0868, 0314, and 0384 to provide the gun collector with an answer to his original question. Show all work and report your answer in the report pad.

7. What is the shortest 95% confidence interval you can report to the gun collector? Hint look at the 2.5\(^{th} \) and 97.5\(^{th} \) percentiles of the simulated distribution for the best estimator.

8. When will the answer you reported to question 7 and the point estimate (the value you reported in step 6) + and − 2 times the standard deviation of the best estimator coincide, and when will they differ? Hint: Look at your best estimators’ simulated sampling distribution.

EXTRA CREDIT: Why did we subtract 1 from the estimator 2 times the median?
Lab 5.3 — Confidence Interval for a Bernoulli Proportion, $\pi$

Objective:

To understand the concept of a confidence interval in the context of a Bernoulli Proportion

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Suppose we actually know the proportion of college students that exercise three or more times a week is 40%. Use simulation to generate 10,000 random samples, each of size 500, and then calculate, for each sample, the proportion of the 500 who exercise three or more times per week. Use the simulated sample proportions to construct 95% confidence intervals for $\pi$. Once you have the 10,000 confidence intervals, count the number of intervals that actually contain $\pi$. Append your contingency table to the report pad and answer all the questions at the end of the lab in complete sentences. The following directions should get you started.

Questions and Directions:

1. To simulate the number of successes out of 500 Bernoulli trials, generate 10,000 binomial random variables with $n = 500$ and $\pi = 0.4$ using the Calc>Random Data>Binomial command. Store the results in a column named BINO. See Figure 5.16 for an example.

2. Divide the 10,000 results in column BINO by 500 to obtain the simulated proportion of successes using MINITAB’s calculator (Calc>Calculator). Store the result in a column named $p$. See Figure 5.17 on the following page for an example.

3. Determine the lower CI limit for each of the 10,000 samples using MINITAB’s calculator. Store the results in a column named LL. The lower CI is determined with formula (5.1).

4. Determine the upper CI limit for each of the 10,000 samples using MINITAB’s calculator. Store the results in a column named UL. The upper CI is determined with formula (5.1).

5. Code all of the values in column LL between 0 and 0.4 to a 1 and values between 0.4 and 1 to a 0. Store the results in a column named CLL.

6. Code all of the values in column UL between 0.4 and 1 to a 1 and values between 0 and 0.4 to a 0. Store the results in a column named CLL.
7. By following the coding scheme in step 5 and 6, any time a row in CLL and CUL contains both 1s, we know the confidence interval has included 0.4.

8. Use the MINITAB™ command `Stat > Tables > Cross Tabulation` to count the number of intervals that contain both 1s. Append this table to your report pad.

9. How many of your simulated intervals contained the value 0.4?

10. Based on your Cross Tabulation table, is the mean for your simulated distribution greater than or less than 0.4? Explain your reasoning in your answer. **Hint:** Look at the number of times the confidence interval did not contain the true parameter. Specifically, which end of the confidence interval failed to contain the true parameter most often.
Lab 5.4 — Confidence Interval for a Population Mean, $\mu$

Objective:

To understand the concept of a confidence interval in the context of a Population Mean

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

The sampling distribution of $\bar{x}$ when sampling from a normal distribution is given by: $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$. From the sampling distribution of $\bar{x}$, we developed confidence intervals for $\mu$ when $\sigma$ is both known and unknown. The confidence interval for $\mu$ when sampling from a normal distribution with known $\sigma$ is found in equation (5.4), while the confidence interval for $\mu$ when sampling from a normal distribution with unknown $\sigma$ is found in equation (5.5).

When calculating a confidence interval, the interval will not always contain the parameter we are trying to estimate. In fact, when we calculate a 95% confidence interval, what we are is 95% confident about the process that produces confidence intervals. In other words, the process will produce intervals that contain the true parameter around 95% of the time. The remaining 5% of the time, the process is in error. The goal of this lab is to illustrate how 95% confidence intervals capture the true parameter, $\mu$, about 95% of the time.

Questions and Directions:

You will generate 1,000 samples of size 9 from a normal distribution with $\mu = 100$ and $\sigma = 30$. Compute 95% confidence intervals for each of the 1,000 samples first by assuming $\sigma = 30$ and then by assuming only that $\sigma$ is finite. When you finish, count how many of the confidence intervals contain the true parameter ($\mu = 100$).

Repeat the experiment with 10,000 samples of size 9 from a normal distribution with $\mu = 100$ and $\sigma = 30$. Create a table that shows how many times the true parameter was and was not contained in the 95% confidence interval.

1. Determine the critical values $Z_{1-\alpha/2}$ and $t_{1-\alpha/2; n-1}$ for a 95% confidence interval with $n = 9$.
2. Store the $Z_{1-\alpha/2}$ value in a constant named $Z$. Store the $t_{1-\alpha/2; n-1}$ value in a constant named $t$.
3. Generate 1,000 samples of size $n = 9$ from a normal distribution with a mean of 100 ($\mu = 100$) and standard deviation of 30 ($\sigma = 30$) using the MINITAB™ commands Calc>Random Data>Normal. See Figure 5.18 on the next page for an example.
4. Use Calc>Row Statistics to calculate the mean of columns C1-C9 and store the results in a column named xbar. Note: The dash used above does not indicate subtraction. Rather C1-C9 indicates column 1 (C1) through column 9 (C9).
5. Use Calc>Row Statistics to calculate the standard deviation of columns C1-C9 and store the results in a column named STDEV.
6. Use Calc>Calculator to calculate the confidence interval’s lower limit for each of the 1,000 samples of size 9 and store the results in a column named LLZ (Lower Limit Z). Note: The lower limit for each confidence interval is computed using the left side of formula (5.4). See Figure 5.19 on the following page for an example.
7. Use Calc>Calculator to calculate the confidence interval’s upper limit for each of the 1,000 samples of size 9 and store the results in a column named ULZ (Upper Limit Z). Note: The upper limit for each confidence interval is computed using the right side of formula (5.4).
8. Use Manip>Code>Numeric to Text to code all the values in LLZ and ULZ. Specifically, code all values in LLZ below 100 to In and values in LLZ above 100 to Out. Store the results in a column named CLLZ for coded lower limit Z. Code all values in ULZ above 100 to In and values in ULZ below 100 to Out. Store the results in a column named CULZ for coded upper limit Z. See Figure 5.20 on the next page for an example of how to code the lower limit.
9. Use the MINITAB™ command Stat>Tables>Cross Tabulation to count the number of times the confidence intervals created using equation (5.4) actually contained the true parameter ($\mu = 100$). See Figure 5.21 for an
5.6. Summary and Review Labs

Figure 5.21: Cross Tabulation Dialog Window for Confidence Interval

example of the Cross Tabulation Dialog Window.

10. Append your results from step 9 on the page before to the report pad.

11. Use Calc>Calculator to calculate the confidence interval’s lower limit for each of the 1,000 samples of size 9 and store the results in a column named LLT (Lower Limit t). Note: The lower limit for each confidence interval is computed using the left side of formula (5.5).

12. Use Calc>Calculator to calculate the confidence interval’s upper limit for each of the 1,000 samples of size 9 and store the results in a column named ULT (Upper Limit t). Note: The upper limit for each confidence interval is computed using the right side of formula (5.5).

13. Use Manip>Code>Numeric to Text to code all the values in LLT and ULT. Specifically, code all values in LLT below 100 to In and values in LLT above 100 to Out. Store the results in a column named CLLT for coded lower limit t. Code all values in ULT above 100 to In and values in ULT below 100 to Out. Store the results in a column named CULT for coded upper limit t.

14. Use the MINITAB™ command Stat>Tables>Cross Tabulation to count the number of times the confidence intervals created using equation (5.5) actually contained the true parameter (μ = 100).

15. Append the results from step 14 to the report pad.

16. Repeat steps 3 through 15 using 10,000 samples instead of 1,000.

17. For each of the four tables appended in the report pad, highlight the number and the percentage of confidence intervals actually containing the true parameter (μ = 100) by changing the font to bold red.

18. To change font color or appearance, first highlight the number you would like to change by holding down your left mouse button as you pass over the number. Right click with your mouse and select Font from the pop up window that appears inside the report pad. Select Bold from the Font Style Box and Red from the Drop Down Color Menu as shown in Figure 5.22 on the following page.

19. Were all of the red bolded percentages from step 18 close to 95%? If not, should they have been close to 95%?

20. If you performed this lab with data generated from a skewed distribution that has a mean of 100, would approximately 95% of your confidence intervals still contain the true parameter μ = 100? If you are not sure of the answer, try the experiment using the exponential distribution with a mean of 100 instead of the normal distribution with a mean of 100. Note: The mean (μ) for the exponential distribution is equal to its standard deviation (σ).

21. Make sure you have your Name, class, date, and Lab 5.4 at the top of your report pad. Save your work to either a disk or a hard drive.
Figure 5.22: Font Dialog Window
Lab 5.5 — Confidence Interval for a Population Median, $\theta$

Objectives:

I. To provide non-trivial opportunities for the student to practice computing probabilities using the binomial distribution.

II. To have the student become familiar with the rationale for creating confidence intervals for the median.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Recall that the confidence interval endpoints are simply the $k^{th}$ and $(n - k + 1)^{st}$ order statistics of the sample and that the specific confidence level for $\theta$ is based on the binomial distribution where the confidence level is given by equation (5.6).

When finding the value of $k$ such that $1 - 2 \times P(X < k) \approx (1 - \alpha) \times 100\%$, it should be noted that $k$ is a discrete value and we will seldom have a confidence interval with an exact $(1 - \alpha) \times 100\%$ confidence level. Without resorting to some type of interpolation, we can either select the value of $k$ such that $1 - 2 \times P(X < k) \geq (1 - \alpha) \times 100\%$, or we can pick the value of $k$ such that $1 - 2 \times P(X < k)$ is a close as possible to the desired $(1 - \alpha) \times 100\%$ confidence level. For example, suppose $n = 17$ and we are looking for the $k^{th}$ and $(n - k + 1)^{st}$ order statistics to construct a 96% interval for $\theta$. Consider the results when $k = 5$ and when $k = 4$. When $k = 5$, $P(X < 5) = P(X \leq 4) = 0.0245209$ and the 5th and the 13th ordered statistics provided the lower and the upper values for a $1 - 2 \times 0.0245209 = 95.0958\%$ confidence interval. If we choose $k = 4$, $P(X < 4) = P(X \leq 3) = 0.00636292$ and the 4th and the 14th ordered statistics provided the lower and the upper values for a $1 - 2 \times 0.00636292 = 98.7274\%$ confidence interval. See Output 5.12 for Session Window commands and code comments.

Output 5.12: Calculation of Median CI

```
MTB > cdf 4 k1;  # $P(X\leq4)$ and store result in k1
SUBC> bino 17 .5.  # Based on X-bin(17, 0.5)
MTB > print k1

Data Display

K1  0.0245209  # $P(X\leq4)=P(X<5)=0.0245209$

MTB > let k2=1-2*k1
MTB > print k2

Data Display

K2  0.950958  # Achieved confidence level for km5

MTB > cdf 3 k1;  # $P(X\leq3)$ and store result in k1
SUBC> bino 17 .5.  # Based on X-bin(17, 0.5)
MTB > print k1

Data Display

K1  0.00636292  # $P(X\leq3)=P(X<4)=0.00636292$

MTB > let k2=1-2*k1  # Based on X-bin(17, 0.5)
MTB > print k2

Data Display

K2  0.997274  # Achieved confidence level for k=4
```
Clearly 95.0958% is closer to the desired 96% than is 98.7274%. However, there may be some drawbacks by claiming a 96% confidence interval when the achieved confidence level is actually 95.0958%.

**Directions:**

Using the two different methods to find the $k^{th}$ order statistic, fill in the missing values for Tables 5.1 and 5.2 that follow.

### Table 5.1: Order Statistic 92% Confidence Levels

<table>
<thead>
<tr>
<th>Sample Size ($n$)</th>
<th>Order Statistic to have at least a 92% confidence level</th>
<th>Achieved Confidence Level for Previous Column</th>
<th>Order Statistic to be closest to a 92% confidence level</th>
<th>Achieved Confidence Level for Previous Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 5.2: Order Statistic 96% Confidence Levels

<table>
<thead>
<tr>
<th>Sample Size ($n$)</th>
<th>Order Statistic to have at least a 96% confidence level</th>
<th>Achieved Confidence Level for Previous Column</th>
<th>Order Statistic to be closest to a 96% confidence level</th>
<th>Achieved Confidence Level for Previous Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>98.7274</td>
<td>5</td>
<td>95.0998</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 6
Hypothesis Testing

6.1 Introduction to Hypothesis Testing

A hypothesis test typically involves two hypotheses. These hypotheses are called the null hypothesis and the alternative hypothesis and are denoted $H_O$ and $H_A$ respectively. In a test of the null hypothesis, data is collected to assess evidence against the null hypothesis. Consequently, the null hypothesis is assumed to be true until evidence is shown to suggest otherwise. Researchers often choose the alternative hypothesis to be what they suspect or hope is true since the null hypothesis is often the status quo. We express the null hypothesis as

$$H_O : \psi = \psi_0.$$ 

Note that $\psi$ is the parameter being tested while $\psi_0$ is some number. The alternative hypothesis can assume one of three forms. When the researcher has reason to believe the results of an experiment may go in a particular direction, the alternative hypothesis becomes a one tailed hypothesis known as a directional hypothesis. A directional hypothesis specifies that the parameter is either less than ($<$) a particular value or greater than ($>$) a particular value. Examples of directional hypotheses are:

$$H_A : \psi < \psi_0$$
$$H_A : \psi > \psi_0.$$ 

When the researcher’s primary focus is detecting a difference from the hypothesized parameter, $\psi$, a two tailed alternative hypothesis is used. The two tailed alternative hypothesis uses the not equal, $\neq$, sign. An example of a two tailed alternative hypothesis is:

$$H_A : \psi \neq \psi_0.$$ 

6.1.1 Choosing Test Statistics

The evidence against the null hypothesis is assessed using a test statistic. A test statistic is a numerical quantity, calculated from sample data, used to assess the evidence against the null hypothesis. The sample mean, $\bar{x}$, is frequently used as a test statistic when testing hypotheses that involve the population mean, $\mu$. Many texts refer to the standardized form of the test statistic as the test statistic. The reason test statistics are converted to standardized forms is so they will have known sampling distributions. Known sampling distributions let us determine which standardized test statistic values are unlikely. Once we know where these unlikely values occur, we can set our rejection regions.
6.1.2 Working with \( p \)-Values

Instead of rejecting or failing to reject the null hypothesis, \( p \)-values are often reported. A \( p \)-value is the probability of observing a value of the test statistic as extreme (inconsistent) or more extreme than that given by the actual sample data under the assumption that the null hypothesis is true. Recall that the alternative hypothesis may take one of three forms. When the alternative hypothesis takes the form \( H_A : \psi > \psi_0 \), the \( p \)-value is found by calculating \( P(\psi \geq \psi_{obs}) \). When the alternative hypothesis takes the form \( H_A : \psi < \psi_0 \), the \( p \)-value is found by calculating \( P(\psi \leq \psi_{obs}) \). Finally, when the alternative hypothesis takes the form \( H_A : \psi \neq \psi_0 \), the \( p \)-value is calculated as \( 2 \times P(\psi \geq \psi_{obs}) \) or \( 2 \times P(\psi \leq \psi_{obs}) \), whichever is less than 1; or, if both are greater than 1, as sometimes occurs with discrete distributions, the \( p \)-value equals 1. If a \( p \)-value is small, we interpret this to mean the observed sample produced a result that is very rare (inconsistent) under the assumption the null hypothesis is true. When the \( p \)-value is large, we conclude that the sample result is consistent with the null hypothesis. In other words, small \( p \)-values indicate samples that are inconsistent with the null hypothesis which lead us to make the statistical conclusion to “reject the null hypothesis” in favor of the alternative hypothesis. Large \( p \)-values indicate samples that are consistent with the null hypothesis and consequently we make the statistical decision to “fail to reject the null hypothesis.” Failure to reject \( H_O \) does not imply \( H_O \) is true. It simply indicates that the experimental results are not inconsistent with the null hypothesis. Although evidence that leads one to reject \( H_O \) is fairly conclusive evidence that \( H_O \) is not true, failure to reject \( H_O \) does not provide conclusive evidence that \( H_O \) is true.

6.1.3 5 Step Procedure For Testing Hypotheses

The following procedures are recommended for solving \textit{test of hypothesis}-type problems. The procedures allow others to follow the steps one takes in reaching a statistical decision.

1. State the null and alternative hypotheses.
2. Select an appropriate test statistic.
3. Determine the sampling distribution of the standardized test statistic under the assumption that the null hypothesis is true.
4. Calculate the \( p \)-value and determine if the evidence warrants rejecting the null hypothesis.
5. State in plain English what the conclusion you reached in step 4 means.

6.1.4 Type I and II Errors

Any time a data analyst rejects a null hypothesis, he risks committing a \textit{type I error}. When the data analyst fails to reject the null hypothesis, he risks committing a \textit{type II error}. The relationship between type I and type II errors is shown in Table 6.1.

<table>
<thead>
<tr>
<th>Decision</th>
<th>Reject ( H_O )</th>
<th>Fail To Reject ( H_O )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null Hypothesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>Type I Error</td>
<td>Correct Decision</td>
</tr>
<tr>
<td></td>
<td>( P(\text{Type I Error}) = \alpha )</td>
<td>( P(\text{Correct Decision}) = 1 - \alpha )</td>
</tr>
<tr>
<td></td>
<td>(Level of Significance)</td>
<td></td>
</tr>
<tr>
<td>False</td>
<td>Correct Decision</td>
<td>Type II Error</td>
</tr>
<tr>
<td></td>
<td>( P(\text{Correct Decision}) = 1 - \beta )</td>
<td>( P(\text{Type II Error}) = \beta )</td>
</tr>
<tr>
<td></td>
<td>(Power of the Test)</td>
<td></td>
</tr>
</tbody>
</table>

If the researcher fails to reject the null hypothesis when the null hypothesis is true, note that no error is committed. Specifically, the correct decision should be reached in roughly \((1 - \alpha) \times 100\%\) of all trials. Using the same logic, approximately \((1 - \beta) \times 100\%\) of the times sample data is evaluated in a test of hypothesis, a false null hypothesis will be rejected.
6.1.5 The Power of a Test

The ability of a test statistic to reject the null hypothesis when the null hypothesis is false is known as the **Power of a test**. Specifically, suppose that $\psi_{\text{obs}}$ is the test statistic and $RR$ is the rejection region for a test of hypothesis involving the value of a parameter $\psi$. Then, we say the power of the test is the probability that the test will lead to rejection of the null hypothesis when the true value of the parameter is $\psi$. In other words, $\text{power}(\psi) = P(\psi_{\text{obs}} \in RR | \text{parameter} = \psi)$. Loosely defined, power is the probability of detecting differences when differences exist. Power can be increased in three fashions. The first way to increase power is to increase the $\alpha$ level of the problem. This solution is not very desirable and is seldom a viable solution for increasing power. The second way to increase power is to decrease the variance of the sample with which one is working. This is seldom a viable solution either. The final way to increase power is to increase sample size. Many fields of study mandate researchers conduct power analyses prior to conducting research and or receiving external funding. A simple power analysis allows the researcher to determine the probability of detecting differences with different sample sizes. Power studies are especially critical when resources are rare or extremely costly. The graph in Figure 6.1 illustrates $\alpha$, $\beta$, and power for testing $H_O : \mu = 100$ versus $H_A : \mu > 100$, with a 5% significance level, when the true value for the mean is 105. It is important to note that the decision rule to either reject or fail to reject the null hypothesis is made under the assumption the null hypothesis is true. Consequently, the area labelled $\alpha$ in Figure 6.1 indicates an area where the null hypothesis would be rejected 5% of the time under the assumption that the null is true if we were to conduct this particular test using the same procedure a large number of times. Since power is the probability of rejecting the null hypothesis when the null hypothesis is false, the area in yellow in Figure 6.1 illustrates the power for this particular test under the alternative hypothesis that $\mu = 105$. Likewise, $\beta$ is the area under alternative distribution ($\mu = 105$) that is shown in cyan. Note that the power of the test when the null hypothesis is true is the same as the level of significance for the test. It is only when the true distribution is shifted from the hypothesized distribution that the power of the test differs from the level of significance for the test. Clearly, the farther the true value for the alternative hypothesis falls from the null hypothesis, the larger the power for the given alternative hypothesis. This functional relationship is illustrated in Figure 6.2 on the following page for a nondirectional alternative hypothesis.

6.1.6 Testing $\mu$ When $\sigma$ Is Known: The $Z$-Test

The $Z$-test statistic is appropriate to use for testing a population mean if the underlying distribution’s standard deviation is known and the underlying distribution’s shape is known to be normal or the sample is of adequate size to appeal to the Central Limit Theorem. However, when making inferences about the population mean, it is highly unlikely that the population standard deviation is a known quantity.
Example 6.1.1: Consider a kindergarten teacher who administers the Stanford Binet IQ test to her class of 19 students and is interested in testing to see if her class’s IQ is different from the population IQ. From many samples, the mean and standard deviation for the Stanford Binet have been determined to be 100 and 16 respectively. The scores for the 19 students on the Stanford Binet IQ test are: 100 101 110 106 103 128 96 96 97 106 88 104 99 102 92 97 100 109 100

Solution: The five step procedure for testing hypotheses is used to answer the teacher’s question.

1. \( H_0 : \mu = 100 \)
2. \( H_A : \mu \neq 100 \)
   
   Note that a two tailed alternative hypothesis is used since the teacher has indicated no reason to believe her class will be either above the mean or below the mean.

3. The test statistic \( \bar{x} \) is selected to test the null hypothesis.

4. Since the standard deviation of the population, \( \sigma \), and the mean under the null hypothesis are known, the sampling distribution of \( \bar{x} \) can be described as normal with mean 100 and standard deviation \( \sigma / \sqrt{n} = 16 / \sqrt{19} = 3.67 \). Note that the standardized test statistic, \( Z_{\text{obs}} \), where \( Z_{\text{obs}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \) follows the standard normal distribution. The value of the test statistic is \( Z_{\text{obs}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{101.79 - 100}{16 / \sqrt{19}} = 0.487 \).

4. The \( p \)-value is calculated as \( 2 \times P(Z_{\text{obs}} \geq 0.487) = 2 \times 0.3131 = 0.6262 \). The large \( p \)-value (0.6262) indicates that observing values as inconsistent as 0.487 or more with the null hypothesis is very likely (62.62% of the time). Consequently, the statistical decision based on the \( p \)-value is to “fail to reject the null hypothesis.”

5. There is no statistical evidence to suggest the IQ scores of the kindergarten class are not 100. (Note that we are not saying the IQ scores are 100. We are simply stating that the sample evidence was not sufficient to conclude the scores were not 100.)
6.1.7 Using MINITAB\textsuperscript{TM} to do a Z-test

To conduct a one sample test of hypothesis using the Z-test statistic with MINITAB\textsuperscript{TM} do the following:

1. Choose \textit{Stat} \textgreater \textit{Basic Statistics} \textgreater 1-Sample Z.
2. Select the variable/s of interest.
3. Enter the value of \( \sigma \) in the \textit{Sigma Box}.
4. Enter the value for the null hypothesis in the \textit{Test mean Box}.
5. Click on \textit{Options} and select the appropriate alternative hypothesis using the drop down menu in the \textit{Alternative Box} of the 1-Sample Z - \textit{Options Dialog Window}. (Note: The default alternative value is “not equal”.)
6. Click \textit{OK} twice.

The 1-Sample Z (Test and Confidence Interval) \textit{Dialog Window} is shown in Figure 6.3. The MINITAB\textsuperscript{TM} results for the 1-Sample Z (Test and Confidence Interval) using the 19 kindergarten students is displayed in the Session Window and illustrated in Output 6.1. The Session Window results do not tell the user to reject the null hypothesis or fail to reject the null hypothesis. However, the \textit{p}-value is reported, which helps the user reach a decision in rejecting or failing to reject the null hypothesis. The sample standard deviation is reported under \textit{StDev} as 8.38, while the standard deviation of the sampling distribution, \( \sigma/\sqrt{n} \), is reported under \textit{SE Mean} as 3.67. Keeping with the convention adopted earlier, the standard deviation of a sampling distribution will only be called a standard error once all unknown parameters have been estimated. The value for the standardized test statistic \( Z_{\text{obs}} \) is given below \( Z \) as 0.49, while the \textit{p}-value for observing a value as extreme or more than 0.49 in a standard normal distribution when conducting a two tailed test of hypothesis is given under \textit{p} as 0.626.
Example 6.1.2: Suppose the teacher in Example 6.1.1 on page 155 considers an IQ shift of 5 points in either direction important. If we assume a 5 point difference in IQ actually exists, how likely are we to detect that difference using a 5% level of significance with our current test? In other words, calculate the power of the test.

We will assume a positive difference for the problem although it makes no difference in the answer if we were to assume a negative difference. Figure 6.4 illustrates two distributions. The distribution to the left in Figure 6.4 is the distribution under the null hypothesis ($\mu = 100$). The distribution to the right in Figure 6.4 is the distribution of IQ scores with a positive 5 point IQ shift, that is $\mu = 105$.

Solution: The power is calculated by first determining the critical values under the null distribution using the specified 5% level of significance. These values are the numbers such that 2.5% and 97.5% of the null distribution are to their left respectively. Before we go any further, what is the null distribution for our problem? Answer: The null distribution is a normal distribution with a mean of 100 and a standard deviation of 3.67065 ($\sigma/\sqrt{n} = 16/\sqrt{19}$). Consequently, the critical values are 92.8057 and 107.1943 respectively. The power of the test is calculated by adding the area in the shifted distribution (under the alternative hypothesis of $\mu = 105$) to the left of 92.8057 and the area to the right of 107.1943 in the shifted distribution. The result is 0.000446699 + 0.274988 = 0.275435. To observe how MINITAB$^\text{TMM}$ determined these areas see Output 6.2. An alternative graphical representation for the power of this problem is given using the Cumulative Distribution Function in Figure 6.5 on the following page. The decision criteria to reject the null hypothesis for values greater than 107.1943 and smaller than 92.8057 is shown in green as the decision criteria in Figure 6.5. The CDF for the alternative distribution depicted in blue in Figure 6.5 has 0.725012 of its area at or below 107.1943. Consequently, the area to the right of the upper critical value (107.1943) in the shifted distribution 1 - 0.725012 = 0.274988 plus the area to the left of the lower critical value (92.8057) 0.000446699 in the shifted distribution is equal to 0.275435. We write Power($\mu = 105$) = 0.275435 to indicate the power of the test under the alternative hypothesis that $\mu = 105$ is 27.5435%.
6.1.8 Using MINITAB\textsuperscript{TM} to calculate Power

MINITAB\textsuperscript{TM} can determine the power of a particular test, solve for sample size to achieve a given power with a test, and solve for minimum differences to obtain a specified power with a test.

To solve for power with a \textit{1-Sample Z} test:

1. Select \textit{Stat>$Power and Sample Size$}>\textit{1-Sample Z}.
2. Type your sample size in the \textit{Sample Sizes Box}.
3. Type the absolute value of the difference between the null hypothesis and the true alternative hypothesis in the \textit{Differences Box}.
4. Enter the value of the population standard deviation in the \textit{Sigma Box}.
5. Click on the \textit{Options Box} and subsequently specify the appropriate alternative hypothesis and significance level for your specific problem in the \textit{Power and Sample Size for 1-Sample Z Dialog Window}.
6. Click \textit{OK} twice.

Results using \textit{Stat>$Power and Sample Size$}>\textit{1-Sample Z} for Example 6.1.2 on the page before are shown in Output 6.3. Note the agreement in answers for Example 6.1.2 in Output 6.2 on the page before and in Output 6.3.

\textbf{Output 6.3: MINITAB\textsuperscript{TM} Results for Example 6.1.2}

\begin{tabular}{|c|c|c|}
\hline
\textbf{Sample} & \textbf{Size} & \textbf{Power} \\
\hline
5 & 19 & 0.2754 \\
\hline
\end{tabular}
6.2 Testing a Population Proportion

When working with a Bernoulli population, it is often desirable to test hypothesis concerning the proportion of successes in the population. MINITAB™ is capable of providing exact an approximate p-values for tests involving the proportion of successes in a Bernoulli population. Suffice it to say that the exact p-values MINITAB™ reports are based on the binomial distribution. The p-values MINITAB™ reports when the user clicks in the Use Test And Interval Based On The Normal Distribution Box are approximations for the exact p-values calculated from the binomial distribution. As we have seen in subsection 5.2.2 and alluded to in Lab 3.3 on page 98, the binomial distribution should only be approximated with a normal distribution when \( n \times \pi \) and \( n \times (1 - \pi) \) are both greater than 5.

To test a hypothesis involving \( \pi \) choose

1. **Stat>Basic Statistics>1 Proportion**

2. Based on your information do one of the following:

   a. If you have summarized data, click in the circle to the left of *Summarized data*. Enter the number of Bernoulli trials in the *Number of trials Box*, and the number of times your Bernoulli random variable was a success in the *Number of successes Box*.

   b. If you have data in a MINITAB™ worksheet, click in the circle to the left of *Samples in columns* and select the column that contains your data. If your data is numeric, MINITAB™ will consider a success the larger of the two values. On the other hand, if you have text data, MINITAB™ will consider the text that comes first in the alphabet a failure and the text that comes last a success. For example, suppose your data consists of the numbers 2 and 5. The 2s would be considered failures and the 5s successes. If you had a column of text such as “go” and “nogo” the “go”’s would be considered failures and the “nogo”’s would be considered successes. If your data is not arranged in a fashion compatible with what you would like to test, you can use *Manip>Code* to recode the values to suit your problem.

3. Click on the **Options Box**.

   a. Enter the value for the null hypothesis in the *Test proportion Box*. Select the appropriate alternative hypothesis using the drop down menu in the *Alternative Box*. Click **OK twice** if you want to calculate an exact p-value based on the binomial distribution.

   b. Enter the value for the null hypothesis in the *Test proportion Box*. Select the appropriate alternative hypothesis using the drop down menu in the *Alternative Box*. Click in the *Use test and interval based on the normal distribution Box* if you want to test your hypothesis based on a normal approximation to the binomial distribution. This is the approximation Basic Statistics and Data Analysis uses.

Consider **Example 6.6** from *Basic Statistics and Data Analysis*. A recent report claimed that 20% of all college graduates find a job in their chosen field of study. A survey of a random sample of 500 graduates found that 110 obtained work in their field. Is there statistical evidence to refute the claim? Following are solutions to this problem using both 3a and 3b from the testing proportions procedure. Note the explanations of exactly what each method is calculating. The Session Window output for each method is shown in Output 6.4 on the next page. Note how the output given in Output 6.4 is in agreement with the values for both methods.

**Solution:** Using 3a

1. \( H_0 : \pi = 0.2 \)
   \( H_A : \pi \neq 0.2 \)

2. Let \( X \) represent the number of college graduates that find work in their chosen field of study. In this case \( X = 110 \).

3. The distribution of \( X \) is binomial with parameters \( n = 500 \) and \( \pi = 0.2 \), written mathematically as
   \( X \sim \text{Bin}(500, 0.2) \).
6.2. Testing a Population Proportion

Output 6.4: $\pi$ Test Using Both Methods

<table>
<thead>
<tr>
<th>Test and CI for One Proportion</th>
<th>Output from using method 3 a.</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Test of } p = 0.2 \text{ vs } p \neq 0.2$</td>
<td>$\text{Sample } X \text{ N Sample } p \text{ 95.0% CI P-Value}$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>110</td>
<td>500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test and CI for One Proportion</th>
<th>Output from using method 3 b.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Test of } p = 0.2 \text{ vs } p \neq 0.2$</td>
<td>$\text{Sample } X \text{ N Sample } p \text{ 95.0% CI Z-Value P-Value}$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>110</td>
<td>500</td>
</tr>
</tbody>
</table>

4. Since we actually observed $X = 110$, we are interested in determining how likely it is to observe a value as extreme as 110 or more given the true proportion is 0.2. To do this, the first step is to find $P(X \geq 110) = 1 - P(X \leq 109) = 0.144354$. Once we have $P(X \geq 110)$, we need to multiply the answer by 2 since the alternative hypothesis is non-directional. Consequently, $2 \times P(X \geq 110) = 0.288708$. In other words, the $p$-value is 0.288708. This value is not small enough to suggest the data is inconsistent with the null hypothesis. Consequently, we fail to reject the null hypothesis.

5. We fail to find evidence to suggest the proportion of college graduates that find work in their chosen field is not 20%.

Solution: Using 3b

1. $H_0 : \pi = 0.2$
   $H_A : \pi \neq 0.2$

2. Let $p$ represent the proportion of college graduates that find work in their chosen field of study. In this case $p = 110/500 = 0.22$.

3. The standardized test statistic is $Z_{\text{obs}} = (p_{\text{obs}} - \pi_0)/\sqrt{\pi_0 \times (1 - \pi_0)/n}$. Which is $Z_{\text{obs}} = (0.22 - 0.2)/\sqrt{0.2 \times 0.8/500} = 1.11803$.

4. We are interested in determining how likely it is to observe a value as extreme as 1.11803 in a standard normal distribution. To do this, we find the area to the right of 1.11803 in a standard normal distribution and multiply the answer times 2 since the alternative hypothesis is non-directional.

   $$2 \times P(Z \geq 1.11803) = 2 \times (1 - P(Z \leq 1.11803)) = 2 \times (1 - 0.868223) = 2 \times (0.131777) = 0.263554$$

   In other words, the $p$-value is 0.263554. This value is not small enough to suggest the data is inconsistent with the null hypothesis. Consequently, we fail to reject the null hypothesis.

5. We fail to find evidence to suggest the proportion of college graduates that find work in their chosen field is not 20%.

Solution: Solution using the normal approximation to the binomial with continuity correction

1. $H_0 : \pi = 0.2$
   $H_A : \pi \neq 0.2$

2. Let $p$ represent the proportion of college graduates that find work in their chosen field of study. In this case $p = (110 - 0.5)/500 = 0.219$. In this problem, it is imperative to remember to subtract 0.5 from 110 as the appropriate continuity correction.

3. The standardized test statistic is $Z_{\text{obs}} = (p_{\text{obs}} - \pi_0)/\sqrt{\pi_0 \times (1 - \pi_0)/n}$. Which is $Z_{\text{obs}} = (0.219 - 0.2)/\sqrt{0.2 \times 0.8/500} = 1.06213$. 
4. We are interested in determining how likely it is to observe a value as extreme as 1.06213 in a standard normal distribution. To do this, we find the area to the right of 1.06213 in a standard normal distribution and multiply the answer times 2 since the alternative hypothesis is non-directional.

\[ 2 \times P(Z \geq 1.06213) = 2 \times (1 - P(Z \leq 1.06213)) = 2 \times (1 - 0.855912) = 2 \times (0.144088) = 0.288176 \]

In other words, the p-value is 0.288176. This value is not small enough to suggest the data is inconsistent with the null hypothesis. Consequently, we fail to reject the null hypothesis.

5. We fail to find evidence to suggest the proportion of college graduates that find work in their chosen field is not 20%.

Note that using a continuity correction when approximating a discrete distribution with a continuous distribution generally produces p-values closer to the exact answer.

Caution: MINITAB\textsuperscript{TM} uses \( p \) for the population proportion of successes and sample \( p \) for the sample proportion of successes where Basic Statistics and Data Analysis uses \( \pi \) for the population proportion of successes and \( p \) for the sample proportion of successes.

### 6.2.1 Equivalence of Confidence Intervals and Two-tailed Tests

The null hypothesis \( H_0 : \pi = \pi_0 \) versus the alternative \( H_A : \pi \neq \pi_0 \) is rejected at an \( \alpha \) level of significance if and only if the hypothesized value \( \pi_0 \) falls outside a \( (1 - \alpha) \times 100\% \) confidence interval for \( \pi \). From Output 6.4 on the preceding page we see that both confidence intervals using methods 3a and 3b contain \( \pi_0 = 0.2 \). Consequently, when we are testing a two-sided alternative hypothesis, we can either look at a \( p \)-value or a confidence interval to make the decision to either reject or fail to reject the null hypothesis.

### 6.2.2 Determining Power with a Bernoulli Population

The basic principles discussed in section 6.1 for determining power when working with a hypothesis about a population mean are applicable when working with a hypothesis about a population proportion. Under the assumption the null hypothesis is true, we develop a decision criterion dealing with when to reject the null hypothesis.

**Example 6.2.1:** Let us develop a decision rule for testing \( H_0 : \pi = 0.2 \) versus \( H_A : \pi > 0.2 \), if we want \( \alpha \) to be as close to 0.05 as possible based on a random sample of size 500.

**Solution:** We say as close to 0.05 as possible because with discrete data, we will most likely not get 0.05 exactly. We will choose as our decision criterion a value that leads us to reject the null hypothesis less than or equal to 0.05 of the time when null hypothesis is true. Since \( X \sim \text{Bin}(500, 0.2) \), this amounts to finding the value \( x \) such that \( P(X > x) \leq 0.05 \). In this problem, \( P(X > 115) = 1 - P(X \leq 115) = 1 - 0.9566 = 0.0434120 \leq 0.05 \). Consequently, our decision rule to reject the null hypothesis if we observe \( X > 115 \) when the true value of \( \pi \) is 0.2 results in a Type I error 4.3412% of the time. Next we find the power of the test when the value for \( \pi \) is 0.22. In this case, the power is simply \( P(Y > 115) \), where \( Y \sim \text{Bin}(500, 0.22) \). Therefore, \( P(Y > 115) = 1 - P(Y \leq 115) = 1 - 0.725826 = 0.274174 \). In other words, there is about a 27% chance of detecting a 2% increase in population proportions. See Output 6.5 on the next page for MINITAB\textsuperscript{TM} commands issued in the Session Window.

Figure 6.6 on the following page illustrates the probability distributions of \( X \sim \text{Bin}(500, 0.2) \) and \( Y \sim \text{Bin}(500, 0.22) \) along with the decision criterion to reject \( H_0 : \pi = 0.2 \) and conclude \( H_A : \pi > 0.2 \) depicted by a vertical black line at the point \( X = 115 \). One should note the resemblance of a normal distribution for both the random variables \( X \) and \( Y \) depicted in Figure 6.6. While Figure 6.6 illustrated the probability distributions for \( X \sim \text{Bin}(500, 0.2) \) and \( Y \sim \text{Bin}(500, 0.22) \), Figure 6.7 on the next page depicts the cumulative distributions for \( X \sim \text{Bin}(500, 0.2) \) and \( Y \sim \text{Bin}(500, 0.22) \) along with the decision criterion to reject \( H_0 : \pi = 0.2 \) and conclude \( H_A : \pi > 0.2 \) depicted by a green vertical line at the point \( X = 115 \).
Output 6.5: Session Window Commands for Example 6.2.1

MTB > invcdf .95; # Requesting the value with 5% to its right.
SUBC> bino 500 .2. # Based on a binomial dist with n=500, and p=0.2.

Inverse Cumulative Distribution Function

<table>
<thead>
<tr>
<th>X</th>
<th>P( X &lt;= x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>114</td>
<td>0.9457</td>
</tr>
<tr>
<td>115</td>
<td>0.9566</td>
</tr>
</tbody>
</table>

MTB > cdf 115 k1;    # Finding P(Y<=115)
SUBC> bino 500 .22.  # Where Y=bin(500, 0.22).
MTB > let k2=1-k1   # Finding P(Y>115)=1-P(Y<=115)
MTB > print k2      # k2 is the Power.

Data Display

| K2 | 0.274174      |

# The power of the test is 27.4174%

Figure 6.6: Probability Distributions of \( X \sim \text{Bin}(500, 0.2) \) and \( Y \sim \text{Bin}(500, 0.22) \) With Decision Criterion

Figure 6.7: Cumulative Probability Distributions of \( X \sim \text{Bin}(500, 0.2) \) and \( Y \sim \text{Bin}(500, 0.22) \) With Decision Criterion

Using the principles of section 6.1 and normal approximations to the binomial random variables \( X \) and \( Y \), we will find the power of testing \( H_0 : \pi = 0.2 \) versus \( H_A : \pi > 0.2 \), when the true value of \( \pi \) is 0.22. Since both \( n \times \pi \) and \( n \times (1 - \pi) \) are both greater than 5 for both random variables \( X \) and \( Y \) we can conclude
6.2. Testing a Population Proportion

X ∼ N(100, 9.94427) and Y ∼ (110, 9.26283) since the mean of a binomial random variable is n × π and the standard deviation of a binomial random variable is \(\sqrt{n \times \pi \times (1 - \pi)}\). Using a normal approximation to the binomial distribution, we first find the value for which \(P(X > x) \leq 0.05\). In this problem, \(P(X > 114.712) = 0.05\).

Next, we find the power of the test when the value for \(\pi\) is 0.22. In this case, the power is simply \(P(Y > 115)\), where \(Y ∼ N(120, 9.26283)\). Therefore, \(P(Y > (115 - 0.5)) = 1 - P(Y < (115 + 0.5)) = 1 - 0.723667 = 0.276333\). It is important to note that the plus and minus 0.5 are used when approximating the binomial distribution with the normal distribution. The power for this test using an exact procedure (27.4174%) and the power for the test using the normal approximation to the binomial distribution with continuity correction (27.6333%) are very close. See Output 6.6 for MINITAB™ Session Window commands.

Output 6.6: Session Window Commands to Calculate Power with Continuity Correction

```
MTB > cdf 115.5 k1; # Finding P(Y<=115.5)
SUBC> norm 110 9.26283. # Y~N(110, 9.26283)
MTB > let k2=1-k1 # Finding P(Y>115.5)=1-P(Y<115.5)
MTB > print k2 # k2 is the power.
```

Data Display

| K2  | 0.276333 | # The power of the test is 27.6333%. |

MINITAB™ has a built in feature for calculating the power for testing \(H_0 : \pi = 0.2\) versus \(H_A : \pi > 0.2\), when the true value of \(\pi\) is 0.22. However, the procedure does not use a continuity correction, so the answers may not be as close as one would like to the “true” answer. To continue using the decision criterion to reject \(H_0 : \pi = 0.2\) when \(X > 115\), we note that the exact significance level of this particular decision criterion is 0.04312. If we assume that \(X ∼ N(100, 8.94427)\) then the criterion value turns out to be 115.316 since \(P(X < 115.316) = 0.04312\). To find the power we calculate \(P(Y > 115.316) = 0.2830\) where \(Y ∼ N(110, 9.26283)\). See Output 6.7 for MINITAB™ Session Window commands.

Output 6.7: Session Window Commands to Calculate Power without a Continuity Correction

```
MTB > invcdf .956588 k1; # Finding x such that P(X=x)=0.04312.
SUBC> norm 100 8.94427. # Where X~N(100, 8.94427)
MTB > cdf k1 k2; # P(Y=k1)=115.316 and storing in k2
SUBC> norm 110 9.26283. # Where Y~N(110, 9.26283)
MTB > let k3=1-k2 # P(Y=115.316)=1-P(Y=115.316)
MTB > print k1-k3 # Printing constants k1, k2, and k3.
```

Data Display

| K1  | 115.316 | # k1 = 115.316 |
| K2  | 0.716989 | # k2 = beta |
| K3  | 0.283011 | # k3 = power = 28.3011% |

To find the power of a test using MINITAB™’s built in function:

1. Choose Stat>Power and Sample Size>1 Proportion
2. Fill in the Sample sizes, Alternative values of p, and the Hypothesized p value boxes.
3. Click on Options then click in the circle next to the appropriate alternative hypothesis and enter the desired significance level in the Significance level Box.
4. Click OK twice.

See Output 6.8 on the following page for the results using MINITAB™’s built in power function. Note that the answer in Output 6.8 agrees with the “by hand” answer calculated in Output 6.7.
Output 6.8: Using MINITAB’s Built In Power Function

**Power and Sample Size**

**Test for One Proportion**

Testing proportion = 0.2 (versus > 0.2)

Alpha = 0.043412

<table>
<thead>
<tr>
<th>Alternative</th>
<th>Sample Proportion</th>
<th>Size</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.220000</td>
<td>500</td>
<td>0.283</td>
<td></td>
</tr>
</tbody>
</table>

### 6.3 Testing a Population Mean When \( \sigma \) Is Unknown: The \( t \)-Test

When the parent distribution is normal, the random variable \( t_{obs} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \) follows the \( t \) distribution with \( n - 1 \) degrees of freedom. The general criteria suggested in 5.4 for using the \( t \) distribution with confidence intervals is applicable to tests of hypotheses. The suggestions for when the \( t \) distribution may be applicable are reiterated in Table 6.2.

<table>
<thead>
<tr>
<th>Sample Size ( n )</th>
<th>Shape of Underlying Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n &lt; 15 )</td>
<td>Normal</td>
</tr>
<tr>
<td>( 15 &lt; n &lt; 30 )</td>
<td>Reasonably symmetric, no outliers</td>
</tr>
<tr>
<td>( n &gt; 30 )</td>
<td>Reasonably symmetric, no extreme outliers or strong skewness</td>
</tr>
</tbody>
</table>

**Example 6.3.1:** Consider an experiment where a medical researcher randomly selects 15 healthy college students and measures their temperatures. The researcher believes that the average temperature for college students is less than the standard 98.6 degrees Fahrenheit. The students' temperatures were all measured at the same time of day and are given to the nearest tenth of a degree: \{98.0, 98.1, 98.2, 98.2, 98.2, 98.3, 98.3, 98.3, 98.3, 98.4, 98.4, 98.4, 98.4, 98.4, 98.6\}. Can we support the researcher’s belief?

**Solution:** The first thing we should do is look at the sample data and try to determine whether it is reasonable to assume the parent population is normal. Since small samples taken from a normal distribution can often yield graphs that do not look normal, we prefer to use a probability plot to assess the normality of the population based on the sample data. The resulting probability plots from choosing **Stat > Basic Statistics > Normality Test** and **Graph > Probability Plot** are shown in Figures 6.8 on the next page and 6.9 on the following page respectively. Neither probability plot causes us to suspect the distribution is anything other than normal.

A \( p \)-value below 0.1 in Figure 6.8 on the next page or values outside of the confidence bands in Figure 6.9 on the following page would be reasons to suspect the parent distribution is not normal.

Consequently, we proceed with the five-step procedure for testing a mean with unknown standard deviation to help the medical researcher reach a conclusion.

1. \( H_0 : \mu = 98.6 \)

2. \( H_A : \mu < 98.6 \)

   Note that a one tailed alternative hypothesis is used since the medical researcher suspects the true mean to be less than 98.6 degrees Fahrenheit.

2. The test statistic \( \bar{x} \) is selected to test the null hypothesis. Prior to selecting \( \bar{x} \) as the test statistic, exploratory data analysis including appropriate probability plots was used to confirm the distribution as relatively symmetrical without outliers.
3. The standardized test statistic is \( t_{\text{obs}} \), where \( t_{\text{obs}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \) follows the \( t \) distribution with \( n - 1 \) degrees of freedom. The value of the standardized test statistic is \( t_{\text{obs}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{98.2933 - 98.6}{0.1438/\sqrt{15}} = -8.26 \).

4. The \( p \)-value is calculated as \( P(t_{14} \leq -8.26) = 0.0000 \). A small \( p \)-value such as 0.0000 indicates that observing values as extreme as –8.26 or more is extremely unlikely when the true mean is 98.6 degrees Fahrenheit. In other words, the observed sample is extremely inconsistent with the null hypothesis. Therefore, we reject the null hypothesis \( H_0 : \mu = 98.6 \) in favor of the alternative hypothesis \( (H_A : \mu < 98.6) \).

5. There is strong statistical evidence to suggest the average temperature of college students is less than 98.6 degrees Fahrenheit.
To conduct a one sample test of hypothesis using the $t$-test statistic do the following:

1. **Stat > Basic Statistics > 1-Sample t.**
2. Select the variable/s of interest.
3. Enter the value for the null hypothesis in the **Test mean Box.**
4. Click on **Options** and select the appropriate alternative hypothesis using the drop down menu in the **Alternative Box of the 1-Sample t - Options Dialog Window.** (Note: The default alternative value is “not equal”.)
5. Click **OK twice.**

The 1-Sample $t$ (Test and Confidence Interval) Dialog Window is very similar to the 1-Sample Z (Test and Confidence Interval) Dialog Window (Figure 6.3 on page 156) discussed in Section 6.1 and is illustrated in Figure 6.10. The only real difference between the 1-Sample $t$ (Test and Confidence Interval) Dialog Window and the 1-Sample Z (Test and Confidence Interval) Dialog Window is that the former does not have a box where the user specifies the value of sigma. The resulting MINITAB output for testing $H_0 : \mu = 98.6$ versus $H_A : \mu < 98.6$ using a $t$ test is displayed in Output 6.9. Note that the value beneath SE Mean, 0.0371, is the standard error of the mean. That is, SE Mean is determined as $s/\sqrt{n} = 0.1438/\sqrt{15} = 0.0371$. Video 6.1 on the following page performs a test of significance to determine whether the mean thickness of varves (in millimeters) of the Green River oil shale deposit is less than 8 millimeters.
6.3. Testing a Population Mean When \( \sigma \) Is Unknown: The \( t \)-Test

To determine the power of a particular alternative hypothesis, we must know the significance level of the desired test as well as the population’s standard deviation. Since the population standard deviation is seldom a known quantity, it is common practice to estimate \( \sigma \) with the sample standard deviation \( s \). Next, we will find the power of Example 6.3.1 under the assumption that the true population mean \( \mu \) is 98.5 degrees Fahrenheit, that the population standard deviation \( \sigma \) is 0.15 degrees Fahrenheit and that the test is carried out at a 5% significance level.

To find the power of a test with MINITAB\textsuperscript{TM},

1. Select Stat>Power and Sample Size>1 Sample-t

2. Fill in the Sample Sizes, Differences, and Sigma boxes of the Power and Sample Size for the 1-Sample \( t \) Dialog Window. See Figure 6.11 for an example. Note that the difference in this problem is 98.5 – 98.6 = –0.1

3. Click on the Options Box and select the appropriate alternative hypothesis and fill in the desired significance level in the Significance level Box of the Power and Sample Size 1-Sample \( t \) - Options Dialog Window.

4. Click OK twice.

Figure 6.11: Power and Sample Size for 1-Sample \( t \)

The power for this particular problem is given in Output 6.10 on the following page. Note that the power for the test is reported as 0.5725. Many agencies would like researchers to conduct tests that have a power of at least 0.70. MINITAB\textsuperscript{TM} can calculate the required sample size for a given alternative hypothesis for a given significance level to attain a given power value.
Output 6.10: Power for One Sample t Test

Power and Sample Size

1-Sample t Test

Testing mean = null (versus < null)
Calculating power for mean = null + difference
Alpha = 0.05  Sigma = 0.15

Sample
Difference  Size  Power
-0.1  9  0.5725

To calculate the required sample size,

1. Select Stat>Power and Sample Size>1 Sample-t.

2. Fill in the Differences, Power values, and Sigma boxes of the Power and Sample Size for the 1-Sample t Dialog Window. Note that the difference in this problem is still 98.5 – 98.6 = −0.1

3. Click on the Options Box then select the appropriate alternative hypothesis and fill in the desired significance level in the Significance level Box of the Power and Sample Size 1- Sample t - Options Dialog Window.

4. Click OK twice.

To attain a power of at least 0.70 with a 5% significance level when the true alternative hypothesis is μ = 98.5 requires a sample size of 13. The actual attained power with a sample of size 13 for this particular test is 0.7324. Recall that the power for a test may be increased by doing one of three things: 1) decreasing the value of σ, 2) increasing the sample size, or 3) increasing the α value. It is vital to understand that power is a function of the alternative hypothesis. Figure 6.12 shows three power curves for a two tailed alternative hypothesis that are functions of the difference between the hypothesized value and the true value for a fixed sample size of 15. The three power curves are for three different values of σ all using a 5% significance level.

Figure 6.12: Three Power Curves for a Two Tailed Alternative Hypothesis

Power Curves for a Two Tailed Hypothesis with n=15

σ = 0.15
σ = 0.10
σ = 0.05

D = (Hypothesized Value - True Value)
To produce a graph similar to Figure 6.12,

1. Select **Stat > Power and Sample Size > 1 Sample-t**

2. Fill in the **Sample Sizes, Differences, and Sigma boxes** of the **Power and Sample Size for the 1-Sample t Dialog Window**. **Note:** The **Sample Sizes Box** will contain a single value for \( n \). We want a range of differences in the **Differences Box**. In Figure 6.12, the value in the **Differences Box** was \(-0.3:0.3/.01\). This is MINITAB\textsuperscript{TM}'s way to create values from -0.3 to 0.3 in increments of 0.01.

3. Click on the **Options Box**, when the **Power and Sample Size for 1-Sample t - Options Dialog Window** opens select the appropriate alternative hypothesis, fill in the desired significance level in the **Significance level Box**, type a name in the **Store differences in Box** such as \( D \), and type a name in the **Store power values in Box** such as \( P \).

4. Click **OK** twice.

5. Select **Graph > Plot**

6. Enter the name of the column where you stored the power values as the **Y variable** for graph 1 and the name of the column where you stored the differences as the **X variable** for graph 1 of the **Plot Dialog Window**.

7. Click on the **Display drop down menu** of the **Plot Dialog Window** and change **Symbol** to **Connect**.

8. Add titles, footnotes, text, and change colors as needed.

9. Click **OK**.

### 6.4 Testing a Population Median

#### 6.4.1 Small-Sample Test Of The Population Median

When the parent distribution is skewed or has long tails, it has been suggested that the median is a better measure of a distribution’s center than is the mean. In this section, we introduce the **sign test**, a procedure to test hypotheses concerning the population median, \( \theta \). To use the sign test, we assume that \( X_1, X_2, \ldots, X_n \) are a random sample of \( n \) observations drawn from a continuous population with unknown median \( \theta \). The sign test statistic, \( S \), is defined as the number of positive differences among the \( X_1 - \theta_0, X_2 - \theta_0, \ldots, X_n - \theta_0 \) where \( \theta_0 \) is the median from the null hypothesis, \( H_0 \). The null distribution of \( S \) is the binomial distribution with the number of trials equal \( n \) and \( \pi = .5 \).

\[
P(S = x) = \binom{n}{x}(.5)^x(.5)^{n-x}
\]  

(6.1)

Since the assumption of a continuous population is a requirement to use the sign test, theoretically we should not observe a value exactly equal to the parameter being tested. However, it is not uncommon to observe sample values equal to \( \theta_0 \), the value under the null hypothesis. In the event the difference between an observed value and the parameter being tested is zero, that particular observation or those observations are eliminated from the sample. The \( p \)-value for a less than or greater than alternative hypothesis is calculated as \( P(S \leq s) \) or \( P(S \geq s) \) respectively. The \( p \)-value for a two-tailed hypothesis test is two times the smaller of \( P(S \leq s) \) or \( P(S \geq s) \).

**Example 6.4.1:** The following numbers are the result of a study of the length in minutes of long distance phone calls for a small business: \{12.8, 3.5, 2.9, 9.4, 8.7, 3.5, 4.8, 7.7, 5.9, 6.2, 2.8, 4.7, 1.7, 2.6, 1.9, 6.5, 2.8, 3.1, 7.2, 15.8\}. Test to see if the median length of a telephone call is 5 minutes.
6.4. Testing a Population Median

**Solution:** First, graphical measures are used to assess the general shape of the underlying distribution. Graphs such as the histogram, the stem-and-leaf plot, and the boxplot all lead us to believe the parent distribution is skewed to the right. Probability plots also provide evidence that the parent population is not normal. See Figure 6.13 for a normal probability plot of the length in minutes of long distance phone calls for the small business. Numerical summaries from using the Stat>Basic Statistics>Display Descriptive Statistics command on the variable Time are also useful in characterizing the underlying distribution and are shown in Output 6.11.

Figure 6.13: Normal Probability Plot of the Length of Long Distance Phone Calls

![Normal Probability Plot](image)

Output 6.11: Descriptive Statistics of the Variable Time

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StdDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>20</td>
<td>5.725</td>
<td>4.750</td>
<td>5.399</td>
<td>3.736</td>
<td>0.835</td>
</tr>
<tr>
<td>Time</td>
<td>1.700</td>
<td>15.800</td>
<td>2.625</td>
<td>7.575</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on several graphs and the normal probability plots, we conclude the underlying population is skewed to the right. Consequently, a test using the median is more appropriate in this situation than is a test of the mean. Next, the five step procedure for testing a hypothesis is illustrated to test the hypothesis that 5 minutes is the median length of a phone call for this particular small business.

1. \( H_0 : \theta = 5 \)
   \( H_A : \theta \neq 5 \)

   Note that a two tailed alternative hypothesis is used since the problem indicated no reason to believe the length of a telephone call would be either above or below the median.

2. The test statistic \( S \), the number of positive differences among the \( X_1 - 5, X_2 - 5, \ldots, X_n - 5 \) differences is selected. \( S = 9 \)

3. The null distribution of \( S \) is the binomial distribution with \( n = 20 \) and \( \pi = .5 \).

4. The value of the test statistic is \( S = 9 \). The \( p \)-value for a two tailed hypothesis is usually reported as twice the smaller one tailed \( p \)-value. The corresponding one tailed \( p \)-values are \( P(\leq 9) = 0.4119 \) and \( P(\geq 9) = 0.7483 \). The \( p \)-value reported for the two tailed test of hypothesis is \( 2 \times 0.4119 = 0.8238 \). The large \( p \)-value of 0.8238 indicates the evidence is not inconsistent with the null hypothesis. Consequently, we fail to reject the null hypothesis.
5. There is no statistical evidence to suggest the median time spent on long distance phone calls is not 5 minutes. (Note that we are not saying the median time spent on long distance phone calls is 5 minutes. We are simply stating that the sample evidence was not sufficient to conclude the median time spent on long distance phone calls is not 5 minutes.)

To conduct a test of hypothesis with the sign test statistic, choose Stat>Nonparametrics>1-Sample Sign. The 1-Sample Sign Dialog Window with all appropriate options selected to test a two tailed test of hypothesis is shown in Figure 6.14, while the Session Window output from using the 1-Sample Sign is displayed in Output 6.12. The numbers directly beneath Below, Equal, and Above in Output 6.12 are the numbers in the sample smaller than, equal to, and larger than the hypothesized median respectively. Note that the p-value (0.8238) is in agreement with the answer calculated in step 4 earlier. Video 6.2 uses the 1-Sample Sign command to test whether schizophrenic patients learn material better when given a particular tranquilizer.

Figure 6.14: 1-Sample Sign Dialog Window for Example 6.4.1

Output 6.12: Session Window Output from 1-Sample Sign

Sign Test for Median: Time

<table>
<thead>
<tr>
<th>Time</th>
<th>N Below</th>
<th>Equal</th>
<th>Above</th>
<th>P Median</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>11</td>
<td>9</td>
<td>0.8236</td>
</tr>
</tbody>
</table>

Video 6.2: For Problem 6.99 — (Duration: 5 minutes 24 seconds)

For optimal video viewing, set your computer's display panel resolution to 1024 x 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

6.4.2 Large-Sample Test Of The Population Median

For large samples (n > 50), MINITAB\textsuperscript{TM} approximates the binomial distribution with the normal distribution. Recall that the mean of the binomial distribution is \(n\pi\) while the variance of the binomial distribution is \(n\pi(1-\pi)\). Consequently, the test statistic \(Z_{\text{obs}} = \frac{S + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0,1)\). The \(\pm 0.5\) used in the numerator is a continuity correction applied to discrete distributions that are being used to approximate continuous distributions. The minus sign is used if \(S > 0.5n\) and the plus sign is used if \(S < 0.5n\).
Example 6.4.2: It is reported that the median annual income for social workers with less than 5 years of experience is $27,500. A random sample of 25 social workers from North Carolina with less than five years experience had the following incomes:

25,200 26,500 26,700 27,900 28,000 25,500 22,200 24,700 27,000 26,500 27,700 21,500 26,000 23,500 28,500 27,000 24,500 27,600 28,100 26,400 27,900 25,900 27,600 26,100 27,400

Is there evidence to suggest North Carolina social workers are underpaid?

Solution: Based on exploratory data analysis techniques we conclude the income distribution for social workers with less than 5 years experience is skewed to the left (See Figure 6.15). Consequently, a test using the median is more appropriate in this situation than is a test of the mean. Next, the five step procedure is used to help decide if North Carolina social workers are underpaid.

1. \( H_0 : \theta = 27,500 \)
2. \( H_A : \theta < 27,500 \)
3. The test statistic \( S \), the number of positive differences among the \( X_1 = 27,500, X_2 = 27,500, ..., X_n = 27,500 \) differences is selected. \( (S = 8) \)
4. Although the distribution of \( S \) is binomial, to illustrate the large sample test of the median, we will use the normal approximation of the binomial distribution,

\[
Z_{obs} = \frac{S \pm 0.5 - 0.5n}{\sqrt{0.25n}}.
\]

The standardized test statistic is computed as \( Z_{obs} = \frac{8 \pm 0.5 - 0.5n}{\sqrt{0.25n}} = \frac{8 + 0.5 - 0.5 \times 25}{\sqrt{0.25 \times 25}} = -1.6 \). The \( p \)-value is the \( P(Z \leq -1.6) = 0.0548 \). The \( p \)-value of 0.0548 is small enough to reject the null hypothesis. However, the evidence is not incredibly compelling, and the results are at best only mildly significant.

5. There is mild statistical evidence to suggest the median salary for North Carolina social workers is below $27,500.

Although Basic Statistics and Data Analysis considers samples greater than 20 large, an exact solution with \( n = 25 \) using the sign test is still possible. In particular, if we use MINITAB’s 1-Sample Sign test for \( n < 50 \) we will get exact answers based on the binomial distribution. The Session Window results for testing \( H_0 : \theta = 27,500 \) versus \( H_A : \theta < 27,500 \) are shown in Output 6.13 on the next page.

Note how similar the \( p \)-value from the normal approximation to the binomial distribution (0.0548) is to the exact \( p \)-value (0.0539) given for the sign test in Output 6.13. Video 6.3 on the following page uses the large sample normal approximation of the binomial distribution to test whether the true parallax of the sun is 8.798 seconds of a degree.
Output 6.13: MINITAB\textsuperscript{TM}'s 1-Sample Sign Test of \( H_O : \theta = 27,500 \) versus \( H_A : \theta < 27,500 \)

Sign Test for Median: NCincome

Sign test of median = 27500 versus < 27500

\begin{array}{cccccc}
\text{N} & \text{Below} & \text{Equal} & \text{Above} & \text{P} & \text{Median} \\
\text{NCincome} & 25 & 17 & 0 & 8 & 0.0539 & 26500 \\
\end{array}

Video 6.3: For Problem 6.65 — (Duration: 6 minutes 43 seconds)

For optimal video viewing, set your computer's display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.
6.5 Summary and Review Labs

Lab 6.1 — Introduction to Hypothesis Testing

Objectives:

I. To create a standardized test statistic
II. To simulate that statistic’s distribution
III. To simulate the power of a hypothesis test

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

The standard normal distribution will be simulated and used to find simulated $Z_{0.025}$ and $Z_{0.975}$ values. The student will also simulate the power of the test $H_0: \mu = 100$ versus $H_A: \mu \neq 100$ when the true value of $\mu$ is 105.

Questions and Directions:

1. Generate 10,000 samples (rows) of size 19 (19 columns, C1–C19) from a normal distribution with mean 100 and standard deviation 16. (Calc>Random Data>Normal)

2. Calculate the mean of the 10,000 samples of size 19, and store the result in a column named xbar. (Calc>Row Statistics>Mean)

3. Create the standardized test statistic ($Z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$) for each of the 10,000 samples of size 19 and store the results in a column named Z. Note that in this problem $\mu_0 = 100$, $\sigma = 16$, and $n = 19$. (Calc>Calculator)

4. Produce a density histogram of the 10,000 Z values. Title the histogram “Simulated Standardized Sampling Distribution” and place your name in a right justified footnote. (Graph>Histogram and Options>Density)

5. Determine the simulated $Z_{0.025}$ and $Z_{0.975}$ values. In other words, determine the 2.5% and 97.5% percentiles of the simulated sampling distribution. One way to do this is to sort the values (Manip>Sort) and locate the 250th and 9750th order statistics.

6. Find the absolute values for the difference between the simulated and theoretical $Z_{0.025}$ and $Z_{0.975}$ values. Report the average for the two absolute values.

7. Generate 10,000 samples (rows) of size 19 (19 columns, C1-C19) from a normal distribution with mean 105 and standard deviation 16. (Calc>Random Data>Normal)

8. Calculate the mean of the 10,000 samples of size 19, and store the result in a column named xbar105. (Calc>Row Statistics>Mean)

9. Create the standardized test statistic ($Z = \bar{x}_{105} - \mu_0)/(\sigma/sqrt(n))$) for each of the 10,000 samples of size 19 and store the results in a column named ZS. Note that in this problem $\mu_0$, $\sigma$, and $n$ are still 100, 16 and 19 respectively. (Calc>Calculator)

10. Produce cumulative percent histograms of both the Z and ZS values on the same page. Title the graph “Simulated Centered and Shifted Z Distributions” and place your name in a right justified footnote. (Graph>Histogram and Options>Cumulative Percent)

A. Select Connect from the Display Drop Down Menu in the Histogram Dialog Window.
B. Click on **Edit Attributes** and make the line type for both graph 1 and 2 **solid** in the **Connect Dialog Window**. Also, change the **Line Size** for graphs 1 and 2 from the default value of 1 to a 2. Change the **Line Color** for graph 1 to **Red** and the **Line Color** for graph 2 to **Blue**. An example of the **Connect Dialog Window** appears in Figure 6.16.

C. To overlay both cumulative percent histograms on the same page select **Frame>Multiple Graphs** and click in the circle to the left of **Overlay graphs on the same page** (to ).

D. Add two vertical lines to the density histogram at the simulated $Z_{0.025}$ and $Z_{0.975}$ values. (**Annotation>Line**) Figure 6.17 illustrates the **Line Dialog Window**. Note that the two simulated critical values $Z_{0.025}$ and $Z_{0.975}$ for this particular simulation are -1.94 and 1.96 respectively. Your simulated $Z_{0.025}$ and $Z_{0.975}$ will most likely be different yet similar in value.

E. An example of the final graph is illustrated in Figure 6.18.
11. The simulated power for the alternative hypothesis of $\mu = 105$ is the area above the cumulative density curve to the right of the right most vertical line in the blue distribution plus the area under the cumulative density curve to the left of the left most vertical line in the blue distribution of Figure 6.18 on the preceding page. Find the exact simulated power by having MINITAB\textsuperscript{TM} count the number of values in the shifted distribution ($Z_S$) that are smaller than the simulated $Z_{0.025}$ and larger than the simulated $Z_{0.975}$. One way to accomplish this task is by first converting all of the values you want to count to 1s (and other values to 0s) using 
\textit{Manip}>\textit{Code}>\textit{Numeric to Numeric}. Finally, use \textit{Stat}>\textit{Tables}>\textit{Tally} to count the 1s.

12. Repeat step 11 with the exception of using the theoretical values for $Z_{0.025}$ and $Z_{0.975}$ in place of the simulated $Z_{0.025}$ and $Z_{0.975}$ values.

13. Use the MINITAB\textsuperscript{TM} commands \textit{Stat}>\textit{Power and Sample Size}>\textit{1-Sample Z} to get an exact answer for the power of this particular test.

14. Was your answer from step 11 or step 12 closer to the answer you reported in step 13? Speculate on why the answers for steps 11–13 are similar but not identical.
Lab 6.2 — Testing a Population Proportion #1

Objectives:

I. To understand population proportion hypothesis tests
II. To calculate power of those tests

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

This lab is designed to help the student understand the various procedures one might use as well as what MINITAB™ is actually doing when one conducts a hypothesis test on a population proportion. It also reinforces the ideas behind calculating power by having the student look at how to calculate power in each of the three testing procedures discussed in this chapter.

Questions and Directions:

1. Consider Problem 6.17 from Basic Statistics and Data Analysis. Use the five step procedure to test $H_0 : \pi = 0.10$ versus $H_A : \pi > 0.10$. Specifically, use the five step procedure with an exact procedure, 3a on page 159, a normal approximation to the binomial procedure, 3b on page 159, and a normal approximation to the binomial with continuity correction. Use a significance level of 10% when reaching your conclusions.

2. For each of the three procedures in step 1, determine the power of the test when the true value of $\pi = .11$.
   a. For the exact procedure, report the value of the decision rule and the attained level of significance using that decision rule.
   b. Use the decision criterion reported in part 2a to find the power of the test using the normal approximation to the binomial by hand. Show all work. Subsequently, use Stat>Power and Sample Size>1 Proportion to find the power for the test. Do the two procedures agree? If not, why not?
   c. Use the normal approximation to the binomial with continuity correction and a significance level of 10% to find the power of the test.
Lab 6.3 — Testing a Population Proportion #2

Objectives:

I. To understand population proportion hypothesis tests
II. To calculate power of those tests

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

This lab is designed to help the student understand the various procedures one might use as well as what MINITAB\textsuperscript{TM} is actually doing when one conducts a hypothesis test on a population proportion. It also reinforces the ideas behind calculating power by having the student look at how to calculate power in each of the three testing procedures discussed in this chapter.

Questions and Directions:

1. Consider Problem 6.17 from Basic Statistics and Data Analysis. Suppose the psychologist actually took a random sample of 1600 students and found that 140 experienced some type of emotional tragedy in their first two years of college instead of the reported 160. Use the five step procedure to test \( H_0 : \pi = 0.10 \) versus \( H_A : \pi < 0.10 \). Specifically, use the five step procedure with an exact procedure, 3a on page 159, a normal approximation to the binomial procedure, 3b on page 159, and a normal approximation to the binomial with continuity correction. Use a significance level of 10% when reaching your conclusions.

2. For each of the three procedures in step 1, determine the power of the test when the true value of \( \pi = .11 \).
   a. For the exact procedure, report the value of the decision rule and the attained level of significance using that decision rule.
   b. Use the decision criterion reported in 2a to find the power of the test using the normal approximation to the binomial by hand. Show all work. Subsequently, use Stat>Power and Sample Size>1 Proportion to find the power for the test. Do the two procedures agree? If not, why not?
   c. Use the normal approximation to the binomial with continuity correction and a significance level of 10% to find the power of the test.
Lab 6.4 — Testing a Population Mean When $\sigma$ is Unknown

Objectives:

I. To answer questions about a population mean with simulation

II. To calculate the power of a test of the population mean

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Professor Gezer has kept records of how his psychology 101 classes have performed on their first exam for the last twenty years. (Professor Gezer has used the same notes and exam for the last twenty years.) Grades have followed a normal distribution with a mean centered at 75. Professor Gezer did not keep records of the standard deviation but estimates the population standard deviation to be 8 based on the distribution of his grades. Last semester, professor Gezer thought his students were slower than normal. He has several theories for why his students may not be performing as well as they have over the last 20 years.

Questions and Directions:

1. If professor Gezer considers a three point shift in the negative direction significant (72 versus 75), how large a sample size does he need to take to have a power of at least 70% when testing at a 5% significance level?

2. What is the actual power for the sample size you reported in question 1?

3. Simulate 10,000 samples of the size reported for question 1 from a normal distribution with mean of 72 and standard deviation 8. Hint: Calc>Random Data>Normal. The size is determined by specifying C1–C(size) in the Store in column(s) of the Normal Distribution Dialog Window.

4. Use the five step procedure to test $H_0 : \mu = 75$ versus $H_A : \mu < 75$, at a 5% significance level for the data in row 1 of your current worksheet using a one sample $t$-test. Show all steps and calculations.

5. If you were to test the hypothesis $H_0 : \mu = 75$ versus $H_A : \mu < 75$ for each of the 10,000 rows using a one sample $t$-test, what percent of the time do you think you would reject the null hypothesis? Justify your answer.

6. Use Calc>Row Statistics to find the mean of the 10,000 samples of the size reported in question 1. Store the answers in a column named $\bar{x}$.

7. Use Calc>Row Statistics to find the standard deviation of the 10,000 samples of the size reported in question 1. Store the answers in a column named $s$.

8. Use Calc>Calculator to create the quantity $(\bar{x} - 75)/(s/\sqrt{n})$ and store the result in a column named $t$. In other words, calculate the observed $t$ value for the 10,000 samples.

9. What is the $t$ value in step 4 below which you reject the null hypothesis?

10. Use Manip>Code>Numeric to Numeric to code all of the values in column $t$ from -100 to the answer reported in step 9 to a 1, and all numbers from the answer reported in step 9 to 100 to a 0 while storing the results in a column named $ct$.

11. Use Stat>Tables>Tally to obtain the count and percent for the number of 1s in column $ct$. Be sure to append your table to the report pad.

12. Explain what it is that we have simulated by counting the number of 1s in column $ct$. Hint: your answers for step 5 and step 11 should be very similar.

13. Produce a histogram of the values in column $t$. Title the histogram “Noncentral t-Distribution”, and report your name in a right justified footnote.

Extra Credit: Color the area in the histogram reported in step 13 blue that represents the power for testing the hypothesis $H_0 : \mu = 75$ versus $H_A : \mu < 75$, at a 5% significance level when the true mean is 72.
Chapter 7

Inference About the Difference Between Two Parameters

7.1 Overview of Selecting Procedures

Numerous scenarios one may encounter when analyzing two samples are presented in sections 7.2 through 7.5. Not all possible scenarios are covered; however, most of the cases that you will confront when dealing with two samples are discussed. Specifically, the procedures discussed are applicable to samples that result from an experimental design. As you read through the various techniques, Figure 7.1 will help you form a mental road map for the procedures that will be appropriate in their respective situations.

Figure 7.1: Flow Chart for Selecting Procedures Covered in Chapter 7
7.2 Inference about the Difference Between Two Population Proportions

Since the sampling distribution of \( p_1 - p_2 \sim N \left( \pi_1 - \pi_2, \sqrt{\pi_1 \times (1 - \pi_1) / n_1 + \pi_2 \times (1 - \pi_2) / n_2} \right) \) provided \( n_i \times p_i > 5 \) and \( n_i \times (1 - p_i) > 5 \) for \( i = 1, 2 \), we are able to construct an approximate confidence interval formula and a standardized test statistic. Recall that all confidence intervals take the form:

\[
\text{Estimator} \pm \text{Margin of Error} \quad (7.1)
\]

The standard deviation of \( p_1 - p_2 \), \( \sigma_{p_1 - p_2} \), is \( \sqrt{\pi_1 \times (1 - \pi_1) / n_1 + \pi_2 \times (1 - \pi_2) / n_2} \), and its standard error, \( \text{SE}(p_1 - p_2) = \sqrt{p_1 \times (1 - p_1) / n_1 + p_2 \times (1 - p_2) / n_2} \). Consequently, the large sample confidence interval formula for the difference in two Bernoulli proportions \( \pi_1 - \pi_2 \) is written as:

\[
(p_1 - p_2) \pm Z_{1 - \alpha/2} \times \sqrt{p_1 \times (1 - p_1) / n_1 + p_2 \times (1 - p_2) / n_2} \quad (7.2)
\]

Likewise, the standardized form of the test statistic to test a hypothesis for the difference between two Bernoulli proportions provided \( n_i \times p_i > 5 \) and \( n_i \times (1 - p_i) > 5 \) for \( i = 1, 2 \) are satisfied takes the form:

\[
Z_{\text{obs}} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{p_1 \times (1 - p_1) / n_1 + p_2 \times (1 - p_2) / n_2}} \quad (7.3)
\]

When the hypothesized difference between \( \pi_1 - \pi_2 \) is 0, often times a pooled estimate of \( \pi \), \( p_c \) is calculated. This pooled estimate is defined as \( p_c = \frac{X_1 + X_2}{n_1 + n_2} \) where \( X_1 \) is the number of successes in sample 1 and \( X_2 \) is the number of successes in sample 2. Using the pooled estimate of \( \pi \), we have a new standard error for the difference in sample proportions, \( \text{SE}(p_1 - p_2) = \sqrt{p_c \times (1 - p_c) \times \frac{1}{n_1} + \frac{1}{n_2}} \). Using this new pooled estimate of the standard deviation, the standardized test statistic when the hypothesized difference between two Bernoulli proportions is 0 becomes:

\[
Z_{\text{obs}} = \frac{(p_1 - p_2)}{\sqrt{p_c \times (1 - p_c) \times \frac{1}{n_1} + \frac{1}{n_2}}} \quad (7.4)
\]

MINITAB\textsuperscript{TM}'s 2 Proportions command can be used to calculate a confidence interval for the difference in two Bernoulli proportions using (7.2), as well as test the difference between two Bernoulli proportions using either (7.3) or (7.4). One should note that Basic Statistics and Data Analysis only discusses the standardized test statistic in (7.4). This particular MINITAB\textsuperscript{TM} command can work on both raw and summarized data. Raw data can be entered as either stacked data or unstacked data. By stacked data, we mean the sample proportions for both columns are in a single column with an identifying code (subscript) in a second column. Unstacked data simply refers to two different columns of sample proportions. That is, each column contains the respective sample proportions for its sample.
To calculate a confidence interval or test a hypothesis for the difference between two Bernoulli proportions,


2. Follow the directions for the letter that corresponds to how your data is stored.

   a. If your data are stored in a single column, click in the circle to the left of Samples in one column (to) and specify the column containing the raw data in the Samples Box, and the column containing the identifying codes in the Subscripts Box of the 2 Proportions (Test and Confidence Interval) Dialog Window. See Figure 7.2 for a view of the 2 Proportions (Test and Confidence Interval) Dialog Window.

   b. If your data are in separate columns (unstacked), click in the circle to the left of Samples in different columns ( to) and enter the column containing the first sample in the First Box and the column containing the second column in the Second Box.

   c. Finally, if you have summarized data, click in the circle to the left of Summarized data ( to) and specify the number of trials and successes for the first and second sample in their respective boxes.

3. Next, click on the Options Box and select the appropriate options in the 2 Proportions - Options Dialog Window. If you are creating a confidence interval, you will want to specify the appropriate confidence level in the Confidence Level Box of the 2 Proportions - Options Dialog Window. If you are testing a hypothesis, you will want to specify the appropriate alternative hypothesis with the drop down box in the Alternative Box of the 2 Proportions - Options Dialog Window.

4. If you want to test a hypothesis based on the standardized test statistic in (7.4) you will need to click in the box to the left of Use pooled estimate of p ( to) for test of the 2 Proportions - Options Dialog Window. Note: by not clicking in the box to the left of Use pooled estimate of p for test of the 2 Proportions - Options Dialog Window, MINITAB™ uses the standardized test statistic of (7.3) to test the difference in two Bernoulli proportions.

5. Click OK.

Figure 7.2: 2 Proportions (Test and Confidence Interval) Dialog Window

Video 7.1 shows how MINITAB™ can be used to verify the answers given in Example 7.7 of Basic Statistics and Data Analysis.

Video 7.1: For Example 7.7 from Basic Statistics and Data Analysis — (Duration: 2 minutes 23 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.
7.3 Comparing Two Population Means

7.3.1 Two Sample Z-procedures

Suppose we have two independent samples with unknown means $\mu_1$ and $\mu_2$, known variances $\sigma_1^2$ and $\sigma_2^2$, and sample sizes $n_1$ and $n_2$, respectively. When both populations are normal, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is normal with mean $\mu_1 - \mu_2$ and standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. The standardized form of the test statistic when the two underlying distributions are assumed to be normal is

$$Z_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

which has a standard normal distribution. The confidence interval for $\mu_1 - \mu_2$ on $\bar{x}_1 - \bar{x}_2$ is written as

$$P \left( (\bar{x}_1 - \bar{x}_2) - Z_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) = (1 - \alpha) \times 100\%$$

However, rarely are the variances $\sigma_1^2$ and $\sigma_2^2$ known in practice. The quantities $\sigma_1^2$ and $\sigma_2^2$ are usually estimated with the quantities $s_1^2$ and $s_2^2$.

7.3.2 Two Sample t Procedures

Again, suppose we have two independent samples with unknown means $\mu_1$ and $\mu_2$, variances $\sigma_1^2$ and $\sigma_2^2$ unknown, and sample sizes $n_1$ and $n_2$, respectively. When both populations are normal, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is actually normal with mean $\mu_1 - \mu_2$ and standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. However, we must use the standard error of $\bar{x}_1 - \bar{x}_2$, $s_{\bar{x}_1-\bar{x}_2}$, which is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. When the unknown variances $\sigma_1^2$ and $\sigma_2^2$ are replaced with $s_1^2$ and $s_2^2$, the standardized test statistic from equation (7.5) no longer follows the standard normal distribution. The exact distribution of the standardized test statistic is not easily written; however, the distribution of the standardized test statistic can be approximated using Satterthwaite’s approximation. The essence of Satterthwaite’s approximation is using the $t$ distribution with $\nu$ degrees of freedom where $\nu$, is calculated by equation (7.7):

$$\nu = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$$

The confidence interval for $\mu_1 - \mu_2$ based on $\bar{x}_1 - \bar{x}_2$ when the variances are both different and unknown is written as

$$P \left( (\bar{x}_1 - \bar{x}_2) - t_{1-\alpha/2;\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{1-\alpha/2;\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) = (1 - \alpha) \times 100\%$$
7.3. Comparing Two Population Means

To test a hypothesis about the difference in two population means or to calculate a confidence interval for the difference between two population means using Satterthwaite’s approximation,


2. If your measurement data is in a single column, click in the circle to the left of Samples in one column (to ) and specify in the Samples Box where the data is contained and specify in the Subscripts Box where the group information is stored. If your measurement data for the two samples is in separate columns, click in the circle to the left of Samples in different columns ( to ) and specify in the First and Second Boxes where the respective information for the first and second samples is located. See Figure 7.3 for an example of the 2-Sample (Test and Confidence Interval) Dialog Window.

3. Click on the Options Box in the 2-Sample (Test and Confidence Interval) Dialog Window and specify the desired confidence level in the Confidence Level Box of the 2-Sample t - Options Dialog Window. If you are conducting hypothesis test, specify the difference value for the null hypothesis in Test Mean Box as well as selecting the appropriate alternative hypothesis with the drop down menu in the Alternative Box.

4. Click the OK button once in each dialog window.

Figure 7.3: 2-Sample (Test and Confidence Interval) Dialog Window Example

Note: by clicking on Graphs in the 2-Sample (Test and Confidence Interval) Dialog Window, then checking the box to the left of Dotplots of data ( ) or the box to the left of Boxplots of data ( ), MINITAB will produce side-by-side dotplots and or side-by-side boxplots. See Figure 7.4 for an example of the 2-Sample t - Graphs Dialog Window.

Figure 7.4: 2-Sample t - Graphs Dialog Window Example

Example 7.3.1: An identical history test is administered to two classes. The professor teaching the two classes would like to determine if the two classes have different mean test scores. The professor suspects Class 2 is sharper than Class 1. The data that follow are the test scores for the two history classes.

Class 1: 74 69 60 79 67 57 59

Class 2: 80 99 77 54 85 49 84 64 100 67 88 89 75 57 74
Solution: To get a better understanding of the two classes’ scores, we produce side-by-side boxplots before conducting any inferential technique. Side-by-side boxplots are shown in Figure 7.5.

Notice from Figure 7.5 that both boxplots are relatively symmetrical without outliers and that there is more variation associated with Class 2 than with Class 1. The Satterthwaite procedure is selected to conduct inferential procedures only after the preliminary analysis is done with the boxplots. Next, the five step procedure is used and explained to test whether or not the evidence suggests class 2 is sharper than class 1.

1. $H_0 : \mu_1 = \mu_2$
   $H_A : \mu_1 < \mu_2$

   Note that a one tailed alternative hypothesis is used since the professor suspects Class 2 is sharper than Class 1. Further, one should be aware that testing $\mu_1 = \mu_2$ is the same thing as testing $\mu_1 - \mu_2 = 0$. When using MINITAB, we will leave the Test Mean Box in the 2-Sample t - Options Box set to its default value of 0 unless we are testing for a certain difference between the two populations.

2. The test statistic $\bar{x}_1 - \bar{x}_2$ is selected to test the null hypothesis.

3. Since the standard deviations of the population, $\sigma_1$ and $\sigma_2$ are unknown and it appears reasonable to assume $\sigma_2 > \sigma_1$, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ can be described as an approximate $t$ distribution with $\nu$ degrees of freedom. Note that the standardized test statistic is $t_{\text{obs}}$ where

   \[
   t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.
   \]

   The degrees of freedom are calculated as

   \[
   \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{n_1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{n_2}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{8.243^2}{7} + \frac{15.547^2}{15}\right)}{7 - 1} = 19.46.
   \]
7.4 Comparing Two Population Centers: Variances Equal

Since MINITAB\textsuperscript{TM} will only evaluate integer degrees of freedom, 19.46 is rounded to 19 as a conservative estimate for the degrees of freedom. The value of the standardized test statistic is

\[ t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}} = \frac{(66.43 - 76.13) - (0)}{\sqrt{\frac{8243^2}{7} + \frac{15547^2}{15}}} = -1.91. \]

4. The \( p \)-value is calculated as \( P(t_{\text{obs}} \leq -1.91) = 0.0357. \) A small \( p \)-value such as 0.0357 indicates that observing values as extreme as -1.91 or more is very unlikely when \( \mu_1 - \mu_2 = 0 \) is in fact true. Consequently, we reject the null hypothesis \( H_0 : \mu_1 - \mu_2 = 0 \) in favor of the alternative hypothesis \( H_A : \mu_1 - \mu_2 < 0 \), which is the same thing as \( H_A : \mu_1 < \mu_2 \).

5. There is strong statistical evidence to suggest the mean for Class 2 is higher than the mean for Class 1.

The MINITAB\textsuperscript{TM} output for this test is given in Output 7.1. Note that MINITAB\textsuperscript{TM} automatically provides a confidence interval even though we were conducting a test of hypothesis. However, when the alternative hypothesis is directional, MINITAB\textsuperscript{TM} does not produce a conventional confidence interval rather MINITAB\textsuperscript{TM} provides a \((1 - \alpha) \times 100\%\) upper or lower bound (depending on the alternative hypothesis) for the difference specified in the null hypothesis. When the alternative hypothesis is not equal, the \((1 - \alpha) \times 100\%\) confidence interval MINITAB\textsuperscript{TM} generates is in agreement with formula (7.8). Video 7.2 calculates a confidence interval for the mean pH difference between low and high elevations for the Great Smoky Mountains using formula (7.8). The video also points out that formula (7.8) is not really the appropriate formula to use with this data and provides an alternative approach discussed in section 7.4.

Video 7.2: For Problem 7.35 — (Duration: 4 minutes 42 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 \times 768 pixels.
To verify or change your display panel resolution, select Start > Settings > Control Panel > Display > Settings.

7.4 Comparing Two Population Centers: Variances Equal

7.4.1 Pooled \( t \) Procedures

Suppose we have two independent samples with unknown means \( \mu_1 \) and \( \mu_2 \), unknown variances \( \sigma_1^2 \) and \( \sigma_2^2 \) with \( \sigma_1^2 = \sigma_2^2 = \sigma^2 \), and sample sizes \( n_1 \) and \( n_2 \) respectively. When both populations are normal, the sampling distribution of \( \bar{x}_1 - \bar{x}_2 \) follows a \( t \) distribution with mean \( \mu_1 - \mu_2 \) and standard error \( s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \), where

\[ s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \]

(7.9)
7.4. Comparing Two Population Centers: Variances Equal

is the pooled standard deviation of \( \bar{x}_1 - \bar{x}_2 \). The standardized test statistic \( t_{\text{obs}} \) follows the \( t \) distribution with \( n_1 + n_2 - 2 \) degrees of freedom. The standardized test statistic is written as

\[
t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.
\] (7.10)

The confidence interval for \( \mu_1 - \mu_2 \) based on \( \bar{x}_1 - \bar{x}_2 \) when the variances are assumed to be equal is written as

\[
P \left( (\bar{x}_1 - \bar{x}_2) - t_{1-\alpha/2;n_1+n_2-2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{1-\alpha/2;n_1+n_2-2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) = (1 - \alpha) \times 100%.
\] (7.11)

To test a hypothesis about the difference in two population means or to calculate a confidence interval for the difference between two population means using the pooled t-test,

2. If your measurement data is in a single column, click in the circle to the left of Samples in one column (to ) and specify in the Samples Box where the data is contained and specify in the Subscripts Box where the group information is stored. If your measurement data for the two samples is in separate columns, click in the circle to the left of Samples in different columns (to ) and specify in the First and Second Boxes where the respective information for the first and second samples is located. See Figure 7.3 on page 184 for an example of the 2-Sample (Test and Confidence Interval) Dialog Window.
3. Click in the box to the left of Assume equal variances ( to ) in the 2-Sample (Test and Confidence Interval) Dialog Window.
4. Click on the Options Box in the 2-Sample (Test and Confidence Interval) Dialog Window and specify the desired confidence level in the Confidence Level Box of the 2-Sample t- Options Dialog Window. If you are conducting a hypothesis test, specify the difference value for the null hypothesis in Test Mean Box as well as selecting the appropriate alternative hypothesis with the drop down menu in the Alternative Box.
5. Click the OK button once in each dialog window.

Note: Graphical features with the 2-Sample t procedure are identical whether variances are assumed to be equal or not.

Example 7.4.1: A questionnaire is devised by the Board of Governors to measure the level of satisfaction for graduates from two competing state schools. Their results are given below.

\[
\begin{align*}
\text{School 1: } & 64 75 76 80 81 82 86 89 91 95 \\
\text{School 2: } & 49 53 54 55 55 60 60 62 63 64 66 69 71 79
\end{align*}
\]

How can the Board of Governors use this information from the two small samples to evaluate the satisfaction level for graduates of the two competing state schools?

Solution: To start the evaluation, side-by-side boxplots are produced. The results are shown in Figure 7.6 on the next page. Note the relatively symmetrical boxplots of both School 1 and School 2. In addition to boxplots, normal probability plots are used to assess the distributional shape of satisfaction level scores for the two schools. Recall that a normal probability plot can be created by selecting either Graph>Probability Plot or by selecting Stat>Basic Statistics>Normality Test. Figure 7.7 on the following page was created using the Graph>Probability Plot command. Note that all of the values plotted in Figure 7.7 for both school 1 and school 2 fall inside their respective 95% confidence interval. Consequently, there is no reason to believe the distribution of satisfaction scores for either school is anything other than normal.
7.4. Comparing Two Population Centers: Variances Equal

Based on the boxplots and normal probability plots, the pooled *t*-test is selected to conduct inferential procedures. Next the five step procedure is used and explained to test whether or not the evidence suggests graduates satisfaction levels from School 1 and School 2 differ.

1. \( H_0 : \mu_1 = \mu_2 \)
   \( H_A : \mu_1 \neq \mu_2 \)

   Note that a two tailed alternative hypothesis is used since the problem gives no reason to suspect graduates from School 1 are any more satisfied with their school than are the graduates of School 2.

2. The test statistic \( \bar{x}_1 - \bar{x}_2 \) is selected to test the null hypothesis.

3. Since the standard deviations of the population, \( \sigma_1 \) and \( \sigma_2 \) are unknown, but it appears reasonable to assume \( \sigma_1 = \sigma_2 = \sigma \), the sampling distribution of \( \bar{x}_1 - \bar{x}_2 \) can be described as a t distribution with \( n_1 + n_2 - 2 \) degrees of freedom. Note that the standardized test statistic is \( t_{\text{obs}} \) where

\[
t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.
\]
The value of the standardized test statistic is

\[ t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(81.90 - 61.933) - (0)}{8.44 \sqrt{\frac{1}{10} + \frac{1}{15}}} = 5.80. \]

4. The \( p \)-value is calculated as \( 2 \times P(t \geq 5.80) = 0.0000 \). A small \( p \)-value such as 0.0000 indicates that observing values as extreme as 5.80 or more is extremely unlikely when the null hypothesis is in fact true.

5. There is strong statistical evidence to suggest the mean satisfaction level for School 1 is different from the mean satisfaction level for School 2.

MINITAB\textsuperscript{TM} output for Example 7.4.1 on page 187 is shown in Output 7.2. Note that a 95\% CI for \( \mu_1 - \mu_2 \) (12.84, 27.09) is also provided in Output 7.2. Since this CI does not contain 0, we can conclude the mean satisfaction level for School 1 is different from the mean satisfaction level for School 2. This conclusion is in agreement with the result reported in the fifth step for testing a hypothesis for the same problem.

Output 7.2: MINITAB\textsuperscript{TM} Output for Example 7.4.1

Another example that uses the pooled \textit{t-test} is provided in Video 7.3 which examines the DBH activity level of 25 schizophrenic patients classified as psychotic or nonpsychotic after being treated with an antipsychotic drug.

Video 7.3: For Problem 7.49 — (Duration: 7 minutes 23 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 \times 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

Just so you do not start thinking decisions about which test statistic to use are always obvious, we provide Video 7.4 which examines the concentration of microparticles from the permanent snowfields of the Greenland ice cap and from Antarctica.

Video 7.4: For Problem 7.87 — (Duration: 9 minutes 38 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 \times 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.
7.4. Comparing Two Population Centers: Variances Equal

7.4.2 The Mann-Whitney-Wilcoxon Test

The Mann-Whitney-Wilcoxon test can originally be traced to Wilcoxon (1945). However, many others have contributed to its development. The widespread use of the test is due in large part to Mann and Whitney. Since the test is attributed to many authors, it is the user’s call as to which name the test is given. However, be aware that many combinations of names with either some or all of Mann, Whitney, and Wilcoxon are all referring to the same test. This manuscript will refer to the test as the Wilcoxon rank sum test. MINITAB™ refers to the same test by the name Mann-Whitney.

The two-sample Wilcoxon rank sum test assumes that data come from two independent random samples \( X_1, X_2, \ldots X_n \) and \( Y_1, Y_2, \ldots Y_m \) of sizes \( n \) and \( m \), respectively, where the underlying distributions of \( X_1, X_2, \ldots X_n \) and \( Y_1, Y_2, \ldots Y_m \) have the same shape. Note that the assumption of identical underlying shapes implies that the variances are also equal. No further assumptions other than continuous data, which is at least on an ordinal scale, are made with the two-sample Wilcoxon rank sum test. The two-sample Wilcoxon rank sum test is used to create confidence intervals and to test hypotheses about the difference between two population medians. Although the test is used to determine differences between two population medians, since the assumption of identical shapes is made, a difference in medians will imply a difference in means.

The test statistic MINITAB™ uses will be slightly different from the test statistic we propose later. MINITAB™’s Mann-Whitney command uses a normal approximation (with an appropriate continuity correction) test statistic to approximate the test statistic \( R(X_i) \). \( R(X_i) \) is defined as the sum of the first sample’s ranks after both samples have been combined and ranked. The Mann-Whitney procedure in MINITAB™ attains its level of significance by using a normal approximation with an appropriate continuity correction. For sufficiently large samples, the distribution of \( R(X_i) \) is approximately normal with mean and variance given by (7.12) and (7.13).

\[
E(R(X_i)) = \frac{n(n+m+1)}{2} \tag{7.12}
\]

\[
Var(R(X_i)) = \frac{nm(n+m+1)}{12} \tag{7.13}
\]

The standardized form of the test statistic is

\[
Z_{obs} = \frac{R(X_i) \pm 0.5 - n(n+m+1)/2}{\sqrt{nm(n+m+1)/12}} \tag{7.14}
\]

The plus sign is used with the continuity correction (±0.5) when \( R(X_i) < n(n + m + 1)/2 \), and the negative sign is used when \( R(X_i) > n(n+m+1)/2 \).

7.4.3 Student’s t Approximation To The Null Distribution Of \( R(X_i) \)

The two-sample Wilcoxon rank sum test procedure discussed in Basic Statistics and Data Analysis is slightly different from the procedure used by MINITAB™. What follows are general steps and a detailed description of how to use the student’s \( t \) approximation to the null distribution of \( R(X_i) \) as described in Basic Statistics and Data Analysis.

1. Assign ranks of 1 to \( n+m \) to each data point in the combined sample. In the case of ties, apply the average rank to the tied observations.
2. Apply the pooled \( t \)-test procedure to the ranked data.

Although the student’s \( t \) approximation to the null distribution of \( R(X_i) \) is not the same as the normal approximation used by MINITAB™, it is slightly more accurate in most cases. The procedure attains good results
provided \( n \geq 10 \) and \( m \geq 10 \). This is not a significant restriction in comparison to the large samples required by the normal approximation to the test statistic \( R(X_i) \). (If exact probabilities are needed for tests where \( n < 10 \) and/or \( m < 10 \), Wilcoxon charts should be consulted.) If the data for the two samples are stored in a single column with a separate column containing subscripts, rank the raw observations and leave the subscripts intact. If the data are stored in two separate columns, stack the data in a single column, and create a column of subscripts to distinguish the samples. Next, rank the column of combined observations. Once the raw observations from the two independent samples have been ranked, apply the pooled \( t \)-test procedures to the ranked data. The following example illustrates Student’s \( t \) approximation to the null distribution of \( R(X_i) \):

**Example 7.4.2:** Thirty-two division I swimmers from the same swim team agree to participate in a year long study to determine whether high (30%) fat diets produce greater drops in swim times than the standard low (10%) fat diets. Times for the thirty-two swimmers for the 200 yard free style were taken right after the swimmers’ conference meet. The group on diet 1 followed a low fat diet the entire year but lost two swimmers along the way. The group on diet 2 followed the high fat diet the entire year and also lost two swimmers. Times for the 200 free style were taken one year later for the remaining 30 swimmers. The swimmers’ drops in seconds for both diets follow.

| Low Fat: | 0.22 | 0.33 | 0.49 | 0.53 | 0.83 | 1.05 | 1.12 | 1.46 | 2.01 | 2.02 | 2.45 | 2.73 | 4.12 | 4.70 | 5.60 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High Fat: | 1.13 | 1.29 | 1.36 | 1.41 | 1.50 | 2.00 | 2.03 | 2.37 | 3.27 | 3.51 | 3.97 | 4.03 | 4.83 | 5.37 | 7.59 |

**Solution:** To start the analysis, side-by-side boxplots are created and shown in Figure 7.8. Based on the side-by-side boxplots in Figure 7.8, it appears reasonable to assume the underlying distributions for swimmers’ time improvements on both the high fat and the low fat diet are similar in shape. However, both distributions are skewed to the right and the decision to compare medians instead of means is made.

Based on the boxplots in Figure 7.8, the Student’s \( t \) approximation to the null distribution of \( R(X_i) \) (two-sample Wilcoxon rank sum test) is selected to conduct inferential procedures. The five step procedure is outlined below. Next, the five step procedure is used and explained to test whether or not the evidence suggests the median drop in swim time for the 200 yard free style for swimmers using a high fat diet is greater than the median drop in swim time for the 200 yard free style for swimmers using a low fat diet.

\[
H_0 : \theta_1 = \theta_2 \\
H_A : \theta_1 < \theta_2
\]

Note that a one tailed alternative hypothesis is used since the coach suspects swimmers on the high fat diet will have greater time drops than swimmers on the low fat diet. (\( \theta_1 \) and \( \theta_2 \) are the median time drops for swimmers on low fat and high fat diets respectively.)
2. The test statistic $\bar{r}_1 - \bar{r}_2$ is selected to test the null hypothesis. ($\bar{r}_1$ and $\bar{r}_2$ are the average ranks for the low fat and high fat diets respectively.)

3. Since the standard deviations of the population, $\sigma_1$ and $\sigma_2$, are unknown, but it appears reasonable to assume $\sigma_1 = \sigma_2 = \sigma$ (since the shapes of both boxplots appear similar), the sampling distribution of $\bar{r}_1 - \bar{r}_2$ can be described as an approximate $t$ distribution with $n_1 + n_2 - 2$ degrees of freedom. Note that the standardized test statistic is $t_{\text{obs}}$, where $S_{\text{pR}}$ is the pooled standard deviation calculated on the ranked data according to equation (7.9).

$$t_{\text{obs}} = \frac{(\bar{r}_1 - \bar{r}_2) - (\theta_1 - \theta_2)}{S_{\text{pR}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}.$$

The value of the test statistic is

$$t_{\text{obs}} = \frac{(\bar{r}_1 - \bar{r}_2) - (\theta_1 - \theta_2)}{S_{\text{pR}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(12.667 - 18.333) - (0)}{8.4656 \sqrt{\frac{1}{15} + \frac{1}{15}}} = -1.833.$$

4. The $p$-value is calculated as $P(t_{28} \leq -1.833) = 0.0387$. A small $p$-value such as 0.0387 indicates that observing a value as extreme as $-1.833$ or more when the null hypothesis is true is very unlikely.

5. There is strong statistical evidence to suggest the median time drop for swimmers using a high fat diet is greater than the median time drop for swimmers using a low fat diet.

The MINITAB$^{\text{TM}}$ Session Window commands and output for the test are provided in Output 7.3. Note that time drops for swimmers on the low fat and high fat diets are initially stored in separate columns. The first things that are done are to stack the swimmers time drops from the low fat and the high fat diets into a joint column (TimeDrop) and to assign subscripts to a column (Diet) using the names of the two stacked columns (Low Fat and High Fat). Next, the raw observations in TimeDrop are ranked and stored in TimeDropRank. Finally, the two-sample pooled $t$-test is applied to the ranked observations stored in TimeDropRank.

Output 7.3: Session Window Commands and Output for Example 7.4.2

```
MINITAB > Stack 'Low Fat' 'High Fat' 'TimeDrop';
SUBC> Subscripts 'Diet';
SUBC> UseNames.
MINITAB > Name c05 = 'TimeDropRank'
MINITAB > Rank 'TimeDrop' 'TimeDropRank'.
MINITAB > TwoT 95 'TimeDropRank' 'Diet';
SUBC> Pooled;
SUBC> Alternative -1.
```

```
Two-Sample T-Test and CI: TimeDropRank, Diet

Two-sample T for TimeDropRank

<table>
<thead>
<tr>
<th>Diet</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Fat</td>
<td>15</td>
<td>12.67</td>
<td>9.49</td>
<td>2.5</td>
</tr>
<tr>
<td>High Fat</td>
<td>15</td>
<td>18.33</td>
<td>7.30</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Difference = mu (Low Fat) - mu (High Fat)
Estimate for difference: -5.67
95% upper bound for difference: -0.41
T-Test of difference = 0 (vs <): T-Value = -1.63  P-Value = 0.039  DF = 28
Both use Pooled StDev = 8.47
```

The order MINITAB$^{\text{TM}}$ processes text values for the information in column Diet was altered using the Editor>Column>Value Order command. Alphabetical ordering is MINITAB$^{\text{TM}}$’s default for processing text values. So, if we had not reordered the values in column Diet before using the two-sample pooled $t$-test, we would have needed to change the direction of the alternative hypothesis since MINITAB$^{\text{TM}}$ would test the null hypothesis $\mu_{\text{High Fat}} - \mu_{\text{Low Fat}}$ instead of the desired $\mu_{\text{Low Fat}} - \mu_{\text{High Fat}}$. 
7.4. Comparing Two Population Centers: Variances Equal

7.4.4 MINITAB™’s Mann-Whitney Test

The five step procedure is applied to the time drop data from Example 7.4.2 on page 191 as before, except now the test is MINITAB™’s normal approximation with continuity correction to the Mann-Whitney test statistic $R(X_i)$. The Session Window output from using MINITAB™’s Mann-Whitney command is shown in Output 7.4.

Output 7.4: Session Window Output From Using MINITAB™’s Mann-Whitney Command

Mann-Whitney Test and CI: Low Fat, High Fat

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Fat</td>
<td>15</td>
<td>1.460</td>
</tr>
<tr>
<td>High Fat</td>
<td>15</td>
<td>2.370</td>
</tr>
</tbody>
</table>

Point estimate for ETA1-ETA2 is -0.960
95.4 Percent CI for ETA1-ETA2 is (-2.980, 0.100)

W = 190.0

Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0.0407

1. $H_0 : \theta_1 = \theta_2$
2. $H_A : \theta_1 < \theta_2$

Note that a one tailed alternative hypothesis is used since the coach suspects swimmers on the high fat diet will have greater time drops than swimmers on the low fat diet. ($\theta_1$ and $\theta_2$ are the median time drops for swimmers on low fat and high fat diets respectively.)

2. The test statistic $R(X_i)$ is selected to test the null hypothesis.

3. Since the standard deviations of the population, $\sigma_1$ and $\sigma_2$ are unknown but it appears reasonable to assume $\sigma_1 = \sigma_2 = \sigma$ since the shapes of both boxplots appear similar, the sampling distribution of $R(X_i)$ can be described as approximately normal. Note that the standardized test statistic is, $Z_{\text{obs}}$, where

$$Z_{\text{obs}} = \frac{R(X_i) \pm 0.5 - n(n + m + 1)/2}{\sqrt{nm(n + m + 1)/12}}.$$  

$R(X_i) = 1+2+3+4+5+6+7+12+15+16+19+20+25+26+29 = 190$ and the value of the test statistic is

$$Z_{\text{obs}} = \frac{R(X_i) \pm 0.5 - n(n + m + 1)/2}{\sqrt{nm(n + m + 1)/12}} = \frac{190 + 0.5 - 15(15 + 15 + 1)/2}{\sqrt{15 \times 15(15 + 15 + 1)/12}} = -1.7421.$$  

4. The $p$-value is calculated as $P(Z_{\text{obs}} \leq -1.7421) = 0.0407$. A small $p$-value such as 0.0407 indicates that observing a value as extreme as -1.7421 or more when the null hypothesis is true is very unlikely.

5. There is strong statistical evidence to suggest the median time drop for swimmers using a high fat diet is greater than the median time drop for swimmers using a low fat diet.

Note that the $p$-value reported by MINITAB™ in Output 7.4 agrees with the $p$-value calculated by hand for the normal approximation with continuity correction to the Mann-Whitney test statistic $R(X_i)$.

1. Choose Stat>Nonparametrics>Mann-Whitney
2. Specify where the first and second samples are stored in the First Sample and Second Sample Boxes respectively of the Mann-Whitney Dialog Window. See Figure 7.9 on the next page for an example.
3. Use the drop down menu in the Alternative Box to specify the appropriate alternative hypothesis.
4. Single click OK.
7.5 Comparing Two Population Centers Using Matched Samples

7.5.1 The Matched Pairs \textit{t}-Test

The \textit{matched pairs \textit{t}-test} is used when two samples are related to one another. This procedure has a smaller variance than does a two sample independent procedure when it is appropriate and is a special case of the experimental design known as the \textit{randomized block design}. The \textit{pooled \textit{t}-test} covered in section 7.4 is also a special case of an experimental design known as the \textit{completely randomized design}. The \textit{matched pairs \textit{t}-test} is also known by other names such as the \textit{dependent \textit{t}-test} and the \textit{paired \textit{t}-test}. The matched differences are known as \textit{blocks}. Blocks should be used anytime the differences within a block are relatively homogeneous compared to the differences within the particular treatment. When blocks are used appropriately, differences noted in the paired observations can subsequently be attributed to differences in treatments and not to differences due to preexisting treatment differences. The scores for the matched pairs are denoted as \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) where \(x_i\) is the score for the \(i\)th subject in the group given treatment one, and \(y_i\) is the score for the \(i\)th subject in the group given treatment two. The difference between the matched pairs is denoted by \(d_i\), where \(d_i = y_i - x_i\), and the mean population difference is denoted as \(\mu_d = \mu_1 - \mu_2\). If the population of differences is distributed normally, then the standardized form of the test statistic \(\bar{d}\), where \(\bar{d}\) is the sample mean of the differences, is distributed as a \textit{t} distribution with \(n_d - 1\) degrees of freedom, where \(n_d\) is the number of matched differences.

\[
t_{\text{obs}} = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n_d}} \sim t_{n_d-1} \quad (7.15)
\]

Although the \textit{t} distribution is based on sampling from a normal distribution, the \textit{t} distribution has been found to work in a wide variety of situations when the sampling distributions are not exactly normal. The recall the general criteria from section 6.3 in Table 6.2 on page 164 that a sample should meet before one uses the \textit{t} distribution to analyze the data.

To conduct a \textit{matched pairs \textit{t}-test} with MINITAB\textsuperscript{TM},

1. Choose \textit{Stat} > \textit{Basic Statistics} > \textit{Paired \textit{t}}

2. Specify the appropriate columns in the \textit{First sample} and \textit{Second sample} boxes of the \textit{Paired \textit{t} (Test and Confidence Interval) Dialog Window}. (Differences will automatically be created according to the values in the \textit{First sample} minus values in the \textit{Second Sample}.)

3. Click on \textit{Options} and specify the appropriate alternative hypothesis using the drop down menu in the \textit{Alternative Box} of the \textit{Paired \textit{t} Options Dialog Window}.

4. Click \textit{OK}.
Example 7.5.1: Recall the Example 7.4.2 on page 191 in section 7.4 that dealt with the thirty-two division I swimmers from the same swim team who agreed to participate in a year long study to determine whether high (30%) fat diets produce greater drops in swim times than the standard low (10%) fat diets. In addition to looking at drops in swim times, the coach decides to track the weights of the swimmers on the high fat diet. The coach suspects that the high fat diet will produce weight loss. Weight measurements for the swimmers on the high fat diet were taken when the diet started and taken again one year later. Following are the before and after weight measurements for the 15 swimmers that completed the study.

<table>
<thead>
<tr>
<th>Swimmer</th>
<th>Before Weight</th>
<th>After Weight</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>152</td>
<td>145</td>
<td>-7</td>
</tr>
<tr>
<td>2</td>
<td>158</td>
<td>160</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>159</td>
<td>156</td>
<td>-3</td>
</tr>
<tr>
<td>4</td>
<td>163</td>
<td>162</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>164</td>
<td>160</td>
<td>-4</td>
</tr>
<tr>
<td>6</td>
<td>165</td>
<td>156</td>
<td>-9</td>
</tr>
<tr>
<td>7</td>
<td>167</td>
<td>160</td>
<td>-7</td>
</tr>
<tr>
<td>8</td>
<td>174</td>
<td>175</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>176</td>
<td>174</td>
<td>-2</td>
</tr>
<tr>
<td>10</td>
<td>179</td>
<td>176</td>
<td>-3</td>
</tr>
<tr>
<td>11</td>
<td>170</td>
<td>164</td>
<td>-6</td>
</tr>
<tr>
<td>12</td>
<td>172</td>
<td>169</td>
<td>-3</td>
</tr>
<tr>
<td>13</td>
<td>168</td>
<td>164</td>
<td>-4</td>
</tr>
<tr>
<td>14</td>
<td>155</td>
<td>151</td>
<td>-4</td>
</tr>
<tr>
<td>15</td>
<td>160</td>
<td>151</td>
<td>-9</td>
</tr>
</tbody>
</table>

Solution: To start the analysis, a boxplot and normal probability plot of the differences in scores are created and shown in Figure 7.10 and Figure 7.11 on the next page respectively. The boxplot in Figure 7.10 reveals a relatively symmetric distribution with no outliers while the normal probability plot in Figure 7.11 on the next page provides no evidence to reject normality. The five step procedure for testing $H_0 : \mu_d = 0$ versus $H_A : \mu_d < 0$ where $d_i = \text{Weight}_{\text{after}} - \text{Weight}_{\text{before}}$ follows.

1. $H_0 : \mu_d = 0$
2. $H_A : \mu_d < 0$

Note that a one tailed alternative hypothesis is used since the coach suspects the high fat diet will produce a weight loss.

2. The test statistic $\overline{d}$ is selected to test the null hypothesis.
3. The relatively symmetric boxplot along with the normal probability plot suggest the underlying distribution may well be normal. The standardized test statistic $t_{\text{obs}}$ where

$$t_{\text{obs}} = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n_d}} \sim t_{n_d - 1}$$

is selected for use. The value of the test statistic is

$$t_{\text{obs}} = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n_d}} = \frac{-3.9333 - 0}{3.2616 / \sqrt{15}} = -4.67.$$

4. The $p$-value is calculated as $P(t_{14} \leq -4.67) = 0.0002$. A small $p$-value such as 0.0002 indicates that observing a value as extreme as -4.67 or more when the null hypothesis is true is extremely unlikely.

5. There is strong statistical evidence to suggest high fat (30%) diets cause division I swimmers to lose weight.

The results from using MINITAB™’s paired t-test with the weight differences from using the high fat diet are shown in Output 7.5. Note that both the test statistic (-4.67) and the $p$-value (0.000) from Output 7.5 agree with the earlier “by hand” calculations.

Output 7.5: Results From Using MINITAB™’s Paired t-Test in Example 7.5.1

<table>
<thead>
<tr>
<th></th>
<th>After Wt.</th>
<th>Before Wt.</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Mean</td>
<td>161.53</td>
<td>165.47</td>
<td>-3.9333</td>
</tr>
<tr>
<td>StDev</td>
<td>9.17</td>
<td>7.86</td>
<td>3.262</td>
</tr>
<tr>
<td>SE Mean</td>
<td>2.37</td>
<td>2.03</td>
<td>0.842</td>
</tr>
</tbody>
</table>

95% upper bound for mean difference: -2.450
T-Test of mean difference = 0 (vs < 0): T-Value = -4.67  P-Value = 0.000

7.5.2 Consequences of Using the Matched Pairs t-Test When It Is Inappropriate

Suppose we use a two-sample pooled t-test with the previous problem when, in fact, a matched pairs t-test is appropriate. In other words, there are block differences. What happens to the standard error? The answer is that by using the two-sample pooled t-test when in fact the matched pairs t-test is appropriate, the standard
error for the \textit{two-sample pooled t-test} is artificially inflated! Consequently, results are too often insignificant (\textit{Type II error}). By using the \textit{two-sample pooled t-test} when the \textit{matched pairs t-test} is appropriate, the probability of committing a Type II error is increased, and subsequently, the power of the test is diminished.

What happens if the reverse is true? That is, what are the consequences of using a \textit{matched pairs t-test} when the \textit{two-sample pooled t-test} is appropriate? In other words, blocks are being used when block differences to not exist. Since there are no block differences, the standard deviations for the \textit{two-sample pooled t-test} and the \textit{matched pairs t-test} are identical. However, since the \textit{matched pairs t-test} has fewer degrees of freedom than does the \textit{two-sample pooled t-test}, the standard error for the \textit{matched pairs t-test} is larger than it should be. This large standard error again causes the user to fail to reject the null hypothesis too often. The end result is an increase in Type II errors and a subsequent decrease in the power of the test.

### 7.5.3 Wilcoxon Signed Rank Test for the Matched Pairs Experiment

The \textit{Wilcoxon signed rank test} is the nonparametric counterpart to the \textit{matched pairs t-test}. This test does not require that the $d_i$s follow a normal distribution. However, the test does assume that the distribution of the $d_i$s is symmetric. The \textit{Wilcoxon signed rank test} assumes the data come from two dependent random samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ where the underlying distributions of $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ have the same shape. Note that the assumption of identical shape implies that the variances are also the same. No further assumptions other than that the data are continuous and at least on an ordinal scale are made with the \textit{Wilcoxon signed rank test}.

The test statistic MINITAB\textsuperscript{TM} uses will be slightly different from the test statistic we propose later. MINITAB\textsuperscript{TM}'s \textit{1-Sample Wilcoxon Test} uses a normal approximation with continuity correction test statistic to approximate the test statistic $T_+$. To carry out the \textit{Wilcoxon signed rank test}, the absolute value of the dependent differences ($d_i$) is taken; and then, ranks are assigned to the $n$ $d_i$s. Next, the original signs of the $d_i$s are multiplied by the ranks of the absolute values of the dependent differences to create signed ranks. The test statistic $T_+$ is defined as the sum of the positive signed ranks.

**Note:** If any of the $d_i$s are zero, they are manually removed from the sample before the ranks are calculated.

For sufficiently large samples ($n \geq 16$), the distribution of $T_+$ is approximately normal with mean and variance given in equation (7.16) and (7.17) respectively

\[
E(T_+) = \frac{n(n+1)}{4} \quad (7.16)
\]

\[
\text{Var}(T_+) = \frac{n(n+1)(2n+1)}{24} \quad (7.17)
\]

The standardized form of the test statistic is shown in equation (7.18)

\[
Z_{obs} = \frac{T_+ \pm 0.5 - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \quad (7.18)
\]

The plus sign is used with the continuity correction ($\pm0.5$) when $T_+ < n(n+1)/4$, and the negative sign is used when $T_+ > n(n+1)/4$. 
To conduct a test of hypothesis using MINITAB™’s 1-Sample Wilcoxon test

1. Create a column of signed ranked differences. This is easily done with MINITAB™’s calculator (Calc>Calculator). Figure 7.12 depicts commands in the Calculator Dialog Window used to create signed ranked differences using a predefined column of differences (Difference).

2. Choose Stat>Nonparameters>1-Sample Wilcoxon.

3. In the Variables Box of the 1-Sample Wilcoxon Dialog Window enter the name of the column where you stored your signed ranked differences from step 1.

4. Use the drop down menu located in the Alternative Box of the 1-Sample Wilcoxon Dialog Window to specify the appropriate alternative hypothesis.

5. Click OK.

Figure 7.12: Calculator Dialog Window Commands to Create Signed Ranked Differences

7.5.4 Student’s $t$ Approximation to the Null Distribution of $T_+$

The procedure we suggest is slightly different from the procedure used by MINITAB™. Assign ranks 1 to $n$ to the $n$ signed rank differences. In the case of ties, apply the average rank to the tied observations. Next, apply the one sample $t$-test procedure to the signed ranks. Although this approximation is not quite the same as the normal approximation used by MINITAB™, it is slightly more accurate in most cases. This procedure attains good results provided $n \geq 10$. (If exact probabilities are needed for tests where $n < 10$, Wilcoxon charts should be consulted.) If the data for the two dependent samples are stored in two separate columns, one must first create a column of differences by subtracting all observations in the first column from all observations in the second column. Next, one should rank the absolute values of the column of differences. Finally, one must multiply the column containing the ranks of the absolute value of the column of differences by the original sign ($\pm$) for each difference. Once a column of signed rank differences is obtained, one should apply the one sample $t$-test procedures to the signed rank data. Example 7.5.2 is used to illustrate and show the similarity in answers from using MINITAB™’s 1-Sample Wilcoxon test for paired data and the Student’s $t$ approximation to the null distribution of $T_+$.

Example 7.5.2: The superintendent of school district B thinks that her students, on the whole, have better study habits than the students in school district A. Fourteen students from each district are paired according to IQ. Then their study habits were scored by an independent party. The results are given in Table 7.1 on the next page.

Solution: To start the analysis, a boxplot and normal probability plot are produced for the $d_i$s (B−A) and shown in Figure 7.13 on the following page and Figure 7.14 on the next page respectively. Based on the boxplot and normal probability plot, the matched pairs $t$-test is ruled out due to the symmetric yet long tailed distribution of differences.
Table 7.1: Study Habits from Example 7.5.2

<table>
<thead>
<tr>
<th>School District A</th>
<th>School District B</th>
<th>B − A</th>
<th>Signed Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>116</td>
<td>5</td>
<td>5.5</td>
</tr>
<tr>
<td>91</td>
<td>91</td>
<td>0 removed</td>
<td>removed</td>
</tr>
<tr>
<td>87</td>
<td>89</td>
<td>2</td>
<td>2.0</td>
</tr>
<tr>
<td>95</td>
<td>100</td>
<td>5</td>
<td>5.5</td>
</tr>
<tr>
<td>87</td>
<td>115</td>
<td>28</td>
<td>13.0</td>
</tr>
<tr>
<td>96</td>
<td>97</td>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>100</td>
<td>108</td>
<td>8</td>
<td>9.0</td>
</tr>
<tr>
<td>79</td>
<td>83</td>
<td>4</td>
<td>4.0</td>
</tr>
<tr>
<td>62</td>
<td>40</td>
<td>-22</td>
<td>-11.0</td>
</tr>
<tr>
<td>93</td>
<td>120</td>
<td>27</td>
<td>12.0</td>
</tr>
<tr>
<td>97</td>
<td>77</td>
<td>-20</td>
<td>-10.0</td>
</tr>
<tr>
<td>97</td>
<td>102</td>
<td>6</td>
<td>7.0</td>
</tr>
<tr>
<td>95</td>
<td>102</td>
<td>7</td>
<td>8.0</td>
</tr>
<tr>
<td>123</td>
<td>126</td>
<td>3</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Figure 7.13: Boxplot for Example 7.5.2

Figure 7.14: Normal Probability Plot for Example 7.5.2
The *one-sample t-test* approximation to the null distribution of \( T_* \) (Wilcoxon signed rank test for matched pairs) is selected to see if the evidence suggests students from school district B have better study habits than the students from school district A. The five step procedure follows:

1. \( H_O : \mu_d = 0 \)
   \( H_A : \mu_d > 0 \)
   Note that a one tailed alternative hypothesis is used since the superintendent from district B suspects her students have better study habits than the students from district A.

2. The test statistic \( \bar{r} \) (the mean of the signed ranks) is selected to test the null hypothesis.

3. The standardized test statistic is \( t_{\text{obs}} \) where
   \[
   t_{\text{obs}} = \frac{\bar{r} - \mu_d}{s_d/\sqrt{n_d}} \sim t_{n_d-1}.
   \]

4. The value of the test statistic is
   \[
   t_{\text{obs}} = \frac{\bar{r} - \mu_d}{s_d/\sqrt{n_d}} = \frac{3.76923 - 0}{7.26755/\sqrt{13}} = 1.87.
   \]
   The \( p \)-value is calculated as \( P(t_{12} \geq 1.87) = 0.043 \). A small \( p \)-value such as 0.043 indicates that observing values as extreme as 1.87 or more when the null hypothesis is true is very unlikely.

5. There is strong statistical evidence to suggest students from school district B have better study habits than students from school district A.

The MINITAB\textsuperscript{TM} Session Window commands to create a column of signed ranked differences along with the subsequent output from testing the null hypothesis \( H_O : \mu_d = 0 \) versus \( H_A : \mu_d > 0 \) is shown in Output 7.6.

Output 7.6: Commands to Create Signed Ranked Differences and Testing \( H_O : \mu_d = 0 \) versus \( H_A : \mu_d > 0 \)

\begin{verbatim}
MTB > Let 'SR(B-A)' = SIGN('B-A')*RANK(ABS0('B-A'))
MTB > OneT 'SR(B-A)';
SUBC> Test 0;
SUBC> Alternative 1.

One-Sample T: SR(B-A)

Test of mu = 0 vs mu > 0

Variable | N  | Mean | StDev | SE Mean
---------|----|------|-------|-------
SR(B-A)  | 13 | 3.77 | 7.27  | 2.02  

Variable | 95.0% Lower Bound | T    | P
---------|-------------------|------|-----
SR(B-A)  | 0.18              | 1.87 | 0.043
\end{verbatim}

*Note:* The difference of 0 from Table 7.1 on the page before is not included in the values of column “B-A” used in Output 7.6.

### 7.5.5 MINITAB\textsuperscript{TM}’s Wilcoxon Signed Rank Test

Note how the five step procedure is applied to the same school district study ratings as those in Example 7.5.2 on page 198 except now the test is MINITAB\textsuperscript{TM}’s normal approximation with continuity correction to the Wilcoxon signed rank test statistic \( T_* \).
7.5. Comparing Two Population Centers Using Matched Samples

1. \( H_0 : \mu_d = 0 \)
\( H_A : \mu_d > 0 \)

Note that a one tailed alternative hypothesis is used since the superintendent from district B suspects her students have better study habits than the students from district A.

2. The test statistic \( T_+ \) (the sum of the positive signed ranks) is selected to test the null hypothesis.
   \[ T_+ = 1 + 2 + 3 + 4 + 5.5 + 5.5 + 7 + 8 + 9 + 12 + 13 = 70. \]

3. The standardized test statistic is \( Z_{\text{obs}} \) where
   \[ Z_{\text{obs}} = \frac{T_+ \pm 0.5 - n(n + 1)/4}{\sqrt{n(n + 1)(2n + 1)/24}}. \]

The value of the test statistic is
   \[ Z_{\text{obs}} = \frac{T_+ \pm 0.5 - n(n + 1)/4}{\sqrt{n(n + 1)(2n + 1)/24}} = \frac{70 - 0.5 - 13(13 + 1)/4}{\sqrt{13(13 + 1)(2 \times 13 + 1)/24}} = 1.677. \]

4. The \( p \)-value is calculated as \( P(Z \geq 1.677) = 0.047 \). A small \( p \)-value such as 0.047 indicates that observing values as extreme as 1.677 or more when the null hypothesis is true is very unlikely.

5. There is strong statistical evidence to suggest students from school district B have better study habits than students from school district A.

To perform the 1-Sample Wilcoxon test with MINITAB\textsuperscript{TM},

1. Create a column of differences between your two matched columns. We will call this column B-A for ease of explanation.

2. Select Calc>Calculator and enter a name for the column where you will store the signed ranked differences in the Store result in variable Box. In the Expression Box type \( \text{SIGN('B-A')} \times \text{RANK(ABSO('B-A'))} \) then click OK.

3. Choose Stat>Nonparametrics>1-Sample Wilcoxon. Enter the name of the column where the signed ranked differences are stored in the Variables Box of the 1-Sample Wilcoxon Dialog Window. Use the drop down menu in the Alternative Box to select the appropriate alternative hypothesis and then click OK.

The 1-Sample Wilcoxon Dialog Window is shown in Figure 7.15 and the subsequent Session Window output for testing the appropriate hypothesis in Example 7.5.2 on page 198 is shown in Output 7.7 on the following page.
Output 7.7: Testing the Appropriate Hypothesis in Example 7.5.2

Wilcoxon Signed Rank Test: SR(B-A)

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{N for Wilcoxon} & \text{N for Wilcoxon} & \text{Estimated} \\
\text{Test Statistic} & \text{P} \text{ Median} \\
\hline
13 & 13 & 70.0 & 0.047 & 4.750 \\
\hline
\end{array}
\]

Note that the \( p \)-value reported by MINITAB\textsuperscript{TM} (0.047) agrees with the \( p \)-value calculated “by hand” (0.047) when rounded to three decimal places for the normal approximation with continuity correction to the Wilcoxon signed rank test statistic \( T_+ \). Video 7.5 examines a classic data set from Charles Darwin’s study of cross-fertilized and self-fertilized plants. The video illustrates how to test for differences between matched subjects using both an approximation to the signed rank statistic as well as using MINITAB\textsuperscript{TM}’s 1-Sample Wilcoxon test.

Video 7.5: For Problem 7.62 — (Duration: 10 minutes 41 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 \( \times \) 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.

7.6 Choosing the Right Procedure

See Figure 7.1 on page 180 for a flow chart of choosing the correct procedure for the scenarios covered in this chapter.
Lab 7.1 — Inference About the Difference Between Two Population Proportions

Objectives:

I. To use the 5-step procedure to test a population proportion
II. To determine the sample size to detect a given difference at a specific $\alpha$ value

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Problem 7.22 from Basic Statistics and Data Analysis states that Time/CNN conducted a telephone poll of 500 adult Americans and reported that 59% of those living in cities versus 57% of those living in suburbs were worried about being a victim of crime.

Questions:

1. Is it mathematically possible for either 57% or 59% of 250 to be an integer?

2. Regardless of your answer to the last question, assume there were 250 adults in each sample; that is, 250 lived in cities and 250 lived in suburbs.
   a. Use the five step procedure to test the hypothesis that there is no difference in the two population proportions.
   b. Based on your answer in step 5 of the hypothesis test from 2a, can these figures be considered statistically significant?
   c. How large a sample do we need in order to declare 57% statistically different from 59% with a power value of at least 0.70? Hint: Use the command Stat $>$ Power and Sample Size $>$ 2 Proportions.
Lab 7.2 — Comparing Two Population Means

Objectives:

I. To use simulation to calculate the power of the Satterthwaite test and the pooled two sample t-test under different conditions

II. To learn why a test for equality of variances is unnecessary

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Recall that the power of a test is the probability that the test leads to rejection of the null hypothesis when the true value of the parameter is $\psi$. In other words, $\text{power}(\psi) = P(\psi_{\text{obs}} \in \text{RR} | \text{parameter} = \psi)$. In this lab, the user will calculate the power of the unpooled (Satterthwaite) and pooled t-test under numerous scenarios. The objective of the first set of simulations is to show the user that the Satterthwaite test attains the same power as does the pooled two sample t-test whenever the sample sizes are equal, the ratio of the variances is close to 1, and sampling is from a normal distribution. The simulated powers are close to each other for both tests even when the two samples have different variances provided the sample sizes are equal. The implications of this result are that tests for equality of variance are not needed whenever the sample sizes are equal because the power of the Satterthwaite test is nearly identical to the power of the pooled two sample t-test. The user may still use a test for equality of variance; however, testing for equality of variance simply adds an extra analysis step.

Directions:

Use the MINITAB™ macro simp.mac in Appendix E.4 with the values provided in Table 7.2 on the next page and record the simulated power for both the unpooled (Satterthwaite) and pooled t-tests. The macro simp.mac by default uses 10,000 simulations and a 5% level of significance. If you want to change either value, you will need to do so in the macro itself. The power values reported using 10,000 simulations have about a ±1% margin of error. (Extra credit if you can explain why this is true.) If you have stored the macro in your MINITAB™ macros directory, simply type $%\text{simp} n_1 n_2 \mu_1 \mu_2 \sigma_1 \sigma_2$ at the MINITAB™ prompt, where $n_1, n_2, \mu_1, \mu_2, \sigma_1$, and $\sigma_2$ are the desired values from Table 7.2.

Output 7.8 on the following page illustrates using macro simp.mac with the first row of values from Table 7.2.

When macros are stored in the MINITAB™’s macro directory you do not need to provide a complete path name for MINITAB™ to know the location of the macro nor add the extension *.mac to the macro name. However, if you are running the macro off a CD, you will need to specify the complete path to where the macro is found after typing % such as %G:\HATforBSDA\MACROS\simp.mac $n_1 n_2 \mu_1 \mu_2 \sigma_1 \sigma_2$.

Questions:

1. Decide whether the unpooled (Satterthwaite) or the pooled t-test has the greater power with the following scenarios:
   a. Larger variance is associated with the smaller sample.
   b. Larger variance is associated with the larger sample.

2. Why should the simulated power be close to 5% when both values for the mean are 100?

3. Given your answer to question 2, explain why you should not pick your testing procedure simply based on a procedure that returns the highest power.

4. Provide recommendations based on your simulations for when to use the unpooled t-test and when to use the pooled t-test.
Table 7.2: Comparison of *Satterthwaite* and *Pooled t-Test*

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
<th>POWER SATTERTHWAITE</th>
<th>POWER POOLED-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>20</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>30</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>20</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>30</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>40</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>20</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>30</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>120</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>120</td>
<td>100</td>
<td>10</td>
<td>20</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>120</td>
<td>100</td>
<td>10</td>
<td>30</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>120</td>
<td>100</td>
<td>10</td>
<td>40</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Output 7.8: Using Macro **simp.mac**

```plaintext
***NIT > simp5 5 100 100 10 10
Executing from file: B:\Program Files\ETHEC\MACROS\simp.mac

Selected Values
n1 5.00000
n2 5.00000
mu 300.000
sigma 100.000
sigma 30.000
sigma 20.000
sigma 3.000
alpha 0.050000

CritT and CritT are the critical values based on an alpha value of 5% for testing the alternative hypothesis H0:mu=mu2 with the pooled t-test and the Satterthwaite approximation respectively.

Critical Values

| CritT | 1.05955 |
| CritT | 1.05955 |

TestT and TestT represent the Satterthwaite and Pooled respectively.

Simulated Power Tables

<table>
<thead>
<tr>
<th>TestT</th>
<th>Count CumCnt</th>
<th>Percent CumCnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>516</td>
<td>516</td>
</tr>
<tr>
<td>Type II Error</td>
<td>9464</td>
<td>10000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TestT</th>
<th>Count CumCnt</th>
<th>Percent CumCnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>516</td>
<td>516</td>
</tr>
<tr>
<td>Type II Error</td>
<td>9464</td>
<td>10000</td>
</tr>
</tbody>
</table>
```
Lab 7.3 — Comparing Two Population Centers: Variances Equal

Objectives:

I. To show that the Wilcoxon Signed Rank test attains close to the same power as does the pooled two sample t-test when sampling from a normal distribution regardless of sample size

II. To show that the Wilcoxon Signed Rank test attains the same power as does the pooled two sample t-test when sampling from a normal distribution regardless of sample size

III. To show that the Wilcoxon Signed Rank test attains higher power than does the pooled two sample t-test when sampling from a Laplace distribution regardless of sample size

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

Recall that the power of a test is the probability that the test leads to rejection of the null hypothesis when the true value of the parameter is $\psi$. In other words, power$(\psi) = P(\psi_{\text{obs}} \in RR | \text{parameter} = \psi)$. In this lab, the user will calculate the power of the pooled t-test and the power of the Wilcoxon Signed Rank test under numerous scenarios. Specifically, you will take random samples from three distributions (Normal, Uniform, and Laplace) with four different sample sizes.

Directions:

Use the MINITAB™ macro simh.mac in Appendix E.5 with the values provided in Table 7.3 on the next page and record the simulated power for both the pooled t-test and the Wilcoxon Signed Rank test. The macro simh.mac uses 1,000 simulations, a 5% level of significance, and the normal distribution by default. You will have to specify using the subcommand (times) the number of times you want to perform the experiment if it is not 1,000, as well as specifying the distribution you want to sample from for the distributions other than the normal distribution (dist 1=Normal, dist 2=Uniform, dist 3=Laplace). The power values reported using 1,000 simulations have about a $\pm 3\%$ margin of error. (Extra credit if you can explain why this is true.)

If you have stored the macro in your MINITAB™ macros directory, simply type %simh $n_1$ $n_2$ $\kappa_1$ $\kappa_2$ $\tau_1$ $\tau_2$ at the MINITAB™ prompt to run the macro, where $n_1$, $n_2$, $\kappa_1$, $\kappa_2$, $\tau_1$, and $\tau_2$ are the desired values from Table 7.3 on the following page.

Output 7.9 on the next page illustrates using macro simh.mac with the tenth row of values from Table 7.3.

It is important to realize that the values in Table 7.3 given for $n_1$, $n_2$, $\kappa_1$, $\kappa_2$, $\tau_1$, and $\tau_2$ are just the parameters needed to specify the distributions of interest rather than being a mean or standard deviation, necessarily. The means and standard deviations for the uniform and Laplace simulation are printed in the Session Window once the simulation has run. The values $\kappa_1$ (0) and $\tau_1$ (1) in the Uniform case for example are the parameters that specify a Uniform $(0, 1)$ distribution. Of course we know the mean and standard deviation of this distribution to be 0.5 and $1/\sqrt{12} = 0.288675$ respectively.

When macros are stored in the MINITAB™’s macro directory you do not need to provide a complete path name for MINITAB™ to know the location of the macro nor add the extension .mac to the macro name. However, if you are running the macro off of the CD, you will need to specify the complete path to where the macro is found after typing % such as %G:\HATforBSDA\MACROS\simh.mac $n_1$ $n_2$ $\kappa_1$ $\kappa_2$ $\tau_1$ $\tau_2$. 
### Table 7.3: Comparison of Wilcoxon Signed Rank and Pooled t-Test

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$k_1$</th>
<th>$k_2$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>POWER TTEST</th>
<th>POWER WTEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (dist 1)</td>
<td>19</td>
<td>19</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Normal (dist 1)</td>
<td>20</td>
<td>20</td>
<td>105</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Normal (dist 1)</td>
<td>95</td>
<td>95</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Normal (dist 1)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Uniform (dist 2)</td>
<td>19</td>
<td>19</td>
<td>.14</td>
<td>1.14</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Uniform (dist 2)</td>
<td>20</td>
<td>20</td>
<td>.14</td>
<td>1.14</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Uniform (dist 2)</td>
<td>95</td>
<td>95</td>
<td>.14</td>
<td>1.14</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Uniform (dist 2)</td>
<td>100</td>
<td>100</td>
<td>.14</td>
<td>1.14</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Uniform (dist 2)</td>
<td>10</td>
<td>10</td>
<td>.7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Uniform (dist 2)</td>
<td>15</td>
<td>15</td>
<td>.7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Uniform (dist 2)</td>
<td>60</td>
<td>60</td>
<td>.7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Uniform (dist 2)</td>
<td>90</td>
<td>90</td>
<td>.7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Uniform (dist 2)</td>
<td>60</td>
<td>60</td>
<td>.7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.14</td>
<td>0.14</td>
</tr>
</tbody>
</table>

### Questions:

1. Why is the simulated power close to 5% for every other row in your completed Table 7.3? Be sure to talk about all three distributions in your answer.
2. Provide recommendations based on your simulations for when to use the *pooled t-test* and when to use the *Wilcoxon Signed Rank test*.

3. In the *Session Window* after running the macro, there is a statement about the asymptotic relative efficiency of one test to the other test. Examine this statement closely and the power values in your table and see if you can explain exactly what this statement means. (*Note:* We have not covered this concept in the manuscript, but you should see a pattern in your completed table mirroring the statements in the *Session Window* that will lead you to a reasonable answer.)

**EXTRA CREDIT:** Explain, line by line, what each command is doing in the macro *SIMH.mac*
Chapter 8
Analysis of Categorical Data

8.0 Introduction

Many times the data we analyze is concerned with how different categories are related to one another. For example, we want to know if people who smoke are more likely to develop cancer than those who do not smoke, or we want to know if there exists a discrepancy among different races and their levels of education. The procedures to analyze categorical data are designed to answer questions like these. The primary statistical tool you will use to analyze categorical data is a \( \chi^2 \) test of some sort. You will be concerned with how well the values you observe match the values you expect. It is important that you understand the different hypotheses that are being tested for goodness-of-fit, independence, and homogeneity tests as you work in this chapter. Though the mechanics of the tests are largely similar, the ideas of what is being tested are quite different for each one.

8.1 The Chi-Square Goodness-of-Fit Test

The \textit{Chi-Square goodness-of-fit test} consists of evaluating how well the frequency counts for a given categorical variable fit a particular hypothesized distribution.

The test is based on the following assumptions regarding the sample:

1. The sample is random.
2. The measurement scale of the variable is at least nominal.

In order to evaluate how well data fits a hypothesized distribution, we first completely specify the distribution of interest. Suppose that \( F(x) \) is the true but unknown distribution function of some random variable \( X \), and that \( F^*(x) \) is some completely specified distribution function.

The null hypothesis we test and its logical alternative are:

- \( H_O : \text{The distribution function of the observed random variable is } F^*(x). \)
- \( H_A : \text{The distribution function of the observed random variable is not } F^*(x). \)

In statistical notation we write the null and alternative hypotheses as:

\[
H_O : F(x) = F^*(x) \\
H_A : F(x) \neq F^*(x)
\] (8.1)

The test statistic used to evaluate the null hypothesis is the difference between the number of observed values \( (O_j) \) falling in category \( j \) and the expected number of outcomes \( (E_j) \) in category \( j \) under the assumption that the null hypothesis is true. The expected number of outcomes in category \( j \) is defined as the sample size, \( n \), multiplied by the hypothesized proportion of successes \( (\pi_j) \) in category \( j \).

\[
E_j = n \times \pi_j
\] (8.2)
Consequently, the test statistic is a measure of the amount of disagreement between the observed data and the hypothesized distribution. When the data agree exactly with the hypothesized distribution, the value of the test statistic is zero. The more differences that exist between the hypothesized distribution and the random sample, the larger the test statistic becomes. The standardized form of the test statistic is known as the Pearson $\chi^2$ test statistic,

$$\chi^2 = \sum_{i=1}^{j} \frac{(O_i - E_i)^2}{E_i} \tag{8.3}$$

which follows an approximate Chi-Square ($\chi^2$) distribution with $j - 1$ degrees of freedom. The Chi-Square distribution is generally a skew right distribution. The shape of the Chi-Square distribution depends on its degrees of freedom ($DOF$). Three different Chi-Square distributions are depicted in Figure 8.1. The distributions depicted in black, red, and blue in Figure 8.1 are Chi-Square distributions with 3, 5, and 8 degrees of freedom respectively.

**Figure 8.1: Three Different Chi-Square Distributions**

![Chi-Square Distributions](Image)

**Note:** If the parameters which completely specify $F^*(x)$ are unknown, then the degrees of freedom for the $\chi^2$ distribution are reduced by the number of parameters ($p$) estimated. In other words, the degrees of freedom become $j - 1 - p$ instead of $j - 1$.

**Example 8.1.1:** Consider an example where the department chair in chemistry is investigating students’ complaints that grades do not follow the Dean’s policy in basic chemistry. The Dean’s policy states that grades in all basic sciences should follow a normal distribution with a mean of 75 and a standard deviation of 16. The chair takes a random sample of 44 students in basic chemistry and records their grades which are stored in worksheet ChemGrad.MTW.

<table>
<thead>
<tr>
<th>100</th>
<th>100</th>
<th>99</th>
<th>96</th>
<th>95</th>
<th>94</th>
<th>91</th>
<th>90</th>
<th>84</th>
<th>84</th>
<th>83</th>
<th>89</th>
<th>88</th>
<th>85</th>
<th>83</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>80</td>
<td>74</td>
<td>74</td>
<td>70</td>
<td>78</td>
<td>76</td>
<td>74</td>
<td>67</td>
<td>66</td>
<td>66</td>
<td>69</td>
<td>68</td>
<td>63</td>
<td>62</td>
</tr>
<tr>
<td>59</td>
<td>56</td>
<td>55</td>
<td>54</td>
<td>40</td>
<td>59</td>
<td>58</td>
<td>58</td>
<td>56</td>
<td>55</td>
<td>54</td>
<td>51</td>
<td>45</td>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>

Based on the sample, is there evidence to suggest grades do not follow a normal distribution with mean 75 and standard deviation 16?

**Solution:** The chair decides to establish five categories: Fs below the 10th percentile, Ds 10th up to the 25th percentile, Cs 25th up to the 75th percentile, Bs 75th up to the 90th percentile, and As 90th to 100th percentile. The MINITAB$^{TM}$ commands in Output 8.1 on the following page first determine the numerical range of values for each category and subsequent commands count the number of students in each category. Next, the five step procedure is used and explained to test the null hypothesis for Example 8.1.1.

1. $H_0$: The distribution of chemistry grades is normal with mean 75 and standard deviation 16.
2. $H_A$: The distribution of chemistry grades is not normal with mean 75 and standard deviation 16.
8.1. The Chi-Square Goodness-of-Fit Test

Output 8.1: Example 8.1.1 Category Calculations

2. The test statistic \((O_j - E_j)\) is selected to test the null hypothesis where \(E_j = n \times \pi\). Specifically \(E_1 = 44 \times 0.1 = 4.4\), \(E_2 = 44 \times 0.15 = 6.6\), \(E_3 = 44 \times 0.5 = 22\), \(E_4 = 44 \times 0.15 = 6.6\), and \(E_5 = 44 \times 0.10 = 4.4\).

3. The sampling distribution of

\[
\chi^2_{\text{obs}} = \sum_{i=1}^{j} \frac{(O_i - E_i)^2}{E_i}
\]

is approximately Chi-Square with \(j - 1\) degrees of freedom. In other words, \(\chi^2_{\text{obs}} \sim \chi^2_{j-1}\). The value of the test statistic is calculated as

\[
\chi^2_{\text{obs}} = \frac{(4 - 4.4)^2}{4.4} + \frac{(6 - 6.6)^2}{6.6} + \frac{(18 - 22)^2}{22} + \frac{(10 - 6.6)^2}{6.6} + \frac{(6 - 4.4)^2}{4.4} = 3.15152
\]

4. The \(p\)-value is calculated as \(P(\chi^2 \geq 3.15152) = 0.532797\). A \(p\)-value as large as 0.532797 indicates that observing values as extreme as 3.15152 or more in a \(\chi^2\) distribution with 4 degrees of freedom is a common occurrence.

5. There is no statistical evidence to suggest the distribution of basic chemistry grades is not normal with a mean of 75 and a standard deviation of 16.

8.1.1 Goodness-of-Fit Global Macro

The global macro GOF in Appendix E.3 calculates the standardized test statistic \(\chi^2_{\text{obs}}\) and returns an appropriate \(p\)-value for the specified null hypothesis. The data from Example 8.1.1 on the page before is used to demonstrate the GOF global macro.

When macros are stored in the MINITAB™’s macro directory you do not need to provide a complete path name for MINITAB™ to know the location of the macro nor add the extension *\.mac to the macro name. However, if you are running the macro off a CD, you will need to specify the complete path to where the macro is found after typing % such as %G:\HATforBSDA\MACROS\gof.mac.

Run the macro and follow the directions on the screen. The Session Window results from using the global macro GOF are shown in Output 8.2 on the following page.
8.2 The Chi-Square Test of Independence

The Chi-Square test of independence consists of evaluating how well the frequency counts for two categorical variables fit the hypothesis of independence. Mathematically, two events $A$ and $B$ are said to independent if $P(A \cap B) = P(A) \times P(B)$. The Chi-Square test of independence evaluates how close the product of the marginal probabilities comes to the joint probabilities for two categorical variables. When the product of the marginal probabilities of a row and a column is equal to the joint probability in each cell of an contingency table, there is no need to conduct a Chi-Square test of independence. However, due to variability in sampling procedures, even when two categorical variables are independent, rarely will the product of their marginal probabilities exactly equal their respective joint probabilities. Consequently, we resort to a statistical test, the Chi-Square test of independence, which evaluates how well the data fit the hypothesis of independence.

The Chi-Square test of independence is based on the following assumptions regarding the sample:

1. The random sample is drawn from a single population.
2. Each observation may fit into only one cell of the contingency table.
3. No expected cell count is less than 1, and no more than 20% of the expected cell counts are less than 5.

Chi-Square Ideas The test’s statistic used to evaluate the hypothesis is the difference between the number of observed values ($O_{ij}$) falling in category $ij$ and the expected number of outcomes ($E_{ij}$) in category $ij$ under the assumption that the null hypothesis is true. The expected number of outcomes in category $ij$ is defined as the row total for category $i$ multiplied by the column total for category $j$ divided by the grand total, $n$, which is the sample size.

$$E_{ij} = \frac{R_i C_j}{n} \quad (8.4)$$

Consequently, the test statistic is a measure of the amount of disparity between the observed data and the hypothesized distribution. When the data agree exactly with the hypothesized distribution, the value of the test
8.2. The Chi-Square Test of Independence

The chi-square statistic is zero. The greater the difference between the hypothesized distribution and the actual distribution, the larger the test statistic becomes. The standardized form of this test statistic is again a Pearson \( \chi^2 \) test statistic,

\[
\chi^2_{\text{obs}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

which follows an approximate Chi-Square \( \chi^2 \) distribution with \((r-1)(c-1)\) degrees of freedom.

Example 8.2.1: Consider the data in Table 8.1 which come from Problem 8.21 in *Basic Statistics and Data Analysis* (SNORE.MTW). The numbers in Table 8.1 are the results of a study reported in the *British Medical Journal*. Determine from the data whether heart disease and snoring are statistically related.

Table 8.1: Results of a Study About Snoring Reported in the *British Medical Journal*

<table>
<thead>
<tr>
<th>History of Heart Disease</th>
<th>Nonsnorer</th>
<th>Occasional Snorer</th>
<th>Snores Nearly Every Night</th>
<th>Snores Every Night</th>
</tr>
</thead>
<tbody>
<tr>
<td>No History of Heart Disease</td>
<td>1355</td>
<td>603</td>
<td>192</td>
<td>224</td>
</tr>
<tr>
<td>History of Heart Disease</td>
<td>24</td>
<td>35</td>
<td>21</td>
<td>30</td>
</tr>
</tbody>
</table>

Solution: Next the five step procedure is illustrated and explained to test the hypothesis of independence.

1. \( H_0 \): Heart disease and snoring are independent.

\( H_A \): Heart disease and snoring are dependent.

2. The test statistic \((O_{ij} - E_{ij})\) is selected to evaluate the null hypothesis. The expected frequency for cell \( E_{ij} \) is calculated as \( E_{ij} = \frac{R_i \times C_j}{n} \). Note: in this problem, \( i = 2 \) and \( j = 4 \) since there are two rows and four columns. \( E_{11} = 2374 \times 1379/2484 = 1317.93, E_{12} = 2374 \times 638/2484 = 609.75, \ldots E_{24} = 110 \times 254/2484 = 11.25 \).

3. The sampling distribution of \( \chi^2_{\text{obs}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \) is approximately Chi-Square with \((4-1)(2-1) = 3\) degrees of freedom. In other words, \( \chi^2_{\text{obs}} \sim \chi^2_3 \). The value of the test statistic is calculated as

\[
\chi^2_{\text{obs}} = \frac{(1355 - 1317.93)^2}{1317.93} + \frac{(603 - 609.7)^2}{609.7} + \frac{(192 - 203.57)^2}{203.57} + \frac{(224 - 242.75)^2}{242.75} + \frac{(24 - 61.07)^2}{61.07} + \frac{(35 - 28.25)^2}{28.25} + \frac{(21 - 9.43)^2}{9.43} + \frac{(30 - 11.25)^2}{11.25} = 72.782
\]

4. The \( p \)-value is calculated as \( P(\chi^2_3 \geq 72.782) = 0.0000 \). A \( p \)-value as small as 0.0000 indicates that observing values as extreme as 72.782 or more in a \( \chi^2 \) distribution with 3 degrees of freedom when the null hypothesis of independence is true is an extremely rare occurrence.

5. There is strong statistical evidence to suggest heart disease is related to snoring.

8.2.1 Chi-Square Tests With MINITAB™

MINITAB™ has two procedures you can use to test a hypothesis of independence. Which one you use depends on how your data is stored. If your data is stored as raw data or frequency data, follow the directions for Procedure 1. If your data is stored as a contingency table, follow the directions for Procedure 2.
8.2. The Chi-Square Test of Independence

All of the MINITAB\textsuperscript{TM} worksheets for Chapter 8 of Basic Statistics and Data Analysis store their information in contingency table format.

Procedure 1: Use the following MINITAB\textsuperscript{TM} commands when your data is stored as raw data or frequency data to test for association in a two-way table.

1. Select Stat>Tables>Cross Tabulation
2. For raw data, enter the columns containing the raw data in the Classification variables Box of the Cross Tabulation Dialog Window.
3. For frequency data, in the Cross Tabulation Dialog Window enter the columns containing the category data in the Classification variables Box; then click in the square to the left of Frequencies are in (to ) and enter the column containing the frequencies.
4. Click in the square to the left of Chi-Square analysis ( to ).
5. Click OK.

Procedure 2: Use the following MINITAB\textsuperscript{TM} commands when your data is stored as contingency data to test for association in a two-way table.

1. Choose Stat>Tables>Chi-Square Test
2. Enter the columns containing the the contingency table data in the Columns containing the table Box of the Chi-Square Test Dialog Window.
3. Click OK.

Figure 8.2 shows the Chi-Square Test Dialog Window with appropriate columns selected to perform a Chi-Square test of independence for Example 8.2.1 while the output from testing Example 8.2.1 is displayed in Output 8.3 on the following page. In Output 8.3, the 1 represents the variable for the first row of information in the contingency table (patients with a history of heart disease) while the 2 represents the variable for the second row of information in the contingency table (patients without a history of heart disease).

Figure 8.2: Chi-Square Test of Independence Dialog Window for Example 8.2.1

Note that the \( \chi^2 \) values and \( p \)-values calculated by hand for Example 8.2.1 on the page before agree with the results in Output 8.3 on the following page from using MINITAB\textsuperscript{TM}’s Stat>Tables>Chi-Square Test command.
8.3. The Chi-Square Test of Homogeneity

The Chi-Square test of independence presented in the last section is used when sampling from a single population and when the question of interest is whether a dependence exists between two classification variables. The Chi-Square test of homogeneity is used when there are \( r \) populations and the user is interested in determining if the \( r \) populations are similar (homogeneous) with respect to some categorical variable. With the Chi-Square test of homogeneity, sample sizes are determined before the data are collected. Consequently, the row (or column) totals are fixed quantities. This is in contrast to the Chi-Square test of independence where neither the row nor the column totals are fixed quantities.

The Chi-Square test of homogeneity is based on the following assumptions regarding the \( r \) samples:

1. The \( r \) samples are drawn at random.
2. Each observation may fit into only one category.
3. The outcomes of the \( r \) samples are all mutually independent.
4. No expected cell count is less than 1, and no more than 20% of the expected cell counts are less than 5.

Example 8.3.1: The numbers given in the Table 8.2 on the next page are the results of a study where random samples of 100 college women and 100 college men were asked to name their favorite sport. The data are stored in a MINITAB™ worksheet named SPORTS.MTW which accompanies Problem 8.32 in Basic Statistics and Data Analysis. Determine from the data whether the favorite sports for women and men are distributed the same.

Solution: The five step procedure is used and explained to test the null hypothesis of homogeneity of gender populations with respect to the classification variable sport.
Table 8.2: Sports Preferences from Example 8.3.1

<table>
<thead>
<tr>
<th></th>
<th>Football</th>
<th>Basketball</th>
<th>Baseball</th>
<th>Tennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>33</td>
<td>38</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>Female</td>
<td>38</td>
<td>21</td>
<td>15</td>
<td>26</td>
</tr>
</tbody>
</table>

1. $H_0: \pi_{1j} = \pi_{2j} = \ldots = \pi_{rj}$, for all $j$ (The proportion of males who like football, basketball, baseball, and tennis is equal to the proportion of females who like football, basketball, baseball, and tennis respectively.)

$H_A: \pi_{ij} \neq \pi_{kj}$, for some $j$ and some pair $i$ and $k$ (The proportion of males who like football, basketball, baseball, and tennis is not equal to the proportion of females who like football, basketball, baseball, and tennis in at least one instance.)

2. The test statistic $(O_{ij} - E_{ij})$ is selected to test the null hypothesis. The expected frequency for cell $E_{ij}$ is calculated according to equation (8.4). Note in this problem that $i = 2$ and $j = 4$ since there are two rows and four columns. $E_{11} = 100 \times 71/200 = 35.50$, $E_{12} = 100 \times 59/200 = 29.50$, $E_{24} = 100 \times 31/200 = 15.50$.

3. The sampling distribution of the test statistic found in equation (8.5) is approximately Chi-Square with $(4 - 1)(2 - 1) = 3$ degrees of freedom. In other words, $\chi^2_{obs} \sim$ approximately $\chi^2_3$. The value of the test statistic is calculated as

$$
\chi^2_{obs} = \frac{(33 - 35.50)^2}{35.50} + \frac{(38 - 35.50)^2}{35.50} + \frac{(24 - 19.50)^2}{19.50} + \frac{(5 - 15.50)^2}{15.50} + \frac{(38 - 35.50)^2}{35.50} + \frac{(21 - 29.50)^2}{29.50} + \frac{(15 - 19.50)^2}{19.50} + \frac{(26 - 15.50)^2}{15.50} = 21.553
$$

4. The $p$-value is calculated as $P(\chi^2_3 \geq 21.553) = 0.0000$. A $p$-value as small as 0.0000 indicates that observing values as extreme as 21.553 or more when the null hypothesis is true is an extremely rare occurrence. Therefore, we reject the null hypothesis.

5. There is strong statistical evidence to suggest the proportion of college males and college females favoring football, basketball, baseball, and tennis are not the same.

### 8.3.1 Chi-Square Test of Homogeneity with MINITAB™

As with the Chi-Square test of independence, MINITAB™ has two procedures you can use to test a hypothesis of homogeneity of populations with respect to a classification variable. Which one you use, as with the Chi-Square test of independence, depends on how your data is stored. The test statistic for a Chi-Square test of independence and the test statistic for a Chi-Square test of homogeneity of populations with respect to a classification variable are identical. Consequently, if your data are stored as raw data or frequency data, follow the directions for Procedure 1 on page 214. If your data are stored as a contingency table, follow the directions for Procedure 2 on page 214.

Figure 8.3 on the following page shows the Chi-Square Test Dialog Window with appropriate columns selected to perform a Chi-Square test of homogeneity for Example 8.3.1 on the page before while the output from testing Example 8.3.1 is displayed in Output 8.4 on the following page. In Output 8.4, the 1 represents the variable for the first row of information in the contingency table (Male) while the 2 represents the variable for the second row of information in the contingency table (Female).

**Note:** The $\chi^2$ values and $p$-values calculated by hand for Example 8.3.1 on the page before agree with the results in Output 8.4 on the following page from using MINITAB™’s Stat>Tables>Chi-Square Test command.

Video 8.1 on the next page uses the Chi-Square test of homogeneity to test homogeneity of categories between two random samples (taken during different time periods) with respect to insurance ratings for Chevrolet vehicles.
8.3. The Chi-Square Test of Homogeneity

Figure 8.3: Chi-Square Test of Homogeneity Dialog Window for Example 8.3.1

Output 8.4: Output from Testing Example 8.3.1

Chi-Square Test: football, basketbl, baseball, tennis

<table>
<thead>
<tr>
<th></th>
<th>football</th>
<th>basketbl</th>
<th>baseball</th>
<th>tennis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.50</td>
<td>29.50</td>
<td>19.50</td>
<td>15.50</td>
<td>100.00</td>
</tr>
<tr>
<td>2</td>
<td>35.50</td>
<td>29.50</td>
<td>19.50</td>
<td>15.50</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>71.00</td>
<td>59.00</td>
<td>39.00</td>
<td>31.00</td>
<td>200.00</td>
</tr>
</tbody>
</table>

Chi-Sq = 0.176 + 2.449 + 1.038 + 7.113 + 0.176 + 2.449 + 1.038 + 7.113 = 21.553
DF = 3, P-Value = 0.000

Video 8.1: For Problem 8.35 — (Duration: 6 minutes 6 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start > Settings > Control Panel > Display > Settings.
8.4 Summary and Review Labs

Lab 8.1 — Introduction to the Chi-Square Distribution

Objectives:

I. To simulate various $\chi^2$ distributions
II. To discover the relationship between degrees of freedom and the mean and variance of $\chi^2$ distributions

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

In lab 5.1 of section 5.6 the Chi-Square distribution was introduced through simulation. Recall that if $x_1, x_2, \ldots, x_n$ denotes a random sample from $\mathcal{N}(\mu, \sigma)$, then the quantity $\frac{(n-1) \times s^2}{\sigma^2} \sim \chi^2_{n-1}$. Another relationship that is quite useful is that if $x_1, x_2, \ldots, x_n$ denotes a random sample from $\mathcal{N}(\mu, \sigma)$, then the quantity $\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \sim \chi^2_n$.

Directions:

Complete the following steps 1 through 6 three times for each of the three sample sizes $n = 3, 4, \text{ and } 9$. Be sure to record all your results for one sample size before moving to the next sample size.

Step 1. Generate 20,000 samples of size $n$ from a normal distribution with a mean of 100 and standard deviation of 10. (Calc>Random data>Normal).

Step 2. Standardize the observations to create standard normal values. (Calc>Standardize).

Step 3. Square each of the standard normal values stored in the $n$ columns (individually) and store the results back in the same columns. (Replace $Z$s with $Z^2$s for each of the $n$ columns.)

Step 4. Add the $n Z^2$s for each of the 20,000 simulated samples and store the results in a column named chi2. (Calc>Row Statistics)

Step 5. Produce a histogram of the 20,000 values stored in column chi2. Since the resulting histogram is a simulated Chi-Square distribution with $n$ degrees of freedom, title the histogram “Simulated Chi-Square Distribution with $n$ DOF”. Be sure to specify the value of $n$.

Step 6. Calculate the mean, standard deviation and variance of chi2 for each value of $n$.

Questions:

1. What happens to the shape of the Monte Carlo simulated Chi-Square distribution (chi2) as $n$ increases?

2. What is the relationship between the degrees of freedom for the Monte Carlo simulated Chi-Square distribution and the mean and variance of the Monte Carlo simulated Chi-Square distribution?

3. Determine the 90th percentile of your Monte Carlo simulated Chi-Square distribution with 9 degrees of freedom. How does your value compare to the theoretical $\chi^2_{0.90, 9}$?
Lab 8.2 — Chi-Square Goodness-of-Fit Test

Objective:

To practice conducting chi-square goodness-of-fit tests

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

The Chi-Square goodness-of-fit test consists of evaluating how well the frequency counts for a given categorical variable fit a particular hypothesized distribution. You will practice this test with three different scenarios.

Questions and Directions:

1. Simulate rolling a fair die 36,000 times (Calc>Random Data>Integer). Then, use the five-step procedure with the Chi-Square goodness-of-fit test to determine if there is any evidence to suggest the die (MINITAB™’s random number generator) is unfair. The number of observed 1s, 2s, ..., 6s is easily determined with the Stat>Tables>Tally command.

2. Simulate rolling two fair dice 36,000 times (Calc>Random Data>Integer). Determine the sum of the two dice for the 36,000 experiments (Calc>Row Statistics). Then, use the five-step procedure with the Chi-Square goodness-of-fit test to determine if there is any evidence to suggest the sum of the two dice (MINITAB™’s random number generator) is unfair. The number of observed 2s, 3s, ..., 12s is easily determined with the Stat>Tables>Tally command.

3. It has been suggested that answer “C” appears more than other answers on standardized tests. If this is true, one might be able to gain an advantage in standardized tests by always guessing “C” when one is not sure of the answer. Use the worksheet GRE.MTW and column one labelled GRE91-18 which contains the actual answers for a GRE (Graduate Record Exam) test selected at random from GRE tests that were actually used in 1991 to see if “C” appears more often than other letters. Specifically, use the five-step procedure with the Chi-Square goodness-of-fit test to determine if there is any evidence to suggest “C” appears more than the other four letters. The number of observed As, Bs, ..., Es is easily determined with the Stat>Tables>Tally command.
Lab 8.3 — Chi-Square Tests of Independence and Homogeneity

Objective:

To practice conducting *Chi-Square tests of independence* and *homogeneity*

*Basic Directions:*

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

*Introduction:*

The *Chi-Square test of independence* evaluates whether two random variables can be declared statistically independent and the *Chi-Square test of homogeneity* evaluates whether the proportions of data that fall into various categories are equal.

*Questions and Directions:*

1. Use the students in your class to see if there is a relationship between gender and grade point average. Have everyone write their current grade point average and gender on a slip of paper. Collect the slips and prepare a two by two contingency table with gender (male/female) and grade point average (below 2.7/2.7 and above). Based on your results, are grade point average and gender independent? Specifically, use the five step procedure to test for independence between grade point average and gender.

2. Worksheet *GRE.MTW* contains actual results for two forms of the Graduate Record Exam (GRE) for two different years (1991, 1992). Each form was selected at random from the forms that were given in 1991 and 1992 respectively. Column one labelled *GRE91-18* and column two labelled *GRE92-2* respectively contain the actual results for the two forms selected at random. Use the five-step procedure to test the hypothesis that the proportions of answers between years are identical.
Chapter 9
Regression Analysis

9.0 Introduction

This chapter on regression analysis will review the ideas of simple linear regression and the assumptions we make to begin our analysis. We will then discuss various estimators of the mean square error and the process of testing hypotheses about the parameters of the model. We will learn how to construct confidence intervals for the parameters of the model, prediction intervals for new observations, and confidence and prediction bands for the entire regression. Finally, we will learn several ways to check the adequacy of the model we have proposed.

9.1 The Linear Regression Model

In Section 2.3, we introduced least squares regression. In this section, we first give a brief review of the material covered in Section 2.3. Next, we discuss the linear regression model’s assumptions. Then, we introduce and discuss the estimator of \( \sigma^2 \), the mean square error (MSE). Simple linear regression attempts to estimate the parameters of the model:

\[
y_i = \beta_0 + \beta_1 x_i + \epsilon_i
\]  

(9.1)

where:

- \( y_i \) is the value of the response (dependent) variable in the \( i \)th trial
- \( \beta_0 \) and \( \beta_1 \) are unknown parameters
- \( x_i \) is a known constant, the level of the predictor variable in the \( i \)th trial and
- \( \epsilon_i \) is a random error term.

The least squares estimators of \( \beta_0 \) and \( \beta_1 \) are \( b_0 \) and \( b_1 \). The equations to calculate \( b_0 \) and \( b_1 \) are given in (9.3) and (9.2) respectively.

\[
b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right) / n}{\sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2 / n}
\]  

(9.2)

\[
b_0 = \bar{y} - b_1 \bar{x}
\]  

(9.3)

Once \( b_0 \) and \( b_1 \) are determined, the fitted regression equation is written:

\[
\hat{y}_i = b_0 + b_1 x_i
\]  

(9.4)

Although no assumptions are necessary to calculate \( b_0 \) and \( b_1 \), assumptions are needed to make correct inferences for the linear models’ parameters \( \beta_0 \) and \( \beta_1 \).
9.1. The Linear Regression Model

9.1.1 Model Assumptions

The standard assumptions for the linear regression model \( y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \) deal with the random error component \( \varepsilon_i \).

### Linear Regression Model Assumptions

Note that each \( \varepsilon_i \) is a random variable, and it is assumed that each \( \varepsilon_i \) adheres to the following conditions:

1. It has mean value 0.
2. It has a standard deviation \( \sigma \), which does not depend on \( x \).
3. It has a normal distribution.
4. The \( \varepsilon_i \)s are mutually independent.

A more concise way to state the linear regression model’s assumptions is to say that the \( \varepsilon_i \)s are normally and independently distributed with mean 0 and standard deviation \( \sigma \). Mathematically, we write the assumptions as follows:

\[
\varepsilon_i \sim N(0, \sigma).
\]

Since \( y_i \) is a linear function of \( \varepsilon_i \), \( y_i \) is also a random variable. This implies that the mean value of \( y_i \) is \( \mu_y = \beta_0 + \beta_1 x \). The second assumption implies that the standard deviation of \( y_i \) is \( \sigma \) regardless of the \( x \) value. The third assumption implies that the distribution of \( y_i \) is normal. The fourth assumption implies that the values for the residuals \( (e_i = y_i - \hat{y}_i) \) should be distributed randomly about 0.

9.1.2 Estimating \( \sigma^2 \)

The **Sum Of Squares Error (SSE)** when divided by its degrees of freedom produces an unbiased estimate of \( \sigma^2 \). We say that the expected value of the mean square error (MSE) is equal to \( \sigma^2 \). Recall that MSE is defined to be SSE divided by its respective degrees of freedom. Mathematically, we write \( E(\text{MSE}) = \sigma^2 \) to show that the MSE is an unbiased estimator of \( \sigma^2 \). The SSE is defined as the sum of the squared residuals and can be represented many ways. Next, we provide five common and equivalent ways SSE can be represented in (9.5).

\[
\begin{align*}
\text{SSE} &= \sum_{i=1}^{n} e_i^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} y_i^2 - b_0 \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i y_i \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 - b_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum_{i=1}^{n} y_i^2 - \left( \frac{\sum_{i=1}^{n} y_i}{n} \right)^2 - \left( \frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} x_i}{n} \cdot \frac{\sum_{i=1}^{n} y_i}{n} \right)^2 - \frac{\left( \sum_{i=1}^{n} x_i \right)^2}{n} + \frac{\sum_{i=1}^{n} x_i^2}{n} \\
&= \sum_{i=1}^{n} x_i^2 - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2 \left( \frac{\sum_{i=1}^{n} y_i}{n} \right)^2 - \left( \frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} x_i}{n} \cdot \frac{\sum_{i=1}^{n} y_i}{n} \right)^2
\end{align*}
\]

(9.5)

The MSE is automatically printed when MINITAB performs a least squares regression. However, the value that is printed in the Session Window is rounded. To get a more precise estimate of MSE, choose **Stat > Regression > Regression**. When the Regression Window opens, select the appropriate Response and
9.2 Inference about the Regression Model

9.2.1 Hypothesis Test For \( \beta_1 \)

In more advanced books the sampling distribution of \( b_1 \) is shown to follow a normal distribution with mean \( \beta_1 \) and variance \( \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \). In other words,

\[
\mu_{b_1} = \beta_1 \tag{9.6}
\]

and

\[
\sigma^2_{b_1} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{9.7}
\]
9.2. Inference about the Regression Model

The standard error of $b_1$ is written as:

$$SE(b_1) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

(9.8)

The standardized test statistic to test a hypothesis concerning $\beta_1$ takes the form

$$t_{\text{obs}} = \frac{b_1 - \beta_1}{SE(b_1)}$$

(9.9)

Under the null hypothesis, the standardized test statistic follows a $t$ distribution with $n - 2$ degrees of freedom. The values for $t_{\text{obs}}$ in equation (9.9) and $SE(b_1)$ in equation (9.8) are given in the Session Window when performing regression with MINITAB™.

The test $H_0 : \beta_1 = 0$ is a test for linear association between $Y$ and $X$ for the simple linear model. In other words, only when we reject the null hypothesis are we able to assume a linear relationship between the dependent ($Y$) and independent ($X$) variable.

9.2.2 Hypothesis Test For $\beta_0$

Like $b_1$, the sampling distribution of $b_0$ is shown in more advanced books to follow a normal distribution with mean $\beta_0$ and variance $\sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$. In other words,

$$\mu_{b_0} = \beta_0$$

(9.10)

and

$$\sigma_{b_0}^2 = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

(9.11)

The standard error of $b_0$ is written as:

$$SE(b_0) = \text{MSE} \times \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

(9.12)

The standardized test statistic to test a hypothesis concerning $\beta_0$ takes the form

$$t_{\text{obs}} = \frac{b_0 - \beta_0}{SE(b_0)}$$

(9.13)

Under the null hypothesis, the standardized test statistic follows a $t$ distribution with $n - 2$ degrees of freedom.

Example 9.2.1: Analyze the regression model of Weight regressed on Chest.G from the worksheet BEARS.MTW located in the MTBWIN\DATA subdirectory. Use MINITAB™’s regression command to test for significant linear associations between two variables as well as to test the significance of the $y$-intercept in the simple linear model.

Solution: Output 9.1 on the following page shows the Session Window output from regressing Weight on Chest.G with MINITAB™’s regression command.
Output 9.1: Session Window Output From Regressing Weight on Chest.G

Regression Analysis: Weight versus Chest.G

The regression equation is
Weight = – 279 + 13.0 Chest.G

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>t</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-278.75</td>
<td>10.88</td>
<td>-25.61</td>
<td>0.000</td>
</tr>
<tr>
<td>Chest.G</td>
<td>12.9680</td>
<td>0.2923</td>
<td>44.36</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 28.68          R-Sq = 93.3%       R-Sq(adj) = 93.3%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1619241</td>
<td>1619241</td>
<td>1968.00</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>141</td>
<td>116012</td>
<td>829</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>142</td>
<td>1739253</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To create the output for yourself,
1. Select Stat>Regression>Regression
2. Select Weight as the Response variable and 'Chest.G' as the Predictor variable.
3. Click OK.

From Output 9.1 we see that \( b_1 \) is calculated to be 12.9680. Likewise, the standard error of \( b_1 \) is 0.2923. Notice that the value for the standardized test statistic when testing the hypothesis \( H_0 : \beta_1 = 0 \) is reported beneath T as 44.36. The standardized test statistic is calculated by hand as:

\[
t_{obs} = \frac{b_1 - \beta_1}{SE (b_1)} = \frac{12.9680 - 0}{0.2923} = 44.36
\]

which is the exact value MINITAB\(^\text{TM}\) created in Output 9.1. The \( p \)-value is reported beneath \( P \) as 0.000 which is calculated by determining \( 2 \times P(t_{140} \geq 44.36) \). Consequently, we conclude the evidence strongly suggests a linear relationship exists.

All \( p \)-values MINITAB\(^\text{TM}\) reports when using regression are for two tailed alternative hypotheses.

From Output 9.1, we also see that MINITAB\(^\text{TM}\) calculates and reports the value of \( b_0 \) and the standard error of \( b_0 \) as -278.75 and 10.88 respectively. Notice that the standardized test statistic for testing the hypothesis \( H_0 : \beta_0 = 0 \) versus \( H_0 : \beta_0 \neq 0 \) is reported beneath T as -25.61. The standardized test statistic is calculated by hand as:

\[
t_{obs} = \frac{b_0 - \beta_0}{SE (b_0)} = \frac{-278.75 - 0}{10.88} = -25.61,
\]

which is exact value MINITAB\(^\text{TM}\) shows in the Output 9.1. The \( p \)-value is reported beneath \( P \) as 0.000 which is calculated by determining \( 2 \times P(t_{140} \geq 25.61) \). Consequently, we conclude the evidence strongly suggests the y-intercept for the simple linear model (\( \beta_0 \)) is not 0.

### 9.2.3 Confidence Intervals For \( \beta_1 \) And \( \beta_0 \)

The confidence intervals for \( \beta_1 \) and \( \beta_0 \) are given by formulas (9.14) and (9.15) respectively.

\[
P \left( b_1 - t_{1-\alpha/2,n-2}SE (b_1) \leq \beta_1 \leq b_1 + t_{1-\alpha/2,n-2}SE (b_1) \right) = (1 - \alpha) \times 100\%
\]  
(9.14)
\[ P \left( b_0 - t_{1-\alpha/2,n-2}SE(b_0) \leq \beta_0 \leq b_0 + t_{1-\alpha/2,n-2}SE(b_0) \right) = (1 - \alpha) \times 100\% \]  
\hfill (9.15)

MINITAB\textsuperscript{TM} currently does not have a built-in function to calculate confidence intervals for the model’s parameters. However, confidence intervals are easily calculated because the standard regression output includes both the estimate of the parameter and its standard error. The degrees of freedom for the \( t \) distribution when using (9.14) and (9.15) will always be the same as those adjacent to Residual Error reported beneath DF in the Analysis of Variance table. With the previous data set, note that the degrees of freedom for error are 141. Output 9.2 shows MINITAB\textsuperscript{TM} commands issued in the Session Window used to calculate 95% confidence intervals for both \( \beta_1 \) and \( \beta_0 \) as well as comments explaining the MINITAB\textsuperscript{TM} commands. From Output 9.2 we see that the 95% confidence interval for \( \beta_1 \) is \[ (12.3901 \leq \beta_1 \leq 13.5459) \], while the 95% confidence interval for \( \beta_0 \) is \[ (-300.259 \leq \beta_0 \leq -257.241) \].

### 9.2.4 Confidence Interval For \( \mu_y \)

When calculating confidence intervals for \( \mu_y \), it is understood that we are calculating an interval for the average value of \( y \) given a particular value of \( x \), denoted \( x' \). In other words, we are calculating an interval to include values of \( y \) that might occur with some degree of confidence when the value of \( x, x' \) is held constant. The estimate of \( \mu_y \) is given by \( b_0 + b_1 x' \). We will denote the quantity \( b_0 + b_1 x' \) as \( \hat{y}_e \). The variance of \( \hat{y}_e, \sigma^2_{\hat{y}_e} \), is given in (9.16).

\[
\sigma^2_{\hat{y}_e} = \sigma^2 \left( \frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \tag{9.16}
\]

The standard error of \( \hat{y}_e \), is calculated by replacing \( \sigma^2 \) with \( \text{MSE} \) and taking the square root of the resulting quantity. The standard error of \( \hat{y}_e \), \( \text{SE}(\hat{y}_e) \), is:

\[
\text{SE}(\hat{y}_e) = \sqrt{\text{MSE} \left( \frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)} \tag{9.17}
\]
9.2. Inference about the Regression Model

The confidence interval for the expected response \( \mu_y \) given \( x' \) is given in (9.18).

\[
P \left( \hat{y}_e - t_{1-\alpha/2 ; n-2} SE(\hat{y}_e) \leq \mu_y \leq \hat{y}_e + t_{1-\alpha/2 ; n-2} SE(\hat{y}_e) \right) = (1 - \alpha) \times 100\% 
\]

(9.18)

To calculate a confidence interval for the expected response \( \mu_y \) given \( x' \) with MINITAB\textsuperscript{TM},

1. Select Stat > Regression > Regression.
2. Enter the appropriate Response (Y) and Predictor (X) variables in the Regression Dialog Window.
3. Click on the Options button. The Regression - Options Dialog Window will open resembling Figure 9.3.
4. Type the value of interest in the Prediction Intervals For New Observations Box.
5. Then click OK twice.

Figure 9.3: Regression - Options Dialog Window

The estimate \( \hat{y}_e \) for the expected response of \( \mu_y \) given \( x' \) and the standard error of \( \hat{y}_e \) along with a confidence interval and prediction interval are given at the bottom of the MINITAB\textsuperscript{TM} regression output. The estimate of \( \hat{y}_e \) is the number printed beneath Fit, while the standard error of \( \hat{y}_e \) is given beneath StDev Fit. The default confidence level is 95%. To change the confidence level, type the desired level in the Confidence Level Box found in the Regression - Options Dialog Window.

9.2.5 Prediction Interval For A New Observation \( y'_{(\text{new})} \) Given \( x' \)

When we calculate a prediction interval, we are calculating an interval for a single future value given a particular \( x' \). This differs from calculating a confidence interval for \( \mu \) given \( x' \), which is an interval for the average value of the \( y \)s given \( x' \). It is more difficult to assess the value of a future observation accurately than it is to assess the average of many future observations. Consequently, one should not use the formulas for a confidence interval for \( \mu_y \) given \( x' \) to determine a prediction interval for a future observation. The greater variability in a prediction interval is reflected in the standard deviation of \( y'_{(\text{new})} \). The estimate of \( y'_{(\text{new})} \), \( \hat{y}'_{(\text{new})} \), is still the quantity \( b_0 + b_1 x' \).
However, the standard error of \( \hat{y}'(\text{new}) \), \( SE(\hat{y}'(\text{new})) \), is slightly different from the standard error of \( \hat{y} \). The standard error of \( y'(\text{new}) \) is given in (9.19).

\[
SE(\hat{y}'(\text{new})) = \sqrt{\frac{\text{MSE}}{n} + \frac{(x' - \bar{x})^2}{\sum (x_i - \bar{x})^2}}
\] (9.19)

Note that the difference between the standard error of \( y'(\text{new}) \) and the standard error of \( \hat{y} \) is an extra 1 under the square root sign. The prediction interval formula for \( y'(\text{new}) \) given \( x' \) is given in (9.20).

\[
P \left( \hat{y}'(\text{new}) - t_{1-\alpha/2; n-2}SE(\hat{y}'(\text{new})) \leq \hat{y}'(\text{new}) \leq \hat{y}'(\text{new}) + t_{1-\alpha/2; n-2}SE(\hat{y}'(\text{new})) \right) = (1 - \alpha) \times 100\%
\] (9.20)

**Example 9.2.2:** Use Weight as the Response variable, and Chest.G. as the Predictor variable with the BEARS.MTW data set located in the MTBWIN\DATA folder to create a 95% confidence interval for the average weight of bears that have a chest girth of 50 inches.

**Solution:** Use formula (9.18) or MINITAB\textsuperscript{TM}. The result from using MINITAB\textsuperscript{TM} is shown in Output 9.3 on the following page. The steps one takes to calculate a prediction interval for \( y'(\text{new}) \) given \( x' \) with MINITAB\textsuperscript{TM} are identical to the procedures used earlier to calculate a confidence interval for \( \mu_y \) given \( x' \). That is, MINITAB\textsuperscript{TM} automatically produces both a confidence interval for \( \mu_y \) given \( x' \) and a prediction interval for \( \hat{y}'(\text{new}) \) given \( x' \) when the user enters the \( x' \) value in the Prediction Intervals for New Observations Box of the Regression - Options Dialog Window. The prediction interval is given to the right of the confidence interval in the MINITAB\textsuperscript{TM} output.

Note that the 95% confidence interval for the average weight of bears with chest girths of 50 inches is reported beneath 95.0% CI as (360.34, 378.87) in Output 9.3 on the next page. Likewise, the 95% prediction interval for the weight of a particular bear given that that bear’s chest girth is 50 inches is reported beneath 95.0% PI as (312.19, 427.11).

To construct several confidence and prediction intervals with MINITAB\textsuperscript{TM},

1. Store the values of the observations for which you want predictions in a single column.
2. Select Stat>Regression>Regression.
3. Select the appropriate Response (Y) and Predictor (X) variables.
4. Click on Options. In the Regression - Options Dialog Window, insert the column containing the new observations (\( x' \)s) in the Prediction Intervals for New Observations Box.
5. Click the OK button once in both the Regression-Options Dialog Window and the Regression Dialog Window.
9.2. Inference about the Regression Model

Output 9.3: Example 9.2.2 Confidence Interval Calculation

The regression equation is
\[ \text{Weight} = -279 + 13.0 \times \text{Chest.G} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-279.75</td>
<td>10.00</td>
<td>-27.61</td>
<td>0.000</td>
</tr>
<tr>
<td>Chest.G</td>
<td>12.9690</td>
<td>0.2923</td>
<td>44.36</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 28.69 \quad \text{R-Sq = 93.3} \quad \text{R-Sq(adj) = 93.3} \%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1619241</td>
<td>1619241</td>
<td>1968.00</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>141</td>
<td>116012</td>
<td>823</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>142</td>
<td>1735253</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicted Values for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95.0% CI</th>
<th>95.0% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>369.65</td>
<td>4.66</td>
<td>(360.43, 378.87)</td>
<td>(312.20, 427.10)</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Chest.G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50.0</td>
</tr>
</tbody>
</table>

9.2.6 Confidence And Prediction Bands

To graph confidence and/or prediction bands with MINITAB\textsuperscript{TM},

1. Select Stat > Regression > Fitted Line Plot. The Fitted Line Plot Dialog Window will open resembling Figure 9.4.

2. Select the appropriate Response (Y) and Predictor (X) variables.

3. Click the Options Box in the Fitted Line Plot Dialog Window. When the Fitted Line Plot - Options Dialog Window opens, click in the Display confidence bands, or Prediction Bands box(es) (to )

4. Click once on the OK in both the Fitted Line Plot - Options Dialog Window and the Fitted Line Plot Dialog Window.

Figure 9.4: Fitted Line Plot Dialog Window

The confidence bands and prediction bands shown in Figure 9.5 on the next page are 95\% confidence bands and 95\% prediction bands for the BEARS.MTW data set using Weight as the RESPONSE variable and Neck.G as the PREDICTOR variable. The information in worksheet SHUTTLE.MTW for Problem 9.20 from Basic Statistics and Data Analysis was collected over a one year period when a free shuttle service was offered from the suburban areas.
9.3 Checking Model Adequacy

In Section 9.2, we discussed how to find and make inferences about the regression coefficients \( b_0 \) and \( b_1 \) for models of the form \( y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i \). However, all of the inferences drawn about the regression coefficients \( b_0 \) and \( b_1 \) are dependent on the random error term \( \varepsilon_i \) adhering to the model assumptions from Section 9.1. Refresh your memory about the assumptions concerning \( \varepsilon_i \) by reviewing the Linear Regression Model Assumptions on page 222.

Briefly, the assumptions state \( \varepsilon_i \sim \mathcal{N}(0, \sigma) \). Since the \( \varepsilon_i \)s are unknown, we use the residuals as estimates of the \( \varepsilon_i \)s. If the assumptions about the \( \varepsilon_i \)s are true, then we can expect the residuals to be centered around 0, have constant variance, display a normal distribution, and show no dependencies. Graphical techniques are used in the residual assessment such as producing a histogram of the residuals, producing normal probability plots of the residuals, plotting the residuals versus predicted values, plotting residuals versus time, plotting residuals versus independent variables, and plotting residuals versus the dependent variable. Since the size of the residuals will depend on the particular problem at hand, it often facilitates residual analysis when the residuals are standardized. Standardized residuals may be found by calculating \( \varepsilon_i / s \), where \( s = \sqrt{\text{MSE}} \) is the estimate of \( \sigma \). MINITAB\textsuperscript{TM} has a feature which calculates standardized residuals; however, what MINITAB\textsuperscript{TM} actually calculates are Studentized residuals. These residuals take the form (residual)/(standard deviation of residual). Specifically, MINITAB\textsuperscript{TM} defines the standard deviation of the residual, \( \text{stdev}(\varepsilon_i) \), as:

\[
\text{stdev}(\varepsilon_i) = \sqrt{\text{MSE} - \text{var}(\hat{Y}_i)}
\]  

(9.21)
9.3. **Checking Model Adequacy**

where

\[
\text{var} \left( \hat{Y}_i \right) = \text{MSE} \cdot x_i \left( X'X \right)^{-1} x_i \tag{9.22}
\]

and \( x_i \) is the \( i \)-th row of the design matrix \( X \). The MINITAB\textsuperscript{TM} defined standardized residuals have variance equal to 1. Consequently, residuals over 2 are usually considered large and warrant further investigation.

To investigate the residuals for a specific regression problem with MINITAB\textsuperscript{TM},

1. First create the regression line by selecting **Stat > Regression > Regression**.
2. Select the appropriate Response \((Y)\) and Predictor \((X)\) variables in the Regression Dialog Window.
3. Click on the **Graphs** button inside the Regression Dialog Window.
4. Select the the type of residual plot of interest (Regular, Standardized, or Deleted) (\(\square\) to \(\bigcirc\)) and the graphs of interest (Histogram of residuals, Normal plot of residuals, Residuals versus fits, Residuals versus order, or Residuals versus another variable) (\(\square\) to \(\square\) for each) from the Regression - Graphs Dialog Window. See Figure 9.6 for a picture of the Regression - Graphs Dialog Window.
5. Examine the graphs you produce for problems.

**Figure 9.6: Regression - Graphs Dialog Window**

Video 9.2 conducts an analysis of the residuals after fitting a line to the data given in worksheet *Hardwood.MTW*.

Video 9.2: For **Problem 9.34** — (Duration: 6 minutes 6 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 \(\times\) 768 pixels.

To verify or change your display panel resolution, select **Start > Settings > Control Panel > Display > Settings**.

The histogram of residuals can be used to see if the **Linear Regression Model Assumptions** 1 and 3 from page 222 are satisfied. The normal plot of residuals can also be used to assess assumption 3 from the **Linear Regression Model Assumptions**. Assumptions 2 and 4 from the **Linear Regression Model Assumptions** can be verified by looking at the residuals versus fits graph and the residual versus order graph. The assumption of
constant standard deviation is usually satisfied when the residuals versus fit graph shows no discernable pattern. (See Figure 2.22 on page 69 for an example of constant standard deviation.) A graph of the residuals versus order is usually created to look for any dependencies in the data. That is, if there is a pattern in the residuals versus order graph, assumption 4 (independent error terms) has been violated.

**Example 9.3.1:** Evaluate the residuals from Example 9.2.2 on page 228 to see if the fitted regression model is adequate.

**Solution:** Figures 9.7, 9.8, 9.9, and 9.10 are standardized residual plots created from following steps outlined to investigate the residuals for a specific regression problem. Figure 9.7 is a relatively symmetric histogram centered at 0. Likewise, the normal probability plot shown in Figure 9.8 indicates a fairly normal distribution for the residuals.

When the residuals are plotted against the fitted values in Figure 9.9 on the next page, one would expect a constant band to appear around 0 if all assumptions are being met and the proper variables are included in the model. However, there appears to be a definite curvature. This curvature indicates that the model is predicting values that are too small for small values of \( \text{Chest.G} \) as well as predicting values that are too small for larger values of \( \text{Chest.G} \). This systematic curvature is a sign that we should reevaluate the proposed model. The final graph MINITAB\textsuperscript{TM} creates (Figure 9.10 on the following page) depicts residuals versus observation and shows no apparent abnormalities. If patterns are visible in a plot of residuals versus observation, it may indicate a time dependency. Since the plot of the residuals versus predicted values (Figure 9.9 on the next page) indicated an inadequate model, a quadratic model is fit using the Fitted Line Plot command by clicking in the circle to the left...
9.3. Checking Model Adequacy

Figure 9.9: Residuals Plotted Against Fitted Values from Example 9.2.2

Figure 9.10: Residuals Plotted Against Observation from Example 9.2.2

Figure 9.11: Checked Circles in Fitted Line Plot Dialog Window

of Quadratic (✓) to (✗) beneath the Type of Regression Model in the Fitted Line Plot Dialog Window as shown in Figure 9.11. Subsequently, the Storage button was selected in the Fitted Line Plot Dialog Window and the Residuals and Fits boxes were selected (✓) to (✗) for both as shown in Figure 9.12 on the next page. The results from the fitted quadratic model are shown in Figure 9.13 on the following page. The quadratic fit appears to be an improvement. To assess the residuals from the quadratic model, we can use either MINITAB TM’s residual diagnostic program or create our own graphs of the residuals as the residuals and fits are stored in the current worksheet.
9.3. Checking Model Adequacy

Figure 9.12: Checked Boxes in *Fitted Line Plot - Storage Dialog Window*

![Fitted Line Plot - Storage Dialog Window](image)

Figure 9.13: Fitted Quadratic Model for Example 9.2.2

![Regression Plot](image)

To use MINITAB™’s residual diagnostics program,

1. Select *Stat > Regression > Residual Plots*.

2. When the *Residual Plots Dialog Window* opens, enter the column where the residuals are stored in the *Residuals Box* and the columns where the predicted values (fits) are stored in the *Fits Box*.

3. Examine the resulting graphs for departures from the **Linear Regression Model Assumptions** (page 222).

The MINITAB™ Residual Plots command can only be used if the residuals and fits have been a priori stored in a MINITAB™ worksheet.

The residual model diagnostics shown in Figure 9.14 on the next page are for the quadratic model shown in Figure 9.13.

Note that the curvature in the residuals versus fits graph of Figure 9.9 on the page before is no longer as pronounced in the residuals versus fits graph (lower right graph in Figure 9.14) for the quadratic model.

Video 9.3 on the following page provides a reminder of how to create plots to verify that a linear regression model is appropriate as well as how to identify outliers.
Figure 9.14: Fitted Quadratic Model Diagnostics for Example 9.2.2

Video 9.3: For Problem 9.62 — (Duration: 4 minutes 14 seconds)

For optimal video viewing, set your computer’s display panel resolution to 1024 × 768 pixels.
To verify or change your display panel resolution, select Start>Settings>Control Panel>Display>Settings.
9.4 Summary and Review Labs

Lab 9.1 — The Linear Regression Model

Objectives:

I. To create a draftsman’s plot to examine multiple relationships at once

II. To use MINITAB’s calculator to compute statistics related to simple linear regression

III. To verify your calculations with MINITAB’s built in regression command

Basic Directions:

Answer all questions with complete sentences. Make sure all graphs and output are copied to the report pad for your final product. Be sure to enable the commands in the Session Window before starting the lab (Editor>Enable Commands). By enabling the Session Window commands, you will be able to show your work by copying the appropriate output from the Session Window and pasting the copied contents into your report pad. However, if you do not want the MINITAB commands to appear in the Session Window, you can still view all MINITAB commands issued during a MINITAB session by viewing the History Folder. To view the contents of the History Folder click on the Project Manager Icon ( ) then click on the History Folder.

Introduction:

The worksheet STA261.MTW located in the MTBWIN\Student1 directory contains the complete grade information for 45 students in an introductory statistics class. Specifically, the worksheet includes the scores from 11 quizzes, the scores from three one hour exams, and the scores on the final exam. The quiz scores are all out of 15 possible points and the four exams are out of 110 possible points (10 points extra credit). Columns C1-C11 labelled Q1-Q11 contain quiz scores, columns C12-C14 labelled T1-T3 contain hour exam scores, and column C15 labelled Final contains the final exam score.

Questions and Directions:

Use row statistics (Calc>Row Statistics) to calculate the quiz mean (Qmean), quiz median (Qmedian), test mean (Tmean), and test median (Tmedian). Then, produce a draftsman plot of Final versus (Qmean, Qmedian, Tmean, and Tmedian).

1. Based on the draftsman plot, which of the four variables seems to be the most linearly related to Final?

2. Verify your answer from 1 by calculating the correlations among Final versus Qmean, Qmedian, Tmean, and Tmedian with MINITAB’s Correlation command.

3. Suppose we are interested in predicting a student’s score on the final (Final) based on knowledge of the mean test scores (Tmean). Use simple linear regression to estimate the parameters for equation (9.1).

4. Use MINITAB’s calculator to compute \( b_1 \) and \( b_0 \) according to the formulas given in equations (9.2) and (9.3) respectively.

5. Use MINITAB’s calculator to compute SSE with any one of the five formulas given in (9.5).

6. Use MINITAB’s regression command (Stat>Regression>Regression) with Final as the response variable and Tmean as the predictor variable. Paste the results into the report pad. Are the values you calculated for \( b_1 \), \( b_0 \), and SSE in steps 4 and 5 in agreement with the values produced from the Stat>Regression>Regression command? If not, why not?

7. Store your current worksheet as STA261A.MTW. Make sure you have access to this worksheet for the next lab or else you will have to recreate columns Qmean, Qmedian, Tmean, and Tmedian.
Lab 9.2 — Inference About the Regression Model

Objectives:

I. To calculate confidence intervals for averages at various \( x \) values

II. To calculate predictions intervals for various \( x \) values

Basic Directions:

Answer all questions with complete sentences. Make sure all graphs and output are copied to the report pad for your final product. Be sure to enable the commands in the Session Window before starting the lab (Editor>Enable Commands). By enabling the Session Window commands, you will be able to show your work by copying the appropriate output from the Session Window and pasting the copied contents into your report pad. However, if you do not want the MINITAB\textsuperscript{TM} commands to appear in the Session Window, you can still view all MINITAB\textsuperscript{TM} commands issued during a MINITAB\textsuperscript{TM} session by viewing the History Folder. To view the contents of the History Folder click on the Project Manager Icon (■) then click on the History Folder.

Introduction:

Open the worksheet STA261A.MTW you created in lab 9.1. Recall that this worksheet includes the scores from 11 quizzes, the scores from three one hour exams, and the scores on the final exam. The quiz scores are all out of 15 possible points and the four exams are out of 110 possible points (10 points extra credit). Columns C1-C11 labelled Q1-Q11 contain quiz scores, columns C12-C14 labelled T1-T3 contain hour exam scores, and column C15 labelled Final contains the final exam score. Columns C16-C19 should contain the quiz mean (Qmean), quiz median (Qmedian), test mean (Tmean), and test median (Tmedian) respectively.

Questions and Directions:

In this lab you will produce confidence intervals for the average Final scores and prediction intervals for the Final scores with the following values for the test mean (Tmean). Tmean =\{65, 70, 75, 80, 85, 90, 95\}.

1. Use simple linear regression to estimate the parameters for model (9.1) by regressing Final on Tmean.

2. Explain the difference between \( \mu_y \) and \( y'(\text{new}) \).

3. Determine the value of \( \hat{y}_e \) for Tmean values of 65, 70, 75, 80, 85, 90, and 95.

4. Determine the value of SE(\( \hat{y}_e \)) for Tmean values of 65, 70, 75, 80, 85, 90, and 95. Explain why the value for SE(\( \hat{y}_e \)) increases as you move in either direction from the Tmean value of 75.53.

5. Determine the value of SE(\( \hat{y}'_{(\text{new})} \)) for Tmean values of 65, 70, 75, 80, 85, 90, and 95. Explain why the value for SE(\( \hat{y}'_{(\text{new})} \)) increases as you move in either direction from the Tmean value of 75.53.

6. Use formulas (9.18) and (9.20) respectively to calculate 90\% confidence intervals and 90\% prediction intervals for Tmean values of 65, 70, 75, 80, 85, 90, and 95.

7. Enter the Tmean values of 65, 70, 75, 80, 85, 90, and 95 in a column named TE. Use MINITAB\textsuperscript{TM}'s built in capability to calculate 90\% confidence and 90\% prediction intervals by typing TE in the prediction intervals for new observations box and changing the default confidence level from 95 to 90 in the Confidence Level Box of the Regression-Options Dialog Window. Are the answers MINITAB\textsuperscript{TM} produced identical to the values calculated in step 6 with formulas (9.18) and (9.20)? If not, why not?
Lab 9.3 — Checking Model Adequacy

Objectives:
I. To assess a linear regression model graphically and numerically
II. To determine how grades should be calculated

Basic Directions:
Answer all questions with complete sentences. Make sure all graphs and output are copied to the report pad for your final product. Be sure to enable the commands in the Session Window before starting the lab (Editor>Enable Commands). By enabling the Session Window commands, you will be able to show your work by copying the appropriate output from the Session Window and pasting the copied contents into your report pad. However, if you do not want the MINITAB™ commands to appear in the Session Window, you can still view all MINITAB™ commands issued during a MINITAB™ session by viewing the History folder. To view the contents of the History Folder click on the Project Manager Icon (Project Manager Icon) then click on the History Folder.

Introduction:
Open the worksheet STA261A.MTW you created in lab 9.1. Recall that this worksheet includes the scores from 11 quizzes, the scores from three one hour exams, and the scores on the final exam. The quiz scores are all out of 15 possible points and the four exams are out of 110 possible points (10 points extra credit). Columns C1-C11 labelled Q1-Q11 contain quiz scores, columns C12-C14 labelled T1-T3 contain hour exam scores, and column C15 labelled Final contains the final exam score. Columns C16-C19 should contain the quiz mean (Qmean), quiz median (Qmedian), test mean (Tmean), and test median (Tmedian) respectively.

In this lab, you will assess your fitted model from lab 9.2 with graphical and numerical procedures to see how well it fits the data. You will also use MINITAB™ to calculate final grades for the nameless students in worksheet STA261A.MTW. Use the simple linear regression model found in equation (9.1).

Questions and Directions:
1. Regress Final on Tmean and create a histogram of the residuals, a normal plot of the residuals, a plot of residuals versus fits, and a plot of residuals versus order. (Stat>Regression>Regression>Graphs)
2. Tile the four graphs created in step 1. Comment on each graph. You may want to review what each graph is depicting and how it is used to assess the fitted line by right clicking on a graph and selecting StatGuide.
3. MINITAB™ identifies two values as outliers and one value as both an outlier and an influential observation. Identify these three values in each of your four graphs with a red pen. Should these three values be removed or do you think they are representative of what actually happens with respect to student grades?
4. The course syllabus specifies that the final grade will be calculated as follows: 30% Quizzes, 45% Tests, and 25% Final. However, it does not specify what measure of center or even if a measure of center will be used with the 30% and 45% for quizzes and tests respectively.
   a. Calculate the students’ final grades assuming the teacher uses 30% of the mean quiz grade, 45% of the mean test grade, and 25% of the final test to determine the final grade. Store the result in a column named FinalMean. (Calc>Calculator)
   b. Calculate the students’ final grades assuming the teacher uses 30% of the median quiz grade, 45% of the median test grade, and 25% of the final test to determine the final grade. Store the result in a column named FinalMedian. (Calc>Calculator)
   c. Regress FinalMedian on FinalMean. What percent of the variability in FinalMedian is explained by FinalMean?
   d. If you were given an option, would you prefer your instructor to grade using means or medians? Explain your answer.
   e. Do you think using the mean or the median produces a better estimate of the students’ performance over the semester? Explain your answer.
Chapter 10
Analysis of Variance

10.1 Introduction

The idea of analysis of variance is that we are comparing the centers of multiple groups together rather than merely two groups as we have before. The One-Way ANOVA test will show you how to compare means and to generate confidence intervals for differences of the means of groups, while the Kruskal-Wallis test deals with medians. MINITAB is used extensively as the equations are calculation intensive.

10.2 The One-Way ANOVA

The name analysis of variance is derived from a partitioning of the total variability into its component parts. In a setting with different treatments and observations per treatment, the total corrected sum of squares is:

$$SS_{Total} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_.)^2$$  \hspace{1cm} (10.1)

In a more advanced class, it is shown that the sum of squares total ($SS_{Total}$) can be broken into two separate components. The first component is known as the sum of squares treatment, ($SS_{Treat}$), and the second component is the sum of squares due to error, ($SS_{Error}$). The dot notation is used to simplify the presentation of sums. The total of the observations in the $i$th treatment is denoted $y_i$, where the dot indicates all of the observations in the $i$th treatment have been summed.

$$y_i. = \sum_{j=1}^{n} y_{ij} = y_{i1} + y_{i2} + \cdots + y_{in}$$  \hspace{1cm} (10.2)

The double dot notation is used to indicate that observations have been added over both subscripts.

$$\bar{y}. = \frac{1}{a} \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}$$  \hspace{1cm} (10.3)

Therefore, we have

$$\sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_.)^2 = \sum_{i=1}^{a} n_i (\bar{y}_i. - \bar{y}.)^2 + \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i.)^2$$  \hspace{1cm} (10.4)

$$SS_{Total} = SS_{Treat} + SS_{Error}$$

Analysis of variance generalizes the 2 sample pooled t-test to $a$ samples. In the one way design, we compare a population means by testing the hypothesis:

$$H_O : \mu_1 = \mu_2 = \cdots = \mu_a \hspace{0.5cm} \text{versus} \hspace{0.5cm} H_A : \mu_i \neq \mu_j \hspace{0.5cm} \text{for at least one pair} \ (i, j).$$  \hspace{1cm} (10.5)
The standardized test statistic used to evaluate the sample information is $F_{obs} = \frac{MST}{MSE}$. Note that if the null hypothesis is true, MST provides an unbiased estimate of $\sigma^2$ and that the ratio of $MST/MSE$ will be close to 1. Large values of $F_{obs}$ provide evidence against the null hypothesis. Under the null hypothesis, $F_{obs} \sim F_{(a-1),(N-a)}$. The $F$ distribution like the $\chi^2$ distribution is skewed to the right. Likewise, the shape of the $F$ distribution also depends on its degrees of freedom. The $F$ distribution has both degrees of freedom in the numerator (the first number: $a - 1$) and degrees of freedom in the denominator (the second number: $N - a$). Figure 10.1 depicts $F$ distributions with (20, 20), and (3, 20) degrees of freedom in blue and red respectively.

The analysis of variance assumes:

1. The $a$ samples are randomly and independently selected from their respective populations.
2. The $a$ sampled populations are normally distributed.
3. The variances of the $a$ populations are all equal.

Checking the one-way analysis of variance assumptions entails evaluating each of the $a$ samples for normality and homogeneity of variance. In practice, the normality assumption and the equality of variance assumptions are not that important provided the number of observations in each treatment are relatively similar. The normality assumption is best tested with a normal probability plot. Equality of variance can be assessed using side by side boxplots of the $a$ samples. Provided the ratio of the largest sample standard deviation to the smallest sample standard deviation is less than 2, it is safe to assume the variances of the $a$ populations are homogeneous.

When working with analysis of variance type problems, it is customary to summarize the information from the data in tabular format often referred to as the analysis of variance table. Table 10.1 on the following page provides this format and formulas for the one-way analysis of variance.

In addition to the ANOVA table for one-way analysis of variance problems, confidence intervals for the $i^{th}$ treatment mean are often provided using equation (10.6).

$$P\left( \bar{y}_i - t_{1-\alpha/2;N-a} \frac{\sqrt{MSE}}{n_i} < \mu_i < \bar{y}_i + t_{1-\alpha/2;N-a} \frac{\sqrt{MSE}}{n_i} \right) = (1 - \alpha) \times 100\%$$

$$(10.6)$$

MINITAB\textsuperscript{TM} commands that refer to unstacked data imply that the measurements for each group/treatment are contained in separate columns. MINITAB\textsuperscript{TM} commands that refer to stacked data on the
### Table 10.1: Format for an ANOVA Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>( F_{\text{obs}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between treatments</td>
<td>( a - 1 )</td>
<td>( SS_{\text{Treat}} = \sum_{i=1}^{a} n_i (\bar{y}<em>i - \bar{y}</em>{\cdot})^2 )</td>
<td>( MS_{\text{Treat}} = \frac{SS_{\text{Treat}}}{a - 1} )</td>
<td>( F_{\text{obs}} = \frac{MS_{\text{Treat}}}{MS_{\text{Error}}} )</td>
</tr>
<tr>
<td>Error (within treatments)</td>
<td>( N - a )</td>
<td>( SS_{\text{Error}} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 )</td>
<td>( MS_{\text{Error}} = \frac{SS_{\text{Error}}}{N - a} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( N - 1 )</td>
<td>( SS_{\text{Total}} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot})^2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The One-Way ANOVA

other hand imply that the measurements/responses for all groups are contained in a single column with a separate column containing group/treatment membership. MINITAB\textsuperscript{TM} will create a one-way analysis of variance table for both unstacked and stacked data using the commands OneWay(Unstacked), and OneWay respectively. OneWay has numerous features including several multiple comparisons procedures that are not present in OneWay(Unstacked). Consequently, there will be times when data is stored in multiple columns, and it will prove beneficial to stack the multiple columns of data into a single column of data in order to use OneWay. Both OneWay(Unstacked) and OneWay automatically produce 95% confidence intervals for the mean of each treatment based on a pooled standard deviation and display the results in the Session Window. Since the user will usually be interested in some form of multiple comparisons if the initial analysis of variance table indicates there are group/treatment differences, we will only present the OneWay command.

If your data are in an unstacked format, create a stacked format by

1. Selecting Manip\textsuperscript{TM}\gt Stack\textsuperscript{TM}\gt Stack Columns and entering the columns that contain the measurement data for the individual groups/treatments in the Stack the following Columns Box of the Stack Columns Dialog Window.
2. Click OK.

To conduct a one-way analysis of variance with MINITAB\textsuperscript{TM},

1. Select Stat\textsuperscript{TM}\gt ANOVA\textsuperscript{TM}\gt OneWay
2. Enter the appropriate columns in the Response and Factor Boxes of the One-way Analysis of Variance Dialog Window. The column containing the measurement data should be entered in the Response Box. The column containing group/treatment membership should be entered in the Factor Box.
3. If you want to conduct multiple comparisons click on the Comparisons Box in the One-way Analysis of Variance Dialog Window. Once the One-way Multiple Comparisons Dialog Window opens, select the multiple comparison procedure(s) and \( \alpha \) level you would like to use then click OK.
4. If you want side by side dotplots and or side-by-side boxplots of the treatments or any of a number of residual diagnostic plots click on the Graphs Box in the One-way Analysis of Variance Dialog Window. When the One-way Analysis of Variance - Graphs Dialog Window opens, click in the square to the left of the type of graph you desire (\( \square \) to \( \square \)). Then, click OK.
5. Click OK.

**Example 10.2.1:** Consider Problem 10.17 from Basic Statistics and Data Analysis. In Problem 10.17, there are four groups of 11 students. Each of the four groups uses a different form of programmed learning to study statistics. A standard test was administered to the four groups and graded on a 15-point scale. Given these results, determine whether there was a significant difference in the results of the four methods. The results for the four methods are presented in Table 10.2 on the next page.
10.2. The One-Way ANOVA

Table 10.2: Test Results from Example 10.2.1

<table>
<thead>
<tr>
<th>Method1</th>
<th>Method2</th>
<th>Method3</th>
<th>Method4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

**Solution:** The five step procedure is used to test whether there is a significant difference among the four methods used to learn statistics.

1. \( H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \)
2. \( H_A : \mu_i \neq \mu_j \) for at least one pair \((i, j)\)
3. Based on side-by-side boxplots (Figure 10.2) and normal probability plots of the four groups (Figure 10.3 on the next page), the test statistic \( MS_{\text{Treat}} / MS_{\text{Error}} \) is selected to test the null hypothesis.
4. The sampling distribution of \( MS_{\text{Treat}} / MS_{\text{Error}} \) follows the central \( F \) distribution with \((a - 1)\) and \((N - a)\) degrees of freedom when the null hypothesis is true.

\[
F_{\text{obs}} = \frac{SS_{\text{Treat}} / (a - 1)}{SS_{\text{Error}} / (N - a)} = \frac{32.4545/3}{60.7273/40} = 7.12575
\]

4. Note that the \( p \)-value is calculated as \( P(F_{3,40} \geq 7.12575) = 0.0006043 \). A small \( p \)-value such as 0.0006 indicates that observing values as extreme as 7.12575 or more when the null hypothesis is true is very unlikely.

5. There is strong statistical evidence to suggest the mean scores for the four methods used to learn statistics are not the same.

Figure 10.2: Side-By-Side Boxplots for Example 10.2.1

Commands to calculate \( SS_{\text{Total}}, SS_{\text{Treat}}, SS_{\text{Error}}, F_{\text{obs}} \) and the \( p \)-value are illustrated in Output 10.1 on the following page.
10.2. The One-Way ANOVA

Figure 10.3: Normal Probability Plots of the Four Groups for Example 10.2.1

Output 10.1: Commands to calculate $SS_{\text{Total}}$, $SS_{\text{Treat}}$, $SS_{\text{Error}}$, $F_{\text{obs}}$ and the $p$-Value for Example 10.2.1

Note that the $p$-value for Example 10.2.1 is found by subtracting from 1 the area to the left of $F_{\text{obs}} = 7.1257$ which is found by using the CDF command applied to an $F$-distribution with 3 and 40 degrees of freedom.

The completed ANOVA table for Example 10.2.1 is shown in Table 10.3. The results from using Stat>ANOVA>One-Way are also provided in Output 10.2 on the following page.

Table 10.3: ANOVA Table for Example 10.2.1

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
<th>$p$-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>3</td>
<td>32.4545</td>
<td>10.8182</td>
<td>7.1257</td>
<td>0.0006</td>
</tr>
<tr>
<td>Error</td>
<td>40</td>
<td>60.7273</td>
<td>1.5182</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>93.1818</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that the numbers from Output 10.2 agree with the by hand results from Table 10.3 when rounded to three decimal places.

95% confidence intervals for the four methods are graphically displayed by default in Output 10.2 on the next page. Output 10.3 on the following page shows how MINITAB™ can be used to calculate 95% confidence intervals for the four methods from Example 10.2.1 on page 241 using formula (10.6).
Output 10.2: Stat\textgreater{}ANOVA\textgreater{}One-Way Results for Example 10.2.1

One-way ANOVA: Grade versus Method

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>3</td>
<td>32.45</td>
<td>10.82</td>
<td>7.13</td>
<td>0.001</td>
</tr>
<tr>
<td>Error</td>
<td>40</td>
<td>60.73</td>
<td>1.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>93.18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% CIs For Mean Based on Pooled StDev

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method1</td>
<td>11</td>
<td>4.818</td>
<td>1.601</td>
<td>4.089</td>
<td>5.547</td>
</tr>
<tr>
<td>Method2</td>
<td>11</td>
<td>6.182</td>
<td>1.160</td>
<td>4.910</td>
<td>7.454</td>
</tr>
<tr>
<td>Method3</td>
<td>11</td>
<td>7.091</td>
<td>1.136</td>
<td>5.965</td>
<td>8.216</td>
</tr>
<tr>
<td>Method4</td>
<td>11</td>
<td>5.364</td>
<td>0.924</td>
<td>4.530</td>
<td>6.198</td>
</tr>
</tbody>
</table>

Pooled StDev = 1.232

Output 10.3: Using MINITAB\textsuperscript{TM} to Calculate 95% Confidence Intervals for Example 10.2.1 on page 241

MTB > let c5(1)=mean(c1)
MTB > let c5(2)=mean(c2)
MTB > let c5(3)=mean(c3)
MTB > let c5(4)=mean(c4)
MTB > Name C6 = 'Sp'
MTB > Let 'Sp' = .1232
MTB > InvCDF .975 'CritT';
SUBC> T 40.
MTB > Name C7 = 'LowerLimit'
MTB > Let 'LowerLimit' = 'ybar'-('CritT'*'Sp'/SQRT(11))
MTB > Name C9 = 'UpperLimit'
MTB > Let 'UpperLimit' = 'ybar'+('CritT'*'Sp'/SQRT(11))
MTB > Print C7-C8

Data Display

<table>
<thead>
<tr>
<th>Row</th>
<th>LowerLimit</th>
<th>UpperLimit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.06743</td>
<td>5.56893</td>
</tr>
<tr>
<td>2</td>
<td>5.43107</td>
<td>6.93257</td>
</tr>
<tr>
<td>3</td>
<td>6.34016</td>
<td>7.84166</td>
</tr>
<tr>
<td>4</td>
<td>4.61288</td>
<td>6.11439</td>
</tr>
</tbody>
</table>
10.2.1 Multiple Comparisons

In the one way design, we compare \( a \) population means by testing the hypothesis:

\[
H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \quad \text{versus} \quad H_A : \mu_i \neq \mu_j \quad \text{for at least one pair } (i, j).
\] (10.7)

When the null hypothesis is rejected, we still do not know which treatments have the larger or smaller means. When differences between the treatments exist, we can further analyze the differences with multiple comparison procedures.

Note that multiple comparison procedures are used only after the null hypothesis has been rejected.

Numerous multiple comparison procedures are in existence. The two multiple comparison procedures addressed here are Tukey’s confidence intervals, and Fisher’s confidence intervals. The choice of Tukey’s or Fisher’s confidence intervals depends on which error rate, individual or family, you wish to control.

**Tukey’s Confidence Intervals**  Tukey’s pairwise confidence interval for \( \mu_i - \mu_j \) based on the studentized range statistic is given by formula (10.8).

\[
\mathbb{P} \left( (\bar{y}_i - \bar{y}_j) - q_{1-a,a,N-a} \sqrt{\frac{MSE}{2}} \left( \frac{1}{n_i} + \frac{1}{n_j} \right) \leq \mu_i - \mu_j \leq (\bar{y}_i - \bar{y}_j) + q_{1-a,a,N-a} \sqrt{\frac{MSE}{2}} \left( \frac{1}{n_i} + \frac{1}{n_j} \right) \right) = (1 - \alpha) \times 100\%
\] (10.8)

where

\[
\bar{y}_i = \text{the sample mean for treatment } a \\
n_i = \text{the number of observations in treatment } i \\
a = \text{the number of treatment levels} \\
q_{1-a,a,N-a} = \text{the upper } 1 - \alpha \text{ point of the studentized range with parameters } a \text{ and } N - a.
\]

Tukey’s pairwise confidence intervals for all differences in treatment means controls **family error rate**. **Family error rate** is the maximum probability of obtaining one or more confidence intervals that do not contain the true difference between treatment means. Family error rate differs from individual error rate since the individual error rate is the probability a specific confidence interval will not contain the true difference in treatment means. The family error rate is also known as the **experimentwise** error rate, while the individual error rate is often referred to as the **comparisonwise** error rate. To request Tukey’s confidence intervals choose **Stat>ANOVA>Oneway**. From the Oneway Window, click on the **Comparisons Box**. When the Comparisons Window opens, select the appropriate multiple comparison procedure. The One-way Multiple Comparisons Dialog Window is shown in Figure 10.4 on the next page. The default error rate for all of the multiple comparison procedures is 5%. To change the default value, type the desired rate in the appropriate window. Example 10.2.1 on page 241 concluded that the evidence from the samples suggests differences exist among the means of the four programmed learning styles. However, which one is the best, next to best etc. was not answered. The analysis of variance table along with Tukey’s confidence intervals for all pairwise differences in treatment means are shown in Output 10.4 on the next page. Recall that confidence intervals are not considered significant when they contain 0. Based on Tukey’s confidence intervals, method 3 is superior to method 4.
10.2. The One-Way ANOVA

Figure 10.4: *One-way Multiple Comparisons Dialog Window*

Output 10.4: ANOVA Table and Tukey’s Confidence Intervals for All $\mu_i - \mu_j$ from Example 10.2.1

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>3</td>
<td>32.45</td>
<td>10.82</td>
<td>7.13</td>
<td>0.001</td>
</tr>
<tr>
<td>Error</td>
<td>40</td>
<td>60.73</td>
<td>1.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>93.18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% CIs For Mean Based on Pooled SDer

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method1</td>
<td>11</td>
<td>4.018</td>
<td>1.001</td>
<td>(-----*-----)</td>
</tr>
<tr>
<td>Method2</td>
<td>11</td>
<td>6.102</td>
<td>1.168</td>
<td>(-----*-----)</td>
</tr>
<tr>
<td>Method3</td>
<td>11</td>
<td>7.091</td>
<td>1.136</td>
<td>(-----*-----)</td>
</tr>
<tr>
<td>Method4</td>
<td>11</td>
<td>5.364</td>
<td>0.924</td>
<td>(-----*-----)</td>
</tr>
</tbody>
</table>

Pooled SDer = 1.232

4.0  6.0  7.2

Tukey’s pairwise comparisons

* Family error rate = 0.0500
* Individual error rate = 0.0106

Critical value = 3.70

Intervals for (column level mean) - (row level mean)

<table>
<thead>
<tr>
<th>Method1</th>
<th>Method2</th>
<th>Method3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method2</td>
<td>-2.772</td>
<td>0.044</td>
</tr>
<tr>
<td>Method3</td>
<td>-3.081</td>
<td>-2.317</td>
</tr>
<tr>
<td></td>
<td>-0.885</td>
<td>0.499</td>
</tr>
<tr>
<td>Method4</td>
<td>-1.953</td>
<td>-0.580</td>
</tr>
<tr>
<td></td>
<td>0.063</td>
<td>2.226</td>
</tr>
</tbody>
</table>
Fisher’s Confidence Intervals  Fisher’s pairwise confidence interval for \( \mu_i - \mu_j \) is given in equation (10.9).

\[
\Pr \left( \bar{y}_i - \bar{y}_j - t_{1-\alpha/2,N-a} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \leq \mu_i - \mu_j \leq \bar{y}_i - \bar{y}_j + t_{1-\alpha/2,N-a} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \right) = (1 - \alpha) \times 100\%
\]

(10.9)

where

- \( \bar{y}_i \) = the sample mean for treatment \( a \)
- \( n_i \) = the number of observations in treatment \( i \)
- \( a \) = the number of treatment levels
- \( t_{1-\alpha/2,N-a} \) = the upper \( 1 - \alpha/2 \) point of the \( t \)-distribution with \( N - a \) degrees of freedom.

Fisher’s pairwise confidence intervals control the individual (comparisonwise) error rate. To request Fisher’s pairwise comparisons choose \textbf{Stat} > \textbf{ANOVA} > \textbf{Oneway}. From the \textit{Oneway Window}, click on the \textit{Comparisons Box}. When the \textit{Comparisons Window} opens, select the appropriate multiple comparison procedure and error rate. Example 10.2.1 on page 241 is used again with the four programmed learning styles to produce pairwise confidence intervals using Fisher’s confidence interval procedures. The analysis of variance table along with Fisher’s confidence intervals for all pairwise differences in treatment means are shown in Output 10.5.

Output 10.5: ANOVA Table and Fisher’s Confidence Intervals for All \( \mu_i - \mu_j \) from Example 10.2.1

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>3</td>
<td>32.55</td>
<td>10.82</td>
<td>7.13</td>
<td>0.001</td>
</tr>
<tr>
<td>Error</td>
<td>40</td>
<td>66.73</td>
<td>1.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>99.28</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% Cls For Mean Based on Pooled SdDev

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>SdDev</th>
<th>-----</th>
<th>------</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method1</td>
<td>11</td>
<td>4.918</td>
<td>1.081</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Method2</td>
<td>11</td>
<td>6.102</td>
<td>1.160</td>
<td>(-+---+)</td>
<td>------</td>
</tr>
<tr>
<td>Method3</td>
<td>11</td>
<td>7.091</td>
<td>1.136</td>
<td>{+-------}</td>
<td>------</td>
</tr>
<tr>
<td>Method4</td>
<td>11</td>
<td>5.384</td>
<td>0.924</td>
<td>{+------}</td>
<td>------</td>
</tr>
</tbody>
</table>

Pooled SdDev = 1.232

4.0 6.0 7.2

Fisher’s pairwise comparisons

- Family error rate = 0.197
- Individual error rate = 0.0909

Critical value = 2.021

Intervals for (column level mean) - (row level mean)

<table>
<thead>
<tr>
<th>Method1</th>
<th>Method2</th>
<th>Method3</th>
<th>Method4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method2</td>
<td>-2.253</td>
<td>-0.302</td>
<td></td>
</tr>
<tr>
<td>Method3</td>
<td>-3.355</td>
<td>-1.971</td>
<td>-3.211</td>
</tr>
<tr>
<td>Method4</td>
<td>-1.607</td>
<td>0.236</td>
<td>0.485</td>
</tr>
</tbody>
</table>

Note that Fisher’s confidence intervals show method 2 to be superior to method 1, and method 3 to be superior to both methods 1 and 4. Although Fisher’s confidence intervals show more significant differences between treatment means than do the Tukey’s confidence intervals, the differences come at the price of a higher experimentwise error rate. The family (experimentwise) error rate for Fisher’s pairwise comparisons is 0.197. A value as high as 0.197 may be an unacceptable error rate in some areas of research.

Video 10.1 on the following page examines the climatic effects around the world associated with El Niño. Assumptions for using ANOVA are scrutinized with various graphical tools. The actual creation and interpretation of the ANOVA table is illustrated and discussed. Post-hoc procedures for declaring differences among factors once the null hypothesis of equality of group means has been rejected is also illustrated.
10.3 The Kruskal-Wallis Test

The Kruskal-Wallis test is an extension of the Wilcoxon test for location with two independent samples (covered in section 7.4.2) to the situation with a mutually independent samples. The null hypothesis is that the a populations are the same. However, the hypothesis is usually written in terms of the population medians as

\[ H_0: \theta_1 = \theta_2 = \cdots = \theta_a \text{ versus } H_A: \theta_i \neq \theta_j \text{ for at least one pair } (i, j) \]

Like the Wilcoxon test (7.4.2), the only assumption the Kruskal-Wallis test requires is that the a populations be continuous and similar in shape. To perform the test, all \( n_1, n_2, \ldots, n_a = N \) observations are pooled into a single column and ranked from 1 to \( N \). The standardized test statistic MINITAB™ uses with the Kruskal-Wallis test is:

\[ H_{\text{obs}} = \frac{12}{N(N+1)} \sum_{i=1}^{a} n_i (R_i - \bar{R})^2 \]

where \( n_i \) is the number of observations in the \( i^{th} \) treatment/group, \( R_i \) is the average of the ranks in the \( i^{th} \) treatment/group, and \( \bar{R} \) is the average of all of the ranks. When ties are present, an adjusted standardized test statistic we denote as \( H' \) is also calculated and reported. The adjusted statistic \( H' \) is defined as:

\[ H' = \frac{H_{\text{obs}}}{f_c} = \frac{H}{1 - \frac{\sum_{j=1}^{r} (t_j^3 - t_j)}{N^3 - N}} \]

where \( t_j \) is the number of times a given rank was tied in the combined sample of size \( N \) and \( r \) is the number of ranks in the combined sample of size \( N \) that were tied. Provided each \( n_i \geq 5 \), the sampling distributions of \( H_{\text{obs}} \) and \( H' \) are both approximately Chi-Square random variables with \( a - 1 \) degrees of freedom (\( \chi^2_{a-1} \)). MINITAB™'s Kruskal-Wallis command also reports a Z score for each treatment/group. Under the null hypothesis, \( z_i \), is approximately normal with mean 0 and standard deviation 1. When using MINITAB™'s Kruskal-Wallis command, the value of \( z_i \) reported in the Session Window indicates how the mean rank for treatment/group differs from the overall mean rank for all \( N \) observations. The formula MINITAB™ uses to determine \( z_i \) is:

\[ z_i = \frac{R_i - \frac{N+1}{2}}{\sqrt{\left\frac{(N+1)}{n_i} - 1\right\}} \]

To perform a Kruskal-Wallis test with MINITAB™, all observations must be stored in a single column with their corresponding treatments stored in a second column. When data for each treatment are stored in a separate column, the information should be stacked using the MINITAB™ Stack Columns command. (Manip>Stack>Stack Columns)
To test a hypothesis using the *Kruskal-Wallis test*,

2. Fill in the Response and Factor boxes of the *Kruskal-Wallis Dialog Window* with the appropriate columns.
3. Click OK.

**Example 10.3.1:** An elementary school gym teacher is interested in evaluating the effectiveness of four free throw teaching techniques. The gym teacher randomly assigns each of 20 students to one of four groups. After two months, the every member of the groups shoots 10 free throws and the gym teacher records the results. The number of successful free throws each student shoots in each of the four groups are presented in Table 10.4. Use the free throw results to decide if differences exist among teaching methods.

| Method 1: 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 3, 2, 1, 1, 3, 1, 1 |
| Method 2: 3, 2, 1, 2, 5, 1, 1, 1, 1, 2, 1, 1, 2, 5, 2, 2, 4, 1, 1, 1 |
| Method 3: 2, 6, 2, 3, 2, 2, 4, 3, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2 |
| Method 4: 2, 1, 1, 5, 1, 6, 1, 1, 1, 1, 1, 1, 1, 2, 10, 1, 1, 4 |

**Solution:** We use and explain the five step procedure to determine if differences exist among teaching methods. Before deciding on an inferential procedure, we first examine side-by-side boxplots for free throws made grouped by teaching method. Based on the skewed boxplots in Figure 10.5 and the ratio of the largest to the smallest standard deviation for methods ($2.368/0.671 = 3.529061$) we conclude that the assumptions for using the one-way analysis of variance are not satisfied. Instead of testing for mean group/treatment differences, we decide to test the null hypothesis that all $\alpha$ populations are the same.

**Figure 10.5:** Side-by-Side Boxplots from Example 10.3.1

\[
H_O : \theta_1 = \theta_2 = \theta_3 = \theta_4 \\
H_A : \theta_i \neq \theta_j \text{ for at least one pair } (i, j)
\]

2. The test statistic $R_i$ is used to evaluate the null hypothesis.

3. The standardized test statistic is

\[
H_{obs} = \frac{12 \sum_{i=1}^{a} n_i (\bar{R}_i - \bar{R})^2}{N (N + 1)}
\]
10.3. **The Kruskal-Wallis Test**

which follows an approximate Chi-Square distribution with \( a - 1 \) degrees of freedom. The value of \( H_{obs} \) is calculated as:

\[
H_{obs} = \frac{12}{(80 \times 81)} \times \left\{ \begin{array}{l}
(20 \times (29.725 - 40.50))^2 + (20 \times (40.425 - 40.50))^2 + \\
(20 \times (57.225 - 40.50))^2 + (20 \times (34.625 - 40.40))^2 \\
\end{array} \right\} 
= 15.9388
\]

The adjusted test statistic \( H' \) is calculated as:

\[
H' = \frac{H_{obs}}{\chi^2_{1-\alpha}} = \frac{H}{1 - \frac{\sum (t^3_j - t_j)}{N^3 - N}} 
= \frac{15.9388}{1 - \frac{(39^3 - 39) + (26^3 - 26) + (6^3 - 6) + (3^3 - 3) + (3^3 - 3) + (2^3 - 2)}{80^3 - 80}} 
= \frac{15.9388}{0.849402} = 18.7647
\]

4. The \( p \)-value for the standardized test statistic without adjustment for ties \( (H) \) and the standardized test statistic adjusted for ties \( (H') \) are calculated as \( P (\chi^2 \geq 15.9388) = 0.001167 \) and \( P (\chi^2 \geq 18.7647) = 0.0003 \) respectively. Small \( p \)-values such as 0.001167 and 0.0003 indicate that observing values as extreme or more than 15.9388 or 18.7647 when the null hypothesis is true is very unlikely.

5. There is strong statistical evidence to suggest differences exist among the median scores for the four free throw teaching methods.

The **Session Window** results from analyzing Example 10.3.1 with MINITAB’s Kruskal-Wallis command are shown in **Output 10.6**.

**Output 10.6: Results from Analyzing Example 10.3.1 with MINITAB’s Kruskal-Wallis Command**

<table>
<thead>
<tr>
<th>Method</th>
<th>N</th>
<th>Median</th>
<th>Ave Rank</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>20</td>
<td>1.000</td>
<td>29.7</td>
<td>-2.39</td>
</tr>
<tr>
<td>Method 2</td>
<td>20</td>
<td>1.500</td>
<td>40.4</td>
<td>-0.02</td>
</tr>
<tr>
<td>Method 3</td>
<td>20</td>
<td>2.000</td>
<td>57.2</td>
<td>3.72</td>
</tr>
<tr>
<td>Method 4</td>
<td>20</td>
<td>1.000</td>
<td>34.6</td>
<td>-1.31</td>
</tr>
<tr>
<td>Overall</td>
<td>80</td>
<td>40.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( H = 15.94 \) \( DF = 3 \) \( P = 0.001 \)
\( H = 18.76 \) \( DF = 3 \) \( P = 0.000 \) (adjusted for ties)

Note that MINITAB reports two values for \( H \) in **Output 10.6**. The first value, 15.94, is the unadjusted value while the second number, 18.76, is an adjusted value to account for ties. The first number in **Output 10.6** corresponds to what we have denoted \( H \) while the second number in **Output 10.6** corresponds to what we have denoted \( H' \).
10.3.1 Kruskal-Wallis Approximation

Just as the pooled \( t \)-test was used on the combined ranked data to yield an approximate Wilcoxon test in section 7.4.3, the standard analysis of variance can be applied to the combined ranked data to yield an approximate Kruskal-Wallis test.

To use the approximate Kruskal-Wallis test, rank the combined data, and perform a regular one-way analysis of variance on the resulting combined ranks.

The results from using the one-way analysis of variance on the combined ranks from Example 10.3.1 are shown in Output 10.7. The analysis is furthered by using Tukey’s and Fisher’s confidence intervals on the ranked data as shown in outputs 10.8 and 10.9 on the following page respectively.

Output 10.7: One-Way Analysis of Variance for the Combined Ranks from Example 10.3.1

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>( F )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatmen</td>
<td>8607</td>
<td>3</td>
<td>2869</td>
<td>7.89</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>27629</td>
<td>76</td>
<td>364</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>36236</td>
<td>79</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% CIs For Mean
Based on Pooled StDev

- Method 1: 29.73 ± 17.74
- Method 2: 40.42 ± 22.27
- Method 3: 57.23 ± 9.64
- Method 4: 34.53 ± 23.04

Pooled StDev = 19.07

Output 10.8: Tukey’s Confidence Intervals from Example 10.3.1

Tukey’s pairwise comparisons

- Family error rate = 0.0500
- Individual error rate = 0.0103
- Critical value = 3.72

Intervals for \((\text{column level mean}) - (\text{row level mean})\)

<table>
<thead>
<tr>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 2</td>
<td>-26.56</td>
<td>5.16</td>
</tr>
<tr>
<td>Method 3</td>
<td>-43.36</td>
<td>-32.66</td>
</tr>
<tr>
<td></td>
<td>-11.64</td>
<td>-0.94</td>
</tr>
<tr>
<td>Method 4</td>
<td>-20.76</td>
<td>-10.06</td>
</tr>
<tr>
<td></td>
<td>10.96</td>
<td>21.66</td>
</tr>
</tbody>
</table>

Based on the results from Tukey’s pairwise comparisons (Output 10.8) and Fisher’s pairwise comparisons (Output 10.9 on the following page), method 3 appears to be the best method for teaching elementary school children to shoot free throws.

Video 10.2 on the next page checks the assumptions required for ANOVA as well as showing how to perform a Kruskal-Wallis test and an \( F \) approximation to the null distribution of \( R_i \) with MINITAB™.
Output 10.9: Fisher’s Confidence Intervals from Example 10.3.1

**Fisher’s pairwise comparisons**

Family error rate = 0.200  
Individual error rate = 0.0500

Critical value = 1.992

Intervals for (column level mean) - (row level mean)

<table>
<thead>
<tr>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 2</td>
<td>-22.71</td>
<td>1.31</td>
</tr>
<tr>
<td>Method 3</td>
<td>-39.51</td>
<td>-28.81</td>
</tr>
<tr>
<td></td>
<td>-15.49</td>
<td>-4.79</td>
</tr>
<tr>
<td>Method 4</td>
<td>-16.91</td>
<td>-6.21</td>
</tr>
<tr>
<td></td>
<td>10.59</td>
<td>34.61</td>
</tr>
</tbody>
</table>
10.4 Summary and Review Labs

Lab 10.1 — The One-Way ANOVA

Objectives:

I. To verify ANOVA assumptions
II. To perform an ANOVA test
III. To create confidence intervals around mean responses

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

The investigators that recorded the information in the worksheet CARTOON.MTW stored in the MINITAB\DATA directory gave all participants the OTIS Quick Scoring Mental Ability Test to get a rough measure of each participant’s natural ability. The participants were preprofessional and professional personnel at three Pennsylvania hospitals involved in an in-service training program, as well as a group of Penn State undergraduates. To read more about the study and the data select Help>Search Help>CARTOON.MTW.

Questions and Directions:

1. Produce side-by-side boxplots and side-by-side normal probability plots of OTIS (C5) versus the Education groups (Ed in C3). Title each appropriately and create a right justified footnote containing your name and the date for the side-by-side boxplots.

2. Based on your graphs from 1, are the analysis of variance assumptions satisfied?

3. Use the formulas given in Table 10.1 to calculate the \(SS_{Treat}\), \(SS_{Error}\), and \(SS_{Total}\) for an analysis of variance table. (Be sure to append all MINITAB™ calculator commands so your instructor can verify your work.)

4. Use MINITAB™’s One-Way command and the five step procedure to test whether there are differences in OTIS scores among the three Education groups. Are the values you calculated in step 3 in agreement with the values in MINITAB™’s ANOVA table? If not, why not?

5. Use equation (10.6) to calculate a 90% confidence interval for the average OTIS score for college students.

6. Use equation (10.9) to calculate 90% confidence intervals for \(\mu_i - \mu_j\).

7. Use MINITAB™’s built in command to create 90% confidence intervals for \(\mu_i - \mu_j\) with Tukey’s procedure. Are the confidence intervals in questions 5 and 6 equivalent? If not, why not?
Lab 10.2 — The Kruskal-Wallis Test

Objectives:

I. To use graphs to decide if the assumptions for conducting the *Kruskal-Wallis test* are satisfied

II. To perform a *Kruskal-Wallis test* by hand, with MINITAB™, and with the *F* approximation to *R*.

Basic Directions:

All graphs and output should be appended to the report pad. Answer all questions with complete sentences in the report pad.

Introduction:

The following description has been taken from the help files of *Poplar1.MTW*, *Poplar2.MTW*, and *Poplar3.MTW*.

Clones are genetically identical cells descended from the same individual. Researchers have identified a single poplar clone that yields fast-growing, hardy trees. These trees may one day serve as an alternative to conventional fuel as an energy resource. Researchers at The Pennsylvania State University planted Poplar Clone 252 on two different sites—one, a rich site by a creek, and the other, a dry, sandy site on a ridge. They measured the diameter in centimeters, height in meters, and the dry weight of wood in kilograms for a sample of three-year-old trees so that they could determine the weight of a tree from its diameter and height measurements. In an effort to determine how to maximize yield, the researchers designed an experiment to determine how two factors, Site and Treatment, influence the weight of four-year-old poplar clones. They applied four different treatments to the trees in Site 1 (the rich, moist soil by a creek) and Site 2 (the dry, sandy soil on a ridge). Treatment 1 was the control (no treatment), Treatment 2 was fertilizer, Treatment 3 was irrigation, and Treatment 4 was both fertilizer and irrigation. To account for a variety in weather conditions, they replicated the data by planting half the trees in Year 1 and the other half in Year 2.

Questions and Directions:

Since we have not covered two-way analysis of variance, we will focus on whether there are differences in weight among the four treatment groups. Note that there are several Diameter, Height, and Weight values of −99 stored in columns C4, C5, and C6 respectively. The problem does not indicate what the researchers used the values −99 to indicate. However, clearly it is impossible to obtain a diameter, height or weight value that is negative. In all likelihood, −99 indicates a missing value. Before starting the lab, convert all of the −99 values in columns C4-C6 to missing (*) using Manip>Code>Numeric to Numeric.

1. Produce side-by-side boxplots and side-by-side normal probability plots of weight versus the treatment groups. Title each appropriately and create a right justified footnote containing your name and the date for the side-by-side boxplots.

2. Based on your graphs from 1, are the analysis of variance assumptions satisfied?

3. Are the assumptions required to use the Kruskal Wallis test satisfied?

4. Use the five step procedure and the approach described next to test whether there are differences in weight among the four treatment groups.

   a. Use the test statistic given in (10.11) and show all calculations to test whether there are differences in weight among the four treatment groups.

   b. Use MINITAB™’s *Kruskal-Wallis* command to test whether there are differences in weight among the four treatment groups.

   c. Use the *F* approximation to *R*, given in section 10.3.1 to test whether there are differences in weight among the four treatment groups.
Appendix A

Glossary

population mean is the average of all the data in a population. It is also called the expected value. The notation that defines the population mean is \( \mu = \frac{\sum_{i=1}^{N} x_i}{N} \) for discrete data or for continuous data with a population density of \( f(x) \) we have \( \mu = \int_{-\infty}^{\infty} x f(x) \, dx \).

population median is the middle of all the data in a population. The symbol for the population median is \( \theta \).

population proportion is the number of successes in a population divided by the size of the population. The notation that defines the population proportion is \( \pi = \frac{X}{N} \) where \( X \) is the number of successes.

population standard deviation is the square root of the population variance. The notation that defines the population standard deviation is \( \sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \) for discrete data and \( \sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx} \) for continuous data.

sample mean the average of the data from a sample. It is calculated by adding all the numbers in the data set and dividing by the number of numbers. The notation that defines the sample mean is \( \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \).

sample median is the middle of all the data in a population. The symbol for the population median is \( M \).

sample proportion is the number of successes in a sample divided by the size of the sample. The notation that defines the sample proportion is \( p = \frac{X}{n} \) where \( X \) is the number of successes.

sample standard deviation the unbiased estimate of the population standard deviation. It is the sum of the deviations from the sample mean squared divided by the number of deviations minus one. The notation that defines the sample standard deviation is \( s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \).
Appendix B

Statistics, Distributions, and Conditions

\[ \bar{x} \sim N \left( \mu, \frac{\sigma}{\sqrt{n}} \right) \] provided \( n \) is sufficiently large.

\[ \bar{x} \sim N \left( \mu, \frac{\sigma}{\sqrt{n}} \right) \] if you are sampling from a \( N \) distribution.

\[ p \sim N \left( \pi, \sqrt{\frac{\pi(1-\pi)}{n}} \right) \] provided \( n\pi \geq 5 \) and \( n(1-\pi) \geq 5 \) and \( n \) is large.

\( X(\#\ \text{successes}) \sim N \left( n\pi, \sqrt{n\pi(1-\pi)} \right) \) provided \( n\pi \geq 5 \) and \( n(1-\pi) \geq 5 \) and \( n \) is large.

\[ \bar{x}_1 - \bar{x}_2 \sim N \left( \mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \] provided \( n_1 \geq 30 \) and \( n_2 \geq 30 \) with both distributions not highly skewed.

\[ p_1 - p_2 \sim N \left( \pi_1 - \pi_2, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}} \right) \] provided \( n_1\pi_1 \geq 5 \), \( n_1(1-\pi_1) \geq 5 \), \( n_2\pi_2 \geq 5 \), and \( n_2(1-\pi_2) \geq 5 \).
Appendix C
Confidence Intervals for Common Parameters

We are \((1 - \alpha) \times 100\%\) confident that \(\mu\) falls in
\[
\left( \bar{x} - t_{1-\alpha/2; n-1} \times \frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2; n-1} \times \frac{s}{\sqrt{n}} \right)
\]
if we are sampling from a normal distribution and \(\sigma\) is unknown.

We are \((1 - \alpha) \times 100\%\) confident that \(\mu\) falls in
\[
\left( \bar{x} - z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \right)
\]
if we are sampling from a normal distribution and \(\sigma\) is known.

We are \((1 - \alpha) \times 100\%\) confident that \(\pi\) falls in
\[
\left( p - z_{1-\alpha/2} \sqrt{\frac{p \times (1 - p)}{n}}, p + z_{1-\alpha/2} \sqrt{\frac{p \times (1 - p)}{n}} \right)
\]
if we are sampling from a binomial distribution when \(n \times p > 5\) and \(n \times (1 - p) > 5\).
Appendix D

Key Words and Problem Hints

D.1 Basic Probability

<table>
<thead>
<tr>
<th>If you see</th>
<th>the probable problem type is a</th>
<th>and you will probably use</th>
</tr>
</thead>
<tbody>
<tr>
<td>given</td>
<td>conditional probability</td>
<td>$P(A</td>
</tr>
<tr>
<td>or</td>
<td>Additive Law</td>
<td>$P(A \cup B) = P(A) + P(B) - P(A \cap B)$</td>
</tr>
<tr>
<td>And and independent both!</td>
<td>Multiplication for Independent Events</td>
<td>$P(A \cap B) = P(A) \times P(B)$</td>
</tr>
</tbody>
</table>

D.2 Sampling Distributions and Probability

<table>
<thead>
<tr>
<th>If you see</th>
<th>the probable problem type is a</th>
<th>and you will probably use</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample proportion or a %</td>
<td>$p$</td>
<td>$p \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$</td>
</tr>
<tr>
<td>normally distributed given a number OR $n &gt; 30$</td>
<td>find $\mathcal{N}$ probability</td>
<td>See Normal Distribution in Appendix D.3</td>
</tr>
<tr>
<td>normally distributed given a % or prob. OR $n &gt; 30$</td>
<td>find $\mathcal{N}$ percentile</td>
<td>See Normal Quantile in Appendix D.3</td>
</tr>
<tr>
<td>Sample Proportion or just % and $n &gt; 100$ given. Note: % s in problem.</td>
<td>$p$</td>
<td>$p \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$</td>
</tr>
<tr>
<td>Average or mean</td>
<td>$\bar{x}$</td>
<td>$\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$</td>
</tr>
<tr>
<td>Two averages given</td>
<td>$\bar{x}_1 - \bar{x}_2$</td>
<td>$\bar{x}_1 - \bar{x}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$</td>
</tr>
<tr>
<td>Two proportion (percents) given</td>
<td>$p_1 - p_2$</td>
<td>$p_1 - p_2 \sim \mathcal{N}(\pi_1 - \pi_2, \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}})$</td>
</tr>
<tr>
<td>Size &lt; 100, and % and $n$ given. Note: # successes given in problem</td>
<td>Binomial</td>
<td>$X \sim \text{Bin}(n, \pi)$</td>
</tr>
<tr>
<td>Size &gt; 100, and % and $n$ given. Note: # successes given in problem</td>
<td>Normal Approximation to the Binomial Distribution</td>
<td>$Y \sim \mathcal{N}(n\pi, \sqrt{n\pi(1 - \pi)})$</td>
</tr>
<tr>
<td>none</td>
<td>general probability</td>
<td>Common sense with multiplicative law</td>
</tr>
</tbody>
</table>
### D.3 Probabilities with Normal Distributions

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>Formula</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than $c$ where $X \sim \mathcal{N}(\mu, \sigma)$</td>
<td>$\mathbb{P}(X &lt; c)$</td>
<td>\textit{Calc}&gt;Probability Distributions&gt;Normal OR&lt;br&gt;MTB &gt; cdf c;&lt;br&gt;SUBC &gt; norm $\mu,\sigma$.</td>
</tr>
<tr>
<td>Greater than, more than $c$ where $X \sim \mathcal{N}(\mu, \sigma)$</td>
<td>$\mathbb{P}(X &gt; c) = 1 - \mathbb{P}(X \leq c)$ and solve $\mathbb{P}(X \leq c)$ with \textit{Calc}&gt;Probability Distributions&gt;Normal OR&lt;br&gt;MTB &gt; cdf $c,k_1$;&lt;br&gt;SUBC &gt; norm $\mu,\sigma$.&lt;br&gt;MTB &gt; let $k_2=1-k_1$&lt;br&gt;MTB &gt; print $k_2$</td>
<td></td>
</tr>
<tr>
<td>Between $a$ and $b$ with $X \sim \mathcal{N}(\mu, \sigma)$</td>
<td>$\mathbb{P}(a &lt; X &lt; b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a)$. Calculate with \textit{Calc}&gt;Probability Distributions&gt;Normal OR&lt;br&gt;MTB &gt; cdf $b,k_1$;&lt;br&gt;SUBC &gt; norm $\mu,\sigma$.&lt;br&gt;MTB &gt; cdf $a,k_2$;&lt;br&gt;SUBC &gt; norm $\mu,\sigma$.&lt;br&gt;MTB &gt; let $k_3=k_1-k_2$&lt;br&gt;MTB &gt; print $k_3$</td>
<td></td>
</tr>
<tr>
<td>Find the $c$ that has $v$ percent below it given $X \sim \mathcal{N}(\mu, \sigma)$</td>
<td>$\mathbb{P}(X &lt; c) = v$</td>
<td>\textit{Calc}&gt;Probability Distributions&gt;Normal OR&lt;br&gt;MTB &gt; invcdf $v$;&lt;br&gt;SUBC &gt; norm $\mu,\sigma$.</td>
</tr>
<tr>
<td>Percentile given $X \sim \mathcal{N}(\mu, \sigma)$</td>
<td>asked for $v^{th}$ percentile</td>
<td>\textit{Calc}&gt;Probability Distributions&gt;Normal OR&lt;br&gt;MTB &gt; let $k_1=v/100$&lt;br&gt;MTB &gt; invcdf $k_1$;&lt;br&gt;SUBC &gt; norm $\mu,\sigma$.</td>
</tr>
</tbody>
</table>
Appendix E

Macros

E.1 NPIE: Macro for finding sample size given error bound for \( \pi \) CI

```
# Required Sample Size Macro
# Author: Alan T. Arnholt
# Date: 12/19/97
# (Revised for version 13 on 4/4/2000)

Gmacro
npie
NOTE Press Enter after typing your answer to each question.
NOTE What is your bound?
set c90;
file "terminal"
; nobs 1.
copy c90 k90 #K90 = Bound
NOTE What is your confidence level?
NOTE Enter your confidence level as a decimal.
set c91;
file "terminal"
; nobs 1.
copy c91 k91
let k99=1-k91
let k98=k99/2
let k97=1-k98
invcdf k97 k92 #k92 = Critical Z-value
NOTE What is your point estimate of the parameter?
set c93;
file "terminal"
; nobs 1.
copy c93 k93 #k93 = \( \pi \)
let c94=(k92**2)*k93*(1-k93)/(k90**2)
let c95=floor(c94)+1
NOTE
NOTE Required Sample Size is \( n \)
name c95 = 'n='
print c95
eendmacro
```
E.2 NMU: Macro for finding sample size given error bound and confidence level

```plaintext
Gmacro
nmu
NOTE Press Enter after typing your answer to each question.
NOTE What is your bound?
set c90;
file "terminal";
nobs 1.
copy c90 k90  #k90 = Bound
NOTE What is your standard deviation?
set c89;
file "terminal";
nobs 1.
copy c89 k89  #k89 = Standard Deviation
NOTE What is your confidence level?
NOTE Enter your confidence level as a decimal.
set c91;
file "terminal";
nobs 1.
copy c91 k91
let k99=1-k91
let k98=k99/2
let k97=1-k98
invcdf k97 k92  #k92 = Critical Z-value
let c94=(k92*k89/k90)**2
let c95=floor(c94)+1
NOTE
NOTE Required Sample Size is n
name c95 = 'n='
print c95
endmacro
```
E.3 GOF: Goodness of Fit Macro

```plaintext
GOF

Gtitle "Chi-Square Goodness Of Fit Test"
NOTE Enter the hypothesized probabilities for each category in decimal
NOTE format at the data prompt (DATA>). When you finish, push the return,
NOTE type END, then push the return again.
note
set c90;
file "terminal".
name c90 = 'Null_p'
print c90
NOTE Enter the observed values for each category at the data prompt (DATA>).
NOTE When you finish, push the return, type END, then push the return again.
note
set c91;
file "terminal".
name c91 = 'Observed'
print c91
let c92 = c90*(sum(c91))
name c92 = 'Expected'
let c93 = ((c91 - c92)**2)/c92
let k90 = sum(c93)
let k91 = n(c91)-1
cdf k90 k94;
chis k91.
let k95 = 1 - k94
name k90 = 'Teststat'
name k95 = 'Pvalue'
Gtitle "Chi-Square Goodness Of Fit Test"
Note Teststat is the standandardized test statistic and Pvalue is the
Note p-value corresponding to the specified null hypothesis.
note
print k90 k95
NOTE
endmacro
```
E.4 SIMP: Simulate Power for Pooled versus Un-pooled $t$ Tests

MACRO
SIMP n1 n2 mu1 mu2 sigma1 sigma2;
times sims;
alpha level.
MCONSTANT n n1 n2 mu1 mu2 sigma1 sigma2 CritTp CritTs dofNU sims yeha alpha level dofpt
MCOLUMN X.1-X.n1 Y.1-Y.n2 xbar1 xbar2 s1 s2 nu s12 s22 dp Tsat Tpool TpoolC TsatC
default sims=10000 level=0.05
let alpha=level
NOTITLE
MTITLE "Selected Values"
Print n1 n2 mu1 mu2 sigma1 sigma2 sims alpha
random sims X.1-X.n1;
normal mu1 sigma1.
rmean X.1-X.n1 xbar1
rstdev X.1-X.n1 s1
Random sims Y.1 - Y.n2;
normal mu2 sigma2.
rmean Y.1 - Y.n2 xbar2
rstdev Y.1 - Y.n2 s2
let dofnu = ((sigma1**2/n1 +sigma2**2 /n2)**2) / 
&((sigma1**2/n1)**2/(n1-1)+(sigma2**2/n2)**2/(n2-1))
let dofpt = n1+n2-2
Let s12 = s1**2
Let s22 = s2**2
let dp = SQRT((1/n1)+(1/n2))*SQRT(((n1-1)*s12+(n2-1)*s22)/(n1+n2-2))
let Tsat = (xbar1 - xbar2)/SQRT((s12/n1)+(s22/n2))
let Tpool = (xbar1 - xbar2)/dp
let level=1-level
InvCDF level CritTp; T dofpt.
InvCDF level CritTs; T dofNU.
NOTE
NOTE CritTp and CritTs are the critical values based on a user supplied alpha value for testing the alternative hypothesis Mu1>Mu2 with the pooled t-test and the Satterthwaite approximation respectively (default alpha value is 5%).

MTITLE "Critical Values"
Print CritTp CritTs

Code (-1000:CritTp) "Type II Error" (CritTp:1000) "Power" Tpool TpoolC
Code (-1000:CritTs) "Type II Error" (CritTs:1000) "Power" Tsat TsatC

NOTE
Note TsatC and TpoolC represent the Satterthwaite and Pooled t respectively

Mtitle "Simulated Power Tables"

Tally TsatC TpoolC;
Counts;
CumCounts;
Percents;
CumPercent.

NOTE Would you like histograms of the simulated noncentral distributions?
YESNO yeha

IF yeha=1
Histogram Tpool;
Density;
MidPoint;
Bar;
Type 1;
Color 4;
Title "Pooled t";
ScFrame;
ScAnnotation;
Axis 1;
Axis 2;
Tick 1;
Tick 2.

Histogram Tsat;
Density;
MidPoint;
Bar;
Type 1;
Color 4;
Title "Satterthwaite t";
ScFrame;
ScAnnotation;
Axis 1;
Axis 2;
Tick 1;
Tick 2.
ENDIF
endmacro
E.5 SIMH: Simulate Power for Pooled t Test versus Wilcoxon Signed Rank Test

# Simulated Power (Mu1 > Mu2)  #
# Pooled T-test Versus WSR test  #
# Author: Alan T. Arnholt  #
# Date: 3/07/02   #
# Written with Minitab release 13  #

macro

simh n1 n2 mu1 mu2 sigma1 sigma2;
times sims;
alpha level;
dist type.

mconstant n1 n2 mu1 mu2 sigma1 sigma2 sims critTW critTP k pn level type
mconstant arl dof CV lower1 lower2 upper1 upper2 sigm1 sigm2
mcolumn x xrank y yrank z zrank sub tp tw TTEST WTEST TI

default sims=100 level=.05 type=1

if type=1
goto 5
elseif type=2
goto 10
else
goto 15
endif

mlabel 5
do k=1:sims
rand n1 x;
norm mu1 sigma1.
rand n2 y;
norm mu2 sigma2.
stack x y z;
subs sub.
rank z zrank
unstack zrank xrank yrank;
subs sub.
let tp(k)=(mean(x)-mean(y))/((((n1-1)*(stdev(x))**2+(n2-1)*(stdev(y))**2)/ & (n1+n2-2))*((1/n1)+(1/n2)))**.5
let tw(k)=(mean(xrank)-mean(yrank))/((((n1-1)*(stdev(xrank))**2+(n2-1)*(stdev(yrank))**2)/ & (n1+n2-2))*((1/n1)+(1/n2)))**.5
endo
disco
MTITLE "User Selected Values For The Normal Distribution"
Print n1 n2 mu1 mu2 sigma1 sigma2 sims level

NOTE
NOTE The asymptotic relative efficiency of the Wilcoxon rank-sum test,
NOTE relative to the t test for normal distributions is 0.955.
goto 25
E.5. SIMH: Simulate Power for Pooled t Test versus Wilcoxon Signed Rank Test

mlabel 10
do k=1:sims
  rand n1 x;
  unif mu1 mu2.
  rand n2 y;
  unif sigma1 sigma2.
  stack x y z;
  subs sub.
  rank z zrank
  unstack zrank xrank yrank;
  subs sub.
  let tp(k)=(mean(x)-mean(y))/((((n1-1)*(stdev(x))**2+(n2-1)*(stdev(y))**2)/
    &(n1+n2-2))*((1/n1)+(1/n2)))**.5
  let tw(k)=(mean(xrank)-mean(yrank))/((((n1-1)*(stdev(xrank))**2+(n2-1)*(stdev(yrank))**2)/
    &(n1+n2-2))*((1/n1)+(1/n2)))**.5
endo
  let lower1=mu1
  let upper1=mu2
  let lower2=sigma1
  let upper2=sigma2
  let mu1=(lower1+upper1)/2
  let mu2=(lower2+upper2)/2
  let sigma1=sqrt((upper1-lower1)**2/12)
  let sigma2=sqrt((upper2-lower2)**2/12)
NOTE
MTITLE "User Selected Values For The Uniform Distribution"
print n1 n2 mu1 mu2 sigma1 sigma2 sims level
NOTE
NOTE The asymptotic relative efficiency of the Wilcoxon rank-sum test,
NOTE relative to the t test for uniform distributions is 1.0.
goto 25

mlabel 15
do k=1:sims
  rand n1 x;
  laplace mu1 mu2.
  rand n2 y;
  laplace sigma1 sigma2.
  stack x y z;
  subs sub.
  rank z zrank
  unstack zrank xrank yrank;
  subs sub.
  let tp(k)=(mean(x)-mean(y))/((((n1-1)*(stdev(x))**2+(n2-1)*(stdev(y))**2)/
    &(n1+n2-2))*((1/n1)+(1/n2)))**.5
  let tw(k)=(mean(xrank)-mean(yrank))/((((n1-1)*(stdev(xrank))**2+(n2-1)*(stdev(yrank))**2)/
    &(n1+n2-2))*((1/n1)+(1/n2)))**.5
endo
  let mu1=mu1
  let sigm1=sqrt(mu2*2)
  let mu2=sigma1
  let sigm2=sqrt(sigma2*2)
NOTE
MTITLE "User Selected Values For The Laplace Distribution"
print n1 n2 mu1 mu2 sigm1 sigm2 sims level
NOTE
NOTE The asymptotic relative efficiency of the Wilcoxon rank-sum test,
NOTE relative to the t test for Laplace distributions is 1.5.
goto 25

mlabel 25
let arl=1-level
let dof=n1+n2-2
invcdf arl CV;
t dof.
code (-10000:CV) "Type II Error" (CV:10000) "Power" tp TTEST
code (-10000:CV) "Type II Error" (CV:10000) "Power" tw WTEST
NOTE
NOTE TTEST and WTEST represent the pooled t-test and the approximated
NOTE Wilcoxon test respectively.
MTITLE "Simulated Power Tables"
tally TTEST WTEST;
counts;
cumcounts;
percents;
cumpercents.
endmacro