

M c G R A W - H I L L

5 STEPS TO A 5

**AP^{*}
Statistics**

- ✓ A unique, 5-Step Plan for acing your Advanced Placement^{*} Exam
- ✓ Sample tests modeled on actual AP^{*} Exams
- ✓ Hundreds of tips and strategies for the AP^{*} Exam in Statistics

*AP, Advanced Placement Program, and College Board are registered trademarks of the College Entrance Examination Board, which was not involved in the production of and does not endorse this product.

Duane C. Hinders

McGRAW-HILL

5 Steps to a 5
AP Statistics

Other books in McGraw-Hill's *5 Steps to a 5* Series include:

AP Biology
AP Calculus AB
AP Chemistry
AP English Language
AP English Literature
AP Psychology
AP Spanish Language
AP U.S. Government and Politics
AP U.S. History
Writing the AP English Essay

McGRAW-HILL

5 Steps to a 5 AP Statistics

Duane C. Hinders

McGRAW-HILL

*New York Chicago San Francisco Lisbon London Madrid Mexico City
Milan New Delhi San Juan Seoul Singapore Sydney Toronto*

The McGraw-Hill Companies

Copyright © 2004 by The McGraw-Hill Companies, Inc. All rights reserved. Manufactured in the United States of America. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

0-07-143145-4

The material in this eBook also appears in the print version of this title: 0-07-141278-6

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps.

McGraw-Hill eBooks are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information, please contact George Hoare, Special Sales, at george_hoare@mcgraw-hill.com or (212) 904-4069.

TERMS OF USE

This is a copyrighted work and The McGraw-Hill Companies, Inc. (“McGraw-Hill”) and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hill’s prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED “AS IS”. McGRAW-HILL AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

DOI: 10.1036/0071431454



Want to learn more?

We hope you enjoy this McGraw-Hill eBook! If you'd like more information about this book, its author, or related books and websites, please [click here](#).

[For more information about this book, click here](#)

Contents

Preface / xi
Acknowledgments / xiii

PART I HOW TO USE THIS BOOK / 1

Chapter 1 The Five-Step Program / 3
How Is this Book Organized? / 3
Introduction to the Five-Step Program / 3
Graphics Used in this Book / 4
Three Approaches to Preparing for the AP Statistics Exam / 5
Calendar for Each Plan / 7

PART II WHAT YOU NEED TO KNOW ABOUT THE AP STATISTICS EXAM / 13

Chapter 2 Introduction to the AP Statistics Exam / 15
Background on the AP Statistics Exam / 15
 What Is Covered in the AP Statistics Exam? / 15
 What Is the Format of the AP Statistics Exam? / 16
 What Are the Advanced Placement Exam Grades? / 16
 How Is the AP Statistics Grade Calculated? / 17
 What Is the Graphing Calculator Policy for the
 AP Statistics Exam? / 18
 What Do I Need to Bring to the Exam? / 18
Tips for Taking the Exam / 18
Getting Started / 20
Diagnostic Test / 21
Answers and Solutions / 32
Scoring and Interpretation / 43

PART III COMPREHENSIVE REVIEW / 45

Chapter 3 Overview of Statistics/Basic Vocabulary / 47
What Is Statistics? / 47
Quantitative versus Qualitative Data / 47
 Discrete and Continuous Data / 48
Descriptive Statistics versus Inferential Statistics / 48
 Parameters versus Statistics / 49

Collecting Data: Surveys, Experiments, Observational Studies / 49
Random Variables / 50
Rapid Review / 51

Chapter 4 **Univariate Data Analysis / 54**

Graphical Analysis / 54
 Shape / 54
 Dotplot / 55
 Stemplot (Stem and Leaf Plot) / 56
 Histogram / 57
Measures of Center / 61
 Mean / 61
 Median / 62
 Resistant / 63
Measures of Spread / 64
 Variance and Standard Deviation / 64
 Interquartile Range / 65
 Outliers / 66
Position of a Term in a Distribution / 67
 5-Number Summary / 67
 Boxplots (Outliers Revisited) / 68
 Percentile Rank of a Term / 69
 z Scores / 69
Normal Distribution / 70
 Empirical Rule / 71
 Standard Normal Distribution / 71
Rapid Review / 75
Practice Problems / 76
Cumulative Review Problems / 81
Solutions to Practice Problems / 82
Solutions to Cumulative Review Problems / 87

Chapter 5 **Bivariate Data Analysis / 88**

Scatterplots / 88
Correlation / 90
 Correlation and Causation / 93
Lines of Best Fit / 94
 Least-Squares Regression Line / 94
Residuals / 98
Coefficient of Determination / 100
Outliers and Influential Observers / 101
Transformations to Achieve Linearity / 102
Rapid Review / 105
Practice Problems / 106
Cumulative Review Problems / 112
Solutions to Practice Problems / 113
Solutions to Cumulative Review Problems / 117

Chapter 6	Design of a Study: Sampling, Surveys, and Experiments / 118
	Samples / 118
	Census / 118
	Probability Sample / 119
	Sampling Bias / 121
	Undercoverage / 121
	Voluntary Response Bias / 121
	Wording Bias / 122
	Response Bias / 122
	Experiments and Observational Studies / 123
	Statistical Significance / 123
	Completely Randomized Design / 125
	Double-Blind Experiments / 126
	Randomization / 126
	Block Design / 127
	Matched Pairs Design / 128
	Rapid Review / 129
	Practice Problems / 130
	Cumulative Review Problems / 136
	Solutions to Practice Problems / 136
	Solutions to Cumulative Review Problems / 141
Chapter 7	Random Variables and Probability / 142
	Probability / 142
	Sample Spaces and Events / 143
	Probabilities of Combined Events / 144
	Mutually Exclusive Events / 145
	Conditional Probability / 145
	Independent Events / 146
	Probability of A and/or B / 147
	Random Variables / 148
	Discrete Random Variables / 148
	Continuous Random Variables / 148
	Probability Distribution of a Random Variable / 149
	Probability Histogram / 150
	Density Curve / 151
	Normal Probabilities / 152
	Simulation and Random Number Generation / 155
	Transforming and Combining Random Variables / 158
	Rules for the Mean and Standard Deviation of Combined Random Variables / 158
	Rapid Review / 159
	Practice Problems / 160
	Cumulative Review Problems / 165
	Solutions to Practice Problems / 166
	Solutions to Cumulative Review Problems / 173

Chapter 8

Binomial Distribution and Sampling Distributions / 174

- Binomial Distribution / 174
 - Normal Approximation to the Binomial / 177
 - Geometric Distribution / 179
- Sampling Distributions / 181
 - Sampling Distribution of a Sample Mean / 181
 - Central Limit Theorem / 182
- Sampling Distribution of a Sample Proportion / 185
- Rapid Review / 186
- Practice Problems / 187
- Cumulative Review Problems / 192
- Solutions to Practice Problems / 193
- Solutions to Cumulative Review Problems / 198

Chapter 9

Confidence Intervals and Introduction to Inference / 200

- Estimation and Confidence Intervals / 200
 - t Procedures / 201
 - General Form of a Confidence Interval / 203
- Confidence Intervals for Means and Proportions / 204
- Sample Size / 209
 - Sample Size for Estimating a Population Mean (Large Sample) / 209
 - Sample Size for Estimating a Population Proportion / 210
- Statistical Significance and P Value / 211
 - Statistical Significance / 211
 - P Value / 212
- The Hypothesis-Testing Procedure / 213
- Type-I and Type-II Errors and the Power of a Test / 215
- Rapid Review / 218
- Practice Problems / 220
- Cumulative Review Problems / 225
- Solutions to Practice Problems / 226
- Solutions to Cumulative Review Problems / 231

Chapter 10

Inference for Means and Proportions / 233

- Significance Testing / 233
 - Large Sample versus Small Sample / 235
 - Using Confidence Intervals for Two-Sided Alternatives / 236
- Inference for a Single Population Mean / 236
- Inference for the Difference Between Two Population Means / 239
- Inference for a Single Population Proportion / 242
- Inference for the Difference Between Two Population Proportions / 245
- Rapid Review / 246
- Practice Problems / 249
- Cumulative Review Problems / 254
- Solutions to Practice Problems / 255
- Solutions to Cumulative Review Problems / 262

Chapter 11	Inference for Regression / 264
	Simple Linear Regression / 264
	Inference for the Slope of a Regression Line / 266
	Significance Test for the Slope of a Regression Line / 266
	Confidence Interval for the Slope of a Regression Line / 268
	Inference for Regression Using Technology / 269
	Rapid Review / 272
	Practice Problems / 274
	Cumulative Review Problems / 278
	Solutions to Practice Problems / 279
	Solutions to Cumulative Review Problems / 284
Chapter 12	Inference for Categorical Data: Chi-square / 286
	Chi-square Goodness-of-Fit Test / 286
	Inference for Two-Way Tables / 290
	Two-Way Tables (Contingency Tables) Defined / 290
	Chi-square Test for Independence / 291
	Chi-square Test for Homogeneity of Proportions / 294
	χ^2 vs. z^2 / 296
	Rapid Review / 296
	Practice Problems / 297
	Cumulative Review Problems / 301
	Solutions to Practice Problems / 302
	Solutions to Cumulative Review Problems / 305
PART IV	PRACTICE EXAMS / 307
	Practice Exam 1 / 309
	Answer Sheet for Section I / 309
	Section I / 311
	Section II / 324
	General Instructions / 324
	Answers to Practice Exam 1, Section I / 328
	Solutions to Practice Exam 1, Section I / 328
	Solutions to Practice Exam 1, Section II, Part A / 333
	Solutions to Practice Exam 1, Section II, Part B / 336
	Scoring Sheet for Practice Exam 1 / 337
	Practice Exam 2 / 339
	Answer Sheet for Section I / 339
	Section I / 341
	Section II / 354
	General Instructions / 354
	Answers to Practice Exam 2, Section I / 358
	Solutions to Practice Exam 2, Section I / 358

Solutions to Practice Exam 2, Section II, Part A / 363
Solutions to Practice Exam 2, Section II, Part B / 366
Scoring Sheet for Practice Exam 2 / 368

PART V APPENDIXES / 369

Appendix A Formulas / 371

Appendix B Tables / 374

Appendix C Bibliography / 379

Appendix D Websites / 380

Appendix E Glossary / 381

Preface

Congratulations, you are now an AP Statistics student. AP Statistics is one of the most interesting and useful subjects you will study in school. Sometimes it has the reputation of being “easy” compared to calculus. However, it is different and challenging in its own way. Unlike calculus, where you are expected to get rather precise answers, in statistics you are expected to learn to become comfortable with uncertainty. Instead of saying things like, “The answer is . . .,” you will more often find yourself saying things like, “We are confident that . . .” It’s a new and exciting way of thinking.

How do you do well on the AP Exam (by well, I mean a “4” or a “5”)? By reading this book; by staying on top of the material during your AP Statistics class; by studying when it is time to study. Note that the questions on the AP Exam are only partially computational—they more often involve thinking about the process you are involved in and communicating your thoughts to the person reading your exam. You can always use a calculator, so the test designers make sure the questions involve more than just button pushing.

This book is self-contained in that it contains all of the material required by the course content description published by the College Board. However, it is not designed to substitute for an in-class experience or for your textbook. Use this book as a supplement to your in-class studies, as a reference for a quick refresher on a topic, and as one of your major aides as you prepare for the AP Exam in May.

You should begin your preparations by reading through Chapters 1 and 2. However, you shouldn’t attempt the Diagnostic Exam in Chapter 2 until you have been through all of the material in the course. Then you can take the exam to help you determine which topics need more of your attention. Note that the Diagnostic Test simulates the AP Exam to a reasonable extent and the Practice Tests are quite similar in style and substance to the AP Exam.

So, how do you get the best possible score on the AP Statistics exam?

- Pick one of the study plans from this book.
- Study the chapters and do the practice problems.
- Take the Diagnostic Test and the Practice Tests.
- Review as necessary based on your performance on the Diagnostic Test and the Practice Tests.
- Be well rested for the exam.

“Statistics are like a bikini. What they reveal is suggestive, but what they conceal is vital.”
—Aaron Levenstein

“The lottery is a tax on people who flunked math.”
—Monique Lloyd

“The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.”
—Stephen Jay Gould

“Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write.”
—H.G. Wells

Acknowledgments

With gratitude, I acknowledge the following people, animals, and organizations:

The *Woodrow Wilson National Fellowship Foundation*, for getting me started seriously thinking about statistics;

The *College Board*, for giving me opportunity to present workshops for teachers in Advanced Placement Statistics;

The people who attended the College Board workshops—I learned as much from them as they did from me;

My AP Statistics classes at Gunn High School in Palo Alto, CA, for being willing subjects as I learned to teach AP Statistics;

Grace Feedson, for giving me the opportunity to write this book;

Don Reis, for the help and communication needed to keep the project going and, more-or-less, on schedule;

Chris Waters, for his help in proofing the problem sets.

My daughter *Gretchen* and her husband *David*, for their support;

My son *Todd* and his wife *Marlys*, for their support;

Petra and *Sophia*, for being the two most beautiful granddaughters in the history of the world;

Baxter, my son's dog, for keeping me company during some of the writing; and

My wife *Normajean* who constantly encouraged me and showed admirable restraint in keeping the household to-do lists manageable during the writing.

This page intentionally left blank.

PART I

HOW TO USE THIS BOOK

This page intentionally left blank.

Chapter 1

The Five-Step Program

HOW IS THIS BOOK ORGANIZED?

Part I contains an introduction to the Five-Step Program and three study plans for preparing for the AP Statistics exam.

Part II discusses the AP exam and contains a full-length diagnostic test. The diagnostic test contains the same number of questions as the full-length AP exam (although the questions are designed more to see if you need more review on a topic rather than as a practice exam).

Part III (Comprehensive Review) contains 10 chapters beginning with a brief overview and proceeding through all parts of the course. At the end of each chapter, you will find a set of five Rapid Review problems that highlights some points in the chapter, a set of 10–25 Practice Problems (including some multiple choice questions), and a set of five Cumulative Review Problems. Complete solutions are provided for each of the problems.

Part IV contains two full-length practice tests as well as the answers, explanations, and worksheets to compute your score.

The main technologies used in this text are Minitab Statistical Software and the TI-83 graphing calculator.

INTRODUCTION TO THE FIVE-STEP PROGRAM

The Five-Step Program is designed to provide you with the skills and strategies vital to the exam and the practice that can help lead you to a 5 on the AP Exam. Each of the five steps will provide you with the opportunity to get closer to earning a 5.



Step One leads you through a brief process to help determine which type of exam preparation you want to commit yourself to.

1. Month-by-month: September through mid-May
2. The calendar year: January through mid-May
3. Basic training: six weeks before the exam



Step Two helps develop the knowledge you need to succeed on the exam.

1. A comprehensive review of the exam
2. A Diagnostic Test you can go through step-by-step and question-by-question to build your confidence level
3. A summary of formulas and tables related to the AP Statistics exam
4. A list of related Websites and a bibliography



Step Three develops the skills necessary to take the exam and do well.

1. Practice multiple-choice questions.
2. Practice free-response questions.



Step Four helps you develop strategies for taking the exam.

1. Learning about the test itself
2. Learning to read multiple-choice questions.
3. Learning how to answer multiple-choice questions, including when not to guess.
4. Learning how to plan and write the free-response questions.



Step Five will help you develop your confidence in using the skills demanded on the AP Statistics exam.

1. The opportunity to take a diagnostic exam.
2. Time management techniques/skills.
3. Two practice exams that test how well-honed your skills are.

GRAPHICS USED IN THIS BOOK

To emphasize particular skills, strategies, and practice, we use seven sets of icons throughout this book.

The first icon is an hourglass, which indicates the passage of time during the school year. This hourglass icon will appear in the margin next to an item that may be of interest to one of the three types of students using this book (Approach A, B, or C students).



For the student who plans to prepare for the AP Statistics exam during the entire school year, September through May, we use an hourglass that is full on the top.



For the student who decides to begin preparing for the exam in January, we use an hourglass that is half full on the top and half full on the bottom.



For the student who wishes to prepare during the final 6 weeks before the exam, we use an hourglass that is almost empty on the top and almost full on the bottom.



The second icon is a footprint, which indicates which step in the five-step program is being emphasized in a given analysis, technique, or practice activity.



Plan



Knowledge



Skills



Strategies



Confidence Building



The third icon is an exclamation point, which points out an exam tip.



The fourth icon points out a calculator tip.

Boldfaced words indicate terms that are included in the glossary at the end of the book.

THREE APPROACHES TO PREPARING FOR THE AP STATISTICS EXAM

No one knows your study habits, likes, and dislikes better than you. So, you are the only one who can decide which approach you want and/or need to adopt to prepare for the AP Statistics exam. Look at the brief profiles below. These may help you to place yourself in a particular prep mode.



You are a full-year prep student (Approach A) if

1. You like to plan for everything far in advance.
2. You arrive at the airport 3 hours before your flight.
3. You like detailed planning and everything in its place.
4. You feel that you must be thoroughly prepared.
5. You hate surprises.



You are a one-semester prep student (Approach B) if

1. You get to the airport 2 hours before your flight.
2. You are willing to plan ahead to feel comfortable in stressful situa-

tions, but are okay with skipping some details.

3. You feel more comfortable when you know what to expect, but a surprise or two is OK.
4. You are always on time for appointments.



You are a 6-week prep student (Approach C) if

1. You get to the airport at the last possible moment.
2. You work best under pressure and tight deadlines.
3. You feel very confident with the skills and background you've learned in your AP Statistics class.
4. You decided late in the year to take the exam.
5. You like surprises.
6. You feel okay if you arrive 10–15 minutes late for an appointment.

CALENDAR FOR EACH PLAN



A Calendar for Approach A: A Year-Long Preparation for the AP Statistics Exam

Although its primary purpose is to prepare you for the AP Statistics Exam you will take in May, this book can enrich your study of statistics, your analytical skill, and your problem-solving techniques.

SEPTEMBER–OCTOBER (Check off the activities as you complete them)

- _____ Determine into which student mode you would place yourself.
- _____ Carefully read Parts I and II.
- _____ Get on the Web and take a look at the AP Websites.
- _____ Skim the Comprehensive Review section (these areas will be part of your year-long preparation).
- _____ Buy a few highlighters.
- _____ Flip through the entire book. Break the book in. Write in it. Toss it around a little bit . . . highlight it.
- _____ Get a clear picture of what your own school's AP Statistics curriculum is.
- _____ Begin to use the book as a resource to supplement the classroom learning.
- _____ Read and study Chapter 3: Overview of Statistics/Basic Vocabulary.
- _____ Read and study Chapter 4: Univariate Data Analysis.
- _____ Read and study Chapter 5: Bivariate Data Analysis.

NOVEMBER

- _____ Read and study Chapter 6: Design of a Study: Sampling, Surveys, and Experiments
- _____ Review Chapters 3 and 4.

DECEMBER

- _____ Read and study Chapter 7: Random Variables and Probability.
- _____ Review Chapters 5 and 6.

JANUARY

- _____ Read and study Chapter 8: Binomial Distribution and Sampling Distributions.
- _____ Review Chapters 5–7.

FEBRUARY

- _____ Read and study Chapter 9: Confidence Intervals and Introduction to Inference.
- _____ Review Chapters 6–8.
- _____ Look over the Diagnostic Exam.

MARCH

- _____ Read and study Chapter 10: Inference for Means and Proportions.
- _____ Read and study Chapter 11: Inference for Regression.
- _____ Review Chapters 7–9.

APRIL

- _____ Read and study Chapter 12: Inference for Categorical Data: Chi-square.

8 • How to Use this Book

- _____ Review Chapters 9–11.
- _____ Take the Diagnostic Exam.
- _____ Evaluate your strengths and weaknesses.
- _____ Study appropriate chapters to correct weaknesses.

MAY

- _____ Take and score Practice Exam 1.
- _____ Study appropriate chapters to correct weaknesses.

- _____ Take and score Practice Exam 2.
- _____ Study appropriate chapters to correct weaknesses.
- _____ Get a good night's sleep the night before the exam.

GOOD LUCK ON THE TEST!



A Calendar for Approach B: A Semester-Long Preparation for the AP Statistics Exam

Working under the assumption that you've complete one semester of statistics studies, the following calendar will use those skills you've been practicing to prepare you for the May exam.

JANUARY

- _____ Carefully read Parts I and II.
- _____ Read and study Chapter 3: Overview of Statistics/Basic Vocabulary.
- _____ Read and study Chapter 4: Univariate Data Analysis.
- _____ Read and Study Chapter 5: Bivariate Data Analysis.
- _____ Read and Study Chapter 6: Design of a Study: Sampling, Surveys, and Experiments.

FEBRUARY

- _____ Read and study Chapter 7— Random Variables and Probability.
- _____ Read and study Chapter 8— Binomial Distributions and Sampling Distributions.
- _____ Review Chapters 3–6.

MARCH

- _____ Read and study Chapter 9: Confidence Intervals and Introduction to Inference.
- _____ Read and study Chapter 10: Inference for Means and Proportions.
- _____ Review Chapters 5–8.

APRIL

- _____ Read and study Chapter 11: Inference for Regression.
- _____ Read and study Chapter 12: Inference for Categorical Data: Chi-square.
- _____ Review Chapters 7–10.
- _____ Take Diagnostic Exam.
- _____ Evaluate your strengths and weaknesses.
- _____ Study appropriate chapters to correct weaknesses.

MAY

- _____ Take and score Practice Exam 1.
- _____ Study appropriate chapters to correct weaknesses.
- _____ Take and score Practice Exam 2.
- _____ Study appropriate chapters to correct weaknesses.
- _____ Get a good night's sleep the night before the exam.

GOOD LUCK ON THE TEST!



A Calendar for Approach C: A 6-Week Preparation for the AP Statistics Exam

At this point we are going to assume that you have been building your statistics knowledge base for more than 6 months. You will, therefore, use this book primarily as a specific guide to the AP Statistics Exam.

Given the time constraints, now is not the time to expand you AP Statistics curriculum. Rather, it is the time to limit and refine what you already do know.

APRIL 1st–15th

- _____ Skim Parts I and II.
- _____ Skim Chapters 3–8.
- _____ Carefully go over the “Rapid Review” sections of Chapter 3–8.

APRIL 16th–30th

- _____ Skim Chapter 9–12.
- _____ Carefully go over the “Rapid Review” sections of Chapters 9–12.
- _____ Take the Diagnostic Exam.
- _____ Evaluate your strengths and weaknesses.
- _____ Study appropriate chapters to correct weaknesses.

MAY

- _____ Take and score Practice Exam 1.
- _____ Study appropriate chapters to correct weaknesses.
- _____ Take and score Practice Exam 2.
- _____ Study appropriate chapters to correct weaknesses.
- _____ Get a good night’s sleep the night before the exam.

GOOD LUCK ON THE TEST!

Summary of the Three Study Plans



Month	Approach A: A Year-Long Plan	Approach B: A Semester-Long Plan	Approach C: A 6-Week Plan
September–October	Chapters 3–5		
November	Chapter 6 Review 3 & 4		
December	Chapter 7 Review 5 & 6		
January	Chapter 8 Review 5–7	Chapters 3–6	
February	Chapter 9 Review 6–8	Chapters 7 & 8 Review 3–6	
March	Chapters 10 & 11 Review 7–9	Chapters 9 & 10 Review 5–8	
April	Chapter 12 Review 9–11 Diagnostic Exam	Chapters 11 & 12 Review 7–10 Diagnostic Exam	Review 3–12 “Rapid Reviews” 3–12 Diagnostic Exam
May	Practice Exam 1 Practice Exam 2	Practice Exam 1 Practice Exam 2	Practice Exam 1 Practice Exam 2

This page intentionally left blank.

PART II

WHAT YOU NEED TO KNOW ABOUT THE AP STATISTICS EXAM

This page intentionally left blank.

Chapter 2

Introduction to the AP Statistics Exam

4 BACKGROUND ON THE AP STATISTICS EXAM

What Is Covered in the AP Statistics Exam

The AP Statistics Exams covers the following four broad themes:

- **Exploring Data:** observing patterns and departures from patterns
- **Planning a Study:** deciding what and how to measure
- **Anticipating Patterns:** producing models using probability theory and simulation
- **Statistical Inference:** confirming models

The broad themes described above cover the following topics:

- Dotplots, stemplots, histograms, cumulative frequency plots, measures of center, measures of spread, position of a term in a distribution, boxplots, the effects of changing units on summary statistics, comparing distributions, scatterplots, correlation and linearity, least square regression, residual plots, outliers and influential points, transformations to achieve linearity, and frequency tables for categorical data
- Census, surveys, experiments, observational studies, sampling, sources of bias, stratification, randomization, blocking, placebo effect, blinding, completely randomized design, and generalizability of results
- Probability, law of large numbers, addition and multiplication rules, conditional probability, discrete and continuous random variables, simulation, binomial and geometric distributions, linear transformations of random variables, combining independent random variables, independence versus dependence, the normal distribution, sampling

distributions, Central Limit Theorem, and simulations of sampling distributions

- Confidence intervals for one and two means, confidence intervals for one and two proportions, significance tests for means and proportions, errors in hypothesis testing, the power of a test, chi-square tests for goodness-of-fit and for 2-way tables, the t -distribution, t -procedures for means and proportions, inference for the slope of the least-squares regression line

For a more detailed description of the topics covered in the AP Statistics exam, visit the College Board's AP Central Website at: <http://apcentral.collegeboard.com>

What Is the Format of the AP Statistics Exam?

The AP Statistics exam has 2 sections:

Section I contains 40 multiple choice questions. The time allowed for this section is 90 minutes.

Section II contains six problems. The time allowed for this section is 90 minutes. Section II is divided into two parts:

- Part A has five free response questions for which the student is expected to take about 65 minutes.
- Part B has one longer investigative task for which the student is expected to take about 25 minutes.

Approved graphing calculators are allowed during all parts of the test. The two sections of the test are completely separate and are administered in separate 90-minute blocks. Please note that you are not expected to be able answer all the questions in order to receive a grade of 5. If you wish to see the specific instructions for each part of the test, visit the College Board's AP Central Website at: <http://apcentral.collegeboard.com>

You will be provided with a set of common statistical formulas and necessary tables. Copies of these materials are in the appendices to this book.

What Are the Advanced Placement Exam Grades?

Advanced placement grades are given on a 5-point scale with 5 being the highest and 1 being the lowest. The grades are described below:

- 5 = Extremely well qualified
- 4 = Well qualified
- 3 = Qualified
- 2 = Possibly qualified
- 1 = No recommendation

There is no official "passing" grade on the exam. Many people consider "3" or better to be passing. Many colleges will give course credit for grades of 3 or better, although many schools require a 4 for credit.

How Is the AP Statistics Grade Calculated?

- The exam has a composite score of 100 points: 50 points for the 40 questions in Section I and 50 points for the 6 problems in Section II.
- In Section I, the weighted score is computed as follows:

$$[(\text{number correct}) - (\frac{1}{4}(\text{number wrong}))] \times 1.25 = \text{Weighted Section I Score (minimum score} = 0).$$
 There is no deduction for blank answers.

- In section II, each problem is scored holistically on a 4, 3, 2, 1, 0 basis. These scores can be interpreted as follows:

4: complete response
 3: substantial response
 2: developing response
 1: minimal response
 0: no credit

The problems in this section are scored holistically both on computational accuracy and on communication of process. Right answers without convincing justification may not receive full credit. A rubric is developed for each question, and the readers are carefully trained to apply the rubric consistently. Unlike calculus, where the top score can be obtained only with a perfect solution, a “complete response” does not mean a “perfect response.” The quality of the complete solution is considered in assigning the grade. You can make small errors (nonstatistical) and still receive a 4 on a problem.

Once a score on each of questions 1–5 has been arrived at, that score is multiplied by 1.875. The score for question 6 is multiplied by 3.125 (not rounded). The effect of this is to make problems 1–5 worth 75% of section II and problem 6 worth 25%. The sum of these six scores is the Weighted Score for Section II (see the scoring sheets after the diagnostic test and after the practice exams).

- The weighted scores for Sections I and II are combined to yield a composite score based on 100 (rounded to the nearest whole number).
- The Chief Reader for the exam has the responsibility of turning the composite score into an AP exam grade. The cut-off points for each grade (1–5) vary from year to year. The following represent the cut-off points in the year 2002:

Composite score	AP grade
68–100	5
53–67	4
40–52	3
29–39	2
0–28	1

These scores change from year to year. It would be misleading to interpret the composite scores as percentage correct, especially given the holistic scoring in Section II of the exam.

What Is the Graphing Calculator Policy for the AP Statistics Exam?

The following is the policy on graphing calculators as stated on the College Board's AP Central Website.

- Students are expected to bring a graphing calculator with statistical capabilities to the exam and to be familiar with its use. The calculator's computational capabilities should include standard statistical univariate and bivariate summaries through linear regression. Graphic capabilities should include common univariate and bivariate displays, such as histograms, boxplots, and scatterplots.
- Calculator memories do not have to be cleared; however, calculator memories may be used only for storing programs, not for storing notes.
- Minicomputers, electronic writing pads or pen input devices (Newton, Palm), pocket organizers, models with QWERTY [i.e., typewriter keyboards (TI-92 and HP-95)], models with paper tapes, models that make noise or "talk," and models that require an electrical outlet are not allowed, but most graphing calculators that are on the market are acceptable.
- Nongraphing scientific calculators are permitted only if they have the statistics computational capabilities described in the AP Statistics Course Description.
- Each student may bring up to two calculators to the exam.

You may use a calculator to do needed computations. However, remember that the person reading your exam needs to see your reasoning in order to score your exam.

What Do I Need to Bring to the Exam?

- Several #2 pencils
- A good eraser and a pencil sharpener
- Two black or blue pens
- One or two approved graphing calculators with fresh batteries
- A watch
- An admissions card or a photo I.D. card if your school requires it
- Your Social Security number
- Your school code number if the test site is not at your school
- A simple snack *if the test site permits it*
- A light jacket if you know that the test site has strong air conditioning
- Do *not* bring *Wite Out* or scrap paper.

TIPS FOR TAKING THE EXAM



- Write legibly.
- Label all diagrams.

- Organize your solutions so that the reader can follow your line of reasoning.
- Use complete sentences whenever possible—communication is very important. Clearly indicate your final answer.
- Do easy questions first
- Write out formulas and indicate all major steps.
- Guess only if you can eliminate some of the choices in a multiple choice question.
- Leave a multiple choice question blank if you have no clue what the answer is.
- Be careful to bubble in the right grid, especially if you skip a question.
- Keep moving—don't linger on a problem too long. You have an average of slightly more than 2 minutes for each question on Section I, 12–13 minutes for each problem in Part A of Section II, and 25–30 minutes for Part B of Section II.
- Indicate units of measurement if required.
- Simplify algebraic or numeric expressions for final answers—that's part of communication. You will not be penalized for making simple arithmetic or algebraic errors.
- Work with full calculator accuracy as you do a problem. Round final answers only to 2 or 3 decimal places unless the problem gives you specific instructions otherwise.
- Read all parts of each question before beginning the question.
- Avoid using calculator syntax in place of statistical syntax in your solutions.
- In Section II, try to answer all parts of every question. You can't get credit for a solution if it is blank; you *might* get some credit if you try to answer it.
- Be familiar with the instructions for the different parts of the exam before the day of the exam. Visit the College Board Website for more information.
- Get a good night's sleep the night before the exam.

GETTING STARTED

Answer Sheet for Diagnostic Test—Section I

- | | |
|-----------|-----------|
| 1. _____ | 21. _____ |
| 2. _____ | 22. _____ |
| 3. _____ | 23. _____ |
| 4. _____ | 24. _____ |
| 5. _____ | 25. _____ |
| 6. _____ | 26. _____ |
| 7. _____ | 27. _____ |
| 8. _____ | 28. _____ |
| 9. _____ | 29. _____ |
| 10. _____ | 30. _____ |
| 11. _____ | 31. _____ |
| 12. _____ | 32. _____ |
| 13. _____ | 33. _____ |
| 14. _____ | 34. _____ |
| 15. _____ | 35. _____ |
| 16. _____ | 36. _____ |
| 17. _____ | 37. _____ |
| 18. _____ | 38. _____ |
| 19. _____ | 39. _____ |
| 20. _____ | 40. _____ |

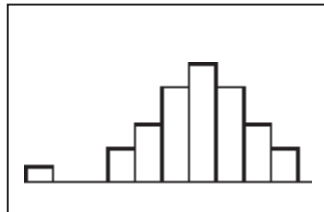
2 DIAGNOSTIC TEST

Section I

Directions: Use the answer sheet provided on the previous page. All questions are given equal weight. There is no penalty for unanswered questions, but $\frac{1}{4}$ of the number of incorrect answers will be subtracted from the number of correct answers. The use of a calculator is permitted in all parts of this test. You have 90 minutes for this part of the test.

1. Eighteen trials of a binomial random variable, X , are conducted. If the probability of success on any one trial is 0.4, write the mathematical expression you would need to evaluate to find $P(X = 7)$. Do not evaluate.
2. Two variables, x and y , seem to be exponentially related. The natural logarithm of each y value is taken and the least-squares regression line of $\ln(y)$ on x is determined to be $\ln(y) = 3.2 + .42x$. What is the predicted value of y when $x = 7$?
3. You need to construct a large sample 94% confidence interval for a population mean. What is the upper critical value of z to be used in constructing this interval?

4.



Describe the shape of the histogram at the left.

5. The probability is .2 that a term selected at random from a normal distribution with mean 600 and standard deviation 50 will be above what number?
6. Which of the following are examples of continuous data?
 - a. The speed your car goes
 - b. The number of outcomes of a binomial experiment
 - c. The average temperature in San Francisco
 - d. The wingspan of a bird
 - e. The jersey numbers of a football team

Use the following computer output for a least-squares regression in questions 7 and 8.

The regression equation is				
Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	22.94	11.79	1.95	.088
<i>x</i>	-0.6442	0.5466	-1.18	_____
s = 2.866		R-sq = 14.8%		R-sq(adj) = 4.1%

7. What is the equation of the least-squares regression line?
8. Given that the analysis is based on 10 datapoints, what is the *P* value for the *t*-test of the hypothesis $H_0: \beta = 0$?
9. A hypothesis test yields a *P* value of .20. Which of the following two statements best describes what is meant by this statement?
 - I. The probability of getting a finding as extreme as obtained by chance alone if the null hypothesis is true is .20.
 - II. The probability of getting a finding as extreme as obtained by chance alone from repeated random samples is .20.
10. A random sample of 25 men and a separate random sample of 25 women are selected to answer questions about attitudes toward abortion. The answers were categorized as “pro-life” or “pro-choice.” Which of the following is the proper null hypothesis for this situation:
 - I. The variables “gender” and “attitude toward abortion” are independent.
 - II. The proportion of “pro-life” men is the same as the proportion of “pro-life” women.
11. A sports talk show asks people to call in and give their opinion of the officiating in the local basketball teams’ most recent loss. What will most likely be the typical reaction?
 - a. They will likely feel that the officiating could have been better, but that it wasn’t the teams’ poor play, not the officiating, that was primarily responsible for the loss.
 - b. The team probably wouldn’t have lost if the officials had been doing their job.
 - c. Because the team has been foul-plagued all year, the callers would most likely support the officials.
12. A major polling organization wants to predict the outcome of an upcoming national election (in terms of the proportion of voters who will vote for each candidate). They intend to use a 95% confidence

interval with margin of error of no more than 2.5%. What is the minimum sample size needed to accomplish this goal?

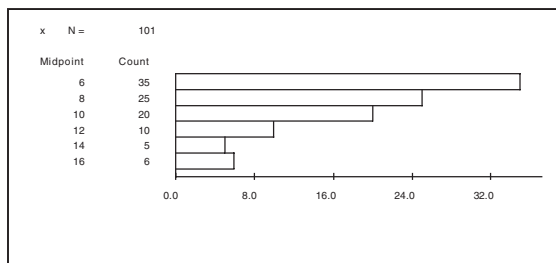
13. A sample of size 35 is to be drawn from a large population. The sampling technique is such that every possible sample of size 35 that could be drawn from the population is equally likely. What name is given to this sampling technique?
14. A teacher's union and a school district are negotiating salaries for the coming year. The teachers want more money, and the district, claiming, as always, budget constraints, wants to pay as little as possible. The district, like most, has a large number of moderately paid teachers and a few highly paid administrators. The salaries of all teachers and administrators are included in trying to figure out, on average, how much the professional staff currently earns. Which of the following would the teachers' union be most likely to quote during negotiations?
 - a. The mean of all the salaries
 - b. The mode of all the salaries
 - c. The median of all the salaries
 - d. The standard deviation of all the salaries
 - e. The ranges of all the salaries
15. Alfred and Ben don't know each other but are each considering asking the lovely Charlene to the school prom. The probability that at least one of them will ask her is .72. The probability that they both ask her is .18. The probability that Alfred asks her is .6. What is the probability that Ben asks Charlene to the prom?
16. A significance test of the hypothesis $H_0: p = .3$ against the alternative $H_A: p > .3$ found a value of $\hat{p} = .35$ for a random sample of size 95. What is the P value of this test?
17. Which of the following best describes the Central Limit Theorem?
 - I. The mean of the sampling distribution of \bar{x} is the same as the mean of the population.
 - II. The standard deviation of the sampling distribution of \bar{x} is the same as the standard deviation of \bar{x} divided by the square root of the sample size.
 - III. If the sample size is large, the shape of the sampling distribution of \bar{x} is approximately normal.
 - a. I only
 - b. II only
 - c. III only
 - d. I and II only
 - e. I, II, and III

18. If three fair coins are flipped, $P(0 \text{ heads}) = .125$, $P(\text{exactly 1 head}) = .375$, $P(\text{exactly 2 heads}) = .375$, and $P(\text{exactly 3 heads}) = .125$. The following results were obtained when three coins were flipped 64 times:

# Heads	Observed
0	10
1	28
2	22
3	4

What is the value of the χ^2 statistic used to test if the coins are behaving as expected, and how many degrees of freedom does the determination of the P value depend on?

- 19.



For the histogram pictured above, what is the class interval (boundaries) for the class that contains the median of the data?

20. Thirteen large animals were measured to help determine the relationship between their length and their weight. The natural logarithm of the weight of each animal was taken and a least-squares regression equation for predicting weight from length was determined. The computer output from the analysis is given below:

The regression equation is				
$\ln(\text{wt}) = 1.24 + 0.0365 \text{ length}$				
Predictor	Coef	St Dev	t ratio	P
Constant	1.2361	0.1378	8.97	.000
Length	0.036502	0.001517	24.05	.000
$s = 0.1318$	$R\text{-sq} = 98.1\%$	$R\text{-sq}(\text{adj}) = 98.0\%$		

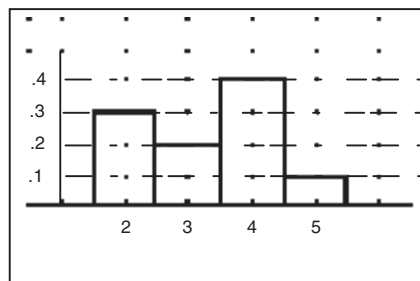
Give a 99% confidence interval for the slope of the regression line.

21. What are the mean and standard deviation of a binomial experiment that occurs with probability of success .76 and is repeated 150 times?
22. Which of the following is the primary difference between an experiment and an observational study?
- Experiments are only conducted on human subjects; observational studies can be conducted on nonhuman subjects.

- b. In an experiment, the researcher manipulates some variable to observe its effect on a response variable; in an observational study, he or she simply observes and records the observations.
- c. Experiments must use randomized treatment and control groups; observational studies also use treatment and control groups, but they do not need to be randomized.
- d. Experiments must be double-blind; observational studies do not need to be.
- e. There is no substantive difference—they can both accomplish the same research goals.
23. The regression analysis of question 20 indicated that “R-sq = 98.1%” Which of the following are true?
- I. There is a strong positive linear relationship between the explanatory and response variables.
- II. There is a strong negative linear relationship between the explanatory and response variables.
- III. About 98% of the variation in the response variable can be explained by the regression on the explanatory variable.
24. A hypothesis test is set up so that $P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) = .05$ and $P(\text{failing to reject } H_0 \text{ when } H_0 \text{ is false}) = .26$. What is the value of the power of the test?
25. For the following observations collected while doing a chi-square test for independence between the two variables A and B , find the expected value of the cell marked with “XXXX.”

5	10(XXXX)	11
6	9	12
7	8	13

26. The following is a probability histogram for a discrete random variable X .



What is μ_x ?

27. A psychologist believes that positive rewards for proper behavior is more effective than punishment for bad behavior in promoting good behavior in children. A scale of “proper behavior” is developed.

μ_1 = the “proper behavior” rating for children receiving positive rewards, and μ_2 = the “proper behavior” rating for children receiving punishment. If $H_0: \mu_1 - \mu_2 = 0$, which of the following is the proper statement of H_A ?

- a. $H_A: \mu_1 - \mu_2 > 0$
 - b. $H_A: \mu_1 - \mu_2 < 0$
 - c. $H_A: \mu_1 - \mu_2 \neq 0$
 - d. Any of the above are an acceptable alternative to the given null.
 - e. There isn't enough information given in the problem for us to make a decision.
28. Estrella wants to become a paramedic and takes a screening exam. Scores on the exam have been approximately normally distributed over the years it has been given. The exam is normed with a mean of 80 and a standard deviation of 9. Only those who score in the top 15% on the test are invited back for further evaluation. Estrella received a 90 on the test. What was her percentile rank on the test, and did she qualify for further evaluation?
29. Which of the following statements are true?
- I. In order to use a χ^2 procedure, the number of expected values for each cell of a one- or two-way table must be at least 5 values.
 - II. In order to use χ^2 procedures, you must have at least 2 degrees of freedom.
 - III. In a 4×2 two-way table, the number of degrees of freedom is 3.
- a. I only
 - b. I and II only
 - c. I and III only
 - d. III only
 - e. I, II, and III.
30. When the point (15,2) is included, the slope of the regression line, $\hat{y} = a + bx$, is $b = -.54$ and $r = -.82$. When the point is removed, the new slope is -1.04 and the new correlation coefficient is $-.95$. What name is given to a point whose removal has this kind of effect on the slope of the regression line and correlation between two variables?
31. A one-sided test of a hypothesis about a population mean, based on a sample of size 14, yields a P value of .075. Which of the following best describes the range of t values that would have given this P value
- a. $1.345 < t < 1.761$
 - b. $1.356 < t < 1.782$
 - c. $1.771 < t < 2.160$
 - d. $1.350 < t < 1.771$
 - e. $1.761 < t < 2.145$

32. Use the following excerpt from a random digits table for assigning 6 people to treatment and control groups:

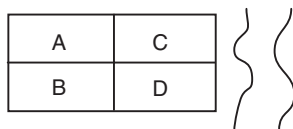
98110 35679 14520 51198 12116 98181 99120 75540
03412 25631

The subjects are labeled: Arnold: 1; Betty: 2; Clive: 3; Doreen: 4; Ernie: 5; Florence: 6. The first three subjects randomly selected will be in the treatment group; the other three in the control group. Assuming you begin reading the table at the extreme left digit, which three subjects would be in the *control* group?

33. A null hypothesis, $H_0: \mu = \mu_0$ is to be tested against a 2-sided hypothesis. A sample is taken, \bar{x} is determined and used as the basis for a C-level confidence interval (e.g., $C = 0.95$) for μ . The researcher notes that μ_0 is not in the interval. Another researcher chooses to do a significance test for μ using the same data. What significance level must the second researcher choose in order to guarantee getting the same conclusion about $H_0: \mu = \mu_0$ (that is, reject or not reject) as the first researcher?
34. Which of the following is not required in a binomial setting?
- Each trial is considered either a success or a failure.
 - Each trial is independent.
 - There are a fixed number of trials.
 - Each trial succeeds or fails with the same probability.
 - The value of the random variable of interest is the number of trials until the first success.
35. X and Y are independent random variables with $\mu_X = 3.5$, $\mu_Y = 2.7$, $\sigma_X = .8$, and $\sigma_Y = .65$. What are μ_{X+Y} and σ_{X+Y} ?
36. A researcher is hoping to find a linear relationship between the explanatory and response variables in her study. Accordingly, as part of her analysis she plans to generate a 95% confidence interval for the slope of the regression line for the two variables. The interval is determined to be $\langle 0.45, 0.80 \rangle$. Which of the following is (are) true?
- She has good evidence of a linear relationship between the variables.
 - It is likely that there is a non-zero correlation (r) between the two variables.
 - It is likely that the true slope of the regression line is 0.
37. In the casino game of roulette, there are 38 slots for a ball to drop into when it is rolled around the rim of a revolving wheel: 18 red, 18 black, and 2 green. What is the probability that the first time a

ball drops into the red slot is on the 8th trial (in other words, suppose you are betting on red every time—what is the probability of losing 7 straight times before you win the first time?)?

38. You are developing a new strain of strawberries (say, Type X) and are interested in their sweetness as compared to another strain (say, Type Y). You have four plots of land, call them A, B, C, and D, which are roughly four squares in one large plot for your study (see the figure below). A river runs alongside of plots C and D. Because you are worried that the river might influence the sweetness of the berries, you randomly plant type X in either A or B (and Y in the other) and randomly plant type X in either C or D (and Y in the other). Which of the following terms best describes this design?



- a completely randomized design.
 - a comparative randomized study.
 - a comparative block design, controlling for the effects of the river.
 - a randomized observational study.
 - a comparative block design, controlling for the strain of strawberry.
39. Grumpy got 38 on the first quiz of the quarter. The class average on the first quiz was 42 with a standard deviation of 5. Dopey, who was absent when the first quiz was given, got 40 on the second quiz. The class average on the second quiz was 45 with a standard deviation of 6.1. Grumpy was absent for the second quiz. After the second quiz, Dopey told Grumpy that he was doing better in the class because they had each taken one quiz, and he had gotten the higher score. Did he really do better? Explain.
40. A random sample of size 45 is obtained for the purpose of testing the hypothesis $H_0 : p = .80$. The sample proportion is determined to be $\hat{p} = .75$. What is the value of the standard error of \hat{p} for this test?

Section II—Part A, Questions 1–5

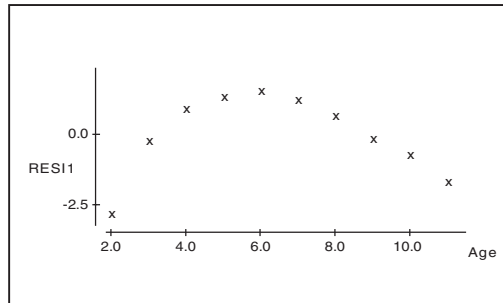
Spend about 65 minutes on this part of the exam. Percentage of Section II grade—75.

Directions: Show all of your work. Indicate clearly the methods you use because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

1. The regression equation is

Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	76.641	1.188	64.52	.000
Age	6.3661	0.1672	38.08	.000

$s = 1.518$ $R\text{-sq} = 99.5\%$ $R\text{-sq}(\text{adj}) = 99.4\%$



The ages (in years) and heights (in cm) of 10 girls, ages 2 through 11, were recorded. Part of the regression output and the residual plot for the data are given above.

- What is the equation of the least-squares regression line for predicting height from age?
 - Interpret the slope of the regression line in the context of the problem.
 - Suppose you wanted to predict the height of a girl 5.5 years of age. Would the prediction made by the regression equation you gave in (a) be too small, too large, or is there not enough information to tell?
2. You want to determine whether a greater proportion of men or women purchase vanilla latte (regular or decaf). To collect data, you hire a person to stand inside the local Scorebucks for 2 hours one morning and tally the number of men and women who purchase the vanilla latte as well as the total number of men and women customers. Sixty-three percent of the women and 59% of the men purchase a Vanilla Latte.
- Is this an experiment or an observational study? Explain.
 - Based on the data collected, you write a short article in the local newspaper claiming that a greater proportion of women than men prefer vanilla latte as their designer coffee of choice. A student in

the local high school AP Statistics class writes a letter to the editor criticizing your study. What might the student have pointed out?

- c. Suppose you wanted to conduct a study less open to criticism, how might you redo the study?
3. Sophia is a nervous basketball player. Over the years she has had a 40% chance of making the first shot she takes in a game. If she makes her first shot, her confidence goes way up, and the probability of her making the second shot she takes rises to 70%. But if she misses her first shot, the probability of her making the second shot she takes doesn't change—it's still 40%.
- a. What is the probability that Sophia makes her second shot?
- b. If Sophia does make her second shot, what is the probability that she missed her first shot?
4. A random sample of 72 seniors taken 3 weeks before the selection of the school Homecoming Queen, identified 60 seniors who planned to vote for Buffy for queen. Unfortunately, Buffy said some rather catty things about some of her opponents, and it got into the school newspaper. A second random sample of 80 seniors taken shortly after the article appeared showed that 55 planned to vote for Buffy. Does this indicate a serious drop in support for Buffy? Use good statistical reasoning to support your answer.
5. Some researchers believe that education influences IQ. One researcher specifically believes that the more education a person has, the higher, on average, will be their IQ. The researcher sets out to investigate this belief by obtaining eight pairs of identical twins reared apart. He identifies the better educated twin as Twin A and the other twin as Twin B for each pair. The data for the study is given in the table below. Do the data give good statistical evidence, at the .05 level of significance, that the twin with more education is likely to have the higher IQ? Give good statistical evidence to support your answer.

Pair	1	2	3	4	5	6	7	8
Twin A	103	110	90	97	92	107	115	102
Twin B	97	103	91	93	92	105	111	103

Section II—Part B, Question 6

Spend about 25 minutes on this part of the exam. Percentage of Section II grade—25.

Directions: Show all of your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

6. A paint manufacturer claims that the average drying time for its best-selling paint is 2 hours. A random sample of drying times for 20 randomly selected cans of paint are obtained to test the manufacturer's claim. The drying times observed were: 123, 118, 115, 121, 130, 127, 112, 120, 116, 136, 131, 128, 139, 110, 133, 122, 133, 119, 135, 109.
- Obtain a 95% confidence interval for the true mean drying time of the paint.
 - Interpret the confidence interval obtained in part (a) in the context of the problem.
 - Suppose, instead, a significance test, at the .05 level, of the hypothesis $H_0: \mu = 120$ was conducted against the alternative $H_A: \mu \neq 120$. What is the P value of the test?
 - Are the answers you got in part (a) and part (c) consistent? Explain.
 - At the .05 level, would your conclusion about the mean drying time have been different if the alternative hypothesis had been $H_A: \mu > 120$? Explain.

ANSWERS AND SOLUTIONS

Answers to Diagnostic Test—Section I

1. $\binom{18}{7}(.4)^7(.6)^{11}$
2. 464.05
3. 1.88
4. Approximately normal with an outlier.
5. 612.6
6. a, c, d
7. $\hat{y} = -.6442 + 22.94x$
8. $.02 < P < .04$ [Calculator answer: .272]
9. I
10. II
11. b.
12. 1537
13. Simple random sample
14. c
15. .3
16. .1446
17. c
18. 3.33, 3
19. $\langle 7, 9 \rangle$
20. $\langle .032, .041 \rangle$
21. 114, 5.23
22. b
23. I or II (not both), III
24. .74
25. 8
26. 3.3
27. a
28. Percentile rank = 86.67; yes
29. c
30. Influential point
31. d.
32. Betty, Doreen, Florence
33. $\alpha - 1 - C$
34. e
35. $\mu_{X+Y} = 6.2, \sigma_{X+Y} = 1.03$
36. I and II are true
37. .0053
38. c
39. No. z_{Dopey} is more negative than z_{Grumpy}
40. .0596

Solutions to Diagnostic Test—Section I

1. From Chapter 8

If X has $B(n, p)$, then, in general,

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

In this problem, $n = 18$, $p = .4$, $x = 7$ so that

$$P(X = 7) = \binom{18}{7} (.4)^7 (.6)^{11}$$

2. From Chapter 5

$$\ln(y) = 3.2 + .42(7) = 6.14 \hat{O} y = e^{6.14} = 464.05$$

3. From Chapter 9

For a 94% z -interval, there will be 6% of the area outside of the interval. That is, there will be 97% of the area less than the upper critical value of z . The nearest entry to 0.97 in the table of standard normal probabilities is 0.9699, which corresponds to a z -score of 1.88.

4. From Chapter 4

If the bar to the far left was not there, this graph would be described as approximately normal. It still has that same basic shape but, because there is an outlier, the best description is: approximately normal with an outlier.

5. From Chapter 7

Let x be the value in question. If there is 0.2 of the area above x , then there is 0.8 of the area to the left of x . This corresponds to a z -score of 0.84 (the nearest table entry is 0.5987). Hence

$$z_x = .84 = \frac{x - 600}{15} \rightarrow x = 612.6.$$

6. From Chapter 3

Discrete data are countable; continuous data correspond to intervals or measured data. Hence, speed (a), average temperature (c), and wingspan (d) are examples of continuous data. The number of outcomes of a binomial experiment (b) and the jersey numbers of a football team (e) are countable and, therefore, discrete.

7. From Chapter 5

The slope of the regression line, -0.6442 , can be found under “Coef” to the right of “ x .” The intercept of the regression line, 22.94 , can be found under “Coef” to the right of “Constant.”

8. From Chapter 11

The t test statistic for $H_0: \beta = 0$ is given in the printout as -1.18 . Because $n = 10$, from the t distribution tables, this value is between 2.449 and 2.896 which, for the two-sided test, corresponds to $0.02 < P < 0.04$. On the TI-83, the answer can be found as $t\text{cdf}(-100, -1.18, 8) = 0.272$

9. From Chapter 10

The statement in I is basically a definition of P value. It is the likelihood of obtaining *by chance* as value as extreme or more extreme by chance alone *if* the null hypothesis is true. A very small P value throws doubt on the truth of the null hypothesis.

10. From Chapter 12

Because the samples of men and women represent different populations, this is a chi-square test of homogeneity of proportions: the proportions of each value of the categorical variable (in this case, “pro-choice” or “pro-life”) will be the same across the different populations. Had there been only one sample of 50 people drawn, 25 of whom happened to be men and 25 of whom happened to be women, this would have been a test of independence.

11. From Chapter 6

This is a voluntary response survey and is subject to voluntary response bias. That is, people who feel the most strongly about an issue are those most likely to respond. Because (a) and (c) are the more rational responses, they are less likely to have been the way most callers would have responded.

12. From Chapter 9

Because we have no basis for estimating the true population proportion in this case, we need to use the formula

$$n = \frac{z^*{}^2}{2m} = \frac{1.96^2}{2(0.025)} = 1536.64.$$

The question asked for the *minimum* number necessary to achieve the goal. Hence, we round up to 1537.

13. From Chapter 6

A random sample from a population is one in which every *member* of the population is equally likely to be selected. A simple random sample is one in which every *sample* of a given size is equally likely to be selected. A sample can be a random sample without being a simple random sample.

14. From Chapter 4

The teachers are interested in showing that the average teacher salary is low. Because the mean is not resistant, it is pulled in the direction of the few higher salaries and, hence, would be higher than the median, which is not affected by a few extreme values. The teachers would choose the median. The mode, standard deviation, and range tell you nothing about the *average* salary.

15. From Chapter 7

$P(\text{at least one of them will ask her}) = P(A \text{ or } B) = .72$. $P(\text{they both ask her}) = P(A \text{ and } B) = .18$. $P(\text{Alfred asks her}) = P(A) = .6$. In general, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. Thus, $.72 = .6 + P(B) - .18 \rightarrow P(B) = .3$.

16. From Chapter 10

$$P \text{ value} = P \left\{ z > \frac{.35 - .30}{\sqrt{\frac{(.3)(.7)}{95}}} = 1.06 \right\} = 1 - .8554 = .1446$$

17. From Chapter 8

Although all three of the statements are true of a sampling distribution, only III is a statement of the Central Limit Theorem.

18.

# Heads	Observed	Expected
0	10	$(.125)(64) = 8$
1	28	$(.375)(8) = 24$
2	22	$(.125)(64) = 8$
3	4	$(.375)(8) = 24$

$$\chi^2 = \frac{(10 - 8)^2}{8} + \frac{(28 - 24)^2}{24} + \frac{(22 - 24)^2}{24} + \frac{(4 - 8)^2}{8} = 3.33.$$

[This calculation can be done on the calculator as follows: L1 = observed values; L2 = expected values; L3 = $(L2 - L1)^2 / L2$; LIST MATH sum(L3)]

In a chi-square goodness-of-fit test, the number of degrees of freedom equals the number of possible outcomes -1 . In this case, $df = n - 1 = 4 - 1 = 3$.

19. From Chapter 4

There are 101 terms, so the median is located at the 56th place in the ordered list of terms. From the counts given, the median must be in the interval whose midpoint is 8. Because the intervals are each of width 2, the class interval for the interval whose midpoint is 8 must be $\langle 7, 9 \rangle$.

20. From Chapter 11

$$0.0365 \pm 3.106(0.0015) = \langle 0.032, 0.041 \rangle$$

21. From Chapter 8

$$\mu_x = 150(0.76) = 114, \sigma_x = \sqrt{150(0.76)(1 - 0.76)} = 5.23$$

22. From Chapter 6

In an experiment, the researcher imposes some sort of treatment on the subjects of the study. Both experiments and observational studies can be conducted on human and nonhuman units; there should be randomization to groups in both to the extent possible; they can both be double blind.

23. From Chapter 5

III is basically what is meant when we say $R\text{-sq} = 98.1\%$. However, $R\text{-sq}$ is the square of the correlation coefficient.

$$\sqrt{R^2} = \pm R = \pm .99 \rightarrow r \text{ could be either positive or negative, but not both. We can't tell direction from } R^2.$$

24. From Chapter 9

The *power* of a test is the probability of correctly rejecting H_0 when H_A is true. You can either fail to reject H_0 when it is false (Type II) or reject it when it is false (Power). Thus, $\text{Power} = 1 - P(\text{Type II}) = 1 - .26 = .74$.

25. From Chapter 12

There are 81 observations total, 27 observations in the second column, 24 observations in the first row. The expected number in the first row and second column equals

$$\sqrt[27]{81}(24) = 8.$$

26. From Chapter 7

$$\mu_x = 2(.1) + 3(.2) + 4(.3) + 5(.1) = 3.3$$

27. From Chapter 10

The psychologist's belief implies that, if she's correct, that $\mu_1 > \mu_2$. Hence, the proper alternative is $H_A: \mu_1 - \mu_2 > 0$.

28. From Chapter 4

$$\begin{aligned} z &= \frac{90 - 80}{9} = 1.33 \rightarrow \text{Percentile rank} = 1 - .1332 \\ &= .8667 \text{ (86.67)}. \end{aligned}$$

Because she had to be in the top 15%, she had to be higher than the 85th percentile, so she was invited back.

29. From Chapter 12

I is true. Another common standard is that there can be no empty cells, and at least 80% of the expected counts are greater than 5. II is not correct because you can have 1 degree of freedom (for example, a 2×2 table). III is correct because $df = (4 - 1)(2 - 1) = 3$.

30. From Chapter 5

An *influential point* is a point whose removal will have a marked effect on the slope of the regression line. Because the slope changes from -0.54 to -1.04 , it is an influential point.

31. From Chapter 10

Because it is a one-sided test, $df = 14 - 1 = 13$. For 13 degrees of freedom, .075 lies between tail probability values of .05 and .10. These correspond, for a one-sided test, to t^* values of 1.771 and 1.350.

32. From Chapter 6

Numbers of concern are 1, 2, 3, 4, 5, 6. We ignore the rest. We also ignore repeats. Reading from the left, the first three numbers we encounter for our subjects are 1, 3, and 5. They are in the treatment group, so numbers 2, 4, and 6 are in the control group. That's Betty, Doreen, and Florence. You might be concerned that the three women were selected and that, somehow, that makes the drawing nonrandom. However, drawing their three numbers had exactly the same probability of occurrence as any other group of three numbers from the six.

33. From Chapter 9

If a significance test at level α rejects a null hypothesis ($H_0: \mu = \mu_0$) against a two-sided alternative, then μ_0 will not be contained in a $C = 1 - \alpha$ level confidence interval constructed using the same value of x . Thus, $\alpha = 1 - C$.

34. From Chapter 8

The statement in (e) describes the random variable for a geometric setting. In a binomial setting, the random variable of interest is the number count of successes in the fixed number of trials.

35. From Chapter 7

$$\mu_{x+y} = \mu_x + \mu_y = 3.5 + 2.7 = 6.2$$

$$\sigma_{x+y} = \sqrt{.8^2 + .65^2} = 1.03$$

A necessary condition for this to be true is that the variables X and Y are independent.

36. Because 0 is not in the interval $\langle .45, .80 \rangle$, it is unlikely that the true slope of the regression line is 0 (III is false). This implies a non-zero correlation coefficient and the existence of a linear relationship between the two variables.

37. This is a geometric setting (independent trials, each succeeding or failing with the same probability).

$$P(\text{1st success is on the 8th trial}) = \left(\frac{18}{38}\right)\left(1 - \frac{18}{38}\right)^7 = .0053$$

[On the calculator this can be found as $\text{geompdf}(18/38, 8)$].

38. The choice is made here to treat plots A and B as a block and plots C and D as a block. That way, we are controlling for the possible confounding effects of the river. Hence the answer is (c). If you answered (e), be careful of confusing the treatment variable with the blocking variable.

$$39. z_{\text{Grumpy}} = \frac{38 - 42}{5} = -.8, \quad z_{\text{Dopey}} = \frac{40 - 45}{6.1} = -.82$$

They are both below average, but Grumpy's z score puts him slightly above Dopey. Note that if Grumpy had been 4 points *above* the mean on the first test and Dopey 5 points above the mean on the second, then Dopey would have done slightly *better* than Grumpy.

$$40. s_{\hat{p}} = \sqrt{\frac{(.8)(.2)}{45}} = .0596$$

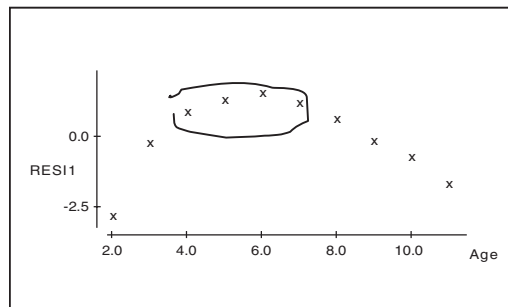
The standard error of \hat{p} for a test of $H_0: p = p_0$ is

$$s_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{N}}$$

If you got an answer of .0645, it means you used the value of \hat{p} rather than the value of p_0 in the formula for $s_{\hat{p}}$.

Solutions to Diagnostic Test—Section II, Part A

1. a. $Height = 76.641 + 6.3661 (Age)$
- b. For each additional year, the height is predicted to increase by 6.36 cm.
- c.



We would expect the residual for 5.5 to be in the same general area as the residuals for 4, 5, 6, and 7 (circled on the graph). The residuals in this area are all positive \rightarrow actual-predicted $> 0 \rightarrow$ actual $>$ predicted. The prediction would probably be too small.

2. a. It is an observational study. The researcher made no attempt to impose a treatment on the subjects in the study. The hired person simply observed and recorded behavior.
- b.
 - The article made no mention of the sample size. Without that you are unable to judge how much sampling variability there might have been. It's possible that the 63–59 split was only attributable to sampling variability.
 - The study was done at *one* Scorebucks, on *one* morning, for a *single* 2-hour period. The population at *that* Scorebucks might differ in some significant way from the patrons at other Scorebucks around the city (and there are many, many of them). It might have been different on a different day or during a different time of the day. A single 2-hour period may not have been enough time to collect sufficient data (we don't know because

the sample size wasn't given) and, again, a 2-hour period in the afternoon might have yielded different results.

- c. You would conduct the study at multiple Scorebucks, possibly blocking by location if you believe that might make a difference (i.e., would a working-class neighborhood have different preferences than the ritziest neighborhood?). You would observe at different times of the day and on different days. You would make sure that the total sample size was large enough to control for sampling variability.

3. From the information given, we have

- $P(\text{hit the first and hit the second}) = (.4)(.7) = .28$
- $P(\text{hit the first and miss the second}) = (.4)(.3) = .12$
- $P(\text{miss the first and hit the second}) = (.6)(.4) = .24$
- $P(\text{miss the first and miss the second}) = (.6)(.6) = .36$

This information can be summarized in the following table:

First shot	Second shot	
	Hit	Miss
Hit	.28	.12
Miss	.24	.36

- a. $P(\text{hit on second shot}) = .28 + .24 = .52$
 b. $P(\text{miss on first/hit on second}) = .24/.52 = 6/13$

4. Let p_1 be the true proportion who planned to vote for Buffy before her remarks. Let p_2 be the true proportion who plan to vote for Buffy after her remarks.

$$H_0: p_1 = p_2$$

$$H_A: p_1 > p_2$$

We want to use a 2-proportion z test for this situation. The problem tells us that the samples are random samples.

$$\hat{p}_1 = \frac{60}{72} = .83, \hat{p}_2 = \frac{56}{80} = .70$$

$$(72)(.83), 72(1 - .83), 80(.70), 80(1 - .70)$$

are all greater than 5, so the conditions for the test are met.

$$\hat{p} = \frac{60 + 56}{72 + 80} = .76, z = \frac{.83 - .70}{\sqrt{(.76)(.24)\left(\frac{1}{72} + \frac{1}{80}\right)}} = \frac{.13}{.069}$$

$$= 1.88 \rightarrow P \text{ value} = .03$$

Because P is very low, we have evidence against the null. We have reason to believe that the level of support for Buffy has declined since her “unfortunate” remarks.

5. The data are paired, so we will use a matched pairs test. Let μ_d = the true mean difference between Twin A and Twin B for identical twins reared apart.

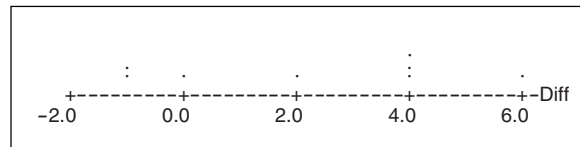
$$H_0: \mu_d = 0$$

$$H_A: \mu_d > 0$$

We want to use a one-sample t -test for this situation. We need the difference scores:

Pair	1	2	3	4	5	6	7	8
Twin A	103	110	90	97	92	107	115	102
Twin B	97	103	91	93	92	105	111	103
d	6	4	-1	4	0	2	4	-1

A dotplot of the difference scores shows no significant departures from normality:



The conditions needed for the one sample t -test are met.

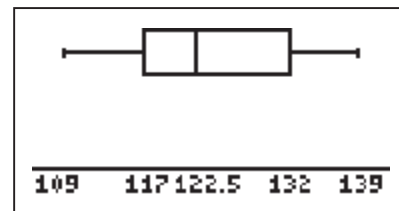
$$\bar{x}_d = 2.25, s = 2.66$$

$$t = \frac{2.25 - 0}{2.66 / \sqrt{8}} = 2.39, df = 8 - 1$$

$$= 7 \rightarrow .02 < P \text{ value} < .025 [.024 \text{ on the TI-83}]$$

Because $P < .05$, reject H_0 . We have evidence that, in identical twins reared apart, that the better educated twin is likely to have the higher IQ score.

6. a. $\bar{x} = 123.85, s = 9.07$. A boxplot of the data shows no significant departures from normality so we are OK to construct a 95% confidence interval.



$$n = 20 \rightarrow df = 19 \rightarrow t^* = 2.093$$

$$123.85 \pm 2.093 \left(\frac{9.07}{\sqrt{20}} \right)$$

$$= 123.85 \pm 2.093(2.03) = < 119.60, 128.10 > .$$

- b. We are 95% confident that the true mean drying time for the paint is between 119.6 minutes and 128.1 minutes. Because 120 minutes is in this interval, we would not consider unusual an average drying time of 120 minutes for the population from which this sample was drawn.

$$c. t = \frac{123.85 - 120}{\frac{9.07}{\sqrt{20}}} = 1.90$$

$$df = 20 - 1 = 19 \rightarrow .05 < P \text{ value} < .10$$

[On the TI-83: .073]

- d. We know that if an α -level significance test rejects (fails to reject) a null hypothesis, then the hypothesized value of μ will not be (will be) in a $C = 1 - \alpha$ confidence interval. In this problem, 120 was not in the $C = .95$ confidence interval and a significance test at $\alpha = .05$ failed to reject the null as expected.
- e. For the one-sided test, $t = 1.90$, $df = 19 \rightarrow .025 < P \text{ value} < .05$
[On the TI-83: .036].

For the two-sided test, we concluded that we did not have evidence to reject the claim of the manufacturer. However, for the one-sided test, we have stronger evidence ($P < .05$) and would conclude that the average drying time is most likely greater than 120 minutes.

SCORING AND INTERPRETATION

Scoring Sheet for Diagnostic Test

Section I: Multiple Choice

$$\left[\frac{\text{Number correct}}{\text{(out of 40)}} - \left(\frac{1}{4} \times \frac{\text{Number wrong}}{\text{Number wrong}} \right) \right] \times 1.25 = \frac{\text{Multiple-Choice Score}}{\text{(if less than zero, enter zero)}} = \frac{\text{Weighted Section I Score}}{\text{Score (Do not round)}}$$

Section II: Free Response

$$\text{Question 1 } \frac{\text{Number correct}}{\text{(out of 4)}} \times 1.875 = \frac{\text{Score}}{\text{(Do not round)}}$$

$$\text{Question 2 } \frac{\text{Number correct}}{\text{(out of 4)}} \times 1.875 = \frac{\text{Score}}{\text{(Do not round)}}$$

$$\text{Question 3 } \frac{\text{Number correct}}{\text{(out of 4)}} \times 1.875 = \frac{\text{Score}}{\text{(Do not round)}}$$

$$\text{Question 4 } \frac{\text{Number correct}}{\text{(out of 4)}} \times 1.875 = \frac{\text{Score}}{\text{(Do not round)}}$$

$$\text{Question 5 } \frac{\text{Number correct}}{\text{(out of 4)}} \times 1.875 = \frac{\text{Score}}{\text{(Do not round)}}$$

$$\text{Question 6 } \frac{\text{Number correct}}{\text{(out of 4)}} \times 3.125 = \frac{\text{Score}}{\text{(Do not round)}}$$

$$\text{Sum} = \frac{\text{Weighted Section II Score}}{\text{(Do not round)}}$$

Composite Score

$$\frac{\text{Weighted Section I Score}}{\text{Score}} + \frac{\text{Weighted Section II Score}}{\text{Score}} = \frac{\text{Composite Score}}{\text{(Round to nearest whole number)}}$$

This page intentionally left blank.

PART III

COMPREHENSIVE REVIEW

This page intentionally left blank.

Chapter 3

Overview of Statistics/ Basic Vocabulary

2 WHAT IS STATISTICS?

Statistics is the study of data. This involves activities such as the collection of data, the organization and analysis of data, and drawing inferences from data. *Statistical methods* and *statistical thinking* can be thought of as using common sense to analyze and draw conclusions from data.

Statistics has been developing as a field of study since the 16th century. Historical figures, some of whom you may have heard of, such as Isaac Newton, Abraham DeMoivre, Carl Gauss, Adolph Quetelet, Florence Nightingale (Yes, Florence Nightingale!), Sir Francis Galton, Karl Pearson, Sir Ronald Fisher, and John Tukey have been major contributors to what we know as the science of statistics today.

Statistics is one of the most practical subjects studied in school. A mathematics teacher may have some trouble justifying the everyday use of algebra for the average citizen, but no statistics teacher ever has that problem. We are bombarded constantly with statistical arguments in the media and, through real-life examples, we can develop the skills to become intelligent consumers of numerical-based knowledge claims.

QUANTITATIVE VERSUS QUALITATIVE DATA

Quantitative data, or **numerical data** are data measured or identified on a numerical scale. **Qualitative data** or **categorical data** are data that can be classified into a group.

Examples of Quantitative (Numerical) Data: The heights of students in an AP Statistics class; the number of freckles on the face of a redhead; the average speed on a busy expressway; the scores on a

final exam; the concentration of DDT in a creek; the daily temperatures in Death Valley; the number of people jailed for marijuana possession each year.

Examples of Qualitative (Categorical) Data: Gender; political party preference; eye color; ethnicity; level of education; socioeconomic level; birth order of a person (first-born, second-born, etc.).

There are times that the distinction between quantitative and qualitative data is somewhat less clear than in the examples above. For example, we could view the variable “family size” as a categorical variable if we were labeling a person based on the size of their family. That is, a woman would go in category “TWO” if she was married but there were no children. Another woman would be in category “FOUR” if she was married and had two children. On the other hand, “family size” would be a quantitative variable if we were observing families and recording the number of people in each family (2, 4, . . .). In situations like this, the context will make it clear whether we are dealing with quantitative or a qualitative data.

Discrete and Continuous Data

Quantitative data can be either **discrete** or **continuous**. **Discrete data** are data that can be listed or placed in order. **Continuous data** can be measured, or take on values in an interval. The number of heads we get on 20 flips of a coin is discrete; the time of day is continuous. We will see more about discrete and continuous data later on.

DESCRIPTIVE VERSUS INFERENCE STATISTICS

Statistics has two main functions: to *describe* data and to make *inferences* from data. **Descriptive statistics** is often referred to as **exploratory data analysis (EDA)**. The components of EDA are *analytical* and *graphical*. When we have collected some **univariate** (one variable) data, we can examine it in a variety of ways: look at measures of center for the distribution (such as the mean and median); look at measures of spread (variance, standard deviation, range, interquartile range); graph it to identify features such as its shape and whether or not it has clusters or gaps (using dotplots, boxplots, histograms, and stemplots).

With bivariate (two variable) data, we look for relationships between variables and ask questions like: “Are these variables related to each other?” Here we consider such analytical ideas as correlation and regression, and graphical techniques such as scatterplots. Chapters 4 and 5 of this book are primarily concerned with exploratory data analysis.

Procedures for collecting data are discussed in Chapter 6. Chapters 7 and 8 are concerned with the probabilistic underpinnings of inference.

Inferential statistics involves using samples to make *inferences* about the population from which the sample was drawn. If we are interested in

the average height of students at a local community college, we could select a random sample of the students and measure their heights. Then we could use the average height of the students in our sample to *estimate* the true average height of the population from which the sample was drawn. In the real world we often are interested in some characteristic of a population (e.g., what percentage of the voting public favors the outlawing of handguns?), but it is too difficult or too expensive to do a census of the entire population. The common technique is to select a *random sample* of the population and, based on an analysis of the sample, make *inferences* about the population from which the sample was drawn. Chapters 9–13 of this book are primarily concerned with inferential statistics.

Parameters versus Statistics

Values that describe a sample are called **statistics**, and values that describe a population are called **parameters**. In *inferential statistics*, we use *statistics* to estimate *parameters*. For example, if we draw a sample of 35 students from a large university and compute their mean GPA (that is, the average grade, usually on a 4-point scale, for each student), we have a *statistic*. If we could somehow compute the mean GPA for *all* students in the university, we would have a *parameter*.

COLLECTING DATA: SURVEYS, EXPERIMENTS, OBSERVATIONAL STUDIES

In the preceding section, we discussed data analysis and inferential statistics. A question not considered in many introductory statistics courses (but considered in detail in AP Statistics) is how the data are collected. Often times we are interested in collecting data in order to make generalizations about a population. One way to do this is to conduct a **survey**. In a well-designed survey, you take a random sample of the population of interest, compute statistics of interest (like the proportion of baseball fans who think Pete Rose should be in the Hall of Fame), and use those to make predictions about the population.

We are often more interested in seeing the reactions of persons or things to certain stimuli. If so, we are likely to conduct an **experiment** or on **observational study**. We discuss the differences between these two types of studies in Chapter 6, but both basically involve collecting comparative data on groups (called **treatment** and **control**) constructed in such a way that the only difference between the groups (we hope) is the focus of the study. Because experiments and observational studies are usually done on volunteers, rather than on random samples from some population of interest (it's been said that most experiments are done on graduate students in psychology), the results of such studies may lack some generalizability to larger populations. Our ability to generalize

involves the degree to which we are convinced that the *only* difference between our groups is the variable we are studying (otherwise some other variable could be producing the responses).

It is extremely important to understand that data must be gathered correctly in order to have analysis and inference be meaningful. You can do all the number crunching you want with bad data, but the results will be meaningless.

In 1936, the magazine, *The Literary Digest*, did a survey of some 10 million people in an effort to predict the winner of the presidential election that year. They predicted that Alf Landon would defeat Franklin Roosevelt by a landslide, and the election turned out just the opposite. *The Literary Digest* had correctly predicted the outcome of the preceding five presidential elections using essentially similar procedures, so this was definitely unexpected. Their problem was in the way they collected the data they based their conclusions on, rather than that they simply didn't have a random sample of the voting population. They had a lot of data (some 2.4 million ballots were returned), but they weren't representative of the voting population. In part because of the fallout from this fiasco, *The Literary Digest* went bankrupt and out of business the following year. If you are wondering why they were wrong this time with essentially the same techniques used in earlier years, understand that 1936 was the heart of the Depression; in earlier years the lists used to select the sample may have been more reflective of the voting public, but in 1936 only the well-to-do, Republicans generally, were in the *Digest's* sample taken from their own subscriber lists, telephone books, etc.

We look more carefully at sources of bias in data collection in Chapter 6, but the point you need to remember as you progress through the next couple of chapters is that conclusions based on data are only meaningful to the extent that the data are representative of the population being studied.

In an experiment or an observational study, the analogous issue to a biased sample in a survey is the danger of treatment and control groups being somehow systematically different. For example, suppose we wish to study the effects of exercise on stress reduction. We let 100 volunteers for the study *decide* if they want to be in the group that exercises or in the group that doesn't. There are many reasons why one group might be systematically different from the other, but the point is that any comparisons between these two groups is confounded by the fact that the two groups could be different in substantive ways.

RANDOM VARIABLES

We consider random variables in some detail in Chapter 7, but it is important at the beginning to understand the role they play in statistics. A **random variable** can be thought of as a numerical outcome of a random phenomenon or an experiment. As an example of a *discrete* random variable, we can toss three fair coins, and let X be the count of heads; we

then note that X can take on the values 0, 1, 2, or 3. An example of a *continuous* random variable might be the number of centimeters a child grows from age 5 to age 6.

An understanding of random variables is what will allow us to use our knowledge of probability (Chapter 7) in statistical inference. Random variables give rise to **probability distributions** (a way of matching outcomes with their probabilities of success), which in turn give rise to our ability to make probabilistic statements about **sampling distributions** (distributions of sample statistics such as means and proportions). This language, in turn, allows us to talk about the probability of a given sample being as different from expected as it is. This ability is the basis for inference. All of this will be examined in detail later in this book, but it's important to remember that random variables are the foundation for inferential statistics.



Exam Tip: There are a number of definitions in this chapter and many more throughout the book (summarized in the Glossary). Although you may not be asked specific definitions on the AP Exam, you are expected to have the working vocabulary needed to understand any statistical situation you might be presented with. In other words, you need to know and understand the vocabulary presented in the course in order to do your best on the AP Exam.

RAPID REVIEW

1. True or False: A study is done in which the data collected are the number of cars a person has owned in his or her lifetime. This is an example of *qualitative* data.
Answer: False. The data is measured on a numerical, not categorical, scale.
2. True or False: The data in the study of question number 1 is *discrete*.
Answer: True. The data is countable (Leroy has owned 8 cars).
3. What are the names given to values that describe *samples* and values that describe *populations*?
Answer: Values that describe samples are called *statistics*, and values that describe populations are called *parameters*.
4. What is a *random variable*?
Answer: A numerical outcome of an experiment or random phenomenon.

5. Why do we need to *sample*?

Answer: Because it is usually too difficult or too expensive to observe every member of the population. Our purpose is to make inferences about the unknown, and probably unknowable, parameters of a population.

6. Why do we need to take care with data collection?

Answer: In order to avoid bias that comes from nonrepresentative samples. It makes no sense to try to predict the outcome of the presidential election if we survey *only* Republicans.

7. Which of the following are examples of *qualitative* data?

- (a) The airline on which a person chooses to book a flight.
- (b) The average number of women in chapters of the Gamma Goo sorority.
- (c) The race (African American, Asian, Hispanic, Pacific Islander, White) of survey respondents.
- (d) The closing Dow Jones average on 50 consecutive market days.
- (e) The number of people earning each possible grade on a statistics test.
- (f) The scores on a given examination.

Answer: a, c, and e are qualitative. f could be either depending on how the data are reported: qualitative if letter grades were given, quantitative if number scores were given.

8. Which of the following are discrete and which are continuous?

- (a) The number of jelly beans in a jar
- (b) The ages of a group of students
- (c) The humidity in Atlanta
- (d) The number of ways to select a committee of three from a group of ten
- (e) The number of people who watched the Super Bowl in 2002
- (f) The lengths of fish caught on a sport fishing trip.

Answer: Discrete: a, d, e
Continuous: b, c, f

Note that (b) could be considered discrete if by age we mean the integer part of the age—a person is considered to be 17, for example, regardless of where they are after their 17th birthday and before their 18th birthday.

9. Which of the following are *statistics* and which are *parameters*?

- (a) The proportion of all voters who will vote Democratic in the next election

- (b) The proportion of voters in a Gallup Poll who say they will vote Democratic in the next election
- (c) The mean score of the home team in all NFL games in 1999
- (d) The proportion of Asian students attending high school in the state of California
- (e) The mean difference score between a randomly selected class taught statistics by a new method and another class taught by an old method
- (f) The speed of a car

Answer: Statistics: b, e, f
Parameters: a, c, d

Chapter 4

Univariate Data Analysis



Main concepts: *shape of a distribution, dotplot, stemplot, histogram, measures of center, measures of spread, 5-number summary, boxplots, z-scores, density curves, normal distribution, the empirical rule*

Often we collect data on a single variable such as height, weight, IQ, miles driven to work, eye color (that one, of course, is *qualitative*), grade point average, braking distance, etc. Our goal in this chapter is to discuss *exploratory data analysis*—looking at data to see what is in it. The approaches are both graphical and analytical. When asked to describe data, three words define what we are looking for: **shape**, **center**, and **spread**. The first three sections of this chapter will cover these three in detail.



Exam Tip: If you are given an instruction to “describe” a set of data, be sure you discuss the *shape* of the data (including gaps and clusters in the data), the *center* of the data (mean, median, mode), and the *spread* of the data (range, interquartile range, standard deviation).

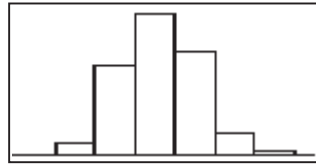
GRAPHICAL ANALYSIS

Our purpose in drawing some sort of graph of data is to get a visual sense of it. We are interested in the **shape** of the data as well as **gaps** in the data, **clusters** of datapoints, and **outliers** (which are datapoints that lie well outside of the general pattern of the data).

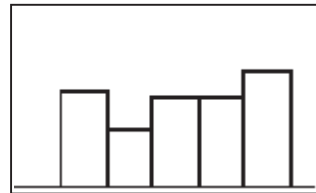
Shape

When we describe **shape**, what we are primarily interested in is the extent to which the graph appears to be **symmetric** (has symmetry around some

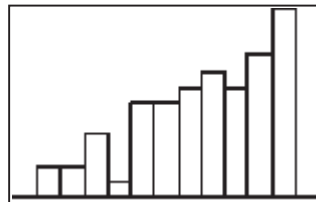
axis), **mound-shaped (bell-shaped)**, **skewed** (data are skewed to the left if the tail is to the left; to the right if the tail is to the right), **bimodal** (has more than one location with many scores), or **uniform** (frequencies of the various values are more-or-less constant).



This graph could be described as *symmetric* and *mound-shaped* (or *bell-shaped*). Note that it doesn't have to be *perfectly* symmetrical to be classified as symmetric.



This graph is of a *uniform* distribution. Again, note that it does not have to be perfectly uniform to be described as *uniform*.



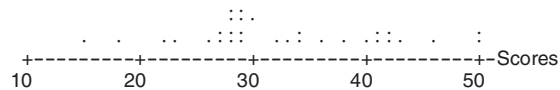
This distribution is **skewed left** because the tail is to the left. If the tail were to the right, the graph would be described as **skewed right**.

There are three types of graph we want to look at in order to help us understand the shape of a distribution (well, actually four, but we don't see the fourth until Section 4.4): dotplot, stemplot, and histogram. We use the following 31 scores from a 50-point quiz in a community college statistics class to illustrate the three plots:

28 38 42 33 29 28 41 40 15 36 27 34 22
 23 28 50 42 46 28 27 43 29 50 29 32 34
 27 26 27 41 18

Dotplot

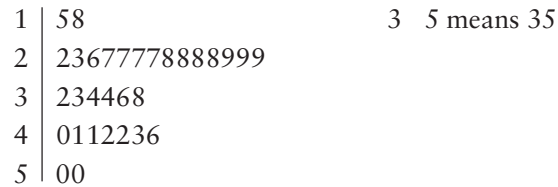
A **dotplot** is a very simple type of graph that involves plotting the data values, with dots, above the corresponding values on a number line. A dotplot of the scores on the statistics quiz, drawn by a statistics computer package, looks like this:



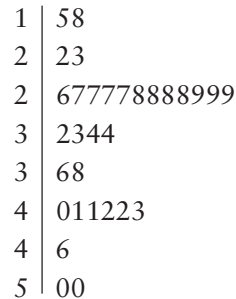
[Calculator note: Most calculators do not have a built-in function for drawing dotplots.]

Stemplot (Stem and Leaf Plot)

A **stemplot** is a bit more complicated than a dotplot. Each data value has a *stem* and a *leaf*. There are no mathematical rules for what constitutes the *stem* and what constitutes the *leaf*. Rather, the nature of the data will suggest reasonable choices for the stem and leaves. With the given score data, we might choose the first digit to be the *stem* and the second digit to be the *leaf*. So, the number 42, in a stem and leaf plot would show up as 4 | 2. All the leaves for a common stem are often on the same line. Often, these are listed in increasing order, so the line with stem 4 could be written as: 4 | 0112236. The complete stemplot of the quiz data looks like this:



Some data lend themselves to breaking the stem into two or more parts. For these data, the stem “4” could be shown with leaves broken up 0–4 and 5–9. Done this way, the stemplot for the scores data would look like this (there is a single “1” because there are no leaves with the values 0–4 for a stem of 1):



The visual image is of data that is slightly skewed to the right (that is, toward the higher scores). We do notice a *cluster* of scores in the high 20s that was not obvious when we used an increment of 10 rather than 5. There is no hard and fast rule about how to break up the stems—it’s easy to try different arrangements on most computer packages.

Sometimes plotting more than one stemplot, side-by-side or back-to-back, can provide us with comparative information. The following stemplot shows the results of two quizzes given for this class (one of them the one discussed above):

Stem-and-leaf of Scores 1

1		33
2		0567799
3		1125
4		0235778889999
5		00000

Stem-and-leaf of Scores 2

1		58
2		23677778888999
3		234468
4		0112236
5		00

Or, as a back-to-back stemplot:

	Quiz 1		Quiz 2
	33		58
	9977650		3677778888999
	5211		234468
	999888775320		0112236
	00000		00

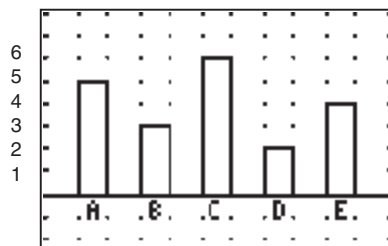
It can be seen from this comparison that the scores on Quiz #1 (on the left) were generally higher than for those on Quiz #2—there are a lot more scores at the upper end. Although both distributions are reasonably symmetric, the one on the left is skewed somewhat toward the smaller scores, and the one on the right is skewed somewhat toward the larger numbers.

[Calculator note: Most calculators do not have a built-in function for drawing stemplots.]

Histogram

A **bar graph** is used to illustrate qualitative data, and a **histogram** is used to illustrate quantitative data. The horizontal axis in a *bar graph* contains the categories, and the vertical axis contains the frequencies, or relative frequencies, of each category. The horizontal axis in a *histogram* contains numerical values, and the vertical axis contains the frequencies, or relative frequencies, of the values (often intervals of values).

example: Twenty people were asked to state their preferences for candidates in an upcoming election. The candidates were Arnold, Betty, Chuck, Dee, and Edward. Five preferred Arnold, three preferred Betty, six preferred Chuck, two preferred Dee, and four preferred Edward. A bar graph of their preferences is shown below:

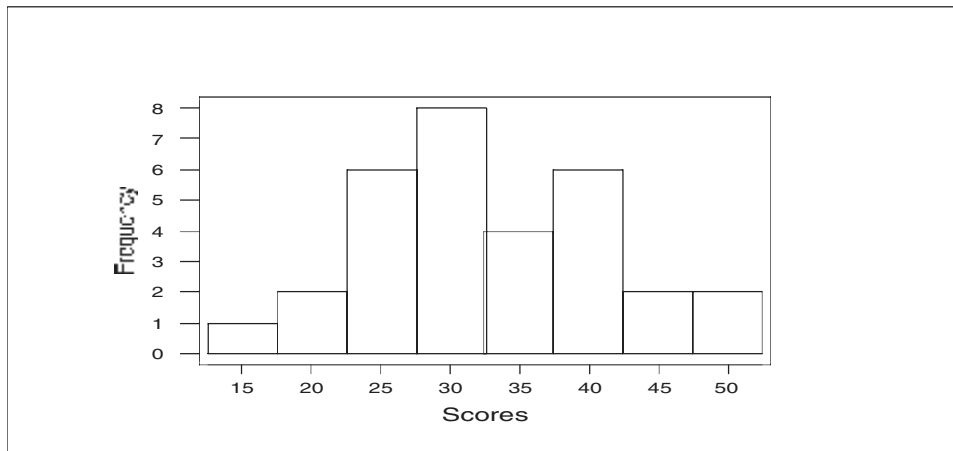
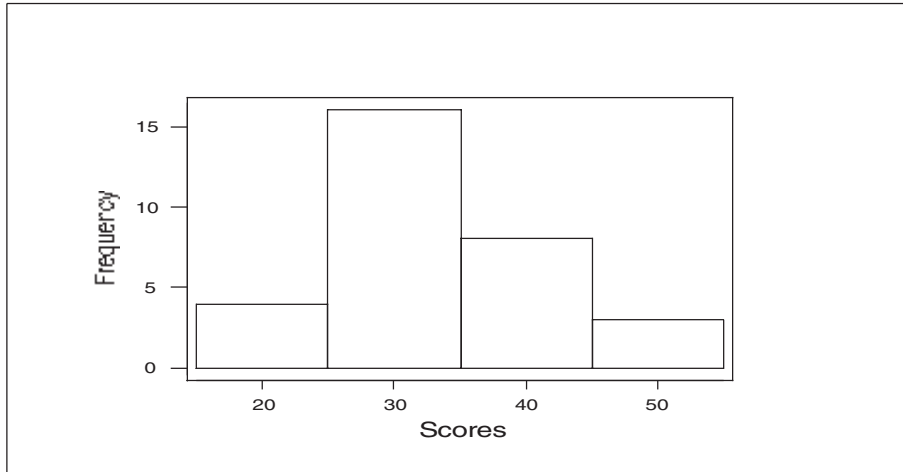


A *histogram* is composed of bars of equal width, usually with common edges. When you choose the intervals, be sure that each of the datapoints fit into a category (it must be clear in which class any given value is). A histogram is much like a stemplot that has been rotated 90 degrees.

Consider again the quiz scores we looked at when we discussed dotplots:

28 38 42 33 29 28 41 40 15 36 27 34 22
 23 28 50 42 46 28 27 43 29 50 29 32 34
 27 26 27 41 18

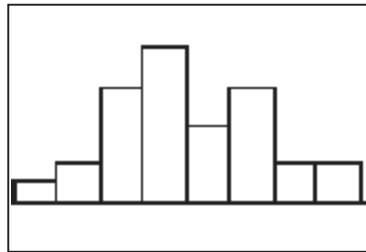
Because the data are integral and range from 15 to 50, reasonable intervals might be of size 10 or of size 5. The graphs below show what happens with the two choices:



Typically, the interval with midpoint 15 would have class boundaries $12.5 < x < 17.5$; the interval with midpoint 20 would have class boundaries $17.5 < x < 22.5$, etc.

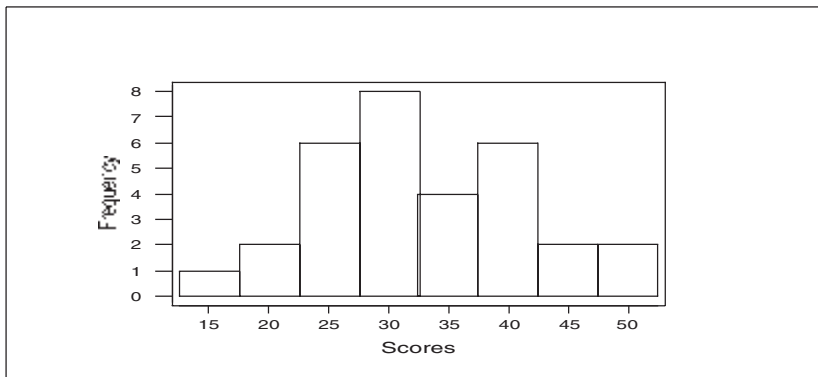
There are no hard and fast rules for how wide to make the bars (called “class intervals”). You should use computer or calculator software to help you find a picture to which your eye reacts. In this case, intervals of size 5 give us a better sense of the data.

The following is the histogram of the same data (with bar width 5) produced by the TI-83 calculator:



Calculator Tip: If you are using your calculator to draw a histogram, the $Xscl$ will be the width of the class interval. The graph above has $Xscl = 5$ and, therefore, a class interval of 5. The $Xmin$ and $Xmax$ were set at 12.5 and 53.5, respectively, so that the middle of each bar would be an integral value.

example: For the histogram below, identify the boundaries for the class intervals.



solution: The midpoints of the intervals begin at 15 and go by 5s. So the boundaries of each interval are 2.5 above and below the midpoints. Hence, the boundaries for the class intervals are 12.5, 17.5, 22.5, 27.5, 32.5, 37.5, 42.5, 47.5, and 52.5

example: For the histogram given in the previous example, what proportion of the scores are less than 27.5?

solution: From the graph we see that there is 1 value in the first bar, 2 in the second, 6 in the third, etc., for a total of 31 altogether. Of these, $1 + 2 + 6 = 9$ are less than 27.5. $9/31 = 0.29$.

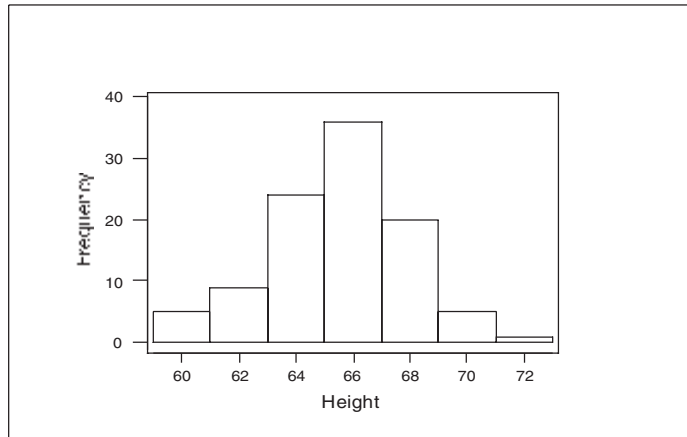
example: The following are the heights of 100 college-age women, followed by a histogram and stemplot of the data. Describe the graph of the data using either the histogram, the stemplot, or both.

Height

63	65	68	62	63	67	66	64	68	64	65	65	67
66	66	65	65	66	65	66	63	64	70	65	66	62
64	65	67	66	62	66	65	68	61	66	63	67	65
63	67	66	61	66	66	61	67	65	63	69	63	65
62	68	63	59	67	62	70	63	69	66	65	66	67
65	63	67	66	60	66	72	67	67	66	68	64	68
60	61	64	65	64	60	69	63	64	65	66	67	64
63	63	68	67	66	65	60	63	63				

This stemplot breaks the heights into increments of 2 inches:

5		9
6		00001111
6		22222333333333333333
6		44444444445555555555555555
6		66666666666666666666777777777777
6		8888888999
7		00
7		2



solution: Both the stemplot and the histogram show a symmetric, bell-shaped distribution centered around a height of 66". The class boundaries are 59, 61, 63, etc. The interval with midpoint 60 would contain heights $59 \leq x < 61$.

MEASURES OF CENTER

In the last example of the previous section, we said that the graph appeared to be *centered* about a height of 66". In this section, we talk about ways to describe the *center* of a distribution. There are two primary measures of center: the **mean** and the **median**. There is a third measure, the **mode**, but it tells where the most frequent values occur, more than it describes the center. In some distributions, the mean, median, and mode will be close in value, but the mode can appear at any point in the distribution.

Mean

Let x be a variable that represents any value in a dataset of n values. The **mean** of the set is the sum of all the x s divided by the sample size n . Symbolically,

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$

Usually, the indices on

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

will be implied, and we simply write

$$\sum_{i=1}^n x_i$$

as

$$\sum x.$$

$\sum x$, pronounced aloud as, "sum of x ," simply means to add up all the x s (think of it as the "add-'em-up" symbol). Sometimes we have access to the entire population of data, we would use the symbol μ for the mean rather than \bar{x} .

(*Note:* in the previous chapter, we made a distinction between *statistics*, which are values that describe sample data, and *parameters*, which

are values that describe populations. Unless we are clear that we have access to an entire population, or that we are discussing a distribution, we use the symbols for *statistics* rather than *parameters*.)

example: During his major league career, Babe Ruth hit the following number of home runs (1914–1935): 0, 4, 3, 2, 11, 29, 54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22, 6. What was the *mean* number of home runs per year for his major league career?

$$\begin{aligned}\text{solution: } \bar{x} &= \frac{\sum x}{n} = \frac{0 + 4 + 3 + 2 + \cdots + 22 + 6}{22} \\ &= \frac{714}{22} = 32.45\end{aligned}$$



Calculator Tip: You should use a calculator to do examples such as the above—it’s a waste of time to do it by hand, or even to use a calculator to add up the numbers and divide by n . To use the TI-83: press *STAT*; select *EDIT*; enter the data in a list, say *L1* (you can clear the list, if needed, by moving the cursor on top of the *L1*, pressing *CLEAR* and *ENTER*). Once the data are in *L1*, press *STAT*, select *CALC*, select *1-Var Stats* and press *ENTER*. *1-Var Stats* will appear on the home screen followed by a blinking cursor—the calculator wants to know where your data are. Enter *L1* (It’s above the 1; enter 2nd 1 to get it). The calculator will return \bar{x} and a lot more.

Median

The **median** of a ordered dataset is the “middle” value in the set. If the dataset has an odd number of values, the *median* is a member of the set and is the middle value. If there are 3 values, the median is the second value. If there are 5, it is the third, etc. If the dataset has an even number of values, the median is the mean of the two middle numbers. If there are 4 values, the median is the mean of the 2nd and 3rd values. In general, if there are n values in the ordered dataset, the *median* is at the

$$\frac{n + 1}{2}$$

position. If you have 28 terms in order, you will find the median at the

$$\frac{28 + 1}{2} = 14.5\text{th}$$

position (that is, between the 14th and 15th terms).

example: Consider once again, the data from Babe Ruth’s career. What was the *median* number of home runs per year he hit during his major league career?

solution: First, put the numbers in order from smallest to largest: 0, 2, 3, 4, 6, 11, 22, 25, 29, 34, 35, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60. There are 22 scores, so the *median* is found at the 11.5th position, between the 11th and 12th scores (35 and 41). So the *median* is

$$\frac{35 + 41}{2} = 38.$$

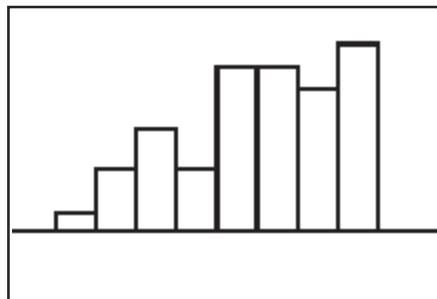
The 1-Var Stats procedure described above will, if you scroll down to the second screen of output, give you the median.

Resistant

Although the mean and median are both measures of center, the choice of which to use depends on the shape of the distribution. If the distribution is symmetric and mound shaped, the mean and median will be close. However, if the distribution has outliers or is strongly skewed, the median is probably the better choice to describe the center. This is because it is a **resistant statistic** (it’s numerical value is not dramatically affected by extreme values), while the mean is not resistant.

example: A group of five teachers in a small school have salaries of \$32,700, \$32,700, \$38,500, \$41,600, and \$44,500. The mean and median salaries for these teachers are \$38,160, and \$38,500, respectively. Suppose the highest paid teacher gets sick, and the school superintendent volunteers to substitute for her. The superintendent’s salary is \$174,300. If you replace the \$44,500 salary with the \$174,300, the median doesn’t change at all (it’s still \$38,500), but the new mean is \$64,120—almost everybody is below average, if by “average,” you mean *mean*. It’s sort of like Lake Woebegone, where all of the children are expected to be above average.

example: For the graph given below, would you expect the mean or median to be larger? Why?



solution: You would expect the median to be larger than the mean. Because the graph is skewed to the left, and the mean is not resistant, you would expect the mean to be pulled to the left (in fact, the data this graph was drawn from have a mean of 5.4 and a median of 6, as expected).

MEASURES OF SPREAD

Simply knowing about the center of a distribution doesn't tell you all you might want to know about the distribution. One group of 20 people earning \$20,000 each will have the same mean and median as a group of 20 where 10 people earn \$10,000 and 10 people earn \$30,000. These two sets of 20 numbers differ not in terms of their center but in terms of their spread, or variability. Just as there were measures of center based on the mean and the median, we also have measures of spread based on the mean and the median.

Variance and Standard Deviation

One measure of spread based on the mean is the **variance**. By definition, the variance is the average squared deviation from the mean. That is, it is a measure of spread because the more distant a value is from the mean, the larger will be the square of the difference between it and the mean.

Symbolically, the variance is defined by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Note that we average by dividing by $n-1$ rather than n as you might expect. This is because there are only $n-1$ independent datapoints, not n , if you know \bar{x} . That is, if you know $n-1$ of the values and you also know \bar{x} , then the n th datapoint is determined.

The problem with using the variance as a measure of spread is that the units for the variance won't match the units of the original data because each difference is squared. For example, if you find the variance of a set of measurements made in inches, the variance will be in square inches. To correct this, we often take the square root of the variance as our measure of spread.

The square root of the variance is known as the **standard deviation**. Symbolically,

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

As discussed earlier, you can leave off the indices and simplify this to

$$s = \sqrt{\frac{1}{n-1} \sum (x - \bar{x})^2}.$$

In practice, you will rarely have to do this calculation by hand because it is one of the values returned when you use your calculator to do 1-Var Stats on a list (it's the Sx near the bottom of the first screen).



Calculator Tip: When you use 1-Var Stats, the calculator will, in addition to Sx , return σx , which is the standard deviation of a distribution. Its formal definition is

$$\sigma = \sqrt{\frac{1}{n} \sum (x - \mu)^2}.$$

This assumes you know μ , the population mean, which you rarely do in practice unless you are dealing with a probability distribution (Chapter 7).

The *standard deviation* has three useful properties when it comes to describing the spread of a distribution:

- *It is independent of the mean.* Because it depends on how far data-points are from the mean, it doesn't matter where the mean is.
- *It is sensitive to the spread.* The more spread we see, the larger will be the standard deviation. For two datasets with the same mean, the one with the larger standard deviation has more variability.
- *It is independent of n .* Because we are averaging squared distances from the mean, the standard deviation will not get larger just because we add more terms.

example: Find the standard deviation of the following 6 numbers: 3, 4, 6, 6, 7, 10.

$$\begin{aligned} \text{solution: } \bar{x} = 6 \rightarrow s &= \sqrt{\frac{(3-6)^2 + (4-6)^2 + (6-6)^2 + (6-6)^2 + (7-6)^2 + (10-6)^2}{6-1}} \\ &= 2.449 \end{aligned}$$

Note that the standard deviation, like the mean, is not *resistant* to extreme values. Because it depends upon distances from the mean, it should be clear that extreme values will have a major impact on the numerical value of the standard deviation.

Interquartile Range

Although the standard deviation works well in situations where the mean works well (reasonably symmetric distributions), we need a measure of spread that works well when we prefer to use the median.

Remember that the median of a distribution divides the distribution in two—it is the middle of the distribution. The medians of the upper and lower halves of the distribution are called **quartiles**. The median of the lower half is called the **lower quartile** or the **first quartile** (Q1 on the calculator). The median of the upper half is called the **upper quartile** or the **third quartile** (Q3 on the calculator). The median could be thought of as the second quartile or Q2 (although we usually don't).

The **interquartile range (IQR)** is the difference between Q3 and Q1. That is, $IQR = Q3 - Q1$. When you do *1-Var Stats*, the calculator will return Q1 and Q3 along with a lot of other stuff. You have to compute the IQR from Q1 and Q3. Note that the IQR comprises the middle 50% of the data.

example: Find Q1 and Q3 for the following dataset: 5, 5, 6, 7, 8, 9, 11, 13, 17.

solution: Because the data are in order, and there is an odd number of values (9), the median is 8. The bottom half of the data comprises 5, 5, 6, 7. The median of the bottom half is the average of 5 and 6, or 5.5 which is Q1. Similarly, Q3 is the medians of the top half, which is the mean of 11 and 13, or 12.

example: Find the standard deviation and IQR for the number of home runs hit by Babe Ruth in his major league career. The number of home runs was: 0, 4, 3, 2, 11, 29, 54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22, 6.

solution: We put these numbers in a list and do *1-Var Stats*. The calculator gives us $Sx = 20.21$, $Q1 = 11$, and $Q3 = 47$. Hence the $IQR = Q3 - Q1 = 47 - 11 = 36$.

The **range** of the distribution is the difference between the maximum and minimum scores in the distribution. For the home run data, the range equals $60 - 0 = 60$. Although this is sometimes used as a measure of spread, it is not very useful for that purpose because we are usually interested in how the data spreads out from the center of the distribution, not in just how far it is from the minimum to the maximum values.

Outliers

We have a pretty good intuitive sense of what an *outlier* is: it's a value far removed from the others. There is no rigorous mathematical formula for determining whether or not something is an outlier, but there are a few conventions that people seem to agree on. Not surprisingly, some of them are based on the mean and some are based on the median!

A commonly agreed upon way to think of outliers based on the mean is to consider how many standard deviations away from the mean a term is. Some texts define an **outlier** as a datapoint that is more than 2 or 3 standard deviations from the mean.

Most texts now define outliers in terms of how far a datapoint is above or below the quartiles in a distribution. To find if a distribution has any outliers, do the following (this is known as the “1.5IQR Rule”):

- find the IQR
- multiply the IQR by 1.5
- Find $Q1 - 1.5(IQR)$ and $Q3 + 1.5(IQR)$
- Any values below $Q1 - 1.5(IQR)$ or above $Q3 + 1.5(IQR)$ is an **outlier**.

Some texts call an outlier defined as above a *mild* outlier. An *extreme* outlier is then one that lies more than 3 IQRs beyond $Q1$ or $Q3$.

example: The following data represent the amount of money, in British pounds, spent weekly on tobacco for 11 regions in Britain: 4.03, 3.76, 3.77, 3.34, 3.47, 2.92, 3.20, 2.71, 3.53, 4.51, 4.56. Do any of the regions seem to be spending a lot more or less than the other regions? That is, are there any outliers in the data?

solution: Using a calculator, we find $\bar{x} = 3.62$, $Sx = s = .59$, $Q1 = 3.2$, $Q3 = 4.03$.

- (using means): $3.62 \pm 2(.59) = \langle 2.44, 4.8 \rangle$. Because there are no values in the data less than 2.44 nor greater than 4.8, there are no outliers using this method.
- (using the 1.5IQR Rule): $Q1 - 1.5(IQR) = 3.2 - 1.5(4.03 - 3.2) = 1.96$, $Q3 + 1.5(IQR) = 4.03 + 1.5(4.03 - 3.2) = 5.28$. Because there are no values in the data less than 1.96 nor greater than 5.28, there are no outliers by this method either.

Outliers are important because they will often tell us that something unusual or unexpected is going on with the data that we need to know about. A manufacturing process that produces products so far out of spec that they are outliers indicates something is wrong with the process. Sometimes outliers are just a natural, but rare, variation. Sometimes, however, an outlier can indicate that the process generating the data is out of control in some fashion.

POSITION OF A TERM IN A DISTRIBUTION

Up until now, we have concentrated on the nature of a distribution as a whole. We have been concerned with the shape, center, and spread of the entire distribution. Now we look briefly at individual terms in the distribution.

5-Number Summary

There are positions in a dataset that give us valuable information about the dataset. The **5-number summary** of a dataset is composed of the

minimum value, the lower quartile, the median, the upper quartile, and the maximum value. On your calculator, these are reported out as *minx*, *Q1*, *Med*, *Q3*, *maxX* when you do 1-Var Stats on a list.

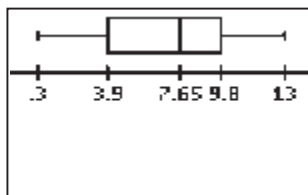
example: The following data are standard of living indices for 20 cities: 2.8, 3.9, 4.6, 5.3, 10.2, 9.8, 7.7, 13, 2.1, .3, 9.8, 5.3, 9.8, 2.7, 3.9, 7.7, 7.6, 10.1, 8.4, 8.3. Find the 5-number summary for the data.

solution: Putting the 20 values into the calculator and doing 1-Var Stats, we find: $\text{min}X = .3$, $Q1 = 3.9$, $\text{Med} = 7.65$, $Q3 = 9.8$, $\text{max}X = 13$.

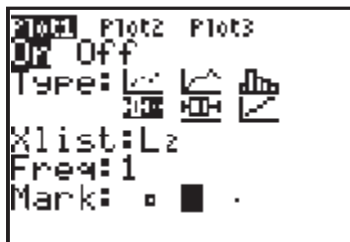
Boxplots (Outliers Revisited)

In section 4.1, we discussed three graphs of a distribution (dotplot, stemplot, and histogram). Using the 5-number summary, we can add a fourth type of univariate graph to this group: the **boxplot**. A *boxplot* is simply a graphical version of the 5-number summary. A box is drawn that contains the middle 50% of the data (from $Q1$ to $Q3$) and “whiskers” extend from the lines at the ends of the box (the lower and upper quartiles) to the minimum and maximum values. The *boxplot* is sometimes referred to as a *box and whisker plot*.

example: Consider again the data from the previous example: 2.8, 3.9, 4.6, 5.3, 10.2, 9.8, 7.7, 13, 2.1, .3, 9.8, 5.3, 9.8, 2.7, 3.9, 7.7, 7.6, 10.1, 8.4, 8.3. A boxplot of this data, done on the TI-83, looks like this (the 5-number summary was [0.3, 3.9, 7.65, 9.8, 13]):

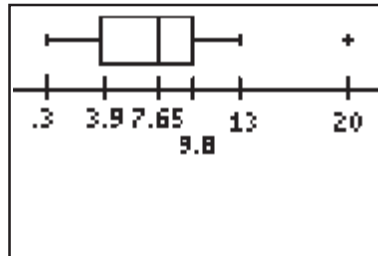


Calculator Tip: To get the graph above on your calculator, go to the STAT PLOTS menu, open one of the plots, say Plot1, and you will see a screen something like this



Note that there are two boxplots available. The one that is highlighted is the one that will show outliers. The calculator determines outliers by the 1.5(IQR) rule.

example: Using the same dataset as the previous problem, but replacing the 10.2 with 20, which would be an outlier in this dataset, we now get the following graph on the calculator:



Percentile Rank of a Term

The **percentile rank** of a term in a distribution equals the proportion of terms in the distribution that that term is greater than. A term that is at the 75th percentile is larger than 75% of the terms in a distribution. If we know the 5-number summary for a set of data, then Q1 is at the 25th percentile, the median is at the 50th percentile, and Q3 is at the 75th percentile.

z Scores

One way to identify the position of a term in a distribution is to note how many standard deviations the term is above or below the mean. The statistic that does this is the **z score**:

$$z_{x_i} = \frac{x_i - \bar{x}}{s}$$

The z score is positive when x is above the mean and negative when it is below the mean.

example: $z_3 = 1.5$ tells us that the value 3 is 1.5 standard deviations above the mean. $z_3 = -2$ tells us that 3 is 2 standard deviations below the mean.

example: For the first test of the year, Harvey got a 68. The class average (mean) was 73, and the standard deviation was 3. What was Harvey's z score on this test?

solution: $z_{68} = \frac{68 - 73}{3} = -1.67$

Suppose we have a distribution with mean \bar{x} and standard deviation s . If we subtract \bar{x} from every term in the distribution, it can be shown that the new distribution will have a mean of $\bar{x} - \bar{x} = 0$. If we divide every term by s , then the new distribution will have a standard deviation of

$s/s = 1$. Conclusion: If you compute the z score for every term in a distribution, the distribution of z scores will have a mean of 0 and a standard deviation of 1.

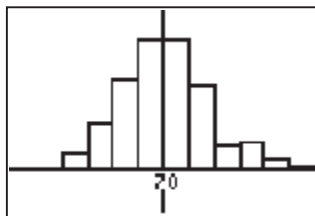


Calculator Tip: We have used *1-Var Stats* a number of times so far. Each of the statistics generated by that command are stored as variables in the VARS menu. To find, say, \bar{x} , after having done *1-Var Stats* on $L1$, press VARS and scroll down to #5, which is STATISTICS. If you enter the STATISTICS menu, you will see several lists. \bar{x} is in the XY column (as is s). Scroll through the other menus to see what they contain.

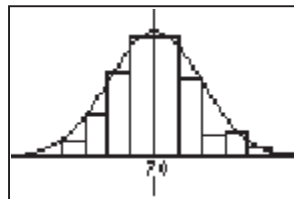
To demonstrate the fact in the previous paragraph, do *1-Var Stats* on, say, $L1$. Then move the cursor to the top of $L2$ and enter $(L2 - \bar{x})/s$ – getting the \bar{x} and s from the VARS menu. This will give you the z -scores for every value in $L1$. Now do *1-Var Stats* $L2$ to verify that $\bar{x} = 0$ and $s = 1$.

NORMAL DISTRIBUTION

We have been discussing characteristics of distributions (shape, center, spread) and of the individual terms (percentiles, z scores) that make up those distributions. Certain distributions have particular interest for us in statistics, in particular those that are known to be symmetric and bell shaped. The following histogram represents the heights of 100 males whose average height is 70" and whose standard deviation is 3".



This is clearly approximately symmetric and mound shaped or bell shaped. We are going to model this with a curve that idealizes what we see in this sample of 100. That is, we will model this with a continuous curve that “describes” the shape of the distribution for very large samples. That curve is the graph of the **normal distribution**. A *normal curve*, when superimposed on the above histogram looks like this.



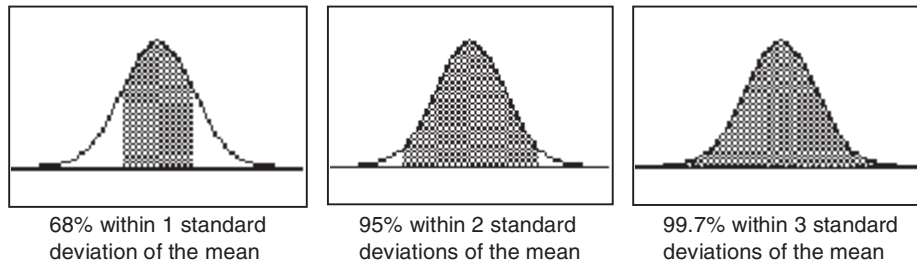
The function that yields the *normal curve* is defined *completely* in terms of its mean and standard deviation. Although you are not required to know it, you might be interested in the function that defines the normal curve:

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

One consequence of this definition is that the total area under the curve is 1. This fact will be of great use to us later when we consider areas under the normal curve as probabilities.

Empirical Rule

The **empirical rule**, or the **68-95-99.7 rule**, states that *approximately* 68% of the terms in a normal distribution are within one standard deviation of the mean, 95% are within two standard deviation of the mean, and 99.7% are within three standard deviations of the mean. The following three graphs illustrate the *empirical rule*.



Standard Normal Distribution

Because we are dealing with a theoretical distribution, we will use μ and σ , rather than \bar{x} and s when referring to the normal curve. If X is a variable that has a normal distribution with mean μ and standard deviation σ (we say “ X has $N(\mu, \sigma)$ ”), There is a related distribution we obtain by **standardizing** the data in the distribution to produce the **standard normal distribution**. To do this, we convert the data to a set of z scores, using the formula

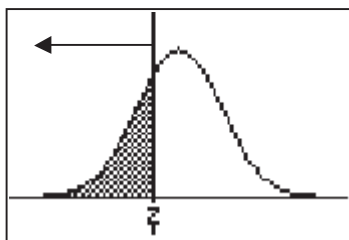
$$z = \frac{x - \mu}{\sigma}.$$

The algebraic effect of this, as we saw in an earlier section, is to produce a normal distribution of z scores with mean 0 and standard deviation 1. We say z has $N(0,1)$. This simplifies the defining density function to

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

For the standardized normal curve, the *empirical rule* says that approximately 68% of the terms lie between $z = 1$ and $z = -1$, 95% between $z = -2$ and $z = 2$, and 99.7% between $z = -3$ and $z = 3$. (Trivia for calculus students: 1 standard deviation from the mean is a *point of inflection*.)

Because many naturally occurring distributions are approximately normal (heights, SAT scores, for example), we are often interested in knowing what proportion of terms lie in a given interval under the normal curve. Problems of this sort can be solved either by use of a calculator or a table of *Standard Normal Probabilities*. In a typical table, the marginal entries are z scores, and the table entries are the areas under the curve to the left of a given z score. All statistics texts have such tables.



The table entry for z is the area to the left of z and under the curve.

example: What proportion of the area under a normal curve lies to the left of $z = -1.37$?

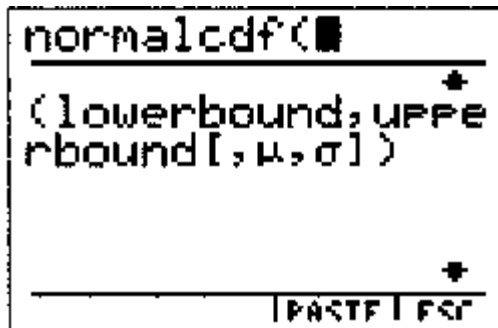
solution: There are two ways to do this problem, and you should be able to do it either way.

- (i) The first way is to use the table of *Standard Normal Probabilities*. To read the table, move down the left column (titled “ z ”) until you come to the row whose entry is -1.3 . The third digit, the 0.07 part, is found by reading across the top row until you come to the column whose entry is 0.07 . The entry at the intersection of the row containing -1.3 and the column containing 0.07 is the area under the curve to the left of $z = -1.37$. That value is 0.0853 .
- (ii) The second way is to use your calculator. Under the DISTR menu, the second entry is “*normalcdf*.” The syntax for a standard normal distribution is *normalcdf(lower bound, upper bound)*. In the present case, the lower bound can be any large negative number. $normalcdf(-100, -1.37) = 0.0853435081$. As you can see, the table entries have been rounded to four decimal places.



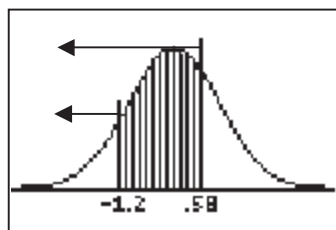
Calculator Tip: It can be difficult to remember the parameters that go with the various functions on your calculator—knowing, for example, that for *normalcdf*, you put *lower bound*, *upper bound*, *mean*, *standard deviation* in the parentheses. The APP “CtlgHelp” can remember for you. It comes on the APP menu of the TI-83 + SE and can be down-

loaded from the TI Web Site for some other TI's. The following is a screen capture of using CtlgHelp for *normalcdf*(:



example: What proportion of the area under a normal curve lies between $z = -1.2$ and $z = .58$

solution: (i) Reading from the table, the area to the left of $z = -1.2$ is 0.1151, and the area to the left of $z = .58$ is 0.7190. The geometry of the situation (see below) tells us that the area between the two values is $0.7190 - 0.1151 = 0.6039$



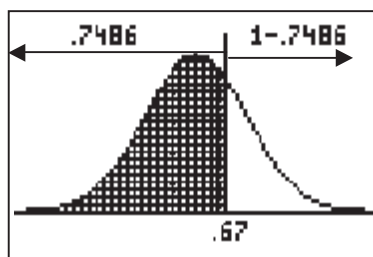
(ii) Using the calculator, we have $normalcdf(-1.2, .58) = 0.603973005$. Round to 0.6040 (difference from the answer in part (i) caused by rounding).

example: In an earlier example, we saw that heights of men are approximately normally distributed with a mean of 70 and a standard deviation of 3. What proportion of men are more than 6' (72") tall? Be sure to include a sketch of the situation.

solution: (i) Another way to state this is to ask what proportion of terms in a normal distribution with mean 70 and standard deviation 3 are greater than 72? In order to use the table of Standard Normal Probabilities, we must first convert to z -scores. The z -score corresponding to a height of 72" is

$$z = \frac{72 - 70}{3} = 0.67$$

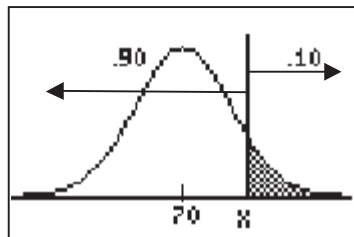
The area to the left of $z = 0.67$ is 0.7486. However, we want the area to the *right* of 0.67, and that is $1 - 0.7486 = 0.2514$.



(ii) Using the calculator, we have $normalcdf(.67, 100) = .2514$. We could use the calculator to get the answer from the raw data as follows: $normalcdf(72, 1000, 70, 3) = .2525$. Simply add the mean and standard deviation of a nonstandard normal curve to the list of parameters for $normalcdf$.

example: For the population of men in the previous example, how tall must a man be to be in the top 10% of all men in terms of height?

solution:



(i) We are looking for the value of x in the drawing. Look through the *Standard Normal Table* to find the nearest table entry equal to 0.90 (because we know an area, we need to read the table from the inside out to the margins). It is 0.8997 and corresponds to a z -score of 1.28.

So $z_x = 1.28$. But also,

$$z_x = \frac{x - 70}{3}.$$

So,

$$z_x = \frac{x - 70}{3} = 1.28 \rightarrow x = 73.84.$$

A man would have to be at least 73.84" tall to be in the top 10% of all men.

(ii) Using the calculator, the solution is given by $invNorm(0.90, 70, 0.3)$

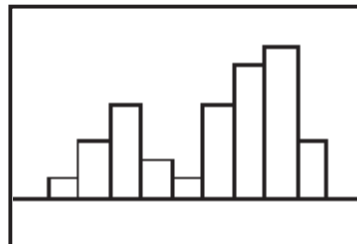


Calculator Tip: $invNorm$ essentially reverses $normalcdf$. That is, rather than reading from the margins in, it reads from the table out (as in the example above) $invNorm(A)$ gives the z -score that corresponds to an area equal to A lying to the left of z . $invNorm(A, \mu, \sigma)$ gives the value of x that has an area of A to the left of x if x has $N(\mu, \sigma)$.

RAPID REVIEW

1. Describe the *shape* of the histogram at the right:

Answer: Bi-modal, somewhat skewed to the left.



2. For the graph of problem #1, would you expect the mean to be larger than the median or the median to be larger than the mean? Why?

Answer: The graph is slightly skewed to the left, so we would expect the mean, which is not resistant, to be pulled slightly in that direction. Hence, we might expect to have the median be larger than the mean.

3. The first quartile (Q1) of a dataset is 12 and the third quartile (Q3) is 18. What is the smallest value above Q3 in the dataset that could possibly be an outlier.

Answer: Outliers lie more than 1.5 IQRs below Q1 or above Q3. $Q3 + 1.5(IQR) = 18 + 1.5(18 - 12) = 27$. Any value greater than 27 would be an outlier.

4. A distribution of quiz scores has $\bar{x} = 35$ and $s = 4$. Beryl got 40. What was her z -score? What information does that give you?

Answer:

$$z = \frac{40 - 35}{4} = 1.25.$$

This means that Beryl's score was 1.25 standard deviations above the mean.

5. In a normal distribution with mean 25 and standard deviation 7, what proportions of terms are less than 20?

Answer: $z_{20} = \frac{20 - 25}{7} = -.71 \rightarrow Area = .2389.$

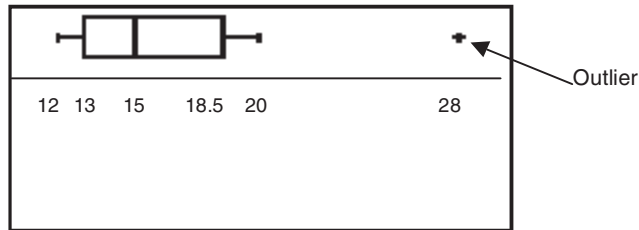
[By calculator: $normalcdf(-100, 20, 25, 7) = .2375$]

6. What are the mean, median, mode, and standard deviation of a *standard normal curve*?

Answer: Mean = median = mode = 0. Standard deviation = 1.

7. Find the 5-number summary and draw the modified box plot for the following set of data: 12, 13, 13, 14, 16, 17, 20, 28

Answer: The 5-number summary is [12, 13, 15, 18.5, 28]. 28 is an outlier, so 20 is the end of the upper whisker, as shown in the following diagram.



8. A distribution is strongly skewed to the right. Would you prefer to use the mean and standard deviation, or the median and interquartile range, to describe the center and spread of the distribution.

Answer: Because the mean is not resistant and is pulled toward the tail of the skewed distribution, you would prefer to use the median and IQR.

3 PRACTICE PROBLEMS

Multiple Choice

- The following list is ordered from smallest to largest: 25, 26, 26, 30, y , y , y , 33, 150. Which of the following statements are true?
 - The mean is greater than the median
 - The mode is 26
 - There are no outliers in the data
 - I only
 - I and II only
 - III only
 - I and III only
 - II and III only
- Jenny is 5'10" tall and is worried about her height. The heights of girls in the school are approximately normally distributed with a mean of 5'5" and a standard deviation of 2.6". What is the percentile rank of Jenny's height?
 - 59
 - 65
 - 74
 - 92
 - 97

3. The mean and standard deviation of a normally distributed dataset are 19 and 4, respectively. 19 is subtracted from every term in the dataset and then the result is divided by 4. Which of the following best describes the resulting distribution.
 - a. It has a mean of 0 and a standard deviation of 1.
 - b. It has a mean of 0, a standard deviation of 4, and its shape is normal.
 - c. It has a mean of 1 and a standard deviation of 0.
 - d. It has a mean of 0, a standard deviation of 1, and its shape is normal.
 - e. It has a mean of 0, a standard deviation of 4, and its shape is unknown.
4. The 5-number summary for a univariate dataset is {5, 18, 20, 40, 75}. If you wanted to construct a modified boxplot for the dataset (that is, one that would show outliers if there are any), what would be the maximum possible length of the right side “whisker?”
 - a. 35
 - b. 33
 - c. 5
 - d. 55
 - e. 53
5. A set of 5000 scores on a college readiness exam are known to be approximately normally distributed with mean 72 and standard deviation 6. To the nearest integer value, how many scores are there between 63 and 75.
 - a. 0.6247
 - b. 4115
 - c. 3650
 - d. 3123
 - e. 3227

Free Response

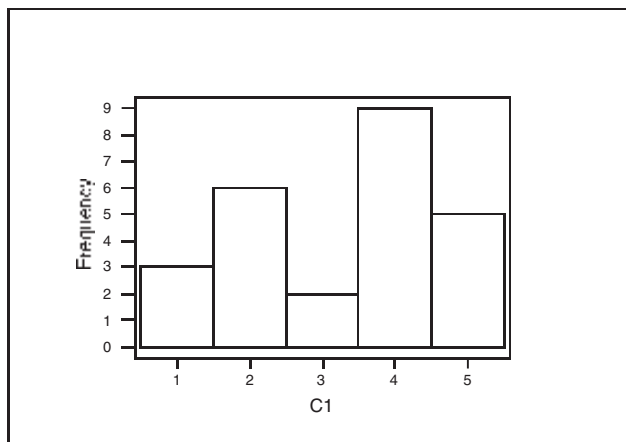
1. Mickey Mantle played with the New York Yankees from 1951 through 1968. He had the following number of home runs for those years: 13, 23, 21, 27, 37, 52, 34, 42, 31, 40, 54, 30, 15, 35, 19, 23, 22, 18. Were any of these years outliers? Explain.
2. Which of the following are properties of the normal distribution?
 - a. It has a mean of 0 and a standard deviation of 1.
 - b. Its mean = median = mode.
 - c. All terms in the distribution lie within 4 standard deviations of the mean.
 - d. It is “bell-shaped”
 - e. The total area under the curve and above the horizontal axis is 1.

3. Make a stemplot for the number of home runs hit by Mickey Mantle during his career (from problem #1, the numbers are: 13, 23, 21, 27, 37, 52, 34, 42, 31, 40, 54, 30, 15, 35, 19, 23, 22, 18). Do it first using an increment of 10, then do it again using an increment of 5. What can you see in the second graph that was not obvious in the first?
4. A group of 15 students were identified as needing supplemental help in basic arithmetic skills. Two of the students were put through a pilot program and achieved scores of 84 and 89 on a test of basic skills after the program was finished. The other 13 students received scores of 66, 82, 76, 79, 72, 98, 75, 80, 76, 55, 77, 68, and 69. Find the z -scores for the students on the pilot program and comment on the success of the program.
5. For the 15 students whose scores were given in question #4, find the 5-number summary and construct a boxplot of the data. What are the most distinguished features of the graph?
6. Assuming that the batting averages in major league baseball over the years has been approximately normally distributed with a mean of 0.265 and a standard deviation of 0.032, what would the percentile rank of a player be who bats 0.370 (as Barry Bonds did in the 2002 season)?
7. In problem #1, we considered the home runs hit by Mickey Mantle during his career. The following is a scatterplot of the number of doubles hit by Mantle during his career. What is the interquartile range (IQR) of this data?

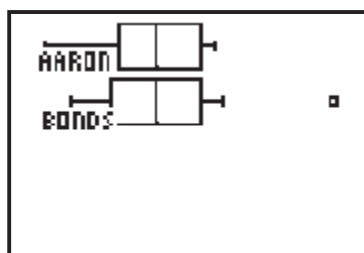
1	0		8
5	1		1224
(5)	1		56777
8	2		1234
4	2		558
1	3		
1	3		7

Note: The column of numbers to the left of the stemplot gives the cumulative frequencies from each end of the stemplot (e.g., there are 5 values, reading from the top, when you finish the second row). The (5) identifies the location of the row that contains the median of the distribution. It is standard for computer packages to draw stemplots in this manner

8. For the histogram pictured below, what proportion of the terms are less than 3.5?



9. The following graph shows boxplots for the number of career home runs for Hank Aaron and Barry Bonds. Comment on the graphs. Which would you rather have on your team *most* seasons? A season in which you needed a *lot* of home runs?

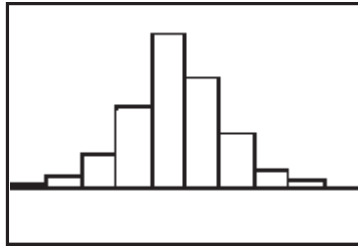


10. Suppose that being in the top 20% of people on blood cholesterol level is considered dangerous. Assume that cholesterol levels are approximately normally distributed with mean 185 and standard deviation 25. What is the maximum cholesterol level you can have and not be in the top 20%?
11. The following are the salaries, in millions of dollars, for members of the 2001–2002 Golden State Warriors: 6.2, 5.4, 5.4, 4.9, 4.4, 4.4, 3.4, 3.3, 3.0, 2.4, 2.3, 1.3, .3, .3. Which gives a better “picture” of these salaries, mean-based or median-based statistics? Explain.
12. The following table gives the results of an experiment in which the ages of 525 pennies from current change recorded. “0” represents the current year, “1” represents pennies one year old, etc.

Age	0	1	2	3	4	5	6	7	8	9	10	11
Count	163	87	52	75	44	24	36	14	11	5	12	2

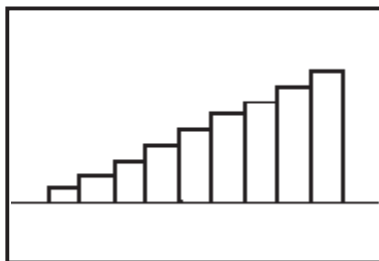
Describe the distribution of ages of pennies (remember that the instruction, “describe” means to discuss center, spread, and shape).

13. Which of the following statements are true?
- I. The median is resistant to extreme values
 - II. The mean is resistant to extreme values
 - III. The standard deviation is resistant to extreme values
- a. I only
 - b. II only
 - c. III only
 - d. II and III only
 - e. I and III only
14. The mean of a set of 150 values is 35, its median is 33, its standard deviation is 6, and its IQR is 12. A new set is created by first subtracting 10 from every term and then multiplying by 5. What are the mean, median, variance, standard deviation, and IQR of the new set?
15. The following graph shows the distribution of the heights of 300 women whose average height is 65" and whose standard deviation is 2.5". Assume that the heights of women are approximately normally distributed. How many of the women would you expect to be less than 5' 2" tall?



16. Which of the following are properties of the *standard deviation*?
- a. It's the square root of the average squared deviation from the mean.
 - b. It's resistant to extreme values.
 - c. It's independent of the number of terms in the distribution.
 - d. If you added 25 to every value in the dataset, the standard deviation wouldn't change
 - e. The interval $\bar{x} \pm 2s$ contains 50% of the data in the distribution.
17. Look again at the salaries of the Golden State Warriors in question 11 (in millions, 6.2, 5.4, 5.4, 4.9, 4.4, 4.4, 3.4, 3.3, 3.0, 2.4, 2.3, 1.3, .3, .3). Erick Dampier was the highest paid player at \$6.2 million. What sort of raise would he need so that his salary would be an *outlier* among these salaries?

18. Given the histogram below, draw, as best you can, the boxplot for the same data.



19. On the first test of the semester, the class average was 72 with a standard deviation of 6. On the second test, the class average was 65 with a standard deviation of 8. Nathan scored 80 on the first test and 76 on the second. Compared to the rest of the class, on which test did Nathan do better?
20. What is the mean of a set of data where $s = 20$,

$$\sum x = 245, \text{ and } \sum (x - \bar{x})^2 = 13,600?$$

3 CUMULATIVE REVIEW PROBLEMS

- Which of the following are examples of quantitative data?
 - The number of years each of your teachers has taught
 - Classifying a statistic as quantitative or qualitative
 - The length of time spent by the typical teenager watching television in a month
 - The daily amount of money lost by the airlines in the 15 months after the 9/11 attacks
 - The colors of the rainbow
- Which of the following are *discrete* and which are *continuous*?
 - The weights of a sample of dieters from a weight-loss program
 - The SAT scores for students who have taken the test over the past 10 years
 - The AP Statistics exam scores for the almost 50,000 students who took the exam in 2002
 - The number of square miles in each of the 20 largest states
 - The distance between any two points on the number line
- Just exactly what *is* statistics and what are its two main divisions?
- What are the main differences between the goals of a *survey* and an *experiment*?

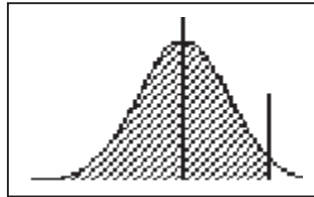
5. Why do we need to understand the concept of a *random variable* in order to do inferential statistics?

3 SOLUTIONS TO PRACTICE PROBLEMS

Multiple Choice

- The correct answer is (a).
- The correct answer is (e).

$$z = \frac{70 - 65}{2.6} = 1.92 \rightarrow \text{percentile} = .9726 \text{ (see drawing below):}$$



- The correct answer is (d). The effect on the mean of a dataset of subtracting the same value is to reduce the old mean by that amount (that is, $\mu_{x-k} = \mu_x - k$). Because the original mean was 19, and 19 has been subtracted from every term, the new mean is 0. The effect on the standard deviation of a dataset of dividing each term by the same value is to divide the standard deviation by that value (that is,

$$\sigma_{\frac{x}{k}} = \frac{\sigma_x}{k}.$$

Because the old standard deviation was 4, dividing every term by 4 yields a new standard deviation of 1. Note that the process of subtracting the mean from each term and dividing by the standard deviation creates a set of *z* scores

$$z_x = \frac{x - \bar{x}}{s}$$

so that any complete set of *z*-scores has a mean of 0 and a standard deviation of 1.

- The correct answer is (b). The maximum length of a “whisker” in a modified boxplot is $1.5(\text{IQR}) = 1.5(40 - 18) = 33$.

5. The correct answer is (d). The area under a normal curve between 63 and 75 is 0.6247. $(0.6247)(5000) = 3123.5$ (when done with calculator accuracy, it is actually slightly less than this which is why the answer is given as 3123 rather than rounding up to 3124).

Free Response

1. Using the calculator, we find that $\bar{x} = 29.78$, $s = 11.94$, $Q1 = 21$, $Q3 = 37$. Using the 1.5(IQR) rule, outliers are values that are less than $21 - 1.5(37 - 21) = -3$ or greater than $37 + 1.5(37 - 21) = 61$. Because no values lies outside of those boundaries, there are no outliers by this rule.

Using the $\bar{x} \pm 2s$ rule, we have $\bar{x} \pm 2s = 29.78 \pm 2(11.94) = \langle 5.9, 53.66 \rangle$. By this standard, the year he hit 54 home runs would be considered an outlier.

2. (a) is a property of the *standard normal distribution*, not a property of normal distributions in general. (b) is a property of the normal distribution. For (c), *almost* all of the terms are within 4 standard deviation of the mean but, at least in theory, there are terms at any given distance from the mean. (d) and (e) are properties of the normal distribution.

3.

1	3589
2	12337
3	01457
4	02
5	24

1	3
1	589
2	1233
2	7
3	014
3	57
4	02
4	
5	24

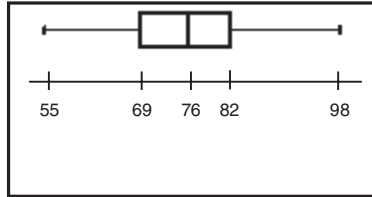
What shows up when done by 5 rather than 10 is the gap between 42 and 52. In 16 out of 18 years, Mantle hit 42 or less home runs. He hit more than 50 more than twice.

4. $\bar{x} = 76.4$ and $s = 10.17$.

$$z_{84} = \frac{84 - 76.4}{10.17} = 0.75. \quad z_{89} = \frac{89 - 76.4}{10.17} = 1.24.$$

Using the *Standard Normal Probability* table, a score of 84 corresponds to the 77.34th percentile, and a score of 89 corresponds to the 89.25th percentile. Both students were in the top quartile of scores after the program and performed better than all but one of the other students. We don't know that there is a cause-and-effect relationship between the pilot program and the high scores (that would require comparisons with a pretest), but it's reasonable to assume that the program had a positive impact. You might wonder how the student who got the 98 did so well!

5.



The most distinguishing feature is that the range (44) is quite large compared to the middle 50% of the scores (13). That is, we can see from the graph that the scores are packed somewhat closely about the median. The shape of a histogram of the data would be symmetric and mound shaped.

$$6. z_{.370} = \frac{.370 - .265}{.032} = 3.28. \rightarrow \text{Area of the left of } 3.28 \text{ is } .9994.$$

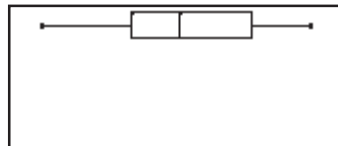
That is, Bond's average in 2002 would have placed him in the 99.94th percentile of batters.

7. There are 18 values in the stemplot. The median is 17 (actually between the last two 7s in the row—it's still 7). Because there are 9 values in each half of the stemplot, the median of the lower half of the data, Q_1 , is the 5th score from the top. So, $Q_1 = 14$. $Q_3 =$ the 5th score counting from the bottom = 24. Thus, $IQR = 24 - 14 = 10$.
8. There are 3 values in the first bar, 6 in the second, 2 in the third, 9 in the fourth, and 5 in the fifth for a total of 25 values in the dataset. Of these, $3 + 6 + 2 = 11$ are less than 3.5. There are 25 terms altogether, so the proportion of terms less than 3.5 is $11/25 = 0.44$.
9. The most obvious thing about these two is, with the exception of the one outlier for Bonds, is just how similar the two are. The medians of the two are almost identical and the IQRs are very similar. The data do not show it, but with the exception of 2001, the year Bonds hit 73 home runs, neither batter ever hit 50 or more home runs in a season. So, for any given season, you should be overjoyed to have either on your team, but there is no good reason to choose one over the other. However, for a single season, you would certainly choose Bonds.
10. Let x be the value in question. Because we do not want to be in the top 20%, the area to the left of x is .8. Hence $z_x = 0.84$ (found by locating the nearest table entry to 0.8, which is 0.7995 and reading the corresponding z -score as 0.84). Then

$$z_x = .84 = \frac{x - 185}{25} \rightarrow x = 206.$$

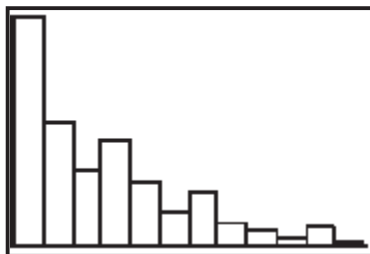
[Using the calculator, the solution to this problem is given by *invNorm* (0.8,185,25)]

11. $\bar{x} = \$3.36$ million, $s = \$1.88$ million, $Med = \$3.35$ million, $IQR = \$2.6$ million. A boxplot of the data looks like this:



The fact that the mean and median are virtually the same, and that the boxplot shows that the data are more or less symmetric, indicates that either set of measures would be appropriate, but that we are certainly justified in using the mean if that is our preference.

12. The easiest way to do this is to use the calculator. Put the age data in L1 and the frequencies in L2. Then do 1-Var Stats L1,L2 (the calculator will read the second list as frequencies for the first list).
- The mean is 2.48 years, and the median is 2 years. This indicates that the mean is being pulled to the right—indicates that the distribution is skewed to the right or has outliers in the direction of the larger values.
 - The standard deviation is 2.61 years. Because one standard deviation to left would yield a negative value, this also indicates that the distribution extends further to the right than the left.
 - A histogram of the data, drawn on the TI-83, is drawn below. This definitely indicates that the ages of these pennies is skewed to the right.



13. (a) is the correct answer. The median is resistant to extreme values, and the mean is not (that is, extreme values will exert a strong influence on the numerical value of the mean but not on the median). II and III involve statistics equal to or dependent upon the mean, so neither of them is resistant.
14. The new mean is $5(35 - 10) = 125$.
 The new median is $5(33 - 10) = 115$.
 The new variance is $5^2(6^2) = 900$.
 The new standard deviation is $5(6) = 30$.
 The new IQR is $5(12) = 60$.

15. First we need to find the *proportion* of women who would be less than 62" tall:

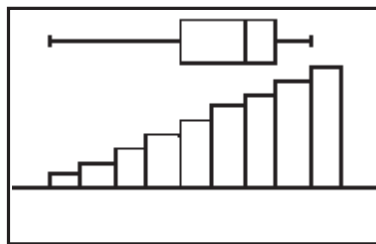
$$z_{ss} = \frac{57 - 60}{2.5} = -1.2 \rightarrow \text{Area} = 0.1151.$$

So 0.1151 of the terms in the distribution would be less than 62". This means that $0.1151(300) = 34.53$, so you would expect that 34 or 35 of the women would be less than 62" tall.

16. a, c, and d are properties of the standard deviation. (a) serves as a definition of the standard deviation. It is independent of the number of terms in the distribution in the sense that simply adding more terms will not necessarily increase or decrease s . (d) is another way of saying that the standard deviation is independent of the mean—it's a measure of spread, not a measure of center.

The standard deviation is *not* resistant to extreme values (b) because it is based on the mean, not the median. (e) is a statement about the interquartile range. In general, unless we know something about the curve, we don't know what proportion of terms are within 2 standard deviations of the mean.

17. For these data, $Q1 = \$2.3 \text{ million}$, $Q3 = \$4.9 \text{ million}$. To be an outlier, Erick would need to make at least $4.9 + 1.5(4.9 - 2.3) = 8.8$ million. In other words, he would need a \$2.6 million dollar raise in order to have his salary be an outlier.
18. You need to estimate the median and the quartiles. Note that the histogram is skewed to the left, so that the scores tend to pack to the right. This means that the median is to the right of center and that the boxplot would have a long whisker to the left. The box plot looks like this:



19. If you standardize both scores, you can compare them on the same scale. Accordingly

$$z_{80} = \frac{80 - 72}{6} = 1.333, \quad z_{76} = \frac{76 - 65}{8} = 1.375.$$

Nathan did slightly, but only slightly, better on the second test.

$$20. \quad s = 20 = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{13600}{n - 1}}$$

$$20^2 = \frac{13600}{n - 1} \rightarrow n = 35$$

$$\bar{x} = \frac{\sum x}{n} = \frac{245}{35} = 7$$

3 SOLUTIONS TO CUMULATIVE REVIEW PROBLEMS

1. a, c, and d are quantitative.
2. a, d, and e are continuous; b and c are discrete. Note that (d) could be considered discrete if what we meant by “number of square miles” was the integer number of square miles.
3. Statistics is the science of data. Its two main divisions are *data analysis* and *inference*. Data analysis (EDA) utilizes graphical and analytical methods to try to see that the data “says.” That is, EDA looks at data in a variety of way in order to understand it. Inference involves using information from samples to make statements or predictions about the population from which the sample was drawn.
4. A survey, based on a sample, from some population, is usually given in order to be able to make statements or predictions about the population. An experiment, on the other hand, usually has as its goal studying the differential effects of some treatment on a sample, which is often comprised of volunteers.
5. Statistical inference is based on being able to determine the probability of getting a particular sample statistic from a population with a hypothesized parameter. For example, we might ask how likely it is to get 55 heads on 100 flips of a fair coin. If it seems unlikely, we might reject the notion that coin we actually flipped is fair. The probabilistic underpinnings of inference can be understood through the language of random variables. In other words, we need random variables to bridge the gap between simple data analysis and inference.

Chapter 5

Bivariate Data Analysis



Main concepts: *scatterplots, lines of best fit, correlation coefficient, least-squares regression line, coefficient of determination, residuals, outliers and influential points, transformations to achieve linearity*

SCATTERPLOTS

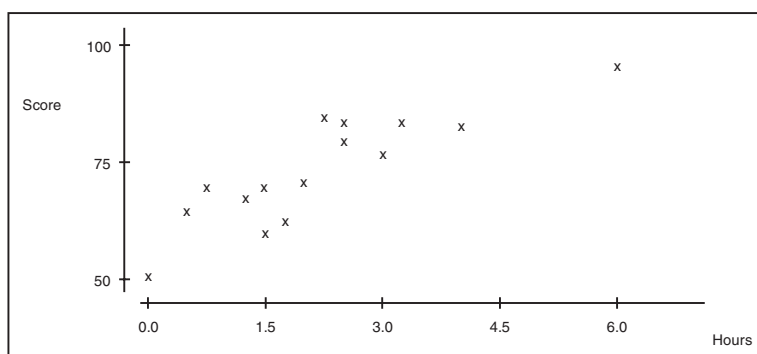
In the previous chapter, we looked at several different ways to graph univariate (one-variable) data. By choosing from dotplots, stemplots, histograms, or boxplots, we were able to examine visually patterns in the data. In this chapter, we consider techniques of data analysis for **bivariate** (two-variable) data. Specifically, our interest is whether or not two variables have a linear relationship and how changes in one variable can predict changes in the other variable.

example: For an AP Statistics class project, a statistics teacher had her students keep diaries of how many hours they studied before their midterm exam. The following are the data for 15 of the students.

Student	Hours Studied	Score on Exam
A	0.5	65
B	2.5	80
C	3.0	77
D	1.5	60
E	1.25	68
F	0.75	70
G	4.0	83

H	2.25	85
I	1.5	70
J	6.0	96
K	3.25	84
L	2.5	84
M	0.0	51
N	1.75	63
O	2.0	71

The teacher wanted to know if additional studying resulted in higher grades and drew the following graph, called a **scatterplot**. It seemed pretty obvious to the teacher that additional hours spent studying generally resulted in higher grades on the exam (do you see the pattern?).



In the previous example, we were interested in seeing whether studying has an effect on test performance. To do this we drew a **scatterplot**, which is just a two-dimensional graph of ordered pairs. We put one variable on the horizontal axis and the other on the vertical axis. In the example, the horizontal axis is for “hours studied” and the vertical axis is for “Score on test.” Each point on the graph represents the ordered pair for one student. If we have an explanatory variable, it should be on the horizontal axis and the response variable should be on the vertical axis.

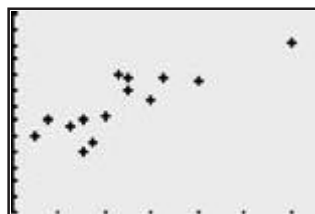
In the example, we saw a situation in which the variable on the vertical axis tended to increase as the variable on the horizontal axis increased. We say that two variables are **positively associated** if one of them increases as the other increases and **negatively associated** if one of them decreases as the other increases.



Calculator Tip: To draw a scatterplot on your calculator, first enter the data in two lists, say the horizontal axis variable in $L1$ and the vertical axis variable in $L2$. Then go to STAT PLOT and choose the scatterplot icon from “Type:” Enter $L1$ for $Xlist:$ and $L2$ for $Ylist:$. Choose whichever *Mark* pleases you. Be sure there are no equations active in the “Y=” list. Then you can do a *ZOOM ZoomStat (Zoom-9)* and the calculator

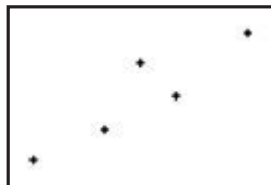
will draw the scatterplot for you. You can then go back, if you want, and adjust the WINDOW in any way you want.

The scatterplot of the data in the example, drawn on the calculator, looks like this:



(The window used was:
[0,6.5,1,40,105,5,1])

example: Which of the following statements best describes the scatterplot pictured?



- I. A line might fit the data pretty well.
 - II. The variables are positively associated.
 - III. The variables are negatively associated
- a. I only
 - b. II only
 - c. III only
 - d. I and II only
 - e. II and III only

solution: d is correct. The data look as though a line might be a good model, and the y -variable increases as the x -variable increases so that they are positively associated.

CORRELATION

We have seen how to graph bivariate data in a scatterplot. Following the pattern we set in the previous chapter, we now want to do some numerical analysis of the data in an attempt to understand better the relationship between the variables.

In AP Statistics, we are primarily interested in determining the extent to which two variables are *linearly* associated. Two variables are *linearly related* to the extent that their relationship can be modeled by a line.

Sometimes, and we look at this more closely later in this chapter, variables may not be linearly related but can be transformed in such a way that the transformed data are linear. Sometimes the data are related but not linearly (e.g., the height of a thrown ball is a quadratic function of the number of elapsed seconds since it was released).

The first statistic we have to determine a linear relationship is the Pearson product moment correlation, or more simply, the **correlation coefficient**, denoted by the letter r . The correlation coefficient gives us information about *strength* of the linear relationship between two variables (how well a line fits the data) as well as the *direction* of the linear relationship (whether the variables are positively or negatively associated).

If we have a sample of size n of paired data, say (x,y) , and assuming that we have computed summary statistics for x and y (means and standard deviations), the **correlation coefficient r** is defined as follows:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left\{ \frac{x_i - \bar{x}}{s_x} \right\} \left\{ \frac{y_i - \bar{y}}{s_y} \right\}$$

Because the terms inside the summation symbol are nothing more than the z -scores of the individual x and y values, an easy way to remember this definition is:

$$r = \frac{1}{n-1} \sum_{i=1}^n z_{x_i} z_{y_i} = \frac{1}{n-1} \sum z_x z_y$$

example: Earlier in the section, we saw some data for hours studied and the corresponding scores on an exam. It can be shown that, for this data, $r = .864$. This indicates a strong positive linear relationship between hours studied and exam score. That is, the more hours studied, the higher the exam score.

The correlation coefficient r has a number of properties you need to be familiar with:

- $-1 \leq r \leq 1$. If $r = -1$ or $r = 1$, the points all lie on a line.
- Although there are no hard and fast rules about how strong a correlation is based on its numerical value, the following guidelines might help you categorize r :

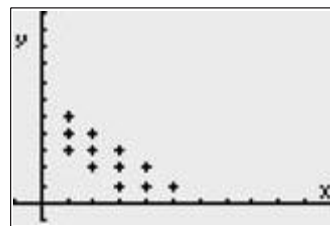
Value of r	Strength of Linear Relationship
$-1 < r < -0.8$ $0.8 < r < 1$	strong
$-0.8 < r < -0.5$ $0.5 < r < 0.8$	moderate
$-0.5 < r < 0.5$	weak

Remember that these are only very rough guidelines. A value of $r = .2$ might well indicate a significant linear relationship (that is, it's unlikely

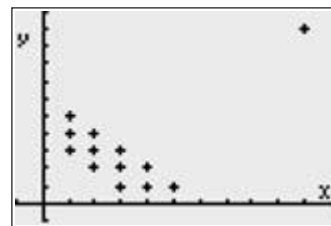
to have gotten 0.2 unless there really was a linear relationship), and an r or 0.8 might be reflective of a single influential point rather than a actual linear relationship between the variables.

- If $r > 0$, it indicates that the variables are positively associated. If $r < 0$, it indicates that the variables are negatively associated.
- If $r = 0$, it indicates that there is no linear association that would allow us to predict y from x . It *does not* mean that there is no relationship—just not a linear one.
- It does not matter which variable you call x and which variable you call y . r will be the same. In other words, r depends only on the paired points, not the *ordered* pairs.
- r is not resistant to extreme values because it is based on the mean. A single extreme value can have a powerful impact on r and may cause us to overinterpret the relationship. You must look at the scatterplot of the data as well as r .

example: To illustrate that r is not resistant, consider the following two graphs. The graph on the left, with 12 points, has a marked negative linear association between x and y . The graph on the right has the same basic visual pattern but, as you can see, the addition of the one outlier has a dramatic effect on r —making what is generally a negative association between two variables appear to have a moderate, positive association.



negative association
 $r = -.80$



negative association (??)
 $r = .51$

example: The following computer output, again for the hours studied versus exam score data, indicates R-sq, which is the square of r . Accordingly, $r = \sqrt{0.747} = 0.864$. There is a lot of other stuff in the box that that doesn't concern us just yet. We learn about other important parts of the output as we proceed through the rest of this book.

The regression equation is
Score = 59.0 + 6.77 Hours

Predictor	Coef	St Dev	t ratio	P
Constant	59.026	2.863	20.62	.000
Hours	6.767	1.092	6.20	.000

$s = 6.135$

R-sq = 74.7%

R-sq(adj) = 72.8%



Calculator Tip: To find r on your calculator, you may first need to change a setting from the factory. Enter CATALOG and scroll down to “Diagnostic On.” Press ENTER twice. Now you are ready to find r .

Assuming you have the x - and y -values in $L1$ and $L2$, enter *STAT* *CALC* *LinReg(a + bx)* [that’s *STAT* *CALC* 8 on the TI-83] and press ENTER. Then enter $L1$, $L2$ and press enter. You should get a screen like this (using the data from the Study Time vs. Score on Test study): Entering *STAT* *CALC* *LinReg(ax + b)* $L2$, $L1$ will change a and b but will not change r because order doesn’t matter in correlation.

```
LinReg
y=a+bx
a=59.02574597
b=6.766833905
r2=.7470313175
r=.8643097347
```

If you compare this with the computer output above, you will see that it contains some of the same data, including both r and r^2 . At the moment, all you care about in this printout is the value of r

Correlation and Causation

Two variables, x and y , may have a strong correlation, but you need to take care not to interpret that as causation. That is, just because two things seems to go together does not mean that one caused the other—some third variable may be influencing them both. Just because you see fire trucks at almost every fire doesn’t mean that fire trucks cause fire.

example: Consider the following dataset that shows the increase in the number of Methodist ministers and the increase in the amount of imported Cuban rum:

Year	Number of Methodist ministers in New England	Number of barrels of Cuban rum imported to Boston	
1860	63	8376	For these data, it turns out that $r = .999986$. Is the increase in number of ministers responsible for the increase in imported rum? Some cynics might want to believe so, but the real reason is that the population was increasing from 1860 to 1940, so the area needed more ministers, and more people drank more rum.
1865	48	6406	
1870	53	7005	
1875	64	8486	
1880	72	9595	
1885	80	10,643	
1890	85	11,265	
1895	76	10,071	
1900	80	10,547	
1905	83	11,008	
1910	105	13,885	
1915	140	18,559	
1920	175	23,024	
1925	183	24,185	
1930	192	25,434	
1935	221	29,238	
1940	262	34,705	

In the above example, there was a *lurking variable*—one we didn't consider when we did the correlation—that caused both of these variables to change the way they did. We will look more at lurking variables in the next chapter, but in the meantime remember, always remember, that **correlation is not causation**.

LINES OF BEST FIT

When we discussed correlation, we saw that the order of the points didn't matter. However, we are often interested in our ability to predict the value of one variable based on the value of the other and, in this situation, order surely matters. Also, our interest is being able to predict given a linear relationship.

Least-Squares Regression Line

Recall again the data from the study that looked at hours studied versus score on test:

Hours Studied	Score on Exam
0.5	65
2.5	80
3.0	77

1.5	60
1.25	68
0.75	70
4.0	83
2.25	85
1.5	70
6.0	96
3.25	84
2.5	84
0.0	51
1.75	63
2.0	71

For these data, $r = .864$, so we have a strong linear association between the variables. Suppose we wanted to predict the score of a person who studied for 2.75 hours. If we had a linear model for the data, that is a line that seems to fit the data well, we could make such a prediction. We are looking for a **line of best fit**. We want to find a **regression line**—a line that can be used for predicting response values from explanatory values. In this situation, we would use the regression line to predict the exam score for a person who studied 2.75 hours.

The line we are looking for is called the **least-squares regression line**. We could draw a variety of lines on our scatterplot trying to determine which has the best fit. Let \hat{y} be the predicted value of y for a given value of x . Then $y - \hat{y}$ represents the error in prediction. We want our line to minimize errors in prediction, so we might first think that

$$\sum (y - \hat{y})$$

would be a good measure ($y - \hat{y}$ is the *actual value* minus the *predicted value*). However, because our line is going to average out the errors in some fashion, we find that

$$\sum (y - \hat{y}) = 0.$$

To get around this problem, we use

$$\sum (y - \hat{y})^2.$$

This expression will vary with different lines and is sensitive to the fit of the line. That is,

$$\sum (y - \hat{y})^2$$

is small when the linear fit is good and large when it is not.

The **least-squares regression line** (LSRL) is the line that minimizes the sum of squared errors. If $\hat{y} = a + bx$ is the LSRL, then \hat{y} minimizes

$$\sum (y - \hat{y})^2.$$

For n ordered pairs (x,y) , we calculate: \bar{x} , \bar{y} , s_x , s_y , and r . Then we have:

$$\text{If } \hat{y} = a + bx \text{ is the LSRL, } b = r \frac{s_y}{s_x}, \text{ and } a = \bar{y} - b\bar{x}$$

example: For the hours studied (x) versus score (y) study, the LSRL is $\hat{y} = 59.03 + 6.77x$. We asked earlier what score would we predict for someone who studied 2.75 hours. Plugging this value into the LSRL, we have $\hat{y} = (2.75) = 59.03 + 6.77(2.75) = 77.63$. Note that this is the predicted value, not the value such a person will necessarily get.

example: Consider once again the computer printout for the data of the preceding example:

The regression equation is				
Score = 59.0 + 6.77 Hours				
Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	59.026	2.863	20.62	.000
Hours	6.767	1.092	6.20	.000
s = 6.135		R-sq = 74.7%		R-sq(adj) = 72.8%

The regression equation is given as “*score = 59 + 6.77 hours*”. The y -intercept, which is the predicted score when the number of hours studied is zero, and the slope of the regression line are listed in the table under the column “*Coef.*”



Exam Tip: An AP Exam question in which you are asked to determine the regression equation from the printout has become almost a yearly event. Be sure you know where the intercept and slope of the regression line are located in the printout (they are under “*Coef*”)

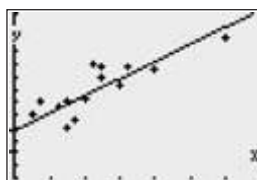
example: We saw earlier that the calculator output for this data was

```
LinReg
y=a+bx
a=59.02574597
b=6.766833905
r^2=.7470313175
r=.8643097347
```

The values of a and b are given. Remember that is was obtained by putting the “hours studied” data in $L1$, the “test score” data in $L2$, and doing $LinReg(ax + b) L1, L2$. When using $LinReg(ax + b)$, the explanatory variable *must* come first and the response variable second.



Calculator Tip: The easiest way to see the line on the graph is to put the regression equation in $Y1$ as you do the regression. This can be done by entering $\text{LinReg}(ax + b) L1, L2, Y1$. The “ $Y1$ ” can be pasted in by entering $\text{VARS } Y\text{-VARS } \text{Function } Y_1$. By creating a scatterplot of $L1$ and $L2$, you then get the following graphic:



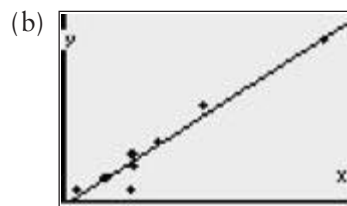
example: An experiment is conducted on the effects of having convicted criminals provide restitution to their victims rather than serving time. The following table gives the data for 10 criminals. The monthly salaries (X) and monthly restitution payments (Y) were as follows:

X	300	880	1000	1540	1560	1600	1600	2200	3200	6000
Y	200	380	400	200	800	600	800	1000	1600	2700

- Find the correlation between X and Y and the regression equation that can be used to predict monthly restitution payments from monthly salaries.
- Draw a scatterplot of the data and put the LSRL on the graph.
- Interpret the slope of the regression line in the context of the problem.
- How much would a criminal earning \$1400 per month be expected to pay in restitution.

solution: Put the X -data in $L1$ and the Y -data in $L2$. Then enter $\text{STAT } \text{CALC } \text{LinReg}(ax + b) L1, L2, Y1$.

- $r = 0.97$, $\text{Payments} = -56.22 + 0.46(\text{Salary})$. [If you answered $\hat{y} = 56.22 + .46x$, you must define x and y so that the regression equation can be understood in the context of the problem.]



- The slope of the regression line is 0.46. This tells us that, for each \$1 increase in the criminal’s salary, the amount of restitution is predicted to increase by \$0.46. Or, you could say that the average increase is \$0.46.
- $\text{Payment} = -56.22 + 0.46(1400) = \587.88 .



Calculator Tip: The fastest, and most accurate, way to perform the computation above, assuming you have stored the LSRL in $Y1$ (or some “ $Y =$ ” location), is to do $Y1(1400)$ on the home screen. To paste $Y1$ to the home screen, enter $VAR\ Y-VARS$ Function $Y1$.

RESIDUALS

When we developed the regression equation, we referred to the *actual value—predicted value*, $y - \hat{y}$, as an error in prediction. The formal name for $y - \hat{y}$ is the **residual**.

example: In the previous example, a criminal earning \$1560/month paid restitution of \$800/month. The predicted restitution for this amount would be $\hat{y} = -56.22 + 0.46(1560) = \661.38 . Thus, the residual for this case is $\$800 - \$661.38 = \$138.62$.



Calculator Tip: The TI-83 will generate a complete set of residuals when you perform a *LinReg*. They are stored in a list called *RESID* which can be found in the *LIST* menu. *RESID* stores only the current set of residuals. That is, a new set of residuals is stored in *RESID* each time you perform a new regression.

Residuals can be useful to us in determining the extent to which a linear model is appropriate for a dataset. If a line is an appropriate model, we would expect to find the residuals more or less randomly scattered about the average residual (which is, of course, 0). In fact, we expect to find them approximately normally distributed about 0. A pattern of residuals that does not appear to be more or less randomly distributed about 0 (that is, there is a systematic nature to the graph of the residuals) is evidence that a line is not a good model for the data. If the residuals are small, the line may predict well even though it isn't a good theoretical model for the data. The usual method of determining if a line is a good model is to examine visually a plot of the residuals plotted against the explanatory variable.



Calculator Tip: To draw a residual plot on your TI-83, and assuming your data is in $L1$ and $L2$, first do *LinReg*($ax + b$) $L1, L2$. Next, you create a *STAT PLOT* scatterplot where *Xlist:* is set to $L1$ and *Ylist:* is set to *RESID*. *RESID* can be retrieved from the *LIST* menu. *ZOOM ZoomStat* (*Zoom 9*) will then draw the residual plot for the current list of residuals. It's a good idea to turn off any equations you may have in the $Y =$ list.

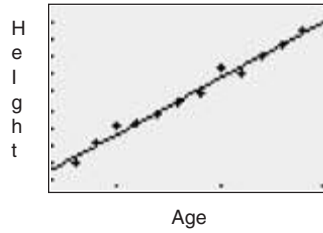
example: The data given below show the height (in cm) at various ages (in months) for a group of children.

- Does a line seem to be a good model for the data? Explain.
- What is the value of the residual for a child of 19 months?

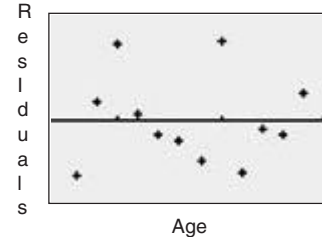
Age	18	19	20	21	22	23	24	25	26	27	28	29
Height	76	77.1	78.1	78.3	78.8	79.4	79.9	81.3	81.1	82.0	82.6	83.5

solution:

- (a) Using the calculator, we find $height = 64.94 + 0.634(age)$, $r = 0.993$. The strong value of r tells us that the points are close to a line. The scatterplot and LSLR are shown below on the graph at the left.



Scatterplot and LSRL for predicting height from age.

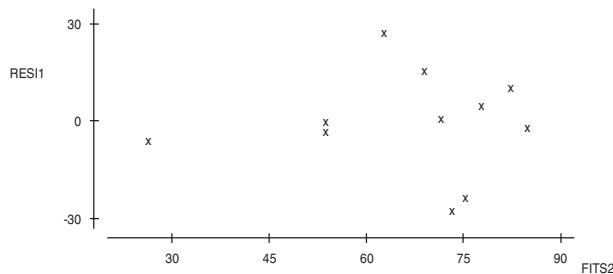


Scatterplot of residuals vs. age.

From the graph on the left, a line appears to be a good fit for the data (the points lie close to the line). The residual plot on the right shows no readily obvious pattern, so we have good evidence that a line is a good model for the data and we can feel good about using the LSRL to predict height from age.

- (b) The residual (actual minus predicted) for $age = 19$ months is $77.1 - (64.94 + 0.634 \cdot 19) = 0.114$.

Digression: Whenever we have graphed a residual plot in this section, the vertical axis has been the residuals and the horizontal axis has been the x -variable. On some computer printouts, you may see the horizontal axis labeled “Fits” or “Predicted Value” as in the graph below:



What you are interested in is the visual image given by the residual plot, and it doesn't matter if the residuals are plotted against the x -variable or something else, like “FITS2”—the scatter of the points above and below 0 stays the same. All that changes are the horizontal distances between points. This is the way it must be done in multiple regression and, as you can see, it can be done in simple linear regression.

If we are trying to predict a value of y from a value of x , it is called **interpolation** if we are trying to predict a value of y from an x -value within the range of x -values. It is called **extrapolation** if we are predicting from a value of x outside of the x -values.

example: Using the age/height data from the previous example, we are *interpolating*

Age	18	19	20	21	22	23	24	25	26	27	28	29
Height	76	77.1	78.1	78.3	78.8	79.4	79.9	81.3	81.1	82.0	82.6	83.5

if we attempt to predict height from an age between 18 and 29 months. It is interpolation if we try to predict the height of a 20.5-month-old baby. We are extrapolating if we try to predict the height of a child less than 18 months old or more than 29 months old.

If a line has been shown to be a good model for the data and if it fits the line well (i.e., we have a strong r and a more or less random distribution of residuals), we can have confidence in interpolated predictions. We can rarely have confidence in extrapolated values. In the example above, we might be willing to go slightly beyond the ages given because of the high correlation and the good linear model, but it's good practice not to extrapolate beyond the data given. If we were to extrapolate the data in the example to a child of 12 years of age (144 months), we would predict the child to be 156.2 inches, or more than 13 feet tall.

COEFFICIENT OF DETERMINATION

In the absence of a better way to predict y -values from x -values, our best guess might well be \bar{y} , the mean value of y .

example: Suppose you had access to the heights and weights of each of the students in your statistics class. You compute the average weight of all the students. You write the heights of each student on a slip of paper and put the slips in a hat, and then draw out one slip. You are asked to predict the weight of the student whose height you now know. What is your best guess as to the weight of the student?

solution: In the absence of any known relationship between height and weight, your best guess would have to be the average weight of all the students. You know the weights vary about the average and that is about the best you could do.

If we guessed at the weight of each student using the average, we would be wrong most of the time. If we took each of those errors and squared them, we would have what is called the *sum of squares total* (SST). It's the total squared error of our guesses when our best guess is simply the mean of the weights of all students, and represents the total variability of y .

Now suppose we have a least-squares regression line that we want to use as a model for predicting weight from height. It is, of course, the LSRL we discussed in detail in section 5.3, and our hope is that there will be less error in prediction than by using \bar{y} . Now, we still have errors from the regression line (called *residuals*, remember?). We call the sum of *those* errors the **sum of squared errors** (SSE). So, SST represents the errors from using \bar{y} as the basis for predicting weight from height, and SSE represents the errors from using the LSRL. SST–SSE represents the benefit of using the regression line rather than \bar{y} for prediction. That is, by using the LSRL rather than \bar{y} , we have *explained* a certain proportion of the total variability by regression.

The proportion of the total variability in y that is explained by the regression of y on x is called the **coefficient of determination**. The *coefficient of determination* is symbolized by r^2 . Based on the above discussion, we note that

$$r^2 = \frac{SST - SSE}{SST}.$$

It can be shown algebraically that r^2 is the square of r , the correlation coefficient. Many computer programs will report r^2 (R-sq) only, so we must take its square root if we want to know r . The TI-83 calculator will report both r and r^2 , as well as the regression coefficients, when you do *LinReg(a + bx)*.

example: Consider the following output for a linear regression:

Predictor	Coef	St Dev	t ratio	P
Constant	-1.95	21.97	-0.09	.931
x	0.8863	0.2772	3.20	.011
s = 16.57		R-sq = 53.2%		R-sq(adj) = 48.0%

We can see that the LSRL for these data is $\hat{y} = -1.95 + 0.8863x$. $r^2 = 53.2\% = 0.532$. This means that 53.2% of the total variability in y can be explained by the regression of y on x . Furthermore, $r = \sqrt{.532}$. We learn more about the other items in the printout later.

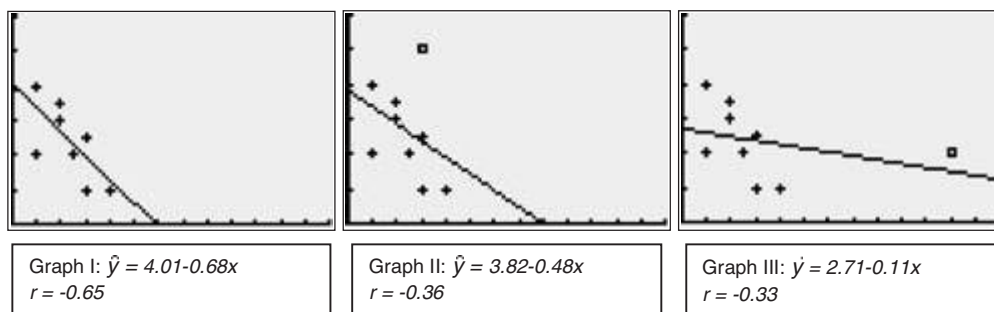
OUTLIERS AND INFLUENTIAL OBSERVERS

Some observations have an impact on correlation and regression. We defined an outlier quite specifically when we were dealing with one-variable data (remember the *1.5 IQR rule*?). We don't have an analogous definition when dealing with bivariate data, but it is the same basic idea: an **outlier** lies outside of the general pattern of the data. An outlier can certainly influence

a correlation and, depending on where it is located, may also exert an influence on the slope of the regression line.

An **influential observation** is often an outlier in the x -direction. Its influence, if it doesn't line up with the rest of the data, is on the slope of the regression line.

example: Graph I below has no outliers or influential points. Graph II has an outlier that is not an influential point. Graph III has an outlier that is an influential point. Compare the correlation coefficients and regression lines for each graph. Note that the outlier in Graph II has some effect on the slope and a significant effect on the correlation coefficient. The influential point in Graph III has about the same effect on the correlation coefficient as the outlier in Graph II, but a major influence on the slope of the regression line.



TRANSFORMATIONS TO ACHIEVE LINEARITY

Until now, we have been concerned with data that can be modeled with a line. Of course, there are many bivariate relationships that are nonlinear. The path of an object thrown in the air is parabolic (quadratic). Population tends to grow exponentially, at least for a while. Even though you could find a LSRL for nonlinear data, it makes no sense to do so. The AP Statistics course deals only with bivariate data that can be modeled by a line OR nonlinear bivariate data that can be *transformed* in such a way that the transformed data can be modeled by a line.

example: Let $g(x) = 2^x$, which is exponential and clearly nonlinear. Let $f(x) = \ln(x)$. Then, $f[g(x)] = \ln(2^x) = x \ln(2)$, which is linear. That is, we can transform an exponential function such as $g(x)$ into a linear function by taking the log of each value of $g(x)$.

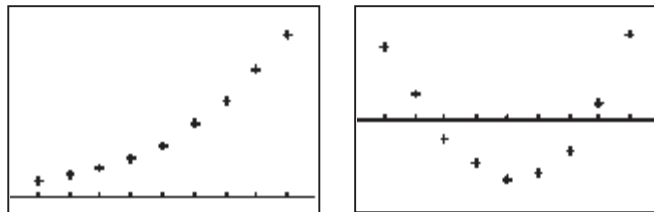
example: Let $g(x) = 4x^2$, which is quadratic. Let $f(x) = \sqrt{x}$. Then $f[g(x)] = \sqrt{4x^2} = 2x$, which is linear.

example: The number of a certain type of bacteria present (in thousands) after a certain number of hours is given in the following chart:

Hours	Number
1.0	1.8
1.5	2.4
2.0	3.1
2.5	4.3
3.0	5.8
3.5	8.0
4.0	10.6
4.5	14.0
5.0	18.0

What would be the predicted quantity of bacteria after 3.75 hours?

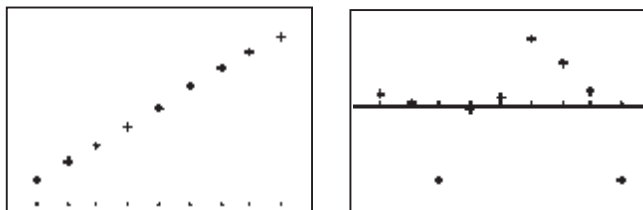
solution: A scatterplot of the data and a residual plot [for $Number = a + b(Hour)$] shows that a line is not a good model for this data:



Now, take $\ln(Number)$ to produce the following data:

Hours	Number	$\ln(Number)$
1.0	1.8	0.59
1.5	2.4	0.88
2.0	3.1	1.13
2.5	4.3	1.46
3.0	5.8	1.76
3.5	8.0	2.08
4.0	10.6	2.36
4.5	14.0	2.64
5.0	18.0	2.89

The scatterplot of $Year$ versus $\ln(population)$ and the residual plot for $\ln(Number) = -0.0048 + 0.586(Hours)$ are as follows:



The scatterplot looks much more linear and the residual plot no longer has the distinctive pattern of the raw data. We have transformed the original data in such a way that the transformed data is well modeled by a line. The regression equation for the transformed data is: $\ln(\text{Number}) = -0.0048 + 0.586(\text{Hours})$.

The question asked for how many bacteria are predicted to be present after 3.75 hours. Plugging 3.75 into the regression equation, we have $\ln(\text{Number}) = -0.0048 + 0.586(3.75) = 2.19$. But that is $\ln(\text{Number})$, not Number . We must back-transform this answer to the original units. Doing so, we have, $\text{Number} = e^{2.19} = 8.94$ thousand bacteria.



Calculator Tip: You do not need to take logarithms by hand in the above example—your calculator is happy to do it for you. Simply put Hours in $L1$ and Number in $L2$. Then let $L3 = \ln(L2)$. The LSRL line is then found by $\text{LinReg}(ax + b) L1, L3, Y1$.

Remember also that the easiest way to find the value of a number substituted into the regression equation is to simply find $Y1(\#)$. $Y1$ is found by entering $\text{VARS } Y\text{-VARS Function } Y1$.

Digression: You will find a number of different regression expressions in the STAT CALC menu: $\text{LinReg}(ax + b)$, QuadReg , CubicReg , QuartReg , $\text{LinReg}(a + bx)$, LnReg , ExpReg , PwrReg , Logistic , and SinReg . While each of these has its use, only $\text{LinReg}(a + bx)$ needs to be used in this course (well, $\text{LinReg}(ax + b)$ gives the same equation, just in algebraic form rather than usual statistical form).



Exam Tip: Also remember, when taking the AP exam, NO calculator-speak. If you do a linear regression on your calculator, simply report the result. The person reading your exam will know that you used a calculator and is NOT interested in seeing something like “ $\text{LinReg } L1, L2, Y1$ ” written on your exam.

It may be worth your while to try several different transformations to see if you can achieve linearity. Some possible transformations are: take the log of both variables, raise one or both variables to a power, take the square root of one of the variables, take the reciprocal of one or both variables, etc.

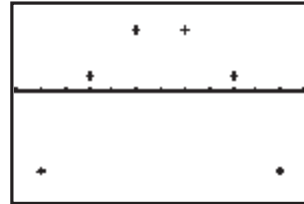
RAPID REVIEW

1. The correlation between two variables x and y is 0.85. Interpret this statement.

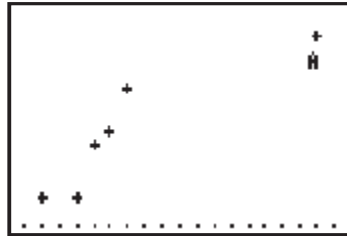
Answer: There is a strong, positive, linear association between x and y . That is, as one of the variables increases, the other variable increases as well.

2. The following is a residual plot of a least-squares regression. Does it appear that a line is a good model for the data? Explain.

Answer: the residual plot shows a definite pattern. If a line was a good model, we would expect to see a more or less random pattern of points about 0. A line is unlikely to be a good model.



3. Consider the following scatterplot. Is the point A an outlier or an influential observation? What effect would its removal have on the slope of the regression line?



Answer: A is an outlier because it is removed from the general pattern of the rest of the points. It is an influential observation because it is an outlier in the x -direction. Its removal would have an effect on the regression line: it would increase the slope of the LSRL.

4. A researcher finds that the LSRL for predicting GPA based on average hours studied per week is $GPA = 1.75 + .11(\text{hours studied})$. Interpret the slope of the regression line in the context of the problem.

Answer: For each additional hour studied, GPA is predicted to increase by 0.11. Alternatively, you could say that the GPA will increase 0.11 on average for each additional hour studied.

5. One of the variables that is related to college success (as measured by GPA) is socioeconomic status. In one study of the relationship, $r^2 = 0.45$. Explain what this means in the context of the problem.

Answer: $r^2 = 0.45$ means that 45% of the variability in college GPA is explained by the regression of GPA on socioeconomic status.

6. Each year of Governor Jones's tenure, the crime rate has decreased in a linear fashion. In fact, $r = -0.8$. It appears that the governor has been effective in reducing the crime rate. Comment.

Answer: Correlation is not causation. The crime rate could have gone down for a number of reasons besides Governor Jones's efforts.

7. What is the regression equation for predicting weight from height in the following computer printout?

The regression equation is				
weight = _____ + _____ height				
Predictor	Coef	St Dev	t-ratio	P
Constant	-104.64	39.19	-2.67	.037
Height	3.4715	0.5990	5.80	.001
s = 7.936		R-sq = 84.8%		R-sq(adj) = 82.3%

Answer: $weight = -104.64 + 3.4715(height)$

3 PRACTICE PROBLEMS

Multiple Choice

1. Given a set of ordered pairs (x,y) so that $s_x = 1.6$, $s_y = .75$, $r = .55$. What is the slope of the least-square regression line for this data?
- 1.82
 - 1.17
 - 2.18
 - .26
 - .78

2.

x	23	15	26	24	22	29	32	40	41	46
y	19	18	22	20	27	25	32	38	35	45

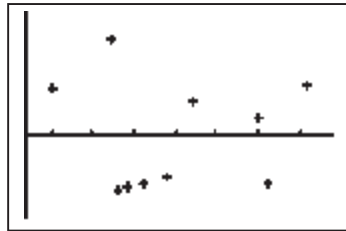
The regression line for the bivariate dataset given above is $\hat{y} = 2.35 + .86x$. What is the residual for the point whose x -value is 29?

- 1.71
 - 1.71
 - 2.29
 - 5.15
 - 2.29
3. A study found a correlation of $r = -0.58$ between hours spent watching television and hours per week spent exercising. That is, the more

hours spent watching television, the less hours spent exercising per week. Which of the following statements is most accurate?

- About one-third of the variation in hours spent exercising can be explained by hours spent watching television.
 - A person who watches less television will exercise more.
 - For each hour spent watching television, the predicted decrease in hours spent exercising is 0.58 hrs.
 - There is a cause-and-effect relationship between hours spent watching television and a decline in hours spent exercising.
 - 58% of the hours spent exercising can be explained by the number of hours watching television.
4. A response variable appears to be exponentially related to the explanatory variable. The natural logarithm of each y -value is taken and the least-squares regression line is found to be $\ln(y) = 1.64 - 0.88x$. Rounded to two decimal places, what is the predicted value of y when $x = 3.1$?
- 1.09
 - 0.34
 - 0.34
 - 0.082
 - 1.09

5. Consider the following residual plot:



Which of the following statements are true?

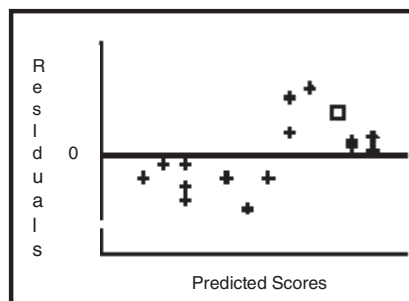
- The residual plot indicates that a line is a reasonable model for the data.
 - The residual plot indicates that there is no relationship between the data.
 - The correlation between the variables is probably non-zero.
- I only
 - II only
 - I and III only
 - II and III only
 - I and II only

Free Response

- Given a bivariate dataset such that $\bar{x} = 14.5$, $\bar{y} = 20$, $s_x = 4$, $s_y = 11$, $r = .80$. Find the least-squares regression line of y on x .
- The data given below give the first and second exam scores of 10 students in a calculus class

Test 1	63	32	87	73	60	63	83	80	98	85
Test 2	51	21	52	90	83	54	73	85	83	46

- Draw a scatterplot of these data.
 - To what extent do the scores on the two tests seem related?
- The following is a residual plot of a linear regression. Clearly, a line would not be a good fit for these data. Why not? Is the regression equation likely to underestimate or overestimate the y -value of the point in the graph marked with the square?



- The regional champion in 10 and under 100 m backstroke has had the following winning times (in seconds) over the past 8 years:

Year	1	2	3	4	5	6	7	8
Time	77.3	80.2	77.1	76.4	75.5	75.9	75.1	74.3

How many years until you expect the winning time to be one minute or less? What's wrong with this estimate?

- Measurements are made of the number of cockroaches present, on average, every 3 days, beginning on the second day, after apartments in one part of town are vacated. The data are as follows:

Days	2	5	8	11	14
# Roaches	3	4.5	6	7.9	11.5

How many cockroaches would you expect to be present after 9 days?

6. A study found a strongly positive relationship between number of miles walked per week and overall health. A local news commentator, after reporting on the results of the study, advised everyone to walk more during the coming year because walking more results in better health. Comment on the reporter's advice.
7. Carla, a young sociologist, is excitedly reporting on the results of her first professional study. The finding she is reporting is that 72% of the variation in math grades for girls can be explained by the girls socioeconomic status. What does this mean, and is it indicative of a strong linear relationship between math grades and socioeconomic status for girls?
8. Which of the following statements are true of a least-squares regression equation?
 - a. It minimizes the sum of the residuals.
 - b. The average residual is 0.
 - c. It minimizes the sum of the squared residuals.
 - d. The slope of the regression line is a constant multiple of the correlation coefficient.
 - e. The slope of the regression line tells you how much the response variable will change for each unit change in the explanatory variable.
9. Consider the following dataset:

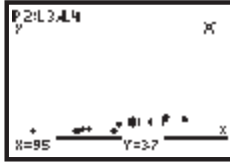
x	45	73	82	91
y	15	7.9	5.8	3.5

Given that the LSRL for this data is $\hat{y} = 26.211 - 0.25x$, what is the value of the residual for $x = 73$?

10. Suppose the correlation between two variables is $r = -0.75$. What is true of the correlation coefficient and the slope of the regression line if
 - (a) each of the y values is multiplied by -1 .
 - (b) the x and y variables are reversed.
 - (c) the x and y variables are each multiplied by -1 .
11. Suppose the regression equation for predicting success on a dexterity task (y) from number of training sessions (x) is $\hat{y} = 45 + 2.7x$ and that $\frac{s_y}{s_x} = 3.33$.

What percentage of the variation in y is not explained by the regression on x ?

12. Consider the following scatterplot. The highlighted point is both an outlier and an influential point. Describe what will happen to the correlation and the slope of the regression line if that point is removed.



13. The computer printout below gives the regression output for predicting *crime rate* (in crimes per 1000 population) from the *number of casino employees* (in 1000s).

The regression equation is				
Rate = _____ + _____ Number				
Predictor	Coef	St Dev	t ratio	P
Constant	-0.3980	0.1884	-2.11	.068
Number	0.118320	0.006804	17.39	.000
s = 0.1499		R-sq = 97.4%		R-sq(adj) = 97.1%

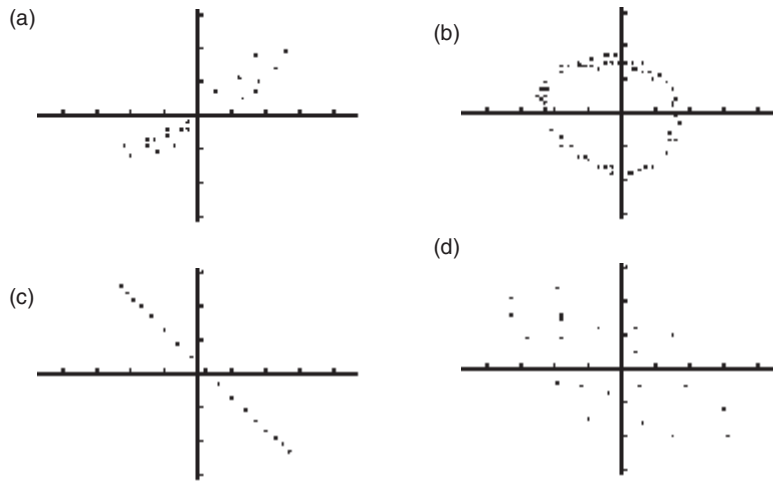
Based on the output,

- give the equation of the LSRL for predicting *crime rate* from *number*.
 - give the value of r , the correlation coefficient.
 - give the predicted *crime rate* for 20,000 casino employees.
14. A study was conducted in a mid-size U.S. city to investigate the relationship between the number of homes built in a year and the mean percentage appreciation for that year. The data for a 5-year period are as follows:

Number	110	80	95	70	55
Percent appreciation	15.7	10	12.7	7.8	10.4

- Obtain the LSRL for predicting appreciation from number of new homes built in a year.
- The following year, 85 new homes are built. What is the predicted appreciation?

- (c) How strong is the linear relationship between number of new homes built and percentage appreciation? Explain.
- (d) Suppose you didn't know the number of new homes built in a given year. How would you predict appreciation?
15. A set of bivariate data has $r^2 = 0.81$.
- (a) x and y are both standardized, and a regression line is fitted to the standardized data. What is the slope of the regression line for the standardized data?
- (b) Describe the scatterplot of the original data.
16. Estimate r , the correlation coefficient, for each of the following graphs:



17. The least-squares regression equation for the given data is $\hat{y} = 3 + x$. Calculate the sum of the squares residuals for the LSRL.

x	7	8	11	12	15
y	10	11	14	15	18

18. Many schools require teachers to have evaluations done by students. A study investigated the extent to which student evaluations are related to grades. Teacher evaluations and grades are both given on a scale of 100. The results for Professor Socrates (y) for 10 of his students are given below together with the average for each student.

x	40	60	70	73	75	68	65	85	98	90
y	10	50	60	65	75	73	78	80	90	95

- (a) Do you think student grades and the evaluations they give their teachers are related? Explain.
 - (b) What evaluation score do you think a student who averaged 80 would give Prof. Socrates?
19. Which of the following statements are true?
- (a) The correlation coefficient, r , and the slope of the regression line, b , always have the same sign.
 - (b) The correlation coefficient is the same no matter which variable is considered to be the explanatory variable and which is considered to be the response variable.
 - (c) The correlation coefficient is resistant to outliers.
 - (d) x and y are measured in inches, and r is computed. Now, x and y are converted to feet, and a new r is computed. The two computed values of r depend on the units of measurement and will be different.
 - (e) The idea of a correlation between height and gender is not meaningful because gender is not numerical.
20. A study of right-handed people found that the regression equation for predicting left-hand strength (measured in kg) from right hand strength is $\text{left hand strength} = 7.1 + 0.35 (\text{right hand strength})$.
- (a) What is the predicted left-hand strength for a right-handed person whose right hand strength is 12 kg.?
 - (b) Interpret the intercept and the slope of the regression line in the context of the problem.

3**CUMULATIVE REVIEW PROBLEMS**

1. Explain the difference between a statistic and a parameter.
2. TRUE–FALSE. The area under a normal curve between $z = 0.1$ and $z = 0.5$ is the same as the area between $z = 0.3$ and $z = 0.7$.
3. The following scores were achieved by students on a statistics test: 82, 93, 26, 56, 75, 73, 80, 61, 79, 90, 94, 93, 100, 71, 100, 60. Compute the mean and median for this data and explain why they are similar or different.
4. Is it possible for the standard deviation of a set of data to be negative? Zero? Explain.
5. For the test scores of problem #3, compute the 5-number summary and draw a boxplot of the data.

3 SOLUTIONS TO PRACTICE PROBLEMS

Multiple Choice

1. The correct answer is (d).

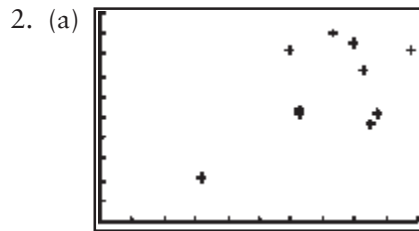
$$b = r \cdot \frac{s_y}{s_x} = (0.55) \left(\frac{0.75}{1.6} \right) = .26$$

2. The correct answer is (e). The value of a residual = actual value – predicted value = $25 - [2.35 + 0.86(29)] = -2.29$.
3. The correct answer is (a). $r^2 = (-0.58)^2 = 0.3364$. This is the *coefficient of determination*, which is the proportion of the variation in the response variable that is explained by the regression on the independent variable. Thus, about one-third (33.3%) of the variation in hours spent exercising can be explained by hours spent watching television.
4. The correct answer is (c). $\ln(y) = 1.64 - 0.88(3.1) = -1.088 \rightarrow y = e^{-1.088} = 0.337$
5. The correct answer is (c). The pattern is more or less random about 0, which indicates that a line would be a good model for the data. If the data are linearly related, we would expect them to have a non-zero correlation.

Free Response

1. $b = r \frac{s_y}{s_x} = (0.80) \left(\frac{11}{4} \right) = 2.2$, $a = \bar{y} - b\bar{x} = 20 - (2.2)(14.5) = -11.9$.

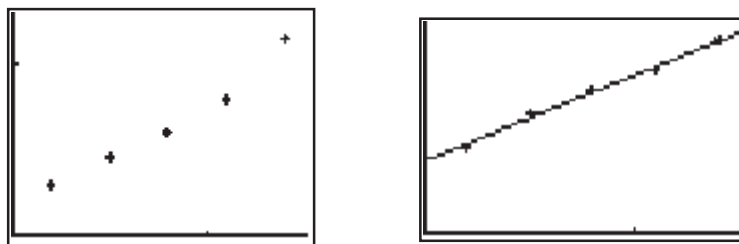
Thus, $\hat{y} = -11.9 + 2.2x$.



- (b) There seems to be a moderate positive relationship between the scores: students who did better on the first test tend to do better on the second, but the relationship isn't very strong; $r = 0.55$.
3. A line is not a good model for the data because the residual plot shows a definite pattern: the first 8 points have negative residuals and the last 8 points have positive residuals. The box is in a cluster

of points with positive residuals. We know that, for any given point, the residual equals actual value minus predicted value. Because $\text{actual} - \text{predicted} > 0$, we have $\text{actual} > \text{predicted}$, so that the regression equation is likely to underestimate the actual value.

4. The regression equation for predicting time from year is $\text{time} = 79.21 - 0.61(\text{year})$. We need $\text{time} = 60$. Solving $60 = 79.1 - 0.61(\text{year})$, we get $\text{year} = 31.3$. So, we would predict that times will drop under one minute in about 31 or 32 years. The problem with this is that we are extrapolating far beyond the data. Extrapolation is dangerous in any circumstance, and especially so 24 years beyond the last know time. It's likely that the rate of improvement will decrease over time.
5. A scatterplot of the data (graph on the left) appears to be exponential. Taking the natural logarithm of each y -value, the scatterplot (graph on the right) appears to be more linear.



Taking the natural logarithm of each y -value and finding the LSRL, we have $\ln(\#Roaches) = 0.914 + 0.108(\text{Days}) = 0.914 + 0.108(9) = 1.89$.
Then $\#Roaches = e^{1.89} = 6.62$.

6. The correlation between walking more and better health may or may not be causal. It may be that people who are healthier walk more. It may be that some other variable, such as general health consciousness, results in walking more and in better health. There may be a causal association, but in general, correlation is not causation.
7. Carla has reported the value of r^2 , the coefficient of determination. If she had predicted each girls grade based on the average grade only, there would have been a large amount of variability. But, by considering the regression of grades on socioeconomic status, she has reduced the total amount of variability by 72%. Because $r^2 = 0.72$, $r = 0.85$, which is indicative of a strong positive linear relationship between grades and socioeconomic status. Carla has reason to be happy.
8. (a) is false. Because

$$\sum (y - \hat{y})$$

equals 0 for the LSRL, there is no unique line for which this is true.

- (b) is true.
 (c) is true. In fact, this is the definition of the LSRL—it is the line that minimizes the sum of the residuals.
 (d) is true.

$$b = r \frac{s_y}{s_x} \text{ and } \frac{s_y}{s_x} \text{ is constant.}$$

- (e) is false. The slope of the regression lines tell you by how much the response variable changes *on average* for each unit change in the explanatory variable.
9. $\hat{y} = 26.211 - 0.25x = 26.211 - 0.25(73) = 7.988$. The residual for $x = 73$ is the actual value at 73 minus the predicted value at 73, or $y - \hat{y} = 7.9 - 7.961 = -0.061$.
10. (a) $r = +0.75$; the slope is positive and is the opposite of the original slope.
 (b) $r = -0.75$; the slope is negative.
 (c) $r = -0.75$; the slope is the same as the original slope.

11. We know that

$$b = r \frac{s_y}{s_x},$$

so that

$$r = b \frac{s_x}{s_y},$$

where $b = 2.7$. Also,

$$\frac{s_x}{s_y} = \frac{1}{\frac{s_y}{s_x}} = \frac{1}{3.33} = 0.30.$$

Thus, $r = (2.7)(0.30) = 0.81$, and $r^2 = 0.66$. The proportion of the variability that is not explained is $1 - r^2 = 1 - 0.66 = 0.33$.

12. Because the linear pattern will be stronger, the correlation coefficient will increase. The influential point pulls up on the regression line so that its removal would cause the slope of the regression line to decrease.
13. (a) $rate = -0.3980 + 0.1183(number)$
 (b) $r = \sqrt{0.974} = 0.987$
 (c) $rate = -0.3980 + 0.1183(20) = 1.97$ crimes per thousand employees.

14. (a) *Percentage appreciation* = $1.897 + 0.115(\text{number})$
 (b) *Percentage appreciation* = $1.897 + 0.115(85) = 11.67$.
 (c) $r = 0.82$, which indicates a strong linear relationship between the number of new homes built.
 (d) If the number of new homes built was unknown, your best estimate would be the average percentage appreciation for the 5 years. In this case, the average percentage appreciation is 11.3%. [For what it's worth, the average error (absolute value) using the mean to estimate appreciation is 2.3; for the regression line, it's 1.3]
15. (a) If $r^2 = 0.81$, then $r = \pm 0.9$. The slope of the regression line for the standardized data is either 0.9 or -0.9 .
 (b) If $r = +0.9$, the scatterplot shows a strong positive linear pattern between the variables. Values above the mean on one variable tend to be above the mean on the other, and values below the mean on one variable tend to be below the mean on the other. If $r = -0.9$, there is a strong negative linear pattern to the data. Values above the mean on one variable are associated with values below the mean on the other
16. (a) $r = 0.8$
 (b) $r = 0.0$
 (c) $r = -1.0$
 (d) $r = -0.5$
17. Each of the points lies on the regression line \rightarrow every residual is 0 \rightarrow the sum of the squared residuals is 0.
18. (a) $r = 0.90$ for these data indicating that there is a strong positive linear relationship between student averages and evaluations of Prof. Socrates. Furthermore, $r^2 = 0.90$, which means that most of the variability in student evaluations can be explained by the regression of student evaluations on student average.
 (b) If y is the evaluation score of Prof. Socrates and x is the corresponding average for the student who gave the evaluation, then $\hat{y} = -29.3 + 1.34x$. If $x = 80$, then $\hat{y} = -29.3 + 1.34(80) = 77.9$, or 78.
19. (a) True, because
- $$b = r \frac{s_y}{s_x} \text{ and } \frac{s_y}{s_x}$$
- is positive.
- (b) True. r is the same if explanatory and response variables are reversed. This is not true, however, for the slope of the regression line.
 (c) False. Because r is defined in terms of the means of the x and y variables, it is not resistant.
 (d) False. r does not depend on the units of measurement.

(e) True. The definition of

$$r, r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

necessitates that the variables be numerical, not categorical.

20. (a) *Left hand strength* = $7.1 + 0.35(12) = 11.3$ kg.
 (b) Intercept: the predicted left-hand strength of a person that has zero right-hand strength is 7.1 kg.
 Slope: On average, left-hand strength increases by 0.35 kg for each 1 kg increase in right-hand strength.

3 SOLUTIONS TO CUMULATIVE REVIEW PROBLEMS

1. A *statistic* is a measurement that describes a sample. A *parameter* is a value that describes a population.
2. FALSE. For an interval of fixed length, there will be a greater proportion of the area under the normal curve if the interval is closer to the center than if it is removed from the center. This is because the normal distribution is mound-shaped, which implies that the terms tend to group more in the center of the distribution than away from the center.
3. The mean is 77.1, and the median is 79.5. The mean is lower than the median because it is not resistant to extreme values—it is pulled down by the 26, but that value does not affect the median.
4. By definition,

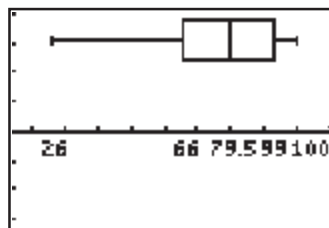
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}.$$

Because $n > 1$ and

$$\sum (x - \bar{x})^2 \geq 0, s$$

cannot be negative. It can be zero, but only if $x = \bar{x}$ for all values of x .

5. The 5-number summary is: [26, 66, 79.5, 99, 100].



Chapter 6

Design of a Study

Sampling, Surveys, and Experiments



Main concepts: *samples and sampling, surveys, sampling bias, experiments and observational studies, statistical significance, completely randomized design, matched pairs design, blocking*

SAMPLES

In the previous two chapters we concentrated on the analysis of data at hand—we didn't worry much about how the data came into our possession. The last part of this book deals with statistical inference—making statements about a population based on samples drawn from the population. In both data analysis and inference, we would like to believe that our analyses, or inferences, are meaningful. If we make a claim about a population based on a sample, we want that claim to be true. Our ability to do meaningful analyses and make reliable inferences is a function of the data we observe. To the extent that the sample data we deal with accurately represent the population of interest, we are on solid ground. No interpretation of data that are poorly collected or systematically biased will be meaningful. We need to understand how to gather quality data before proceeding on to inference. In this chapter, we study techniques for gathering data in such a way that we have reasonable confidence that they are representative of our population of interest.

Census

The alternative to sampling to gather data about a population is to conduct a **census**, a procedure by which every member of a population is

selected for study. Doing a census, especially when the population of interest is quite large, is often impractical, too time consuming, or too expensive. Interestingly enough, relatively small samples can give quite good estimates of population values if the samples are selected properly. For example, it can be shown that approximately 1500 randomly selected voters can give reliable information about the entire voting population of the United States.

The goal of sampling is to produce a **representative sample**, one that has the essential characteristics of the population being studied, and is free of any type of systematic bias. One of the main ways we attempt to make a sample representative is to use some sort of random process in selecting our sample.

Probability Sample

A list of all members of the population from which we can draw a sample is called a **sampling frame**. A **probability sample** is one in which each member of the population has a known probability of being in the sample. Each member of the population may or may not have an equal chance of being selected. Probability samples are used to avoid the bias that can arise in a nonprobability sample (such as when a researcher selects the subjects she will use). Probability samples use some sort of random mechanism to choose the members of the sample. Following are some types of probability samples.

- **random sample:** Each member of the population is equally likely to be included.
- **simple random sample (SRS):** A sample of a given size chosen is in such a way that every possible sample of that size is equally likely to be chosen. A sample can be a *random sample* and yet not be a *simple random sample*.
- **systematic sample:** The first member of the sample is chosen according to some random procedure, and then the rest are chosen according to some well-defined pattern. For example, if you wanted 100 people in your sample to be chosen from a list of 10,000 people, you could randomly select one of the first 100 people and then select every 100th name on the list after that.
- **stratified random sample:** This is a sample in which subgroups of the sample, *strata*, appear in approximately the same proportion in the sample as they do in the population. For example, assuming males and females were in equal proportion in the population, you might structure your sample to be sure that males and females were in equal proportion in the sample. For a sample of 100 individuals, you would select an SRS of 50 females from all the females and an SRS of 50 males from all the males.

example: You are going to conduct a survey of your senior class concerning plans for graduation. You want a 10% sample of the class.

Describe a procedure by which you could use a systematic sample to obtain your sample and explain why this sample isn't a simple random sample. Is this a random sample?

solution: One way would be to obtain first an alphabetical list of all the seniors. Use a random number generator (such as a table of random digits or a scientific calculator with a random digits function) to select one of the first 10 names on the list. Then proceed to select every 10th name on the list after the first.

Note that this is not an SRS because not every possible sample of 10% of the senior class is equally likely. For example, people next to each other in the list can't both be in the sample. Theoretically, the first 10% of the list could be the sample if it were an SRS. This clearly isn't possible.

Before to the first name being randomly selected, every member of the population has an equal chance to be selected for the sample. Hence, this is a random sample, although it is not a simple random sample.

example: You are sampling from a population with mixed ethnicity. The population is 45% Caucasian, 25% Asian American, 15% Latino, and 15% African American. How would a *stratified random sample* of 200 people be constructed.

solution: You want your sample to mirror the population in terms of its ethnic distribution. Accordingly, from the Caucasians, you would draw an SRS of 90 (that's 45%), an SRS of 50 from the Asian Americans, an SRS of 30 from the Latinos, and an SRS of 30 from the African Americans.

Of course, not all samples are probability samples. At times, people try to obtain samples by processes that are nonrandom but still hope, through design or faith, that the resulting sample is representative. The danger in all nonprobability samples is that some (unknown) bias may affect the degree to which the sample is representative. That isn't to say that random samples can't be biased, just that we have a better chance of avoiding *systematic* bias. Some types of nonrandom samples are:

- **self-selected sample** or **voluntary response sample:** People choose whether or not to participate in the survey. A radio call-in show is a typical voluntary response sample.
- **convenience sampling:** The pollster obtains the sample any way he can, usually with the ease of obtaining the sample in mind. For example, handing out questionnaires to every member of a given class at school would be a convenience sample. The key issue here is that the surveyor makes the decision who to include in the sample.
- **quota sampling:** The pollster attempts to generate a representative sample by choosing sample members based on matching individual characteristics to known characteristics of the population. This is sim-

ilar to a stratified random sample, only the process for selecting the sample are nonrandom.

SAMPLING BIAS

Any sample may contain bias. What we are trying to avoid is **systematic bias**, the tendency for our results to favor systematically one outcome over another. This can occur through faulty sampling techniques or through faults in the actual measurement instrument.

Undercoverage

One type of bias results from **undercoverage**. This happens when some part of the population being sampled is somehow excluded. This can happen when the sampling frame (the list from which the sample will be drawn) isn't the same as the target population. It can also occur when part of the sample selected fails to respond for some reason.

example: A pollster conducts a telephone survey to gather opinions of the general population about welfare. Persons on welfare too poor to be able to afford a telephone are certainly interested in this issue, but will be systematically excluded from the sample. The resulting sample will be biased because of the exclusion of this group.

Voluntary Response Bias

Voluntary response bias occurs with self-selected samples. Persons who feel most strongly about an issue are most likely to respond. **Non-response bias**, the possible biases of those who choose not to respond, is a related issue.

example: You decide to find out how your neighbors feel about the neighbor who seems to be running a car repair shop on his front lawn. You place a questionnaire in every mailbox within sight of the offending home and ask the people to fill it out and return it to you. About $\frac{1}{2}$ of the neighbors return the survey, and 95% of those who do say that they find the situation intolerable. We have no way of knowing the feelings of the 50% of those who didn't return the survey—they may be perfectly happy with the “bad” neighbor. Those who have the strongest opinions are those most likely to return your survey—and they may not represent the opinions of all. Most likely they do not.

example: In response to a question once posed by in Ann Landers's advice column, some 70% of respondents wrote that they would choose to not have children if they had the choice to do it over again.

This is most likely representative only of those parents who were having a *really* bad day with their children when they decided to respond to the question.

Wording Bias

Wording bias occurs when the wording of the question itself influences the response in a systematic way. A number of studies have demonstrated that welfare gathers more support from a random sample of the public when it is described as “helping people until they can better help themselves” than when it is described as “allowing people to stay on the dole.”

example: Compare the probable responses to the following ways of phrasing a question.

- (i) “Do you support a woman’s right to make medical decisions concerning her own body?”
- (ii) “Do you support a woman’s right to kill an unborn child?”

It’s likely that (i) is designed to show that people are in favor of a woman’s right to choose an abortion and that (ii) is designed to show that people are opposed to that right. Whoever wrote either question would probably argue that their response reflects society’s attitudes toward abortion.

Response Bias

Response bias arises in a variety of ways. The respondent may not give truthful responses to a question (perhaps they are ashamed of the truth); the respondent may fail to understand the question (you ask if a person is educated but fail to distinguish between levels of education); the respondent desires to please the interviewer (questions concerning race relations may well solicit different answers depending on the race of the interviewer); the ordering of the question may influence the response (“Do you prefer A to B?” may get different responses than “Do you prefer B to A?”)

example: What form of bias do you suspect in the following situation? You are a school principal and want to know students’ level of satisfaction with the counseling services at your school. You direct one of the school counselors to ask her next 25 counsees how favorably they view the counseling services at the school.

solution: A number of things would be wrong with the data you get from such a survey. First, the sample is nonrandom—it is a sample of convenience obtained by selecting 25 consecutive counsees. They may or may not be representative of students who use counseling service. You don’t know.

Second, you are asking people who are seeing their counselor about their opinion of counseling. You will probably get a more favorable view of the counseling services than you would if you surveyed the general population of the school (would students really unhappy with the counseling services be seeing their counselor?). Also, because the counselor is administering the questionnaire, the respondents would have a tendency to want to please the interviewer. The sample certainly suffers from undercoverage—only a small subset of the general population is actually being interviewed. What do those *not* being interviewed think of the counseling?

EXPERIMENTS AND OBSERVATIONAL STUDIES

Statistical Significance

One of the main purposes of a study is to help us determine cause and effect. We do this by looking for differences between groups that are so great that we cannot reasonably attribute the difference to chance. We say that a difference between what we would expect to find if there were no treatment and what we actually found is **statistically significant** if the difference is too great to attribute to chance. We discuss numerical methods of determining *significance* in Chapters 9–13.

An **experiment** is a study in which the researcher imposes some sort of treatment on the **experimental units** (which can be human—usually called **subjects** in that case). The idea is to see to what extent, if any, the treatment (the **explanatory variable**) has on the **response variable**. For example, a researcher might vary the reward to a work group to see how that affects the subject’s ability to perform a particular task.

An **observational study**, on the other hand, simply observes and records behavior but does not attempt to impose a treatment in order to manipulate the response.



Exam Tip: The distinction between an experiment and an observational study is an important one. There is a reasonable chance that you will be asked to show you understand this distinction on the exam. Be sure this section makes sense to you.

example: A group of 60 exercisers are classified as “walkers” or “runners.” A longitudinal study (one conducted over time) is conducted to see if there are differences between the groups in terms of their scores on a wellness index. This is an *observational study* because, although the two groups differ in an important respect, the researcher is not manipulating any treatment. “Walkers” and “runners” are simply observed and measured. Note that the groups in this study are self-selected. That is, they were already in their groups before the

study began. The researchers just noted their group membership and proceeded to make observations. There may be significant differences between the groups in addition to the variable under study.

example: A group of 60 volunteers who do not exercise are randomly assigned a fitness program. One group of 30 is enrolled in a daily walking program, and the other group is put into a running program. After a period of time, the two groups are compared based on their scores on a wellness index. This is an *experiment* because the researcher has imposed the treatment (walking or running) and then measured the effects of the treatment on a defined response.

It may be, even in a controlled experiment, that the measured response is a function of variables present in addition to the treatment variable. A **confounding variable** is one that has an effect on the outcomes of the study but whose effects cannot be separated from those of the treatment variable. A **lurking variable** is one that has an effect on the outcomes of the study but whose influence was not part of the investigation.

example: A study is conducted to see if Yummy Kibble dog food results in shinier coats on Golden Retrievers. It's possible that the dogs with shinier coats have them because they have owners who are more conscientious in terms of grooming their pets. Both the dog food and the conscientious owners could contribute to the shinier coats. The variables are **confounded** because their effects cannot be separated.

A well-designed study attempts to anticipate confounding variables in advance and **control** for them. **Statistical control** refers to a researcher holding constant variables not under study that might have an influence on the outcomes.

example: You are going to study the effectiveness of SAT preparation courses on SAT score. You know that better students tend to do well on SAT tests. You could control for this by running your study with groups of "A" students, "B" students, etc.

Control is often considered to be one of the three basic principles of experimental design. The other two basic principles are **randomization** and **replication**.

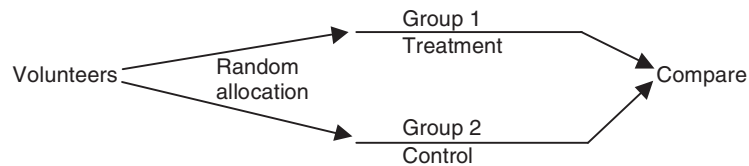
The purpose of *randomization* is to equalize groups so that the effects of lurking variables are equalized among groups. *Randomization* involves the use of chance (like a coin flip) to assign subjects to treatment and control groups. The hope is that the groups being studied will differ systematically *only* in the effects of the treatment variable. Although individuals within the groups may vary, the idea is to make the groups as alike as possible except for the treatment variable.

Replication involves repeating the experiment on enough subjects (or units) to reduce the effects of chance variation on the outcomes. For

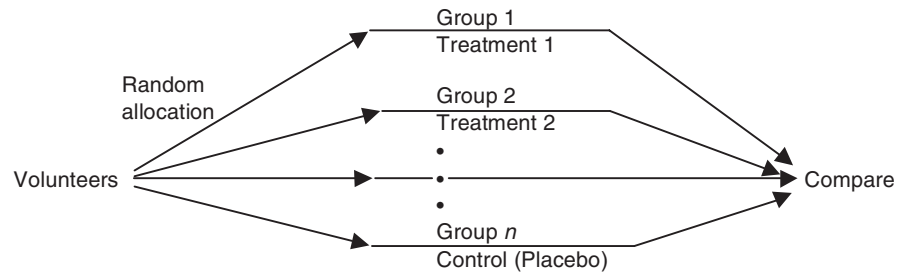
example, we know that the number of boys and girls born in a year are approximately equal. A small hospital with only 10 births a year is much more likely to vary dramatically from 50% each than a large hospital with 500 births a year.

Completely Randomized Design

A **completely randomized design** for a study involves three essential elements: random allocation of subjects to treatment and control groups; administration of different treatments to each randomized group (in this sense we are calling a control group a “treatment”); and some sort of comparison of the outcomes from the various groups. A standard diagram of this situation is the following:



There may be several different treatment groups (different levels of a new drug, for example) in which case the diagram could be modified to have several lines instead of just two. The control group can either be a older treatment (like a medication currently on the market) or a **placebo**, a dummy treatment. A diagram for this situation might look like this:

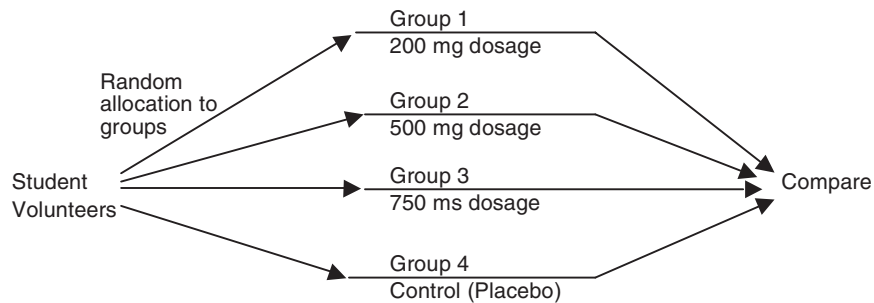


Remember that each group must have enough subjects so that the replication condition is met. The purpose of the *placebo* is to separate genuine treatment effects from possible subject responses due to simply being part of an experiment. Placebos are *not* necessary if a new treatment is being compared to a treatment whose effects have been previously experimentally established. In that case, the old treatment can serve as the control.

example: Three hundred graduate students in psychology (by the way, a huge percentage of subjects in published studies are psychology

graduate students) volunteer to be subjects in an experiment whose purpose is to determine what dosage level of a new drug has the most positive effect on a performance test. There are three levels of the drug to be tested: 200 mg, 500 mg, and 750 mg. Design a completely randomized study to test the effectiveness of the various drug levels.

solution: There are three levels of the drug to be tested: 200 mg, 500 mg, and 750 mg. A placebo control can be included although, strictly speaking, it isn't necessary if our purpose is to compare the three dosage levels. We need to randomly allocate the 300 students to each of four groups of 75 each: one group will receive the 200 mg dosage; one will receive the 500 mg dosage; one will receive the 750 mg dosage; and one will receive a placebo (if included). No group will know which treatment they are receiving (all the pills look the same), nor will the test personnel who come in contact with them know which subjects received which pill. Each group will complete the performance test and the results of the various groups will be compared. This design can be diagrammed as follows:



Double-Blind Experiments

In the example above, it was explained that neither the subjects nor the researchers knew who was receiving which dosage, or the placebo. A study is said to be **double-blind** when neither the subjects (or experimental units) nor the researchers know which group(s) is/are receiving each treatment or control. The reason for this is that, on the part of subjects, simply knowing that they are part of a study may affect the way they respond, and, on the part of the researchers, knowing which group is receiving which treatment can influence the way in which they evaluate the outcomes. Our worry is that the treatment and control groups will differ by something other than the treatment unless the study is double-blind.

Randomization

There are two main procedures for performing a *randomization*. They are:

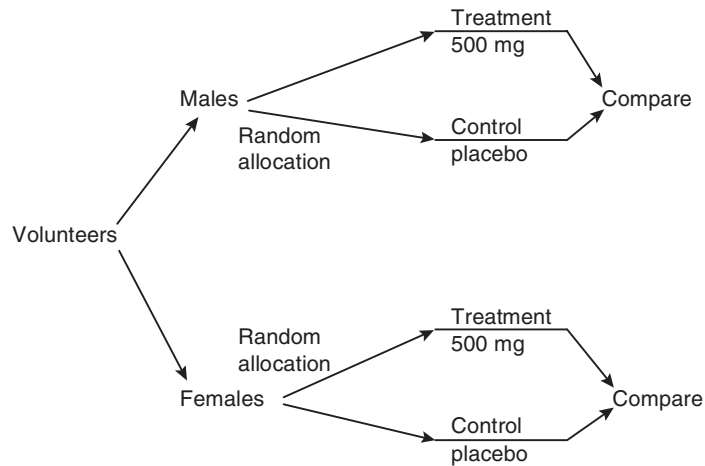
- **Tables of random digits.** Most textbooks contain tables of random digits. These are usually tables where the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 appear in random order (well, as random as most things get, anyhow). That means that, as you move through the table, each digit should appear with probability 1/10, and each entry is independent of the others (knowing what came before doesn't help you make predictions about what comes next).
- **Calculator “rand” functions.** The TI-83 calculator has several random functions: *rand*, *randInt*, *randNorm*, and *randbin*. If you wanted to generate a set of random digits similar to the random digit table described above, you would enter *randInt(0,9)* and press ENTER. If you wanted 10 random integers stored in a list (say *L1*), you would enter *randInt(0,9,10)STO→L1*.

Digression: Although the calculator is an electronic device, it is just like a random digit table in that if two different people enter the list in the same place, they will get the same sequence of numbers. You “enter” the list on the calculator by “seeding” the calculator as follows: “(Some number) *STO→rand*”. If five different people using the same model of calculator entered, say, “18 *STO→rand*” and then began to press ENTER repeatedly, they would all generate *exactly* the same list of random digits.

We use tables of random digits and/or the calculator in Chapter 7 when we discuss simulation.

Block Design

Earlier we discussed the need for **control** in a study and identified **randomization** as the main method to control for lurking variables—variables that might influence the outcomes in some way but are not considered in the design of the study. Another type of control involves variables we know might influence the outcome of a study. Suppose we suspect, as in our previous example, that the performance test varies by gender as well as by dosage level of the test drug. That is, we suspect that gender is a *confounding variable* (its effects cannot be separated from the effects of the drug). To control for the effects of gender on the performance test, we utilize what is known as a **block design**. A **block design** involves doing a completely randomized experiment *within* each block. In this case, that means that each level of the drug would be tested within the group of females and within the group of males. To simplify the example, suppose that we were only testing one level (say 500 mg) of the drug versus a placebo. The experimental design, blocked by gender, could then be diagrammed as follows.



Randomization and block designs each serve a purpose. It's been said that you *block* to control for the variables you know about and *randomize* to control for the ones you don't. Note that your interest here is in studying the effect of the treatment within the population of males and within the population of females, *not* to compare the effects on men and women.

Matched Pairs Design

A particular block design of interest is the **matched pairs** design. One possible matched pairs design involves before and after measurements on the same subjects. In this case, each subject becomes a block in which the experiment is conducted. Another type of matched pairs involves pairing the subjects in some way (matching on, say, height, race, age, etc).

example: A study is instituted to determine the effectiveness of training teachers to teach AP Statistics. A pretest is administered to each of 23 prospective teachers who subsequently undergo a training program. When the program is finished, the teachers are given a post-test. A score for each teacher is arrived at by subtracting their pretest score from their post-test score. This is a matched pairs design because two scores are paired for each teacher.

example: One of the questions on the 1997 AP Exam in Statistics asked students to design a study to compare the effects of differing formulations of fish food on fish growth. Students were given a room with eight fish tanks. The room had a heater at the back of the room, a door at the front center of the room, and windows at the front sides of the room. The most correct design involved blocking so that the two tanks nearest the heater in the back of the room were in a single

block, the two away from the heater in a second block, the two in the front near the door in a third, and the two in the front near the windows in a fourth. This matching had the effect of controlling for known environmental variations in the room caused by the heater, the door, and the windows. Within each block, one tank was randomly assigned to receive one type of fish food and the other tank received the other. The blocking controlled for the known effects of the environment in which the experiment was conducted. The randomization controlled for unknown influences that might be present in the various tank locations.



Exam Tip: You need to be clear on the distinction between the purposes for “blocking” and randomizing. If you are asked to describe an experiment involving blocking, be sure to remember to randomize treatments within blocks.

RAPID REVIEW

1. You are doing a research project on attitudes toward fast food and decide to use as your sample the first 25 people to enter the door at the local FatBurgers restaurant. Which of the following are true of this sample?
 - a. It is a systematic sample.
 - b. It is a convenience sample.
 - c. It is a random sample.
 - d. It is a simple random sample.
 - e. It is a self-selected sample.

Answer: Only (b) is correct. (a), (c), and (d) are all probability samples, which rely on some random process to select the sample, and there is nothing random about the selection process in this situation. (e) is incorrect because, although the sample members voluntarily entered FatBurgers, they haven’t volunteered to respond to a survey.

2. How does an experiment differ from an observational study?

Answer: In an experiment, the researcher imposes some treatment on the subjects (or experimental units) in order to observe a response. In an observational study, the researcher simply observes and compares, but does not impose a treatment.

3. What are the three key components of an experiment? Explain each.

Answer: The three components are randomization, control, and replication. You randomize to be sure that the response does not systematically favor one outcome over another. The idea is to equalize groups as much as possible so that differences in response are attrib-

utable to the treatment variable alone. Control is designed to hold confounding variables constant (such as the placebo effect). Replication ensures that the experiment is conducted on sufficient numbers of subjects to minimize the effects of chance variation.

4. Your local pro football team has just suffered a humiliating defeat at the hands of their arch rival. A local radio sports talk show conducts a call-in poll on whether or not the coach should be fired. What is the poll likely to find?

Answer: The poll is likely to find that, overwhelmingly, respondents think the coach should be fired. This is a *voluntary* response poll, and we know that such a poll is most likely to draw a response from those who feel most strongly about the issue being polled. Fans who bother to vote in a call-in poll such as this are most likely upset at their team's loss and are looking for someone to blame—this gives them the opportunity.

5. It is known that exercise and diet both influence weight loss. Your task is to conduct a study of the effects of diet on weight loss. Explain the concept of *blocking* as it relates to this study.

Answer: If you did a completely randomized design for this study using diet as the treatment variable, it's very possible that your results would be confounded by the effects of exercise. Because you are aware of this, you would like to control for the effects of exercise. Hence, you *block* by exercise level. You might define, say, three blocks by level of exercise (very active, active, not very active) and do a completely randomized study within each of the blocks. Because exercise level is held constant, you can be confident that differences between treatment and control groups within each block are attributable to diet, not exercise.

6. Explain the concept of a *double-blind* study and why it is important.

Answer: A study is *double-blind* if neither the subject of the study nor the researchers are aware of who is in the treatment group (or at what level) and who is in the control group. This is to control for the well-known effect of people to (subconsciously) attempt to respond in the way they think they should.

3 PRACTICE PROBLEMS

Multiple Choice

1. Data were collected in 20 cities on the percentage of women in the workforce. Data were collected in 1990 and again in 1994. Gains, or losses, in this percentage were the measurement upon which the studies' conclusions were to be based. What kind of design was this?

- I. A matched pairs design
 - II. An observational study
 - III. An experiment using a block design
 - a. I only
 - b. II only
 - c. III only
 - d. I and III only
 - e. I and II only
2. You want to do a survey of members of the senior class at your school and want to select a *simple random sample*. You intend to include 40 students in your sample. Which of the following approaches will generate a simple random sample?
- a. Write the name of each student in the senior class on a slip of paper and put the papers in a container. Then randomly select 40 slips of paper from the container.
 - b. Assuming the classes are unlined, select two classes at random and include those students in your sample.
 - c. From a list of all seniors, select one of the first 10 names at random. The select every n th name on the list until you have 40 people selected.
 - d. Select the first 40 seniors to pass through the cafeteria door at lunch.
 - e. Randomly select 10 students from each of the four senior calculus classes.
3. Which of the following are important in designing an experiment?
- I. Control of all variables that might have an influence on the response variable.
 - II. Randomization of subjects to treatment groups.
 - III. Use a large number of subjects to control for small-sample variability.
 - a. I only
 - b. I and II only
 - c. II and III only
 - d. I, II, and III
 - e. II only
4. Your company has developed a new treatment for acne. You think men and women might react differently to the medication, so you separate them into two groups. Then the men are randomly assigned to two groups and the women are randomly assigned to two groups. One of the two groups is given the medication and the other is given a placebo. The basic design of this study is
- a. completely randomized
 - b. comparative randomized, blocked by gender
 - c. completely randomized, stratified by gender

- d. randomized, blocked by gender and type of medication
 - e. a matched pairs design
5. A *double-blind* design is important in an experiment because
- a. There is a natural tendency for subjects in an experiment to want to please the researcher.
 - b. It helps control for the placebo effect.
 - c. Evaluators of the responses in a study can influence the outcomes if they know which subjects are in the treatment group and which are in the control group.
 - d. Subjects in a study might react differently if they knew they were receiving an active treatment or a placebo.
 - e. All of the above are reasons why an experiment should be *double-blind*.

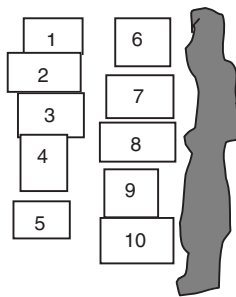
Free Response

1. You are interested in the extent to which ingesting vitamin C inhibits getting a cold. You identify 300 volunteers, 150 of which have been taking more than 1000 mg of vitamin C a day for the past month, and 150 of whom have not taken vitamin C at all during the past month. You record the number of colds during the following month for each group and find that the vitamin C group had significantly less colds. Is this an experiment or an observational study? Explain. What do we mean in this case when we say that the finding was *significant*?
2. Design an experiment that employs a *completely randomized design* to study the question of whether or not taking large doses of vitamin C is effective in reducing the number of colds.
3. A survey of physicians found that some doctors gave a placebo rather than an actual medication to patients who experience pain symptoms for which no physical reason can be found. If the pain symptoms were reduced, the doctors concluded that there was no real physical basis for the complaints. Do the doctors understand *the placebo effect*? Explain.
4. Explain how you would use a table of random digits to help obtain a systematic sample of 10% of the names on an alphabetical list of voters in a community. Is this a random sample? Is it a simple random sample?
5. The *Literary Digest* magazine, in 1936, predicted that Alf Landon would defeat Franklin Roosevelt in the presidential election that year based on a sample of some 10 million people who were encouraged to return their preferences. They were wrong by something like

- 19 percentage points. They received back some 2.3 million of the 10 million ballots sent out. What might have caused their error in prediction?
6. Interviewers, after the 9/11 attacks, ask a group of Arab Americans if they trust the administration to make efforts to counter anti-Arab activities. If the interviewer was of Arab descent, 42% responded “yes” and if the interviewer was of non-Arab descent, 55% responded “yes.” What seems to be going on here?
 7. There are three classes of statistics at your school, each with 30 students. You want to select a simple random sample of 15 students from the 90 students as part of an opinion-gathering project for your social studies class. Describe a procedure for doing this.
 8. Question #1 above stated, in part: “You are interested in the extent to which ingesting vitamin C inhibits getting a cold. You identify 300 volunteers, 150 of which have been taking more than 1000 mg of vitamin C a day for the past month, and 150 of whom have not taken vitamin C at all during the past month. You record the number of colds during the following month for each group and find that the vitamin C group had significantly less colds.” Explain the concept of *confounding* in the context of this problem and give an example of how it might have affected the finding that vitamin C group has less colds.
 9. A shopping mall wants to know about the attitudes of all shoppers who visit the mall. On a Wednesday morning, the mall places 10 interviewers at a variety of places in the mall and asks questions of shoppers as they pass by. Comment on any bias that might be inherent with this approach.
 10. Question #2 above asked you to design a *completely randomized experiment* for the situation presented in question #1. That is, to design an experiment that uses treatment and control groups to see if the groups differed in terms of the number of colds suffered by users of 1000 mg a day of vitamin C and those that didn’t use vitamin C. Question #8 asked you about possible *confounding variables* in this study. Given that you believe that both general health habits and use of vitamin C might explain a reduced number of cold, design an experiment to determine the effectiveness of vitamin C taking into account general health habits. You may assume your volunteers vary in their history of vitamin C use.
 11. You have developed a weight-loss treatment that involves a combination of exercise and diet pill. The treatment has been effective with subjects who have used a regular dose of the pill of 200 mg, when exercise level is held constant. There is some indication that higher

doses of the pill will promote even better results, but you are worried about side effects if the dosage becomes too great. Assume you have 400 overweight volunteers for your study, who have all been on the same exercise program, but who have not been taking any kind of diet pill. Design a study to evaluate the relative effects of 200-mg, 400-mg, 600-mg, and 800-mg daily dosage of the pill.

12. You are going to study the effectiveness of three different SAT preparation courses. You obtain 60 high school juniors as volunteers to participate in your study. You want to assign each of the 60 students, at random, to one of the three programs. Describe a procedure for assigning students to the programs if
- you want there to be an equal number of students taking each course.
 - you want each student to be assigned independently to a group. That is, each student should have the same probability of being in any of the three groups.
13. A researcher want to obtain a sample of 100 teachers who teach in high schools at various economic levels and has access to a list of teachers in several schools for each of the levels. She has identified four such levels (A, B, C, and D) that comprise 10%, 15%, 45%, and 30% of the schools in which the teachers work. Describe what is meant by a *stratified random sample* in this situation and discuss how she might obtain it.
14. You are testing for sweetness in five varieties of strawberry. You have 10 plots available for testing. The 10 plots are arranged in two side-by-side groups of five. A river runs along the edge of one of the groups of five plots something like the following (the available plots are numbered 1–10).



You decide to control for the possible confounding effect of the river by deciding to plant one of each type of strawberry in plots 1–5 and one of each type in plots 6–10 (that is, you block to control for the river). Then, within each block, you randomly assign one type of strawberry to each of the five plots within the block. What is the purpose of randomization in this situation?

15. Look at problem 14 again. It is the following year, and you now have only two types of strawberries to test. Faced with the same physical conditions you had in problem 14, and given that you are concerned that differing soil conditions (as well as proximity to the

river) might affect sweetness, how might you block the experiment to produce the most reliable results?

16. A group of volunteers, who had never been in any kind of therapy, were randomly separated into two groups, one of which received an experimental therapy to improve self-concept. The other group, the control group, received the traditional therapy. The subjects were not informed of which therapy they were receiving. Psychologists who specialize in self-concept issues evaluated both groups after training for self-concept, and the self-concept scores for the two groups were compared. Could this experiment have been *double-blind*? Explain. If it wasn't *double-blind*, what might have been the impact on the results?
17. You want to determine how students in your school feel about a new dress code for school dances. One faction in the student council, call them group A, wants to word the question, "As one way to help improve student behavior at school sponsored events, do you feel that there should be a dress code for school dances?" Another group, group B, prefers, "Should the school administration be allowed to restrict student rights by imposing a dress code for school dances?" Which group do you think favors a dress code and which opposes it? Explain.
18. A study of reactions to different types of billboard advertising is to be carried out. Two different types of ads (call them Type I and Type II) for each product will be featured on numerous billboards. The organizer of the campaign is concerned that communities representing different economic strata will react differently to the ads. There are three communities where billboards will be placed and they have been identified as Upper Middle, Middle, and Lower Middle. Four billboards are available in each of the three communities. Design a study to compare the effectiveness of the two types of advertising taking into account the communities involved.
19. In 1976, Shere Hite published a book entitled *The Hite Report on Female Sexuality*. The conclusions reported in the book were based on 3000 returned surveys from some 100,000 sent out to, and distributed by, various women's groups. The results were that women were highly critical of men. In what way might the author's findings have been biased.
20. You have 26 women available for a study: Annie, Betty, Clara, Darlene, Edie, Fay, Grace, Helen, Ina, Jane, Koko, Laura, Mary, Nancy, Ophelia, Patty, Quincy, Robin, Suzy, Tina, Ulla, Vivien, Wanda, Xena, Yolanda, and Zoe. The women need to be divided into four groups for the purpose of the study. Explain how you could use a table of random digits to make the needed assignments.

3 CUMULATIVE REVIEW PROBLEMS

- The 5-number summary for a set of data is [52,55,60,63,85]. Is the mean most likely to be *less than* or *greater than* the median?
- Pamela selects a random sample of 15 of her classmates and computes the mean and standard deviation of their pulse rates. She then uses these values to predict the mean and standard deviation of the pulse rates for the entire school. Which of these measures are *parameters* and which are *statistics*?
- Consider the following set of values for a dataset: 15, 18, 23, 25, 25, 27, 28, 29, 35, 46, 55. Does this data set have any *outliers* if we use an outlier rule that
 - is based on the median?
 - is based on the mean?
- For the dataset of problem #3 above, what is z_{55} ?
- A study examining factors that contributes to a strong college GPA finds that 62% of the variation in college GPA can be explained by SAT score. What name is given to this statistic and what is the correlation (r) between SAT score and college GPA?

3 SOLUTIONS TO PRACTICE PROBLEMS

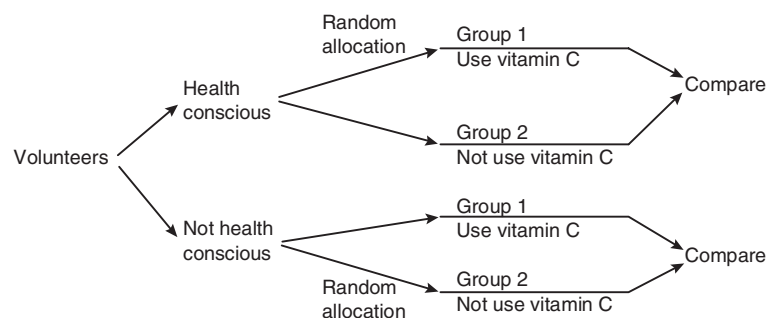
Multiple Choice

- The correct answer is (e). The data are paired because there are two measurements on each city so the data are not independent. There is no treatment being applied, so this is an observational study. Matched pairs is one type of block design, but this is NOT an experiment, so III is false.
- The answer is (a). In order for this to be a SRS, all samples of size 40 must be equally likely. None of the other choices do this [and choice (d) isn't even random]!
- The correct answer is (d). These three items represent the three essential parts of an experiment: control, randomization, and replication.
- The correct answer is (b). You block men and women into different groups because you are concerned that differential reactions to the medication may confound the results. It is not completely randomized because it is blocked.
- The correct answer is (e).

Free Response

1. It's an observational study because the researcher didn't provide a treatment, but simply observed different outcomes from two groups with at least one different characteristic. Participants self-selected themselves into either the vitamin C group or the nonvitamin C group. To say that the finding was significant in this case means that the difference between the number of colds in the vitamin C group and in the nonvitamin C group was too great to attribute to chance—it appears that something besides random variation may have accounted for the difference.
2. Identify 300 volunteers for the study, preferably none of whom have been taking vitamin C. Randomly split the group into two groups of 150 participants each. One group can be randomly selected to receive a set dosage of vitamin C each day for a month and the other group to receive a placebo. Neither the subjects nor those that administer the medication will know which subjects received the vitamin C and which received the placebo (that is, the study should be *double blind*). During the month following the giving of pills, you can count the number of colds within each group. Your measurement of interest is the difference in the number of colds between the two groups.
3. The doctors probably did not understand the placebo effect. We know that, sometimes, a real effect can occur even from a placebo. If people believe they are receiving a real treatment, they will often show a change. But without a control group, we have no way of knowing if the improvement would not have been even more significant with a real treatment. The *difference* between the placebo score and the treatment score is what is important, not one or the other.
4. If you want 10% of the names on the list, you need every 10th name for your sample. Number the first ten names on the list 0,1,2, . . . , 9. Pick a random place to enter the table of random digits and note the first number. The first person in your sample is the person among the first 10 on the list corresponds to the number chosen. Then pick every 10th name on the list after that name. This is a random sample to the extent that, before the first name was selected, every member of the population had an equal chance to be chosen. It is not a simple random sample because not all possible samples of 10% of the population are equally likely.
5. This is an instance of *voluntary response bias*. This poll was taken during the depths of the depression, and people felt strongly about national leadership. Those who wanted a change were more likely to respond than those who were more or less satisfied with the current administration.

6. Almost certainly, respondents are responding in a way they feel will please the interviewer. This is a form of response bias—in this circumstance, people may just not give a truthful answer.
7. Many different solutions possible. One possible way would be to put the names of all 90 students on slips of paper and put the slips of paper into a box. Then draw out 15 slips of paper at random. The names on the paper are your sample. Another way would be to identify each student by a two digit number 01,02, . . . , 90 and use a table of random digits to select 15 numbers. Or you could use the *randInt* function on your calculator to select 15 number between 1 and 90 inclusive. What you *cannot* do, if you want it to be a SRS, is to employ a procedure that selects five students randomly from each of the three classes.
8. Because the two groups were not selected randomly, it is possible that the fewer number of colds in the vitamin C group could be the result to some variable whose effects cannot be separated from the effects of the vitamin C. That would make this other variable a *confounding variable*. A possible confounding variable in this case might be that the group who takes vitamin C might be, as a group, more health conscious than those who do not take vitamin C. This could account for the difference in the number of colds but would not be detected by the given study.
9. The study suffers from *undercoverage* of the population of interest, which was declared to be all shoppers at the mall. By restricting their interview time to a Wednesday morning, they effectively exclude most people who work. They essentially have a sample of the opinions of nonworking shoppers. There may be other problems with randomness, but without more specific information about how they gathered their sample, talking about it would only be speculation.
10. We could first administer a questionnaire to all 300 volunteers to determine differing levels of health consciousness. For simplicity, let's just say that the two groups identified are "health conscious" and "not health conscious." Then you would block by "health conscious" and "not health conscious" and run the experiment within each block. A diagram of this experiment might look like this:



11. Because exercise level seems to be more or less constant among the volunteers, there is no need to block for its effect. Furthermore, because the effects of a 200-mg dosage are known, there is no need to have a placebo (although you could)—the 200-mg dosage will serve as the control. Randomly divide your 400 volunteers into four groups of 100 each. Randomly assign each group to one of the four treatment levels: 200 mg, 400 mg, 600 mg, or 800 mg. The study can be and should be double blind. After a period of time, compare the weight loss results for the four groups.
12. (a) Many answers are possible. One solution involves putting the names of all 60 students on slips of paper, then randomly selecting the papers. The first student goes into program 1, the next into program 2, etc. until all 60 students have been assigned. Note that this is not an independent assignment because the probability that a student will end up in any one group depends on which groups the previous students have been assigned to.
(b) Use a random number generator to select a program from 1 to 3 (like $\text{randInt}(1,3)$ on the TI-83; or use a table of random numbers assigning each of the programs a range of values such as 1–3, 4–6, 7–9, and ignore 0). Pick any student and generate a random number from 1–3. The student enters the program that corresponds to the number. In this way, the probability of a student ending up in any one group is $1/3$, and the selections are independent. It would be unlikely to have the three groups come out completely even in terms of the numbers in each, but we would expect it to be close.
13. In this situation, a *stratified random sample* would be a sample in which the proportion of teachers from each of the four levels reflects that of the population from which the sample was drawn. That is, in the sample of 100 teachers, 10 should be from level A, 15 from level B, 45 from level C, and 30 from level D. For level A, she could accomplish this by taking a SRS of 10 teachers from a list of all teachers who teach at that level. SRSs of 15, 45, and 30 would then be obtained from each of the other lists.
14. Remember that you block to control for the variables that might affect the outcome that you know about, and you randomize to control for the effect of those you don't know about. In this case, then, you randomize to control for any unknown systematic differences between the plots that might influence sweetness. An example might be that the plots on the northern end of the rows (plots 1 and 6) have naturally richer soil than those plots on the south side.
15. The idea is to get plots that are most similar in order to run the experiment. One possibility would be to match the plots the following way: close to the river north (6 and 7); close to the river south (9 and 10); away from the river north (1 and 2); and away from the river south

(4 and 5). This pairing controls for both the effects of the river and possible north–south differences that might affect sweetness. Within each pair, you would randomly select one plot to plant one variety of strawberry, planting the other variety in the other plot.

This arrangement leaves plots 3 and 8 unassigned. One possibility is simply to leave them empty. Another possibility is to assign randomly each of them to one of the pairs they adjoin. That is, plot 3 could be randomly assigned to join either plot 2 or plot 4. Similarly, plot 8 would join either plot 7 or plot 9.

16. The study could have been double-blind. The text indicates that the subjects did not know which treatment they were receiving. If the psychologists did not know which therapy the subjects had received before being evaluated, then the basic requirement of a double-blind study were met: neither the subjects nor the researchers who come in contact with them are aware of who is in the treatment and who is in the control group.

If the study weren't double-blind, it would be because the psychologists were aware of which subjects had which therapy (we were told the subjects were unaware). In this case, the attitudes of the psychologists toward the different therapies might influence their evaluations—probably because they might read more improvement into a therapy of which they approve.

17. Group A favors a dress code, group B does not. Both groups are hoping to bias the response in favor of their position by the way they have worded the question.
18. You probably want to block by community. That is, you will have three blocks: Upper Middle, Middle, and Lower Middle. Within each, you have four billboards. Randomly select two of the billboards within each block to receive the Type I ads, and put the Type II ads on the other two. After a few weeks, compare the differences in reaction to each type of advertising within each block.
19. With only 3000 of 100,000 surveys returned, *voluntary response bias* is most likely operating. That is, the 3000 women represented those who felt strongly enough (negatively) about men and were the most likely to respond. We have no way of knowing if the 3% that returned the survey were representative of the 100,000 that received it, but they most likely were not.
20. Assign each of the 26 women a 2-digit number, say 01, 02, . . . , 26. Then enter the table at a random location and note two digit numbers. Ignore numbers outside of the 01–26 range. The first number chosen assigns the corresponding woman to the first group, the second to the second group, etc. until all 26 have been assigned. This method roughly equalizes the numbers in the group (not quite because 4 doesn't go evenly into 26), but does not assign them independently.

If you wanted to assign the women independently, you would consider only the digits 1,2,3, or 4, which correspond to the four groups. As one of the women steps forward, one of the random digits is identified, and that woman goes into the group that corresponds to the chosen number. Proceed in this fashion until all 26 women are assigned a group. This procedure yields independent assignments to groups, but the groups most likely will be somewhat unequal in size. In fact, with only 26 women, group sizes might be quite unequal (a TI-83 simulation of this produced 4 1s, 11 2s, 4 3s, and 7 4s).

3 SOLUTIONS TO CUMULATIVE REVIEW PROBLEMS

- The dataset has an outlier at 85. Because the mean is not resistant to extreme values, it tends to be pulled in the direction of an outlier. Hence, we would expect the mean to be larger than the median.
- Parameters* are values that describe populations, and *statistics* are values that describe samples. Hence, the mean and standard deviation of the pulse rates of Pamela's sample are *statistics*, and the predicted mean and standard deviation for the entire school are *parameters*.
- Putting the numbers in the calculator and doing 1-Var Stats, we find that $\bar{x} = 29.64$, $s = 11.78$, $Q1 = 23$, $Med = 27$, and $Q3 = 35$.
 - The interquartile range (IQR) = $35 - 23 = 12$, $1.5(\text{IQR}) = 1.5(12) = 18$. So, the boundaries beyond which we find outliers are $Q1 - 1.5(\text{IQR}) = 23 - 18 = 5$ and $Q3 + 1.5(\text{IQR}) = 35 + 18 = 53$. Because 55 is beyond the boundary value of 53, it is an outlier, and it is the only outlier.
 - The usual rule for outliers based on the mean is $\bar{x} \pm 3s$. $\bar{x} \pm 3s = 29.64 \pm 3(11.78) = -5.7, 64.98$. Using this rule, there are no outliers because there are no values less than -5.7 nor greater than 64.98 . Sometimes $\bar{x} \pm 2s$ is used to determine outliers. In this case, $\bar{x} \pm 2s = 29.64 \pm 2(11.78) = 6.08, 53.2$. Using this rule, 55 would be an outlier.
- For the given data, $\bar{x} = 29.64$ and $s = 11.78$. Hence,

$$z_{55} = \frac{55 - 29.64}{11.78} = 2.15.$$

Note that in doing problem #3, we could have computed this z -score and observed that because it is larger than 2; that is, represents an outlier by the $\bar{x} \pm 2s$ rule that is sometimes used.

- The problem is referring to the *coefficient of determination*—the proportion of variation in one variable that can be explained by the regression of that variable on another. $r = \sqrt{\text{coefficient of determination}} = \sqrt{.62} = 9.79$.

Chapter 7

Random Variables and Probability



Main concepts: *probability, random variables, discrete and continuous random variables, probability distributions, normal probability problems, simulation, transforming and combining random variables*

PROBABILITY

The second major part of a course in statistics involves making inferences about populations based on sample data (the first was exploratory data analysis). The ability to do this is based on being able to make statements such as, “The probability of getting a finding as different, or more different, from expected, as we got by *chance alone*, under the assumption that the null hypothesis is true, is .6.” To make sense of this statement, you need to have a understanding of what is meant by the term “probability” as well as an understanding of some of the basics of probability theory.

An **experiment or chance experiment (random phenomenon)**: An activity whose outcome we can observe or measure but we do not know how it will turn out on any single trial. Note that this is a different meaning of the word “experiment” than we developed in the last chapter.

example: if we roll a die, we know that we will get a 1, 2, 3, 4, 5, or 6, but we don’t know *which* one of these we will get on the next trial. Assuming a fair die, however, we *do* have a good idea of approximately what proportion of each possible outcome we will get over a large number of trials.

Outcome: One of the possible results of an experiment (random phenomenon)

example: the possible outcomes for the roll of a single die are 1, 2, 3, 4, 5, 6. Outcomes are sometimes called **simple events**.

Sample Spaces and Events

Sample space: The set of all possible outcomes, or simple events, of an experiment.

example: For the roll of a single die, $S = \{1,2,3,4,5,6\}$

Event: A collection of outcomes or simple events. That is, an event is a subset of the sample space.

example: For the roll of a single die, let event $A =$ “the face value is less than 4.” Then, $A = \{1,2,3\}$.

example: Consider the experiment of flipping two coins and noting whether each coin lands heads or tails. The sample space, $S = \{HH, HT, TH, TT\}$. Let event $B =$ “at least one coin shows a head.” Then $B = \{HH, HT, TH\}$. Event B is a subset of the sample space S .

Probability of an event: the relative frequency of the outcome. That is, it is the fraction of time that the outcome would occur if the experiment were repeated indefinitely. If we let $E =$ the event in question, $s =$ the number of ways an outcome can succeed, and $f =$ the number of ways an outcome can fail, then

$$P(E) = \frac{s}{s + f}.$$

Note that $s + f$ equals the number of outcomes in the sample space. Another way to think of this is that the probability of an event is the sum the probabilities of all outcomes that make up the event.

For any event A , $P(A)$ ranges from 0 to 1, inclusive. That is: $0 \leq P(A) \leq 1$. This is an algebraic result from the definition of probability when success is guaranteed ($f = 0$, $s = 1$) or failure is guaranteed ($f = 1$, $s = 0$).

The sum of the probabilities of all possible outcomes in a sample space is one. That is, if the sample space is composed of n possible outcomes,

$$\sum_{i=1}^n p_i = 1.$$

example: In the experiment of flipping two coins, consider the event $A =$ obtain at least one head. The sample space contains four elements ($\{HH, HT, TH, TT\}$). $s = 3$ because there are three ways for our outcome to be considered a success ($\{HH, HT, TH\}$) and $f = 1$.

Thus

$$P(A) = \frac{3}{3+1} = \frac{3}{4}.$$

example: Consider rolling two fair die and noting the sum of the two dice. A sample space for this event can be given in table form as follows:

Face	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Let B = “the sum of the two dice is greater than 4.” There are 36 outcomes in the sample space, 30 of which are greater than 4. Thus,

$$P(B) = \frac{30}{36} = \frac{5}{6}$$

Furthermore,

$$\sum p_i = P(2) + P(3) + \dots + P(12) = \frac{1}{36} + \frac{2}{36} + \dots + \frac{1}{36} = 1.$$

Probabilities of Combined Events

$P(A \text{ or } B)$: The probability that **either** event A **or** event B occurs. Using set notation, $P(A \text{ or } B)$ can be written $P(A \cup B)$. $A \cup B$ is spoken as, “ A union B .”

$P(A \text{ and } B)$: The probability that **both** event A **and** event B occur. Using set notation, $P(A \text{ and } B)$ can be written $P(A \cap B)$. $A \cap B$ is spoken as, “ A intersection B .”

example: Roll two dice and consider the sum (see table). Let A = “one die shows a 3,” B = “the sum is greater than 4.” Then $P(A \text{ or } B)$ is the probability that *either* one die shows a 3 *or* the sum is greater than 4. Of the 36 possible outcomes in the sample space, there are 32 possible outcomes that are successes [30 outcomes greater than 4 as well as (1,3) and (3,1)], so

$$P(A \text{ or } B) = \frac{32}{36}.$$

There are nine ways in which a sum has one die showing a 3 and has a sum greater than 4 [(3,2), (3,3), (3,4), (3,5), (3,6), (2,3), (4,3), (5,3), (6,3)], so

$$P(A \text{ and } B) = \frac{9}{36}.$$

Complement of an event A: events in the sample space that are not in event A. Symbolized by A^c . $P(A^c) = 1 - P(A)$.

Mutually Exclusive Events

Mutually exclusive (disjoint) events: Two events are *mutually exclusive* if they have no outcomes in common. If A and B are mutually exclusive, then $P(A \text{ and } B) = 0$.

example: in the two-dice rolling experiment, $A =$ “face shows a 1” and $B =$ “sum of the two dice is 8” are mutually exclusive because there is no way to get a sum of 8 if one die shows a 1.

Conditional Probability

Conditional Probability: “The probability of A given B ” assumes we have knowledge of an event B having occurred before we compute the probability of event A . This is symbolized by $P(A|B)$. Also,

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

Although this formula will work, it’s often easier to think of a condition as reducing, in some fashion, the original sample space. The following example illustrates this “shrinking sample space.”

example: Once again consider the possible sums on the roll of two dice. Let $A =$ “the sum is 7,” $B =$ “one die shows a 5.” We note, by counting outcomes in the table, that $P(A) = 6/36$. Now, consider a slightly different question: what is $P(A|B)$ (that is, what is the probability of the sum being 7 *given that* one die shows a 5)?

solution: Look again at the table:

Face	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

The condition has effectively reduced the sample space from 36 outcomes to only 11 (note you do not count the “10” twice because it is

in both the row and column for 5). Of those, two are 7s. Thus, the $P(\text{the sum is } 7 | \text{one die shows a } 5) = 2/11$.

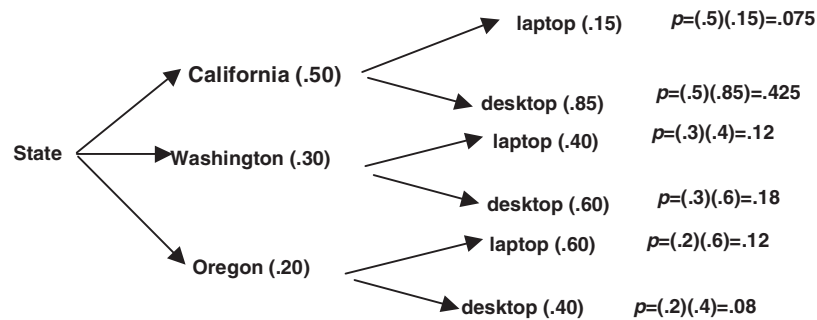
alternate solution: If you insist on using the formula for conditional probability, we note that $P(A \text{ and } B) = P(\text{the sum is } 7 \text{ and one die shows a } 5) = 2/36$, and $P(B) = P(\text{one die shows a } 5) = 11/36$. By formula

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{2/36}{11/36} = \frac{2}{11}.$$

Some conditional probability problems can be solved by using a **tree diagram**. A tree diagram is a schematic way of looking at all possible outcomes.

example: Suppose a computer company has manufacturing plants in three states. 50% of their computers are manufactured in California, and 85% of these are desktops, 30% of computers are manufactured in Washington, and 40% of these are laptops, and 20% of computers are manufactured in Oregon, and 40% of these are desktops. All computers are first shipped to a distribution site in Nebraska before being sent out to stores. If you picked a computer at random from the Nebraska distribution center, what is the probability that it is a laptop?

solution:



Note that the final probabilities add to 1 so we know we have considered all possible outcomes. Now, $P(\text{laptop}) = .075 + .12 + .12 = .315$.

Independent Events

Independent Events: Events A and B are said to be *independent* if and only if $P(A) = P(A | B)$ or $P(B) = P(B | A)$. That is, A and B are independent if the knowledge of one event having occurred does not change the probability that the other event occurs.

example: Consider drawing one card from a standard deck of 52 playing cards.

Let A = “the card drawn is an ace.” $P(A) = 4/52 = 1/13$.
 Let B = “the card drawn is a 10, J, Q, K, or A.” $P(B) = 20/52 = 5/13$
 Let C = “the card drawn is a diamond.” $P(C) = 13/52 = 1/4$.

(i) Are A and B independent?

solution: $P(A \cap B) = P(\text{the card drawn is an ace and the card is a 10, J, Q, K, or A}) = 4/20 = 1/5$ (there are 20 cards to consider, 4 of which are aces). Because $P(A) = 1/13$, knowledge of B has changed what we know about A . That is, in this case, $P(A) \neq P(A \cap B)$, so events A and B are not independent.

(ii) Are A and C independent?

solution: $P(A \cap C) = P(\text{the card drawn is an ace and the card drawn is a diamond}) = 1/13$ (there are 13 diamonds, one of which is an ace). So, in this case, $P(A) = P(A \cap C)$, so that the events “the card drawn is an ace” and “the card drawn is a diamond” are independent.

Probability of A and/or B

The Addition Rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Special case of *The Addition Rule*: If A and B are mutually exclusive, $P(A \text{ and } B) = 0$, so $P(A \text{ or } B) = P(A) + P(B)$

The Multiplication Rule: $P(A \text{ and } B) = P(A) \cdot P(B | A)$

Special case of *The Multiplication Rule*: If A and B are independent, $P(B | A) = P(B)$, so $P(A \text{ and } B) = P(A) \cdot P(B)$.

example: If A and B are two mutually exclusive events for which $P(A) = .3$, $P(B) = .25$. Find $P(A \text{ or } B)$.

solution: $P(A \text{ or } B) = .3 + .25 = .55$

example: A basketball player has a .6 probability of making a free. What is his probability of making two consecutive free throws if

(a) he gets very nervous after making the first shot and his probability of making the second shot drops to .4.

solution: $P(\text{making the first shot}) = .6$, $P(\text{making the second shot he made the first}) = .4$. So, $P(\text{making both shots}) = (.6)(.4) = .24$

(b) the events “he makes his first shot” and “he makes the succeeding shot” are independent

solution: Because the events are independent, his probability of making each shot is the same. Thus, $P(\text{he makes both shots}) = (.6)(.6) = .36$.

RANDOM VARIABLES

Recall our earlier definition: **experiment (random phenomenon)**: An activity whose outcome we can observe or measure but we do not know how it will turn out on any single trial. A **random variable, X** , is a numerical value assigned to an outcome of a random phenomenon. Particular values of the random variable X are often given small case names, such as x . It is common to see expressions of the form $P(X = x)$, which refers to the probability that the random variable X takes on the particular value x .

example: If we roll a fair die, the random variable X could be the face-up value of the die. The possible *values* of X are $\{1, 2, 3, 4, 5, 6\}$. $P(X=2) = 1/6$.

example: The score a college-hopeful student gets on her SAT test can take on values from 200 to 800. These are the possible values of the random variable X , the score a randomly selected student gets on their his/her test.

There are two types of random variables: **discrete random variables** and **continuous random variables**.

Discrete Random Variables

A **discrete random variable (DRV)**: is a random variable with a countable number of outcomes. That is, it has values that correspond to individual points on a number line.

example: the number of votes earned by different candidates in an election.

example: the number of successes in 25 trials of an event whose probability of success on any one trial is known to be .3.

Continuous Random Variables

A **continuous random variable (CRV)** is a random variable that assumes values associated with one or more intervals on the number line. The continuous random variable X has an infinite number of outcomes.

example: Consider the *uniform* distribution $y = 3$ defined on the interval $1 \leq x \leq 5$. The area under $y = 3$ and above the x axis for any interval corresponds to a continuous random variable. For example, if $2 \leq x \leq 3$, then $X = 3$. If $2 \leq x \leq 4.5$, then $X = (4.5 - 2)(3) = 7.5$. Note that there are an infinite number of possible outcomes for X .

Probability Distribution of a Random Variable

A **probability distribution for a random variable** is the possible values of the random variable X together with the probabilities corresponding to those values.

A **probability distribution for a discrete random variable** is a list of all possible values of the DRV together with their respective probabilities.

example: Let X be the number of boys in a three-child family. Assuming that the probability of a boy on any one birth is .5, the probability distribution for X is

x	0	1	2	3
$p(x)$	1/8	3/8	3/8	1/8

The probabilities p_i of a DRV satisfy two conditions:

- (1) $0 \leq p_i \leq 1$ (that is, every probability is between 0 and 1)
- (2) $\sum p_i = 1$ (that is, the sum of all probabilities is 1)

(Are these conditions satisfied in the above example?)

The **mean** of a discrete random variable, also called the **expected value** is given by

$$\mu_X = \sum_{i=1}^n x_i p_i.$$

The **variance of a discrete random variable** is given by

$$\sigma_X^2 = \sum_{i=1}^n (x_i - \mu_X)^2 p_i.$$

The **standard deviation of a discrete random variable** is given by

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\sum_{i=1}^n (x_i - \mu_X)^2 p_i}.$$

example: Given that the following is the probability distribution for a DRV, find $p(x = 3)$.

x	2	3	4	5	6
$p(x)$.15		.2	.2	.35

solution: Because

$$\sum p_i = 1, p(3) = 1 - (.15 + .2 + .2 + .35) = .1$$

example: For the probability distribution given above, find μ_x and σ_x .

solution:

$$\mu_x = 2(.15) + 3(.1) + 4(.2) + 5(.2) + 6(.35) = 4.5$$

$$\begin{aligned} \sigma_x &= \sqrt{(2 - 4.5)^2(.15) + (3 - 4.5)^2(.1) + (4 - 4.5)^2(.2) + (5 - 4.5)^2(.2) + (6 - 4.5)^2(.35)} \\ &= 1.432 \end{aligned}$$



Calculator Tip: Although it's important to know the formulas given above, in practice it's easier to use your calculator to do the computations. The TI-83 (but not the TI-82) can do this easily by putting the x values in, say, $L1$, and the y values in, say, $L2$. Then entering *1-Var Stats L1,L2* and pressing ENTER will return the desired mean and standard deviation. Note that the only standard deviation given is σ_x —the S_x is blank. This is because the calculator recognizes the relative frequencies in $L2$ and assumes you are dealing with a probability distribution. If you are using a TI-82, you can still do this by first multiplying the relative frequencies by 100 (let $L3 = 100 \bullet L2$; then entering *1-Var Stats L1,L3*).

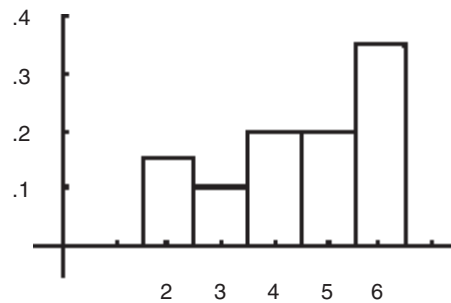
example: Redo the previous example using the TI-83, or equivalent, calculator.

solution: Enter the x values in a list (say $L1$) and the y values in another list (say $L2$). Then enter “*1-Var Stats L1, L2*” and press ENTER. The calculator will read the probabilities in $L2$ as relative frequencies and return 4.5 for the mean and 1.432 for the standard deviation.

Probability Histogram

A **probability histogram** of a DRV is a way to picture the probability distribution. The following is a TI-83 histogram of the probability distribution we used in a couple of the examples above.

x	2	3	4	5	6
$p(x)$.15	.1	.2	.2	.35



Probability Distribution for a Continuous Random Variable (CRV). The probability distribution of a continuous random variable has several properties.

Density Curve

- There is a smooth curve, called a **density curve** (defined by a **density function**), that describes the probability distribution of a CRV. A density curve is always on or above the horizontal axis (that is, it is always non-negative) and has a total area of 1 underneath the curve and above the axis.
- The probability of any individual event is 0. That is, if a is a point on the horizontal axis, $P(X = a) = 0$.
- The probability of a given event is the probability that x will fall in some given interval on the horizontal axis and equals the area under the curve and above the interval. That is, $P(a < X < b)$ equals the area under the graph of the curve and above the horizontal axis between $X = a$ and $X = b$.

In this course, there are several CRVs for which we know the *probability density functions* (a probability distribution defined in terms of some density curve). The *normal distribution* (introduced in Chapter 4) is one whose probability density function is the **normal probability distribution**. Remember that the normal curve is “bell-shaped” and is symmetric about the mean μ of the population. The tails of the curve extend to infinity, although there is very little area under the curve when we get more than, say, three standard deviations away from the mean (the *Empirical Rule* stated that about 99.7% of the terms in a normal distribution are within 3 standard deviations of the mean. Thus, only about .3% lie beyond 3 standard deviations of the mean).

Areas between two values on the number line and under the normal probability distribution correspond to probabilities. In Chapter 4, we found the proportion of terms falling within certain intervals. Because the total area under the curve is 1, in this chapter we will consider those proportions to be probabilities.

Remember that we *standardized* the normal distribution by converting the data to z scores

$$\left\{ z = \frac{x - \bar{x}}{s_x} \right\}$$

We learned in Chapter 4 that a standardized distribution has a mean of 0 and a standard deviation of 1.

A table of **Standard Normal Probabilities** for this distribution is included in this book and in any basic statistics text. We used these tables when doing some normal curve problems in Chapter 4. Standard normal probabilities, and other normal probabilities, are also accessible on many calculators. We will use a table of standard normal probabilities as well as technology to solve probability problems, which are very similar to the problems we did in Chapter 4, involving the normal distribution in the next section.

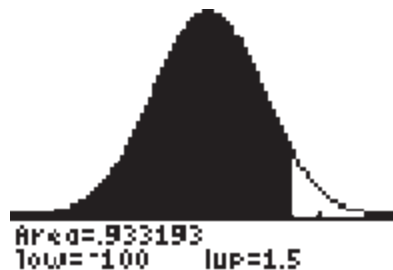
Using the tables, we can determine that the percentages in the 68–95–99.7 rule are more precisely .6827–.9545–.9973. The TI-83 syntax is *normalcdf(lower bound, upper bound)*. Thus, the area between $z = -1$ and $z = 1$ in a standard normal distribution is *normalcdf(-1,1) = .6826894809*.

NORMAL PROBABILITIES

When we know a distribution is approximately normal, we can solve many types of problems.

example: In a standard normal distribution, what is the probability that $z < 1.5$? (note that because Z is a CRV, $p(z = a) = 0$, so this problem could have been equivalently stated “what is the probability that $z \leq 1.5$?”)

solution: The standard normal table gives areas to the left of a specified z score. From the table, we determine that the area to the left of $z = 1.5$ is .9332. That is, $p(z < 1.5) = .9332$. This can be visualized as follows:





Calculator Tip: the above image was constructed on a TI-83 graphing calculator using the “DISTR DRAW ShadeNorm” (The syntax is $\text{ShadeNorm}(\text{lower bound}, \text{upper bound}, [\text{mean}, \text{standard deviation}])$ —only the first two parameters need be included if we want standard normal probabilities. In this example, the lower bound is given as -100 (any large negative number will do—there are very few values more than 3 or 4 standard deviations from the mean.). You *will* need to set the window to match the mean and standard deviation of the normal curve being drawn.

example: It is known that the heights (X) of students at Downtown College are approximately normally distributed with a mean of 68 inches and a standard deviation of 3 inches. That is, X has $N(68,3)$. Determine

(a) $P(X < 65)$.

solution:

$$z = \frac{65 - 68}{3} = -1 \rightarrow$$

Area to the left of -1 is $.1587 = P(x < 65)$.

If you use a TI-83, the corresponding calculation is $\text{normalcdf}(-100, -1)$ or $\text{normalcdf}(-1000, 65, 68, 3) = .1586552596$.

(b) $P(X > 65)$.

solution: From part (a), we know that the area to the left of 65 is $.1587$. Thus, $P(X > 65) = 1 - .1587 = .8413$.

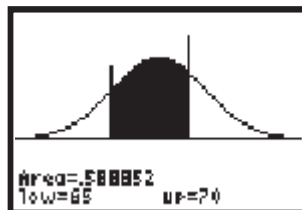
[On the TI-83, the calculation would be $\text{normalcdf}(-1100) = \text{normalcdf}(65, 1000, 68, 3) = .8413447404$].

(c) $P(65 < X < 70)$.

solution: From part (a), the area to the left of 65 is $.1587$.

$$z_{70} = \frac{70 - 68}{3} = .67 \rightarrow A = .7486 \text{ (the area to the left of } z = .67\text{)}.$$

Thus, $P(65 < X < 70) = .7486 - .1587 = .5899$. The situation is pictured below:

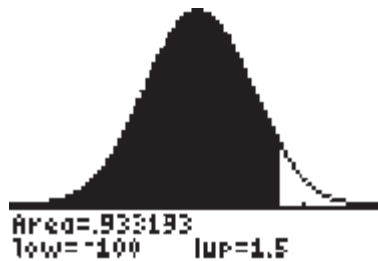


[The corresponding TI-83 calculation would be $normalcdf(-1,.67)$, or $normalcdf(65,70,68,3)$]

Note that there is some rounding error when using the table. In part (c), $z = .66667$, but we must use $.67$ to use the table. Using the greater accuracy of the calculator, the actual area is $.5889$.

(d) $P(70 < X < 75)$

solution: Now we need the area between 70 and 75. The geometry of the situation dictates that we subtract the area to the left of 70 from the area to the left of 75. This is pictured below:



We saw from part (c) that the area to the left of 70 is $.7486$. Similarly we determine that the area to the left of 75 is $.9901$ (based on $z = 2.33$). Thus, $P(70 < X < 75) = .9901 - .7486 = .2415$. [The TI-83 solution to the problem is given by $normalcdf(70,75,68,3) = .2427$.]

example: SAT scores are approximately normally distributed with a mean of about 500 and a standard deviation of 100. Laurie needs to be in the top 15% on the SAT in order to ensure her acceptance by Giant U. What is the minimum score she must earn to be able to start packing her bags for college?

solution: This is somewhat different than the previous examples. Up until now, we have been given, or have figured out, a z score, and have needed to determine an area. This time we are given an area and are asked to determine a particular score. If we are using the table of normal probabilities, it is a situation in which we must read from inside the table out to the z scores rather than from the outside in. If the particular value of X we are looking for is the lower bound for the top 15% of scores, then there are 85% of the scores to the left of x . We look through the table and find the closest entry to $.8500$ and determine it to be $.8508$. This corresponds to a z score of 1.04 . Another way to write the z score of the desired value of X is

$$z = \frac{x - 500}{100}.$$

Thus,

$$z = \frac{x - 500}{100} = 1.04.$$

Solving for x , we get $x = 500 + 1.04(100) = 604$. So, Laurie must achieve an SAT score of at least 604. This problem can be done on the calculator as follows: $InvNorm(.85,500,100)$.

SIMULATION AND RANDOM NUMBER GENERATION

Sometimes probability situations do not lend themselves easily to analytical solutions. In some situations, an acceptable solution might be achieved by doing a **simulation**. A *simulation* utilizes some random process to conduct numerous trials of the situation and then count the number of successful outcomes to arrive at an estimated probability. In general, the more trials, the more confidence we can have that the relative frequency of successes accurately approximates the desired probability. The **law of large numbers** states that the proportion of successes in the simulation should become, over time, close to the true proportion in the population.

One interesting example of the use of *simulation* has been in the development of certain “systems” for playing Blackjack. The number of possible situations in Blackjack is large but countable. A computer was used to conduct thousands of simulations of each possible playing decision for each of the possible hands. In this way, certain situations favorable to the player were identified and formed the basis for the published systems.

example: Suppose there is a small Pacific Island society that places a high value on families having a baby girl. Suppose further that *every* family in the society decides to keep having children until they have a girl and then they stop. If the first child is a girl, they are a one-child family, but it may take several tries before they succeed. Assume that when this policy was decided on that the proportion of girls in the population was 0.5 and the probability of having a girl is .5 for each birth. Would this behavior change the proportion of girls in the population. Design a simulation to answer this question.

solution: Use a random number generator, say a fair coin, to simulate a birth. Let heads = “have a girl” and tails = “have a boy.” Flip the coin and note whether it falls heads or tails. If it falls heads, the trial ends. If it falls tails, flip again because this represents having a boy. The outcome of interest is the number of trials (births) necessary until a girl is born (if the 3rd flip gives the first head, then $x = 3$). Repeat this many times and determine how many girls and how many boys have been born.

The following represents a few trials of this simulation (actually done using a random number generator on the TI-83 calculator):

Trial #	Trial Results (H = "girl")	# Flips until first girl	Total # of girls after trial is finished	Total # of boys after trial is finished
1	TH	2	1	1
2	H	1	2	1
3	TTH	4	3	4
4	H	1	4	4
5	TH	2	5	5
6	H	1	6	5
7	H	1	7	5
8	H	1	8	5
9	TH	2	9	6
10	H	1	10	6
11	TTH	4	11	9
12	H	1	12	9
13	H	1	13	9
14	TTTTH	5	14	13
15	TTH	3	15	15

This limited simulation shows that the number of boys and girls in the population are equal. In fairness, it should be pointed out that you usually won't get exact results in a simulation such as this, especially with only 15 trials, but this time the simulation gave the correct answer: the behavior would not change the proportion of girls in the population.



Exam Tip: If you are asked to do a simulation on the AP Statistics exam (and there have been such questions), you will have to be able to read a table of random numbers rather than using the random number generator on your calculator. This is to make your solution understandable to the person reading your solution. A table of random numbers is simply a list of the whole numbers 0,1,2,3,4,5,6,7,8,9 appearing in a random order. This means that each digit should appear approximately an equal number of times in a large list and the next digit should appear with probability 1/10 no matter what sequence of digits has preceded it.

The following gives 200 outcomes of a typical random number generator separated into groups of 5:

79692 51707 73274 12548 91497 11135 81218 79572 06484 87440
 41957 21607 51248 54772 19481 90392 35268 36234 90244 02146
 07094 31750 69426 62510 90127 43365 61167 53938 03694 76923
 59365 43671 12704 87941 51620 45102 22785 07729 40985 92589

example: A coin is known to be biased in such a way that the probability of getting a head is .4. If the coin is flipped 50 times, how many heads would you expect to get?

solution: Let 0,1,2,3 be a head and 4,5,6,7,8,9 be a tail. If we look at 50 digits beginning with the first row, we see that there are 18 heads (bold-faced below), so the proportion of heads is $18/50 = 0.36$. This is close to the expected value of 0.4.

79692 51707 73274 12548 91497 11135 81218 79572 06484 87440

Sometimes the simulation will be a **wait-time simulation**. In the example above, we could have asked how long it would take, on average, until we get, say, five heads. In this case, using the same definitions for the various digits, we would proceed through the table until we noted five even numbers. We would then write down how many digits we had to look at. Three trials of that simulation might look like this (individual trials are separated by \):

79692 51707 732\74 12548 91497 11\135 8121\ So, it took 13, 14, and 7 trials to get our five heads, or an average of 11.3 trials (the theoretical expected number of trials is 12.5)



Calculator Tip: There are several random generating functions built into your calculator, all in the MATH PRB menu: [*rand*, *randInt*(*randNorm*, and *randBin* *rand*(*k*)] will return *k* random numbers between 0 and 1; *randInt*(*lower bound*, *upper bound*, *k*) will return *k* random integers between *lower bound* and *upper bound*; *randNorm*(*mean*, *standard deviation*, *k*) will return *k* values from a normal distribution with mean *mean* and standard deviation *standard deviation*; *randBin*(*n,p,k*) returns *k* values from a binomial random variable with *n* trials and probability of success *P*.

Again, you will not be able to use these to do an assigned simulation on the AP exam although you can use them to do a simulation you create.



Exam Tip: You may see probability questions on the AP Exam that you choose to do by a simulation rather than by traditional probability methods. As long as you explain your simulation and provide the results for a few trials, this approach is usually acceptable. If you do design a simulation for a problem where a simulation is not required, you *can* use the random number generating functions on your calculator. Just explain clearly what you have done—clearly enough that the reader could replicate your simulation if needed.

TRANSFORMING AND COMBINING RANDOM VARIABLES

If X is a random variable, we can transform the data by adding a constant to each value of X , multiplying each value by a constant, or some linear combination of the two. We may do this to make numbers more manageable. For example, if values in our dataset ranged from 8500 to 9000, we could subtract, say, 8500 from each value to get a dataset that ranged from 0 to 500. We would then be interested in the mean and standard deviation of the new dataset as compared to the old dataset.

Some facts from algebra can help us out here. Let μ_x and σ_x be the mean and standard deviation of the random variable X . Each of the following statements can be algebraically verified if we add or subtract the same constant, a , to each term in a dataset ($X \pm a$), or multiply each term by the same constant b (bX), or some combination of these ($a \pm bX$):

- $\mu_{a \pm bX} = a \pm b\mu_x$
- $\sigma_{a \pm bX} = b\sigma_x$ ($\sigma_{a \pm bX}^2 = b^2\sigma_x^2$)

example: Consider a distribution with $\mu_x = 14$, $\sigma_x = 2$. Multiply each value of X by 4 and then add 3 to each. Then $\mu_{3+4X} = 3 + 4(14) = 59$, $\sigma_{3+4X} = 4(2) = 8$

Rules for the Mean and Standard Deviation of Combined Random Variables

Sometimes we need to combine two random variables. For example, suppose one contractor can finish a particular job, on average, in 40 hours ($\mu_x = 40$). Another contractor can finish a similar job in 35 hours ($\mu_y = 35$). If they work on two separate jobs, how many hours, on average will they bill for completing both jobs. It should be clear that the average of $X + Y$ is just the average of X plus the average for Y . That is,

- $\mu_{X \pm Y} = \mu_x \pm \mu_y$

The situation is somewhat less clear when we combine variances. In the contractor example above, suppose that

$$\sigma_x^2 = 5 \text{ and } \sigma_y^2 = 4.$$

Does the variance of the sum equal the sum of the variances. Well, yes and no. Yes, if the random variables X and Y are independent (that is, one of them has no influence on the other—i.e., the correlation between X and Y is zero). No, if the random variables are not independent, but are dependent in some way. Furthermore, it doesn't matter if the random variables are added or subtracted, we are still combining the variances. That is,

- $\sigma_{X \pm Y}^2 = \sigma_x^2 + \sigma_y^2$, if and only if X and Y are independent.
- $\sigma_{X \pm Y} = \sqrt{\sigma_x^2 + \sigma_y^2}$, if and only if X and Y are independent.

Digression: if X and Y are *not* independent, then $\sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$ where ρ is the correlation between X and Y . You do *not* need to know this for the AP exam.



Exam Tip: The rules for means and variances when you combine random variables may seem a bit obscure, but there have been questions on more than one occasion that depend on your knowledge of how this is done.

The rules given above generalize. That is, no matter how many random variables you have: $\mu_{X_1 \pm X_2 \pm \dots \pm X_n} = \mu_{X_1} \pm \mu_{X_2} \pm \dots \pm \mu_{X_n}$ and, if X_1, X_2, \dots, X_n are all independent, $\sigma_{X_1 \pm X_2 \pm \dots \pm X_n}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \dots + \sigma_{X_n}^2$.

example: A prestigious private school offers an admission test on the first Saturday of November and the first Saturday of December each year. In 2002, the mean score for hopeful students taking the test in November (X) was 156 with a standard deviation of 12. For those taking the test in December (Y), the mean score was 165 with a standard deviation of 11. What are the mean and standard deviation of the total score $X + Y$ of all students who took the test in 2002?

solution: We have no reason to think that scores of students who take the test in December are influenced by the scores of those students who took the test in November. Hence, we will assume that X and Y are independent. Accordingly,

$$\begin{aligned}\mu_{X+Y} &= \mu_X + \mu_Y = 156 + 165 = 321, \\ \sigma_{X+Y} &= \sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{12^2 + 11^2} = \sqrt{265} = 16.28.\end{aligned}$$

RAPID REVIEW

1. A bag has eight green marbles and 12 red marbles. If you draw one marble from the bag, what is $P(\text{draw a green marble})$

Answer:

$$P(E) = \frac{s}{s+f} = \frac{8}{8+12} = \frac{8}{20} = \frac{2}{5}$$

2. A married couple has three children. At least one of their children is a boy. What is the probability that the couple has exactly two boys?

Answer: The sample space for having three children is {BBB, BBG, BGB, GBB, BGG, GBG, GGB, GGG}. Of these, there are seven outcomes that have at least one boy. Of these, three have two boys and one girl. Thus, $P(\text{the couple has exactly two boys} | \text{they have at least one boy}) = 3/7$.

3. Does the following table represent the probability distribution for a discrete random variable?

X	1	2	3	4
$p(X)$.2	.3	.3	.4

Answer: No, because

$$\sum p_i = 1.2$$

4. In a standard normal distribution, what is $P(z > .5)$?

Answer: From the table, we see that $P(z < .5) = .6915$. Hence, $P(z > .5) = 1 - .6915 = .3085$. By calculator, $\text{normalcdf}(.5, 100) = .3085375322$.

5. A random variable X has $N(13, .45)$. Describe the distribution of $2 - 4X$ (that is, each datapoint in the distribution is multiplied by 4, and that value is subtracted from 2)

Answer: We are given that the distribution of X is normal with $\mu_X = 13$ and $\sigma_X = .45$. Because $\mu_{a \pm bX} = a \pm b\mu_X$, $\mu_{2-4X} = 2 - 4\mu_X = 2 - 4(13) = -50$. Also, because $\sigma_{a \pm bX} = b\sigma_X$, $\sigma_{2-4X} = 4\sigma_X = 4(.45) = 1.8$

3 PRACTICE PROBLEMS

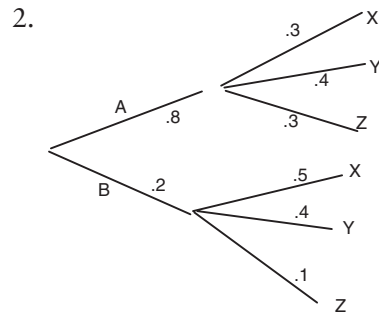
Multiple Choice

1.

	D	E	Total
A	15	12	27
B	15	23	38
C	32	28	60
Total	62	63	125

In the table above what are $P(A \text{ and } E)$ and $P(C \text{ and } E)$?

- 12/125, 28/125
- 12/63, 28/60
- 12/125, 28/63
- 12/125, 28/60
- 12/63, 28/63



For the tree diagram pictured above, what is $P(B \cap X)$?

- $1/4$
 - $5/17$
 - $2/5$
 - $1/3$
 - $4/5$
3. It turns out that 25 seniors at Fashionable High School took both the AP Statistics exam and the AP Spanish Language exam. The mean score on the statistics exam for the 25 seniors was 2.4 with a standard deviation of 0.6 and the mean score on the Spanish Language exam was 2.65 with a standard deviation of 0.55. We want to combine the scores into a single score. What are the correct mean and standard deviation of the combined scores?
- 5.05; 1.15
 - 5.05; 1.07
 - 5.05; 0.66
 - 5.05; 0.81
 - 5.05; you cannot determine the standard deviation from this information.
4. The GPA (grade point average) of students who take the AP Statistics exam are approximately normally distributed with a mean of 3.4 with a standard deviation of 0.3. What is the probability that a student selected at random from this group has a GPA lower than 3.0?
- 0.0918
 - 0.4082
 - 0.9082
 - 0.0918
 - 0
5. The 2000 Census identified the ethnic breakdown of the state of California to be approximately as follows: White: 46%, Latino: 32%, Asian: 11%, Black: 7%, and Other: 4%. Assuming that these are mutually exclusive categories (not a realistic assumption, by the way),

what is the probability that a random selected person from the state of California is of Asian or Latino descent?

- 46%
- 32%
- 11%
- 43%
- 3.5%

Free Response

- Find μ_X and σ_X for the following discrete probability distribution:

X	2	3	4
$p(X)$	1/3	5/12	1/4

- Given that $P(A) = 0.6$, $P(B) = 0.3$, and $P(B|A) = 0.5$.
 - $P(A \text{ and } B) = ?$
 - $P(A \text{ or } B) = ?$
 - Are events A and B independent?
- Consider a set of 9000 scores on a national test that is known to be approximately normally distributed with a mean of 500 and a standard deviation of 90.
 - What is the probability that a randomly selected student has a score greater than 600?
 - How many scores are there between 450 and 600?
 - Rachel needs to be in the top 1% of the scores on this test to qualify for a scholarship. What is the minimum score Rachel needs?
- Consider a random variable X with $\mu_X = 3$, $\sigma_X^2 = 0.25$. Find
 - μ_{3+6X}
 - σ_{3+6X} .
- Harvey, Laura, and Gina take turns throwing spit-wads at a target. Harvey hits the target 1/2 the time, Laura hits it 1/3 of the time, and Gina hits the target 1/4 of the time. Given that somebody hit the target, what is the probability that it was Laura?
- Consider two discrete, independent, random variables X and Y with $\mu_X = 3$, $\sigma_X^2 = 1$, $\mu_Y = 5$, and $\sigma_Y^2 = 1.3$. Find μ_{X+Y} and σ_{X+Y} .
- Which of the following statements are true of a normal distribution?
 - Exactly 95% of the data are within 2 standard deviations of the mean.

- II. The mean = the median = the mode
 III. The area under the normal curve between $z = 1$ and $z = 2$ is greater than the area between $z = 2$ and $z = 3$.
8. Consider the experiment of drawing two cards from a standard deck of 52 cards. Let event $A =$ “draw a face card on the first draw,” $B =$ “draw a face card on the second draw,” $C =$ “the first card drawn is a diamond”
- (a) Are the events A and B *independent*?
 (b) Are the events A and C *independent*?
9. A normal distribution has mean 700 and standard deviation 50. The probability is .6 that a randomly selected term from this distribution is above x . What is x ?
10. Suppose 80% of the homes in Lakeville have a desktop computer and 30% have both a desktop computer and a laptop computer. What is the probability that a randomly selected home will have a laptop computer given that they have a desktop computer?
11. Consider a probability density curve defined by the line $y = 2x$ on the interval $[0,1]$ (the area under $y = 2x$ on $[0,1]$ is 1). Find $P(.2 \leq X \leq .7)$.
12. Half Moon Bay, California, has an annual pumpkin festival at Halloween. A prime attraction to this festival is a “largest pumpkin” contest. Suppose that the weights of these giant pumpkins is approximately normally distributed with a mean of 125 pounds and a standard deviation of 18 pounds. Farmer Harv brings a pumpkin that is at the 90% percentile of all the pumpkins in the contest. What is the approximate weight of Harv’s pumpkin?
13. Consider the following two probability distributions for independent discrete random variable X and Y :

X	2	3	4
$P(X)$.3	.5	?

Y	3	4	5	6
$P(Y)$?	.1	?	.4

If $P(X = 4 \text{ and } Y = 3) = .03$, what is $P(Y = 5)$?

14. A contest is held to give away a free pizza. Contestants pick an integer at random from the integers 1 through 100. If the picked number is divisibly by 24 or by 36, the contestant wins the pizza. What is the probability that a contestant wins a pizza?

Use the following excerpt from a random number table for questions 15 and 16:

79692 51707 73274 12548 91497 11135 81218 79572 06484 87440
 41957 21607 51248 54772 19481 90392 35268 36234 90244 02146
 07094 31750 69426 62510 90127 43365 61167 53938 03694 76923
 59365 43671 12704 87941 51620 45102 22785 07729 40985 92589
 91547 03927 92309 10589 22107 04390 86297 32990 16963 09131

15. Men and women are about equally likely to earn degrees at City U. However, there is some question whether or not women have equal access to the prestigious School of Law. This year, only 4 of the 12 new students are female. Describe and conduct five trials of a simulation to help determine if this is evidence that women are under represented in the School of Law.
16. Suppose that, on a planet far away, the probability of a girl being born is .6, and it is socially advantageous to have three girls. How many children would a family have to have, on average, until they have three girls? Describe and conduct five trials of a simulation to help answer this question.
17. Consider a random variable X with the following probability distribution:

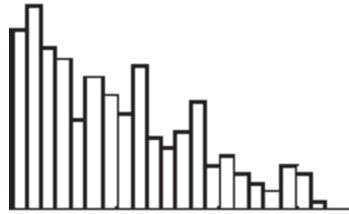
x	20	21	22	23	24
$p(x)$.2	.3	.2	.1	.2

- (a) Find $P(x \leq 22)$
 (b) Find $P(x > 21)$
 (c) Find $P(21 \leq x < 24)$
 (d) Find $P(x \leq 21 \text{ or } x > 23)$
18. In the casino game of roulette, a ball is rolled around the rim of a circular bowl while a wheel containing 38 slots into which the ball can drop is spun in the opposite direction from the rolling ball; 18 of the slots are red, 18 are black, and 2 are green. A player bets a set amount, say \$1, and wins \$1 if the ball falls into the color slot the player has wagered on. Assume a player decides to bet that the ball will fall into one of the red slots.
- (a) What is the probability that the player will win.
 (b) What is the expected return on a single bet of \$1 on red?

19. A random variable X is normally distributed with mean μ , and standard deviation s (that is, X has $N(\mu, s)$). What is the probability that a term selected at random from this population will be more than 2.5 standard deviations from the mean?
20. The normal random variable X has a standard deviation of 12. We also know that $P(x > 50) = .90$. Find the mean μ of the distribution.

3 CUMULATIVE REVIEW PROBLEMS

1. Consider the following histogram:



Which of the following statements is true and why?

- I. The mean and median are approximately the same value
 - II. The mean is probably greater than the median
 - III. The median is probably greater than the mean.
2. You are going to do an opinion survey in your school. You can sample 100 students and desire that the sample accurately reflects the ethnic composition of your school. The school data clerk tells you that the student body is 25% Asian, 8% African American, 12% Latino, and 55% Caucasian. How could you sample the student body so that your sample of 100 would reflect this composition and what is such a sample called?
3. The following data represent the scores on a 50-point AP Statistics quiz:

46, 36, 50, 42, 46, 30, 46, 32, 50, 32, 40, 42, 20, 47, 39, 32, 22, 43, 42, 46, 48, 34, 47, 46, 27, 50, 46, 42, 20, 23, 42

Determine the 5-number summary for the quiz and draw a box plot of the data.

4. The following represents some computer output that relates the number of Manatee deaths to the number of powerboats registered in Florida.

Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	-41.430	7.412	-5.59	.000
Boats	0.12486	0.01290	9.68	.000

- (a) Write the least-square regression line for predicting the number of manatee deaths from the number of powerboat registrations.
 (b) Interpret the slope of the line in the context of the problem.
5. Use the *empirical rule* to state whether it seems reasonable that the following sample data could have been drawn from a normal distribution: 12.3, 6.6, 10.6, 9.4, 9.1, 13.7, 12.2, 9, 9.4, 9.2, 8.8, 10.1, 7.0, 10.9, 7.8, 6.5, 10.3, 8.6, 10.6, 13, 11.5, 8.1, 13.0, 10.7, 8.8

3 SOLUTIONS TO PRACTICE PROBLEMS

Multiple Choice

1. The correct answer is (c). There are 12 values in the *A* and *E* cell and this is out of the total of 125. When we are given column *E*, the total is 63. Of those, 28 are *C*.
2. The correct answer is (b).

$$P(X) = (.8)(.3) + (.2)(.5) = .34.$$

$$P(B|X) = \frac{(.2)(.5)}{(.8)(.3) + (.2)(.5)} = \frac{10}{34} = \frac{5}{17}$$

3. The correct answer is (e). If you knew that the variables “Score on Statistics Exam” and “Score on Spanish Language Exam” were independent, then the standard deviation would be given by

$$\sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{(0.6)^2 + (0.55)^2} = 0.82.$$

However, you cannot assume that they are independent in this situation. In fact, they aren’t because we have two scores on the same people. Hence, there is not enough information.

4. The correct answer is (a).

$$P(X < 3.0) = P\left\{z < \frac{3 - 3.4}{0.3} = -1.33\right\} = .0918$$

5. The correct answer is (d). Because ethnic group categories are assumed to be mutually exclusive, $P(\text{Asian or Latino}) = P(\text{Asian}) + P(\text{Latino}) = 32\% + 11\% = 43\%$.

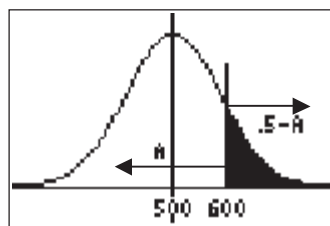
Free Response

$$1. \mu_x = 2\left(\frac{1}{3}\right) + 3\left(\frac{5}{12}\right) + 4\left(\frac{1}{4}\right) = \frac{35}{12} \approx 2.92$$

$$\sigma_x = \sqrt{\left(2 - \frac{35}{12}\right)^2 \left(\frac{1}{3}\right) + \left(3 - \frac{35}{12}\right)^2 \left(\frac{5}{12}\right) + \left(4 - \frac{35}{12}\right)^2 \left(\frac{1}{4}\right)} = .759$$

This can also be done on the TI-83 by putting the X values in $L1$ and the probabilities in $L2$. Then $1\text{-Var Stats } L1, L2$ will give the above values for the mean and standard deviation.

2. (a) $P(A \text{ and } B) = P(A) \cdot P(B|A) = (.6)(.5) = .30$
 (b) $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = .6 + .3 - .3 = .6$
 [note that the .3 that is subtracted came from part (a)]
 (c) $P(B) = .3, P(B|A) = .5$. Because $P(B) \neq P(B|A)$, events A and B are *not* independent.
3. (a) Let X represent the score a student earns. We know that X has approximately $N(500, 90)$. What we are looking for is shown in the following graph



$$P(X > 600) = 1 - .8667 = .1333$$

- (b) We already know, from part (a), that the area to the left of 600 is .8667. Similarly we determine the area to the left of 450 as follows:

$$z_{.450} = \frac{450 - 500}{90} = -.56 \rightarrow A = .2877.$$

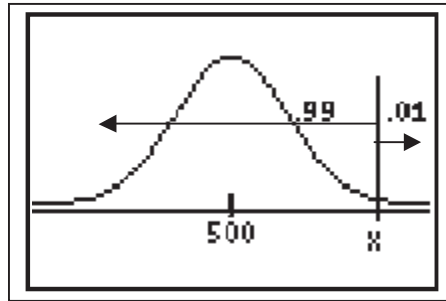
Doing the geometry, we get

$$P(450 < X < 600) = .8667 - .2877 = .5790.$$

There are $.5790(9000) = 5197$ scores.

[This could be done on the calculator as follows: *normalcdf*(450,600,500,90) = .5775]

(c) This situation could be pictured as follows



The z score corresponding to an area of .99 is 2.33. So, $z_x = 2.33$. But, also,

$$z_x = \frac{x - 500}{90}.$$

Thus,

$$\frac{x - 500}{90} = 2.33$$

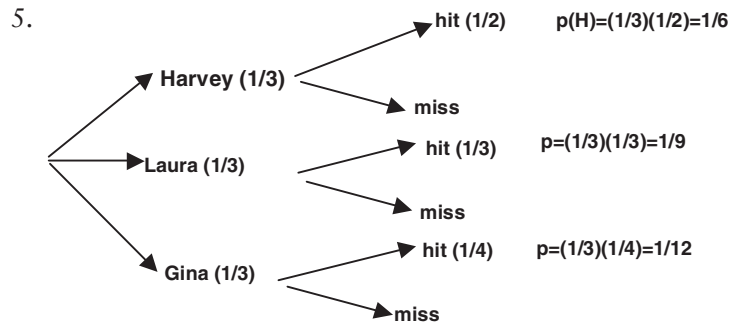
Solving algebraically for x , we get $x = 709.7$. Rachel needs a score of 710 or higher.

(On the calculator, we would find the z score corresponding to this area to be *invNorm*(.99). The complete problem could be done as follows: *invNorm*(.99,500,90) = 709.37.

4. (a) $\mu_{3+6X} = 3 + 6\mu_X = 3 + 6(3) = 21$
 (b) Because $\sigma_{a+bx}^2 = b^2\sigma^2$, $\sigma_{3+6x}^2 = 6^2\sigma_x^2 = 36(.25) = 9$.

Thus,

$$\sigma_{3+6x} = \sqrt{\sigma_{3+6X}^2} = \sqrt{9} = 3.$$



Because we are given that the target was hit, we only need to look at those outcomes. $P(\text{the person who hit the target was Laura} \mid \text{the target was hit})$

$$= \frac{\frac{1}{9}}{\frac{1}{6} + \frac{1}{9} + \frac{1}{12}} = \frac{4}{13}.$$

6. $\mu_{X+Y} = \mu_X + \mu_Y = 3 + 5 = 8;$

Because X and Y are independent, we have

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{1 + 1.3} = 1.52.$$

7. I is not true. This is an approximation based on the *empirical rule*. The actual value proportion within two standard deviations of the mean, to 4 decimal places is 0.9545.

II is true. This is a property of the normal curve.

III is true. This is because the bell-shape of the normal curve means that there is more area under the curve for a given interval length for intervals closer to the center.

8. (a) No, the events are not independent. The probability of B changes depending on what happens with A . Because there are 12 face cards, if the first card drawn is a face card, then $P(B) = 11/51$. If the first card is not a face card, then $P(B) = 12/51$. Because the probability of B is affected by the outcome of A , A and B are not independent.

(b) $P(A) = 12/52 = 3/13$. $P(A \cap C) = 3/52$ (3 of the 52 cards are face cards). Because these are the same, the events, “draw a face card on the first draw” and “the first card drawn is a diamond” are independent.

9. The area to the right of x is 0.6, so the area to the left is 0.4. From the table of *Standard Normal Probabilities*, $A = 0.4 \rightarrow z_x = -0.25$. Also

$$z_x = \frac{x - 700}{50}$$

So,

$$z_x = \frac{x - 700}{50} = -0.25 \rightarrow x = 675.$$

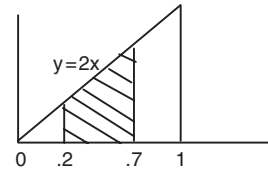
60% of the area is to the right of 675.

10. Let D = “a home has a desktop computer,”; L = “a home has a laptop computer.” We are given that $P(D) = 0.8$ and $P(D \text{ and } L) = 0.3$. Thus,

$$P(L|D) = \frac{P(D \cap L)}{P(D)} = \frac{0.3}{0.8} = \frac{3}{8}.$$

11. The situation can be pictured as follows:
The shaded area is a trapezoid whose area is

$$\frac{1}{2} (0.7 - 0.2)[2(0.2) + 2(0.7)] = 0.45$$



12. The fact that Harv’s pumpkin is at the 90th percentile means that it is larger than 90% of the pumpkins in the contest. From the table of *Standard Normal Probabilities*, the area to the left of a term with a z score of 1.28 is about 0.90. Thus,

$$z_x = 1.28 = \frac{x - 125}{18} \rightarrow x = 148.04.$$

So, Harv’s pumpkin weighed about 148 lbs.

13. For X , $P(X = 4) = 1 - 0.3 - 0.5 = .2$

X	2	3	4
$P(X)$.3	.5	.2

Now, because X and Y are independent, $P(X = 4 \text{ and } Y = 3) = P(X = 4) \cdot P(Y = 3) = (.2) \cdot P(Y = 3) = .03 \rightarrow P(Y = 3) = .15$. Finally, $P(Y = 5) = 1 - P(Y = 3) - P(Y = 4) - P(Y = 6) = 1 - .15 - .1 - .4 = .35$

Y	3	4	5	6
$P(Y)$.15	.1	.35	.4

14. Let $A =$ “the number is divisible by 24” $= \{24,48,72,96\}$
 Let $B =$ “the number is divisible by 36” $= \{36,72\}$

Note that $P(A \text{ and } B) = \frac{1}{100}$ (72 is the only number divisible by both 24 and 36)

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = \frac{4}{100} + \frac{2}{100} - \frac{1}{100} = \frac{5}{100} = 0.05.$$

15. Because the numbers of men and women in the school are about equal (that is, $P(\text{women}) = .5$), let an even number represent a female, and an odd number represent a male. Begin on the first line of the table and consider groups of 12 digits. Count the even numbers among the 12. This will be the number of females among the group. Repeat five times. The relevant part of the table is shown below, with even numbers underlined and groups of 12 separated by two slanted bars ($\backslash\backslash$):

79692 51707 73 $\backslash\backslash$ 274 12548 9149 $\backslash\backslash$ 7 11135 81218
 7 $\backslash\backslash$ 9572 06484 874 $\backslash\backslash$ 40 41957 21607 $\backslash\backslash$

In the five groups of 12 people, there were 3, 6, 3, 8, and 6 women. So, in 40% of the trials there were 4 or fewer women in the class even though we would expect the average to be 6 (the average of these 5 trials is 5.2). Hence, it seems that getting only 4 women in a class when we expect 6 really isn't too unusual because it occurs 40% of the time. (It is shown in the next chapter that the theoretical probability of getting 4 or fewer women in a group of 12 people, assuming that men and women are equally likely, is about .19).

16. Because $P(\text{girl}) = .6$, let the random digits 1,2,3,4,5,6 represent the birth of a girl and 0,7,8,9 represent the birth of a boy. Start on the second row of the random digit table and move across the line until you found the third digit that represents a girl. Note the number of digits needed to get three successes. Repeat 5 times and compute the average. The simulation is shown below (each success; i.e., girl, is underlined and separate trials are delineated by $\backslash\backslash$)

79692 51707 73274 12548 91497 11135 81218 79572 06484 87440
 Start: 4195 $\backslash\backslash$ 7 21607 5 $\backslash\backslash$ 124 $\backslash\backslash$ 8 54772 $\backslash\backslash$ 19481 $\backslash\backslash$ 90392 35268 36234

It took 4,7,3,6, and 5 children before they got their three girls. The average wait was 5. (The theoretical average is exactly 5—we got lucky this time!)

17. (a) $P(x \leq 22) = P(x = 20) + P(x = 21) + P(x = 22) = .2 + .3 + .2 = .7$
 (b) $P(x > 21) = P(x = 22) + P(x = 23) + P(x = 24) = .2 + .1 + .2 = .5$
 (c) $P(21 \leq x < 24) = P(x = 21) + P(x = 22) + P(x = 23) = .3 + .2 + .1 = .6$
 (d) $P(x \leq 21 \text{ or } x > 23) = P(x = 20) + P(x = 21) + P(x = 24) = .2 + .3 + .2 = .7$

18. (a) 18 of the 38 slots are winners, so $P(\text{win if bet on red})$

$$= \frac{18}{38} = .474$$

- (b) The probability distribution for this game is

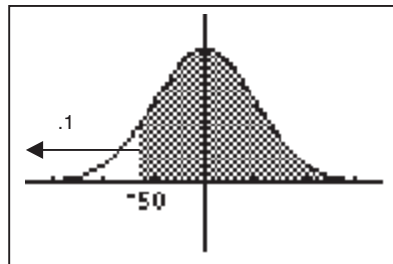
Outcome	Win	Lose
X	1	-1
$p(X)$	$\frac{18}{38}$	$\frac{20}{38}$

$$E(X) = \mu_X = (1) \left(\frac{18}{38} \right) + (-1) \left(\frac{20}{38} \right) = 0.052 \text{ or } -5.2\text{¢}.$$

The player will lose 5.2¢, on average, for each dollar bet.

19. From the tables, we see that $P(z < -2.5) = P(z > 2.5) = .0062$. So the probability that we are more than 2.5 standard deviations from the mean is $2(.0062) = .0124$. [This can also be found on the calculator as follows: $2[\text{normalcdf}(2.5, 1000)]$.]

20. The situation can be pictured as follows:

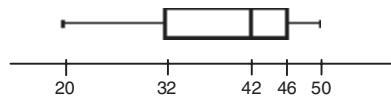


If 90% of the area is to the *right* of 50, then 10% of the area is to the left. So,

$$z_{50} = -1.28 = \frac{50 - \mu}{12} \rightarrow \mu = 65.36$$

3 SOLUTIONS TO CUMULATIVE REVIEW PROBLEMS

1. It is true: the mean is most likely greater than the median. This is because the mean, being nonresistant, is pulled in the direction of outliers or skewness. Because the given histogram is clearly skewed to the right, the mean is likely to be to the right of (that is, greater than) the median.
2. The kind of sample you want is a *stratified random sample*. The sample should have 25 Asian students, 8 African American students, 12 Latino students, and 55 Caucasian students. You could get a list of all Asian students from the data clerk and randomly select 25 students from the list. Repeat this process from lists of the African American, Latino, and Caucasian students. Now the proportion of each ethnic group in your sample is the same as their proportion in the population.
3. The 5-number summary is 20–32–42–46–50. The box plot looks like this:



4. (a) The LSRL line is: #Manatee deaths = $-41.430 + 0.12486(\#boats)$.
 (b) For each increase of one registered powerboat, the number of manatee deaths is predicted to increase by 0.12.
5. For this set of data, $x = 9.9$ and $s = 2.0$. Examination of the 25 points in the dataset yields the following.

	Percentage expected in each interval by the empirical rule	# of terms from dataset in each interval
$9.9 \pm 1(2) = \langle 7.9, 11.9 \rangle$	68% (17/25)	$16/25 = 64\%$
$9.9 \pm 2(2) = \langle 5.9, 13.9 \rangle$	95% (23.8/25)	$25/25 = 100\%$
$9.9 \pm 3(2) = \langle 3.9, 15.9 \rangle$	99.7% (24.9/25)	$25/25 = 100\%$

The actual values of in the dataset (16,25,25) are quite close to the expected values (17,23.8,24.9) if this truly were data from a normal population. Hence, it seems reasonable that the data could have been a random sample drawn from a population that is approximately normally distributed.

Chapter 8

Binomial Distribution and Sampling Distributions



Main Concepts: *Binomial distribution, normal approximation to the binomial, geometric distribution, sampling distributions, Central Limit Theorem*

BINOMIAL DISTRIBUTION

A **binomial experiment** has the following properties:

- The experiment consists of a fixed number, n , of identical trials.
- There are only two possible outcomes (that's the "bi" in "binomial"): success (S) or failure (F).
- The probability of success, p , is the same for each trial.
- The trials are independent (that is, knowledge of the outcomes of earlier trials does not affect the probability of success of the next trial).
- Our interest is in a **binomial random variable X** , which is the count of successes in n trials. The probability distribution of X is the **binomial distribution**.

(Trials of an experiment are called Bernoulli Trials if each trial is independent, and the probability of success remains the same from trial to trial. A binomial experiment is one in which we count the number of successes in n Bernoulli Trials)

The short version of this is to say that a *binomial experiment* consists of n independent trials of an experiment that has two possible outcomes (success or failure), each trial having the same probability of success (p). The *binomial random variable X* is the count of successes.

In practice, we may consider a situation to be binomial when, in fact, the independence condition is not quite satisfied. This occurs when the

probability of occurrence on a given trial is affected only slightly by prior trials. For example, suppose that the probability of a defect in a manufacturing process is .0005. That is, there is, on average, only 1 defect in 2000 items. Suppose we check a sample of 10,000 items for defects. When we check the first item, the proportion of defects remaining changes slightly for the remaining 9,999 items in the sample. We would expect 5 out of 10,000 (.0005) to be defective. But if the first one we look at is *not* defective, the probability of the next one being defective has changed to $5/9999$ or .00050005. It's a small change but it means that the trials are not, strictly speaking, independent. A common rule of thumb is that we will consider a situation to be binomial if the population size is more than 10 times the sample size.

Symbolically, for the *binomial random variable* X , we say X has $B(n,p)$.

example: Suppose Dolores is a 65% free throw shooter. If we assume that that repeated shots are independent, we could ask, "What is the probability that Dolores makes exactly 7 of her next 10 free throws?" If X is the binomial random variable that gives us the count of successes for this experiment, then we say that X has $B(10, .65)$. Our question is then: $P(X = 7) = ?$.

We can think of $B(n,p,x)$ as a particular binomial probability. In this example, then, $B(10, .65, 7)$ is the probability that there are exactly 7 success in 10 repetitions of a binomial experiment where $P = .65$. This is handy because it is the same syntax used by the TI-83 calculator (*binompdf*(n,p,x)) when doing binomial problems.

If X has $B(n,p)$, then X can take on the values $0, 1, 2, \dots, n$. Then,

$$B(n, p, x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

gives the *binomial probability* of exactly x successes for a binomial random variable X that has $B(n,p)$.

Now,

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

On the TI-83,

$$\binom{n}{x} = {}_n C_x,$$

and this is found in the *MATH PRB* menu. $n!$ (" n factorial") means $n(n-1)(n-2) \dots (2)(1)$, and the factorial symbol can be found in the *MATH PRB* menu.

example: Find $B(15, .3, 5)$. That is, find $P(X = 5)$ for a 15 trials of a binomial random variable X that succeeds with probability .3.

solution:

$$\begin{aligned} P(X = 5) &= \binom{15}{5} (.3)^5 (1 - .3)^{15-5} = {}_{15}C_5 (.3)^5 (.7)^{10} \\ &= \frac{15!}{5! 10!} (.3)^5 (.7)^{10} = .206 \end{aligned}$$



Calculator Tip: On the TI-83, the solution is given by $\text{binompdf}(15, .3, 5)$. The “ binompdf ” function is found in the *DISTR* menu of your calculator. The syntax for this function is $\text{binompdf}(n, p, x)$. The function $\text{binomcdf}(n, p, x) = P(X = 0) + P(X = 1) + \dots + P(X = x)$. That is, it adds up the binomial probabilities from 0 through x .

example: Consider once again our free throw shooter (Dolores) from an earlier example. Dolores is a 65% free throw shooter and each shot is independent. If X is the count of free throws made by Dolores, then X has $B(10, .65)$ if she shoots 10 free throws. What is $P(X = 7)$?

solution:

$$\begin{aligned} P(X = 7) &= \binom{10}{7} (.65)^7 (.35)^3 = \frac{10!}{7! 3!} (.65)^7 (.35)^3 \\ &= \text{binompdf}(10, .65, 7) = .252 \end{aligned}$$

example: What is the probability that Dolores makes *no more than* 5 free throws? That is, what $P(X \leq 5)$?

solution:

$$\begin{aligned} P(X \leq 5) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &+ P(X = 4) + P(X = 5) = \binom{10}{0} (.65)^0 (.35)^{10} + \binom{10}{1} (.65)^1 (.35)^9 \\ &+ \dots + \binom{10}{5} (.65)^5 (.35)^5 = .249. \end{aligned}$$

There is about a 25% chance that she will make 5 or less free throws. The solution to this problem using the calculator is given by $\text{binomcdf}(10, .65, 5)$.

example: What is the probability that Dolores makes at least 6 free throws?

solution: $P(X \geq 6) = P(X = 6) + P(X = 7) + \dots + P(X = 10) = 1 - \text{binomcdf}(10, .65, 5) = .751$.

(Note that $P(X > 6) = 1 - \text{binomcdf}(10, .65, 6)$).

The mean and standard deviation a binomial random variable X are given by $\mu_X = np$; $\sigma_X = \sqrt{np(1-p)}$. A binomial distribution for a given n and P (meaning you have all possible values of x along with their corresponding probabilities) is an example of a *probability distribution* as defined in Chapter 7. The mean and standard deviation of a binomial random variable X could be found by using the formulas from Chapter 7

$$\mu_x = \sum_{i=1}^n x_i p_i \text{ and } \sigma_x = \sqrt{\sum_{i=1}^n (x - \bar{x})^2 p_i}$$

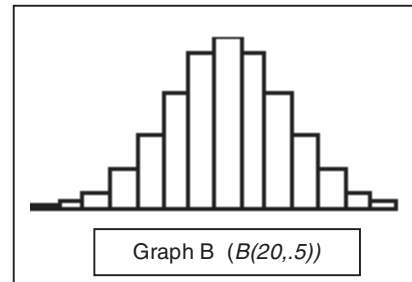
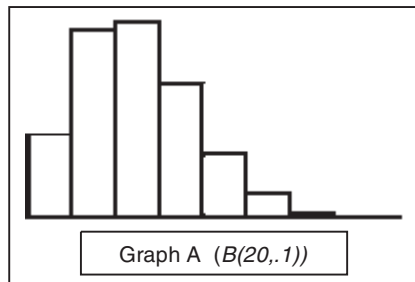
but clearly the formulas for the binomial are easier to use. Be careful that you don't try to use the formulas for the mean and standard deviation of a binomial random variable and a discrete random variable that is *not* binomial.

example: Find the mean and standard deviation of a binomial random variable X that has $B(85, .6)$

solution: $\mu_x = (85)(.6) = 51$, $\sigma_x = \sqrt{85(.6)(.4)} = 4.52$

Normal Approximation to the Binomial

Under the proper conditions, the shape of a binomial distribution is approximately normal, and binomial probabilities can be estimated using normal probabilities. Generally, this is true when $np \geq 10$ and $n(1-p) \geq 10$ (some books use $np \geq 5$ and $n(1-p) \geq 5$; that's OK). These conditions are not satisfied in graph A (X has $B(20, .1)$) below, but are satisfied in graph B (X has $B(20, .5)$)



It should be clear that Graph A is noticeably skewed to the right, and Graph B is approximately normal in shape, so it is reasonable that a normal curve would approximate Graph B better than Graph A.

When np and $n(1 - p)$ are sufficiently large (that is, they are both greater than or equal to 5 or 10), the binomial random variable X has approximately a normal distribution with

$$\mu = np \text{ and } \sigma = \sqrt{np(1 - p)}.$$

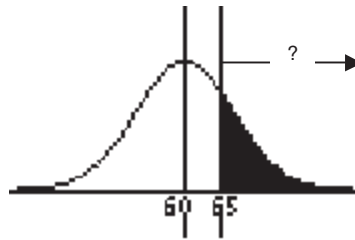
Another way to say this is: If X has $B(n, p)$, then X has approximately $N(np, \sqrt{np(1 - p)})$, provided that $np \geq 10$ and $n(1 - p) \geq 10$ (or $np \geq 5$ and $n(1 - p) \geq 5$).

example: Nationally there are data that 15% of community college students live more than 6 miles from campus. Data from a simple random sample of 400 students at one community college is analyzed.

- What are the mean and standard deviation for the number of students in the sample that live more than 6 miles from campus?
- Use a normal approximation to calculate the probability that at least 65 of the students in the sample live more than 6 miles from campus.

solution: If X is the number of students that live more than 6 miles from campus, then X has $B(400, .15)$.

- $\mu = 400(.15) = 60$, $\sigma = \sqrt{400(.15)(.85)} = 7.14$.
- Because $400(.15) = 60$ and $400(.85) = 340$, we can use the normal approximation to the binomial with mean 60 and standard deviation 7.14. the situation is pictured below:



Using standard normal distribution procedures, we have

$$z_{65} = \frac{65 - 60}{7.14} = .70 \rightarrow P(X > 65) = 1 - .7580 = .242.$$

By calculator, this can be found as $normalcdf(65, 1000, 60, 7.14) = .242$.

The exact binomial solution to this problem is given by $1 - binomcdf(400, .15, 64) = .261$ (you use 64 because the problem stated “at least 65”).

In reality, you will need to use a normal approximation to the binomial only in limited circumstances. In the example above, the answer can be arrived at quite easily using the exact binomial capabilities of your calculator. The only time you might want to use a normal approximation is if the size of the binomial exceeds the capacity of your calculator [try, for example, $\text{binomcdf}(50000000, .7, 325000)$], and you didn't have access to a computer. The real reason you need to understand this is that another way of looking at binomial data is in terms of the *proportion* of successes rather than the count of successes. We *will* approximate a distribution of sample proportions with a normal distribution and the concepts and conditions for it are the same.

Geometric Distribution

In section 8.1, we defined a binomial setting as an experiment in which the following exist.

- The experiment consists of a fixed number, n , of identical trials.
- There are only two possible outcomes: success (S) or failure (F).
- The probability of success, p , is the same for each trial.
- The trials are independent (that is, knowledge of the outcomes of earlier trials does not affect the probability of success of the next trial).
- Our interest is in a **binomial random variable X** , which is the count of successes in n trials. The probability distribution of X is the **binomial distribution**.

There are times we are interested not in the count of successes out of n fixed trials, but in the probability that the first success occurs on a given trial, or in the average number of trials until the first success. A **geometric setting** is defined as follows.

- There are only two possible outcomes: success (S) or failure (F).
- The probability of success, p , is the same for each trial.
- The trials are independent (that is, knowledge of the outcomes of earlier trials does not affect the probability of success of the next trial).
- Our interest is in a **geometric random variable X** , which is the number of trials necessary to obtain the first success.

Note that if X is a *binomial*, then X can take on the values $0, 1, 2, \dots, n$. If X is *geometric*, then it takes on the values $1, 2, 3, \dots$. There can be zero successes in a binomial, but the earliest a first success can come in a geometric setting is on the first trial.

If X is geometric, the probability that the first success occurs on the n th trial is given by $P(X = n) = p(1 - p)^{n-1}$. This is in the *DISTR* menu on the TI-83 as $\text{geompdf}(p, n)$.

example: Remember Dolores, the basketball player whose free throw shooting percentage was .65? What is the probability that the first free throw she manages to hit is on her fourth attempt?

solution: $P(X = 4) = (.65)(1 - .65)^{4-1} = (.65)(.35)^3 = .028$.
[This can be done on the TI-83 as follows: *geometpdf(p,n) = geometpdf(.65,4)*]

example: In a standard deck of 52 cards, there are 12 face cards. So the probability of drawing a face card from a full deck is $12/52 = .231$.

- (a) If you draw cards with replacement (that is, you replace the card in the deck before drawing the next card), what is the probability that the first face card you draw is the 10th card?
 (b) If you draw cards without replacement, what is the probability that the first face card you draw is the 10th card?

solution:

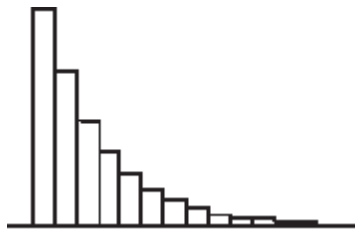
- (a) $P(X = 10) = (.231)(1 - .231)^9 = .022$, or, on the calculator, *geometpdf(.231,10)*.
 (b) If you don't replace the card each time, the probability of drawing a face card on each trial is different because the proportion of face cards in the deck changes each time a card is removed. Hence, this is not a geometric setting and cannot be answered by the techniques of this section.

Rather than the probability that the first success occurs on a specified trial, we may be interested in the average wait until the first success. The average wait until the first success of a geometric random variable is $1/p$.

example: On average, how many free throws will Dolores have to take before she makes one (remember, $p = .65$)?

solution: $1/.65 = 1.54$

Because the probabilities for the geometric distribution are given by $P(X = n) = P(1 - p)^{n-1}$, the probabilities become less likely on each trial because we multiply the previous result by $1 - p$, a number less than one. The geometric distribution looks like this:



SAMPLING DISTRIBUTIONS

Suppose we drew a sample of size 10 from a normal population with unknown mean and standard deviation and got $\bar{x} = 18.87$. Two questions arise: (1) what does this sample tell us about the population from which the sample was drawn, and (2) what would happen if we drew more samples?

Suppose we drew 5 more samples of size 10 from this population and got: $\bar{x} = 20.35$, $\bar{x} = 20.04$, $\bar{x} = 19.20$, $\bar{x} = 19.02$, $\bar{x} = 20.35$. In answer to question (1), we might believe that the population from which these samples was drawn had a mean around 20 because these averages tend to group there (in fact, the six samples were drawn from a normal population whose mean is 20 and whose standard deviation is 4). The mean of the six samples is 19.64, which supports our feeling that the mean of the original population might have been 20.

The standard deviation of the 6 samples is 0.68 and you might not have any intuitive sense about how that relates to the population standard deviation, although you might suspect that the standard deviation of the samples should be less than the standard deviation of the population because the chance of an extreme value for an average should be less than that for an individual term (it just doesn't seem very likely that we would draw a *lot* of extreme values in a single sample).

Suppose we continued to draw samples of size 10 from this population until we were exhausted or until we had drawn *all possible samples of size 10*. If we did succeed in drawing all possible samples of size 10, and computed the mean of each sample, the distribution of these sample means would be the **sampling distribution of \bar{x}** .

Remembering that a “statistic” is a value that describes a sample, the **sampling distribution of a statistic** is the distribution of that statistic for all possible samples of a given size. It's important to understand that a dot plot of a few samples drawn from a population is not a distribution (it's a *simulation* of a distribution)—it becomes a distribution only when all possible samples of a given size are drawn.

Sampling Distribution of a Sample Mean

Suppose we have the sampling distribution of \bar{x} . That is, we have formed a distribution of the means of all possible samples of size n from an unknown population (that is, we know little about its shape, center, or spread). Let $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ represent the mean and standard deviation of the sampling distribution of \bar{x} , respectively.

Then

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

for any population with mean μ and standard deviation σ .

(Note: the value given for $\sigma_{\bar{x}}$ above is generally considered correct only if the sample size (n) is small relative to N , the number in the population. A general rule is that n should be no more than 5% of N to use the value given for $\sigma_{\bar{x}}$. If n is more than 5% of N , the correct value for the standard deviation of the sampling distribution is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

In practice this usually isn't a major issue because

$$\sqrt{\frac{N-n}{N-1}}$$

is close to one whenever N is large in comparison to n)

example: A large population is known to have a mean of 23 and a standard deviation of 2.5. What are the mean and standard deviation of the sampling distribution of means of samples of size 20 drawn from this population?

solution:

$$\mu_{\bar{x}} = \mu = 23, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{20}} = .559.$$

Central Limit Theorem

The discussion above gives us measures of center and spread for the sampling distribution of \bar{x} but tells us nothing about the *shape* of the sampling distribution. It turns out that the shape of the sampling distribution is determined by (a) the shape of the original population and (b) n , the sample size. If the original population is normal, then it's easy: *The shape of the sampling distribution will be normal if the population is normal.*

If the shape of the original population is not normal, or unknown, and the sample size is small, then the shape of the sampling distribution will be similar to that of the original population. For example, if a population is skewed to the right, we would expect the sampling distribution of the mean for small samples to also be somewhat skewed to the right, although not as much as the original population.

When the sample size is large, we have the following result, known as the **Central Limit Theorem**: *For n large the sampling distribution of \bar{x} will be approximately normal.*

A rough rule-of-thumb for using the central limit theorem is that n should be at least 30, although the sampling distribution may be

approximately normal for much smaller values of n if the population doesn't depart markedly from normal. The central limit theorem allows us to use normal calculations to do problems involving sampling distributions without having to have knowledge of the original population. Given that the population size (N) is large in relation to the sample size (n), the information presented in this section can be summarized in the following table:

	Population	Sampling Distribution
Mean	μ	$\mu_{\bar{x}} = \mu$
Standard Deviation	σ	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Shape	Normal	Normal
	Undetermined (skewed, etc.)	<p>If n is "small" \rightarrow shape is similar to shape of original graph</p> <p>OR</p> <p>If n is "large" (rule of thumb: $n \gtrsim 30$) \rightarrow shape is approximately normal (central limit theorem)</p>

example: Describe the sampling distribution of \bar{x} for samples of size 15 drawn from a normal population with mean 65 and standard deviation 9.

solution: Because the original population is normal, \bar{x} is normal with mean 65 and standard deviation

$$\frac{9}{\sqrt{15}} = 2.32. \text{ That is, } \bar{x} \text{ has } N\left(65, \frac{9}{\sqrt{15}}\right).$$

example: Describe the sampling distribution of \bar{x} for samples of size 15 drawn from a population that is strongly skewed to the left (like the scores on a very easy test) with mean 65 and standard deviation 9.

solution: $\mu_{\bar{x}} = 65$ and $\sigma_{\bar{x}} = 2.32$ as in the above example. However this time the population is skewed to the left. The sample size is reasonably large, but not large enough to argue, based on our rule-of-thumb ($n \gtrsim 30$) that the sampling distribution is normal. The best we can say is that the sampling distribution is probably more mound-shaped than the original but might still be somewhat skewed to the left.

example: The average adult has completed an average of 11.25 years of education with a standard deviation of 1.75 years. A random sam-

ple of 90 adults is obtained. What is the probability that the sample will have a mean

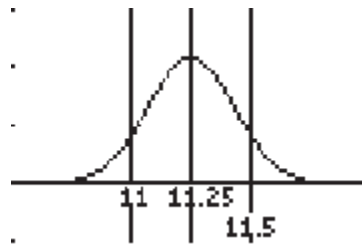
- (a) greater than 11.5 years?
 (b) between 11 and 11.5 years?

solution: The sampling distribution of \bar{x} has $\mu_{\bar{x}} = 11.25$ and

$$\sigma_{\bar{x}} = \frac{1.75}{\sqrt{90}} = .184.$$

Because the sample size is large ($n = 90$), the central limit theorem tells us that large sample techniques are appropriate. Accordingly,

- (a) The graph of the sampling distribution is shown below:



$$P(\bar{x} > 11.5) = P\left(z > \frac{11.5 - 11.25}{1.75/\sqrt{90}} = \frac{.25}{.184} = 1.36\right) = .0869$$

- (b) From part (a), the area to the left of 11.5 is $1 - .0869 = .9131$. Because the sampling distribution is approximately normal, it is symmetric. Because 11 is the same distance to the left of the mean as 11.5 is to the right, we know that $P(\bar{x} < 11) = P(\bar{x} > 11) = .0869$. Hence, $P(11 < \bar{x} < 11.5) = .9131 - .0869 = .8262$.

[by calculator: $normalcdf(11, 11.5, 11, .184) = .8258$]

example: Over the years, the scores on the final exam for AP Calculus have been normally distributed with a mean of 82 and a standard deviation of 6. The instructor thought that this year's class was quite dull and, in fact, they only averaged 79 on their final. Assuming that this class is a random sample of 32 students from AP Calculus, what is the probability that the average score on the final for this class is no more than 79? Do you think the instructor was right?

solution:

$$P(\bar{x} < 79) = P\left(z < \frac{79 - 82}{6/\sqrt{32}} = \frac{-3}{1.06} = -2.83\right) = .0023.$$

If this group really were typical, there is less than a 1% chance of getting an average this low by random chance alone. That seems unlikely, so we have some pretty good evidence that the instructor was correct.

[The calculator solution for this problem is: $\text{normalcdf}(-1000, 79, 82, 1.06)$.]

SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION

If X is the count of successes in a sample of n trials of a binomial random variable, then the **proportion of success** is given by $\hat{p} = X/n$. \hat{p} is what we use for the sample proportion (a statistic). The true population proportion would then be given by p .

Digression: Before introducing \hat{p} , we have used \bar{x} and s as statistics, and μ and σ as parameters. Often we represent statistics with English letters and parameters with Greek letters. However, we depart from that convention here by using \hat{p} as a statistic and p as a parameter. There are texts that are true to the English/Greek convention by using P for the sample proportion and Π as the population proportion.

We learned in Section 8.1 that, if X is a binomial random variable, that the mean and standard deviation of the sampling distribution of X are given by

$$\mu_X = np, \sigma_X = \sqrt{np(1-p)}.$$

We know that if we divide each term in a dataset by the same value n , then the mean and standard deviation of the transformed data set will be the mean and standard deviation of the original data set divided by n . Doing the algebra, we find that the mean and standard deviation of the sampling distribution of \hat{p} are given by:

$$\mu_{\hat{p}} = p \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Like the binomial, the sampling distribution of \hat{p} will be approximately normally distributed if n and p are large enough. The test is exactly the same as it was for the binomial: If X has $B(n, p)$, and $\hat{p} = X/n$, then \hat{p} has approximately

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right),$$

provided that $np \geq 10$ and $n(1-p) \geq 10$ (or $np \geq 5$ and $n(1-p) \geq 5$).

example: Harold fails to study for his statistics final. The final has 100 multiple choice questions, each with 5 choices. Harold has no choice but to guess randomly at all 100 questions. What is the probability that Harold will get at least 30% on the test?

solution: Because $100(.2)$ and $100(.8)$ are both greater than 10, we can use the normal approximation to the sampling distribution of \hat{p} . Because $p = .2$, the sampling distribution of \hat{p} has $\mu = .2$ and

$$\sigma_{\hat{p}} = \sqrt{\frac{.2(1 - .2)}{100}} = .04.$$

Therefore,

$$P(\hat{p} > .3) = P\left\{z > \frac{.3 - .2}{.04} = 2.5\right\} = .0062.$$

Harold should have studied.

RAPID REVIEW

- A coin is known to be unbalanced in such a way that heads only comes up 0.4 of the time.
 - What is the probability the first head appears on the 4th toss?
 - How many tosses would it take, on average, to flip two heads?

Answer:

- $P(\text{first head appears on 4th toss}) = .4(1 - .4)^{4-1} = .4(.6)^3 = .0864$
- Avg. wait to flip two heads = 2 (average wait to flip one head)

$$= 2 \left(\frac{1}{.4} \right) = 5.$$

- The coin of problem #1 is flipped 50 times. Let X be the number of heads. What is;
 - the probability of *exactly* 20 heads?
 - the probability of at *least* 20 heads?

Answer:

- $P(X = 20) = \binom{50}{20} (.4)^{20} (.6)^{30} = .115$ [or *binompdf*(50, .4, 20)]
- $P(X \geq 20) = \binom{50}{20} (.4)^{20} (.6)^{30} + \binom{50}{21} (.4)^{21} (.6)^{29}$
 $+ \dots + \binom{50}{50} (.4)^{50} (.6)^0 =$ [or *1-binomcdf*(50, .4, 19) = .554]

3. A random variable X has $B(300, .2)$. Describe the sampling distribution of \hat{p} .

Answer: Because $300(.2) = 60 \geq 10$ and $300(.8) = 240 \geq 10$, \hat{p} has approximately a normal distribution with $\mu_{\hat{p}} = .2$ and $\sigma_{\hat{p}} = \sqrt{\frac{.2(1-.2)}{300}} = .023$

4. A distribution is known to be highly skewed to the left with mean 25 and standard deviation 4. Samples of size 10 are drawn from this population and the mean of each sample is calculated. Describe the sampling distribution of \bar{x} .

Answer: $\mu_{\bar{x}} = 25$, $\sigma_{\bar{x}} = \frac{4}{\sqrt{10}} = 1.26$.

Since the samples are small, the shape of the sampling distribution would probably show some left-skewness but would be more mound-shaped than the original population.

5. What is the probability that a sample of size 35 drawn from a population with mean 65 and standard deviation 6 will have a mean less than 64?

Answer: The sample size is large enough that we can use large-sample procedures. Hence,

$$P(\bar{x} < 64) = P\left(z < \frac{64 - 65}{6/\sqrt{35}} = -.99\right) = .1611.$$

[Calculator solution: $normalcdf(-100, 64, 65, 6/\sqrt{35})$]

3 PRACTICE PROBLEMS

Multiple Choice

1. A binomial event has $n = 60$ trials. The probability of success on each trial is .4. Let X be the count of successes of the event during the 60 trials. What are μ_x and σ_x ?
- 24, 3.79
 - 24, 14.4
 - 4.90, 3.79
 - 4.90, 14.4
 - 2.4, 3.79

2. Consider repeated trials of a binomial random variable. Suppose the probability of the first success occurring on the second trial is .25, what is the probability of success on the first trial?
- $\frac{1}{4}$
 - 1
 - $\frac{1}{2}$
 - $\frac{1}{8}$
 - $\frac{3}{16}$
3. To use a normal approximation to the binomial, which of the following does *not* have to be true?
- $np \geq 5, n(1-p) \geq 5$
 - The individual trials must be independent.
 - The sample size in the problem must be too large to permit doing the problem on a calculator.
 - For the binomial, the population size must be at least 10 times as large as the sample size.
 - All of the above are true.
4. You form a distribution of the means of all samples of size 9 drawn from an infinite population that is skewed to the left (like the scores on an easy Stat quiz!). The population from which the samples are drawn has a mean of 50 and a standard deviation of 12. Which one of the following statements is true of this distribution?
- $\mu_{\bar{x}} = 50, \sigma_{\bar{x}} = 12$, the sampling distribution is skewed somewhat to the left.
 - $\mu_{\bar{x}} = 50, \sigma_{\bar{x}} = 4$, the sampling distribution is skewed somewhat to the left.
 - $\mu_{\bar{x}} = 50, \sigma_{\bar{x}} = 12$, the sampling distribution is approximately normal.
 - $\mu_{\bar{x}} = 50, \sigma_{\bar{x}} = 4$, the sampling distribution is approximately normal.
 - $\mu_{\bar{x}} = 50, \sigma_{\bar{x}} = 4$, the sample size is too small to make any statements about the shape of the sampling distribution.
5. A 12-sided die has faces numbered from 1–12. Assuming the die is fair (that is, each face is equally likely to appear each time), which of the following would give the exact probability of getting at least 10 3s out of 50 rolls?
- $$\binom{50}{0}(.083)^0(.917)^{50} + \binom{50}{1}(.083)^1(.917)^{49}$$

$$+ \dots + \binom{50}{9}(.083)^9(.917)^{41}$$
 - $$\binom{50}{11}(.083)^{11}(.917)^{39} + \binom{50}{12}(.083)^{12}(.917)^{38}$$

$$+ \dots + \binom{50}{50}(.083)^{50}(.917)^0$$

$$\text{c. } 1 - \binom{50}{0}(.083)^0(.917)^{50} + \binom{50}{1}(.083)^1(.917)^{49} \\ + \cdots + \binom{50}{10}(.083)^{10}(.917)^{40}$$

$$\text{d. } 1 - \binom{50}{0}(.083)^0(.917)^{50} + \binom{50}{1}(.083)^1(.917)^{49} \\ + \cdots + \binom{50}{10}(.083)^9(.917)^{41}$$

$$\text{e. } \binom{50}{0}(.083)^0(.917)^{50} + \binom{50}{1}(.083)^1(.917)^{49} \\ + \cdots + \binom{50}{10}(.083)^{10}(.917)^{40}$$

Free Response

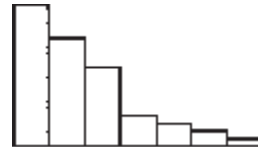
1. A factory manufacturing tennis balls determines that the probability that a single can of three balls will contain at least one defective ball is .025. What is the probability that a case of 48 cans will contain at least two cans with a defective ball?
2. A sampling distribution is highly skewed to the left. Describe the shape of the sampling distribution of \bar{x} if the sample size is (a) 3 or (b) 30.
3. Suppose you had gobs of time on your hands and decided to flip a coin 1,000,000 times and note whether each flip was a head or a tail. Let X be the count of heads. What is the probability that there are at least 1000 more heads than tails? (*Note:* this is a binomial but your calculator will not be able to do the binomial computation because the numbers are too large for it).
4. In Section 7.4, we had an example in which we asked if it would change the proportion of girls in the population (assumed to be .5) if families continued to have children until they had a girl and then they stopped. That problem was to be done by simulation. How could you use what you know about the geometric distribution to answer this same question?
5. At a school better known for football than academics (a school its football team can be proud of), it is known that only 20% of the scholarship athletes graduate within 5 years. The school is able to give 55 scholarships for football. What are the expected mean and standard deviation of the number of graduates for a group of 55 scholarship athletes?

6. Consider a population consisting of the numbers 2,4,5, and 7. List all possible samples of size two from this population and compute the mean and standard deviation of the sampling distribution of \bar{x} . Compare this with the values obtained by relevant formulas for the sampling distribution of \bar{x} . Note that the sample size is large relative to the population—this may affect how you compute $\sigma_{\bar{x}}$ by formula.
7. Approximately 10% of the population of the United States is known to have blood type B. If this is correct, what is the probability that between 11% and 15% of a random sample of 50 adults will have type B blood?
8. Which of the following is/are true of the central limit theorem? (More than one might be true.)
 - I. $\mu_{\bar{x}} = \mu$
 - II. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
 - III. The sampling distribution of a sample mean will be approximately normally distributed for sufficiently large samples, regardless of the shape of the original population.
 - IV. The sampling distribution of a sample mean will be normally distributed if the population from which the samples are drawn is normal.
9. A brake inspection station reports that 15% of all cars tested have brakes in need of replacement pads. For a sample of 20 cars that come to the inspection station,
 - (a) What is the probability that exactly 3 have defective breaks?
 - (b) What is the mean and standard deviation of cars that need replacement pads?
10. A tire manufacturer claims that his tires will last 40,000 miles with a standard deviation of 5000 miles.
 - (a) Assuming that the claim is true, describe the sampling distribution of the mean lifetime of a random sample of 160 tires. “Describe” means discuss center, spread, and shape.
 - (b) What is the probability that the mean life time of the sample of 160 tires will be less than 39,000 miles? Interpret the probability in terms of the truth of the manufacturer’s claim.
11. The probability of winning a bet on red in roulette is .474. The binomial probability of winning money if you play 10 games is .31 and drops to .27 if you play 100 games. Use a normal approximation to the binomial to estimate your probability of coming out ahead (that is, winning more than $\frac{1}{2}$ of your bets) if you play 1000 times. Justify being able to use a normal approximation for this situation.

12. Crabs off the coast of Northern California have a mean weight of 2 lbs. with a standard deviation of 5 oz. A large trap captures 35 crabs.
- Describe the sampling distribution for the average weight of a random sample of 35 crabs taken from this population.
 - What would the mean weight of a sample of 35 crabs have to be in order to be in the top 10% of all such samples?
13. The probability that a person recovers from a particular type of cancer operation is .7. Suppose 8 people have the operation. What is the probability that
- exactly 5 recover?
 - they all recover?
 - at least one of them recovers?
14. A certain type of light bulb is advertised to have an average life of 1200 hours. If, in fact, light bulbs of this type only average 1185 hours with a standard deviation of 80 hours, what is the probability that a sample of 100 bulbs will have an average life of at least 1200 hours?
15. Your task is to explain to your friend Gretchen, who knows virtually nothing (and cares even less) about statistics, just what a sampling distribution of the mean is. Explain the idea of a sampling distribution in such a way that even Gretchen, if she pays attention, will understand.
16. Consider the distribution shown at the right:

Describe the sampling distribution of \bar{x} for samples of size n if

- $n = 3$
- $n = 40$



17. After the Challenger disaster of 1986, it was discovered that the explosion was caused by defective O-rings. The probability that a single O-ring was defective and would fail (with catastrophic consequences) was .003 and there were 12 of them (6 outer and 6 inner). What was the probability that at least one of the O-rings would fail (as it actually did)?
18. Your favorite cereal has a little prize in each box. There are 5 such prizes. Each box is equally likely to contain any one of the prizes. So far, you have been able to collect 2 of the prizes. What is:
- the probability that you will get the third different prize on the next box you buy?
 - the average number of boxes of cereal you will have to buy before getting the third prize?

19. We wish to approximate the binomial distribution $B(40, .8)$ with a normal curve $N(\mu, \sigma)$. Is this an appropriate approximation and, if so, what are μ and σ for the approximating normal curve?
20. Opinion polls in 2002 showed that about 70% of the population had a favorable opinion of President Bush. That same year, a simple random sample of 600 adults living in the San Francisco Bay Area showed found only 65% that had a favorable opinion of President Bush. What is the probability of getting a rating of 65% or less in a random sample of this size if the true proportion in the population was .70?

3 CUMULATIVE REVIEW PROBLEMS

1. Toss three fair coins and let X be the count of heads among the three coins. Construct the probability distribution for this experiment.
2. You are doing a survey for your school newspaper and want to select a sample of 25 seniors. You decide to do this by randomly selecting 5 students from each of the 5 senior-level classes, each of which contains 28 students. The school data clerk assures you that students have been randomly assigned, by computer, to each of the 5 classes. Is this sample
 - (a) a random sample?
 - (b) a simple random sample?
3. Data are collected in an experiment to measure a person's reaction time (in seconds) as a function of the number of milligrams of a new drug. The least-squares regression line for the data is $Reaction\ Time = -.2 + .8(mg)$. Interpret the slope of the regression line in the context of the situation.
4. If $P(A) = .5$, $P(B) = .3$, and $P(A \text{ or } B) = .65$, are events A and B independent?
5. Which of the following examples of *quantitative data* and which are examples of *qualitative data*?
 - a. The height of an individual, measured in inches.
 - b. The color of the shirts in my closet.
 - c. The outcome of a flip of a coin described as "heads" or "tails."
 - d. The value of the change in your pocket.
 - e. Individuals, after they are weighed, are identified as thin, normal, or heavy.
 - f. Your pulse rate.
 - g. Your religion.

3 SOLUTIONS TO PRACTICE PROBLEMS

Multiple Choice

- The correct answer is (a). $\mu_x = (60)(.4)$, $\sigma_x = \sqrt{60(.4)(.6)} = \sqrt{14.4} = 3.79$
- The correct answer is (c). If it is a binomial random variable, the probability of success, p , is the same on each trial. The probability of not succeeding on the first trial and then succeeding on the second trial is $(1-p)(p)$. Thus, $(1-p)p = .25$. Solving algebraically, $p = 1/2$.
- The correct answer is (c). Although you probably wouldn't need to use a normal approximation to the binomial for small sample sizes, there is no reason (except perhaps accuracy) that you couldn't.
- The answer is (b).

$$\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

For small samples (typically for $n < 30$ or so), the shape of the sampling distribution of \bar{x} will resemble the shape of the sampling distribution of the original population. The shape of the sampling distribution of \bar{x} is approximately normal for n sufficiently large.

- The correct answer is (d). Because the problem stated “at least 10,” we must include the term where $x = 10$. If the problem has said “more than 10,” the correct answer would have been (b) or (c) (they are equivalent). The answer could also have been given as

$$\binom{50}{10}(.083)^{10}(.917)^{40} + \binom{50}{11}(.083)^{11}(.917)^{39} + \dots + \binom{50}{50}(.083)^{50}(.917)^0.$$

Free Response

- If X is the count of cans with at least one defective ball, then X has $B(48, .025)$.

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) + P(X = 1) = 1 - \binom{48}{0}(.025)^0(.975)^{48} \\ &\quad - \binom{48}{1}(.025)^1(.975)^{47} = .338. \end{aligned}$$

[Calculator solution: $1 - \text{binomcdf}(48, .025, 1)$.]

2. We know that the sampling distribution of \bar{x} will be similar to shape of the original population for small n and approximately normal for large n (that's the central limit theorem). Hence,
- If $n = 3$, the sampling distribution would probably be somewhat skewed to the left.
 - if $n = 30$, the sampling distribution would be approximately normal.

Remember that using $n \geq 30$ as a rule of thumb for deciding whether to assume normality is just that: a rule of thumb. This is probably a bit conservative. Unless the original population differs markedly from mound-shaped and symmetric, we would expect to see the sampling distribution of \bar{x} be approximately normal for considerably smaller values of n .

3. Because the “*binomcdf*” function can't be used, we will use a normal approximation to the binomial with $\mu_x = (1,000,000)(.5) = 500,000$ (assuming a fair coin) and $\sigma_x = \sqrt{(1,000,000)(.5)(.5)} = 500$. If we are to have at least 1000 more heads than tails, then there must be at least 500,500 heads (and, of course, no more than 495,500 tails). Thus, $P(\text{there are at least 1000 more heads than tails})$

$$= P(X \geq 500,500) = P\left(z > \frac{500,500 - 500,000}{500} = 1\right) = .1587.$$

4. The average wait for the first success to occur in a geometric setting is $1/p$, where p is the probability of success on any one trial. In this case, the probability of a girl on any one birth is $p = .5$. Hence, the average wait for the first girl is $1/.5 = 2$. So, we have one boy and one girl, on average, for each two children. The proportion of girls in the population would not change.
5. If X is the count of scholarship athletes that graduate from any sample of 55 players, then X has $B(55, .20)$. $\mu_x = 55(.20) = 11$ and $\sigma_x = \sqrt{55(.20)(.80)} = 2.97$
6. Putting the numbers 2,4,5, and 7 into a list in a calculator and doing 1-*Var Stats*, we have $\mu = 4.5$ and $s = 1.802775638$. The set of all samples of size 2 is $\{(2,4), (2,5), (2,7), (4,5), (4,7), (5,7)\}$ and the means of these samples are $\{3, 3.5, 4.5, 4.5, 5.5, 6\}$. Putting the means into a list and doing 1-*Var Stats* to find $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$, we get $\mu_{\bar{x}} = 4.5$ (which agrees with the formula) and $\sigma_{\bar{x}} = 1.040833$ (which does not agree with

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.802775638}{\sqrt{4}} = 1.27475878 \}$$

Because the sample is large compared with the population (that is, the population isn't at least 20 times as large as the sample), we use

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{1.802775638}{\sqrt{4}} \sqrt{\frac{4-2}{4-1}} = 1.040833,$$

which does agree with the computed value.

7. If

$$p = .10, \mu_{\hat{p}} = .10 \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{.10(1-.10)}{50}} = .042.$$

Then

$$\begin{aligned} P(.11 < \hat{p} < .15) &= P\left(\frac{.11 - .10}{.042} < z < \frac{.15 - .10}{.042}\right) \\ &= P(.238 < z < 1.19) = .289. \end{aligned}$$

8. All four of these statements are true. However, only III is a statement of the central limit theorem. The others are true of sampling distributions in general.

9. If X is the count of cars with defective pads, then X has $B(20, .15)$.

$$(a) P(X = 3) = \binom{20}{3} (.15)^3 (.85)^{17} = .243$$

[Calculator solution: $\text{binomcdf}(20, .15, 3)$.]

10. $\mu_{\bar{x}} = 40,000$ miles and $\sigma_{\bar{x}} = \frac{5000}{\sqrt{160}} = 395.28$ miles.

(a) With $n = 160$, the sampling distribution of \bar{x} will be approximately normally distributed with mean equal to 40,000 miles and standard deviation 395.28 miles.

$$(b) P(\bar{x} < 39,000) = P\left(z < \frac{39,000 - 40,000}{395.28} = -2.53\right) = .006.$$

If the manufacturer is correct, there is only about a .6% chance of getting an average this low or lower. That makes it unlikely to be just a chance occurrence and we should have some doubts about the manufacturer's claim.

11. If X is the number of times you win, then X has $B(1000, .474)$. To come out ahead, you must win more than half your bets. That is, you are being asked for $P(X > 500)$. Because $(1000)(.474) = 474$ and $1000(1 - .474) = 526$ are both greater than 10, we are justified in using a normal approximation to the binomial. Furthermore, we find that

$$\mu_X = 1000(.474) = 474 \text{ and } \sigma_X = \sqrt{1000(.474)(.526)} = 15.79.$$

Now,

$$P(X > 500) = P\left\{z > \frac{500 - 474}{15.79} = 1.65\right\} = .049.$$

That is, you have slightly less than a 5% chance of making money if you play 1000 games of roulette.

[The normal approximation solution using the calculator is $normalcdf(500, 1000, 474, 15.79) = .0498$; the exact binomial solution using the calculator is $1-binomcdf(1000, .474, 500) = .0467$.]

12. $\mu_{\bar{x}} = 2 \text{ lbs.} = 32 \text{ oz.}$ and $\sigma_{\bar{x}} = \frac{5}{\sqrt{35}} = .338 \text{ oz.}$

- (a) With samples of size 35, the central limit theorem tells us that the sampling distribution of \bar{x} is approximately normal with mean 32 oz. and standard deviation .338 oz.
- (b) In order for \bar{x} to be in the top 10% of samples, it would have to be at the 90th percentile, which tells us that its z -score is 1.28 [that's $InvNorm(.9)$ on your calculator]. Hence,

$$z_{\bar{x}} = 1.28 = \frac{\bar{x} - 32}{.338}.$$

Solving, we have $\bar{x} = 32.43$ oz. A crab would have to weigh at least 32.43 oz., or about 2 lb. 7 oz., to be in the top 10% of samples of this size.

13. If X is the number that recover, then X has $B(8, .7)$

(a) $P(X = 5) = \binom{8}{5} (.7)^5 (.3)^3 = .254$

[Calculator solution: $binompdf(8, .7, 5)$.]

$$(b) P(X = 8) = \binom{8}{8} (.7)^8 (.3)^0 = .058$$

$$(c) P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{8}{0} (.7)^0 (.3)^8 = .9999$$

$$14. \mu_{\bar{x}} = 1185 \text{ hours, and } \sigma_{\bar{x}} = \frac{80}{\sqrt{100}} = 8 \text{ hours.}$$

$$P(\bar{x} > 1200) = P\left(z > \frac{1200 - 1185}{8}\right) = P(z > 1.875) = .03$$

15. The first thing Gretchen needs to understand is that a distribution is just the set of all possible values of some variable. For example the distribution of SAT scores for the current senior class is just a list of all the SAT scores. We can draw samples from that population if, say, we want to estimate the average SAT score for the senior class but don't have the time or money to get all the data. Suppose we draw samples of size n and compute \bar{x} for each sample. Imagine drawing ALL possible samples of size n from the original distribution (that was the list of SAT scores for everybody in the senior class). Now consider the distribution (all the values) of means for those samples. That is what we call the sampling distribution of \bar{x} . (The short version: the sampling distribution of \bar{x} is the set of all possible values of \bar{x} computed from samples of size n .)

16. The distribution is skewed to the right.

- (a) If $n = 3$, the sampling distribution of \bar{x} will have some right skewness, but will be more mound-shaped than the parent population.
- (b) If $n = 40$, the central limit theorem tells us that the sampling distribution of \bar{x} will be approximately normal.

17. If X is the count of O-rings that failed, then X has $B(12, .003)$.

$$\begin{aligned} P(\text{at least one fails}) &= P(X = 1) + P(X = 2) + \dots + P(X = 12) \\ &= 1 - P(X = 0) = 1 - \binom{12}{0} (.003)^0 (.997)^{12} = .035. \end{aligned}$$

The clear message here is that even though the probability of any one failure seems remote (.003), the probability of at least one failure is large enough to be worrisome.

18. Because you already have 2 of the 5 prizes, the probability that the next box contains a prize you don't have is $3/5 = .6$. If n is the number of trials until the first success, then $P(X = n) = (.6)(.4)^{n-1}$.

(a) $P(X = 1) = (.6)(.4)^{2-1} = (.6)(.4) = .24$

(b) The average number of boxes you will have to buy before getting the third prize is $1/.6 = 1.67$.

19. $40(.8) = 32$ and $40(.2) = 8$. The rule we have given is that both nP and $n(1 - p)$ must be greater than 10 to use a normal approximation. However, as noted in Section 8.2, many texts allow the approximation when $np \geq 5$ and $n(1 - p) \geq 5$, so we will allow it. Also, we have

$$\mu = 40(.8) = 32, \text{ and } \sigma = \sqrt{40(.8)(.2)} = 2.53$$

20. If

$$p = .70, \mu_{\hat{p}} = .70 \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{.70(1 - .70)}{600}} = .019.$$

Then

$$P(\hat{p} < .65) = P\left\{z < \frac{.65 - .70}{.019} = -2.63\right\} = .004.$$

It appears that the San Francisco Bay Area may not be representative of the United States as a whole.

SOLUTIONS TO CUMULATIVE REVIEW PROBLEMS

1. The sample space for this event is {HHH, HHT, HTH, THH, HTT, HTH, THH, TTT}. The possible values of X are 0, 1, 2, or 3. Observation of the sample space yields the follows probability distribution:

x	0	1	2	3
$p(x)$	1/8	3/8	3/8	1/8

2. (a) Yes, it is a random sample because each student in any of the 5 classes is equally likely to be included in the sample.

(b) No, it is not a simple random sample (SRS) because not all samples of size 25 are equally likely. For example, in an SRS, one possible sample is having all 25 come from the same class. Because we only take 5 from each class, this isn't possible

3. The slope of the regression line is 0.8. The interpretation is that for each additional milligram of the drug, reaction time is *predicted* to increase by 0.8 seconds. Or you could say: for each additional milligram of the drug, reaction will increase by 0.8 seconds, *on average*.
4. $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = .5 + .3 - P(A \cap B) = .65 \rightarrow P(A \cap B) = .15$. Now, A and B are independent if $P(A \cap B) = P(A) \cdot P(B)$. So, $P(A) \cdot P(B) = (.3)(.5) = .15 = P(A \cap B)$. Hence, A and B are independent.
5. a. Quantitative
 b. Qualitative
 c. Qualitative
 d. Quantitative
 e. Qualitative
 f. Quantitative
 g. Qualitative

Chapter 9

Confidence Intervals and Introduction to Inference



Main concepts: *estimation, confidence intervals, t procedures; sample size; P value; statistical significance; hypothesis testing procedure; errors in hypothesis testing; power of a test*

ESTIMATION AND CONFIDENCE INTERVALS

As we proceed in statistics, our interest turns to estimating unknown population values. We have previously described a *statistic* as a value that describes a sample and a *parameter* as a value descriptive of a population. Now we want to think of a *statistic* as an **estimate** of a *parameter*. We know that if we draw multiple samples and compute some statistic of interest, say \bar{x} , that we will likely get different values each time even though the samples are all drawn from a population with a single mean, μ . What we now do is to develop a process by which we will use our *estimate* to generate a range of likely population values for the parameter. The statistic itself is called a **point estimate**, and the range of likely population values is a **confidence interval**.

example: We do a sample survey and find that 42% of the sample plans to vote for Normajeau for student body treasurer. That is, $\hat{P} = .42$. Based on this, we generate an interval of likely values (the confidence interval) for the true proportion of students who will vote for Normajeau and find that between 38% and 46% of the students are likely to vote for Normajeau. The interval $\langle .38, .46 \rangle$ is a confidence interval for the true proportion who will vote for Normajeau.

Note that saying a confidence interval is likely to contain the true population is not to say that it necessarily does. It may or may not—we will see ways to quantify just how “confident” we are in our interval.

In this chapter, we construct confidence intervals for a single mean, the difference between two means, a single proportion, and the difference between two proportions. Our ability to construct confidence intervals depends on our understanding of the sampling distributions for each of the parameters. In Chapter 8, we discussed the concept of sampling distribution for sample means and sample proportions. Similar arguments exist for the difference between two means or the difference between two proportions.

***t* Procedures**

When we discussed the sampling distribution of \bar{x} in Chapter 8, we assumed that we knew the population standard deviation. This is a big and questionable assumption because if we know the population standard deviation, we would probably also know the population mean and, if so, why are we drawing sample estimates of μ ? What saves us, of course, is the central limit theorem, which tells us that the sampling distribution of \bar{x} is approximately normal when the sample size, n , is large enough (roughly, $n \geq 30$), so we can use z procedures in these situations. We simply use s as an estimate of σ in this case. That is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} =: \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

When we estimate a standard deviation from data, we call the estimator the **standard error**. In this case, then,

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

is the standard *error* for

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

We will need the standard error for each different statistic we will use to generate a confidence interval. (A mnemonic device for remember what *standard error* stands for is: we are estimating the *standard* deviation but, because we are estimating it, there will probably be some *error*.) We will use this term from now on as we study inference because we will always be estimating the unknown standard deviation.

When n is small, we cannot safely assume the sampling distribution of \bar{x} is approximately normal. Under certain conditions (see below), the sampling distribution of \bar{x} follows a ***t* distribution**, which is similar in many respects to the normal distributions but which, because of the error

involved in using s to estimate σ , is more variable. How much more variable depends on the sample size. The t statistic is given by

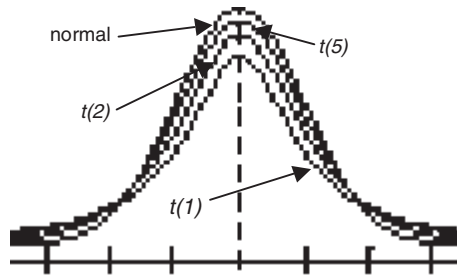
$$t = \frac{x - \mu}{s / \sqrt{n}}$$

This statistic follows a t distribution if the following are true.

- The population from which the sample was drawn is approximately normal, or the sample is large enough ($n \geq 30$).
- The sample is a SRS from the population.

There is a different t distribution for each n . The distribution is determined by the number of **degrees of freedom**, $df = n - 1$. We will use the symbol $t(k)$ to identify the t distribution with k degrees of freedom.

As n increases, the t distribution gets closer to the normal distribution. We can see this in the following graphic:



The table used for t values is set up differently than the table for z . In the table for z , the marginal entries are z scores, and the table entries are the corresponding areas under the normal curve to the left of z . In the t table, the left-hand column is *degrees of freedom*, the top margin gives upper tail probabilities, and the table entries are the corresponding **critical values** of t required to achieve the probability. In this book, we will use t^* (or z^*) to indicate critical values.

example: For 12 df , and an upper tail probability of .05, we see that the critical value of t is 1.782 ($t^* = 1.782$). For an upper tail probability of .02, the corresponding critical value is 2.303 ($t^* = 2.303$).

example: For 1000 df , the critical value of t for an upper tail probability of .025 is 1.962 ($t^* = 1.962$). This is very close to the critical z value for an upper tail probability of .025 is 1.96 ($z^* = 1.96$).

General Form of a Confidence Interval

A confidence interval is composed of two parts: an estimate of a population value and a margin of error. We identify confidence intervals by how confident we are that they contain the true population value.

A **C-Level Confidence Interval** has the following form:

$$(\text{estimate}) \pm (\text{margin or error})$$

The **margin of error** is also composed of two parts: the critical value of z or t (which is dependent on C), and the standard error. It has the following form:

$$\text{margin of error} = (\text{critical value})(\text{standard error}) = z^{**} \left(\frac{s}{\sqrt{n}} \right) \text{ OR } t^{**} \left(\frac{s}{\sqrt{n}} \right)$$

Putting the previous statements together, a *C-level confidence interval* has the following form: $(\text{estimate}) \pm z^{**} \left(\frac{s}{\sqrt{n}} \right)$ OR $(\text{estimate}) \pm t^{**} \left(\frac{s}{\sqrt{n}} \right)$.

example: A t confidence interval for μ would take the form:

$$\bar{x} \pm t^{**} \left(\frac{s}{\sqrt{n}} \right)$$

t^{**} is dependent on the *C level*, s is the sample standard deviation, and n is the sample size.

The *C level* is often expressed as a percent: a 95% confidence interval means that $C = .95$, or a 99% confidence interval means that $C = .99$. Although any value of C can be used as a confidence level, typical levels are .90, .95, and .99.

IMPORTANT: When we say that “We are 95% confident that the true population value lies in an interval,” we mean that the process used to generate the interval will capture the true population value 95% of the time. We are not making any probability statement about the interval. Our “confidence” is in the process that generated the interval. We do not know whether the interval we have constructed contains the true population value or not—it either does or it doesn’t. All we know for sure is that, on average, 95% of the intervals so constructed *will* contain the true value.



Exam Tip: For the exam, be VERY clear on the discussion above. Many students (not you, of course, but perhaps someone you know) seem to think that we can attach a probability to a confidence interval. We cannot.

example: Floyd told Betty that the probability was .95 that the 95% confidence interval he had constructed contained the mean of the population. Betty corrected him by saying that his interval either does contain the value ($P = 1$) or it doesn't ($P = 0$). This interval could be one of the 95 out of every 100 on average that contain the population mean, or it might be one out of the 5 out of every 100 that don't.

example: Find the critical value of t required to construct a 99% confidence interval for a population mean based on a sample of size 15.

solution: To use the t distribution table, we need to know the upper tail probability. Because $C = .99$, and confidence intervals are two-sided, the upper-tail probability is

$$\frac{1 - .99}{2} = .005.$$

Looking in the row for $df = 15 - 1 = 14$, and the column for .005, we find $t^* = 2.977$. Note that the table is set up so that if you look at the *bottom* of the table and find 99%, you are in the same column.

example: Find the critical value of z required to construct a 95% confidence interval for a population proportion.

solution: We are reading from the table of *Standard Normal Probabilities*. Remember that table entries are areas to the left of a given z score. With $C = .95$, we want

$$\frac{1 - .95}{2} = .025$$

in each tail, or .975 to the left if z^* . Finding .975 in the table, we have $z^* = 1.96$.

CONFIDENCE INTERVALS FOR MEANS AND PROPORTIONS

In the previous section we discussed the concept of a confidence interval. In this section, we get more specific by actually constructing confidence intervals for each of the parameters under consideration. The chart below lists each parameter for which we will construct confidence intervals, the conditions under which we are justified in constructing the interval, and the formula for actually constructing the interval. We are assuming that the population standard deviations are unknown.

• Parameter • Estimator	Conditions	Formula
• Population mean: μ • Estimator: \bar{x}	• SRS • Large sample ($n \geq 30$) or normal population • SRS • Small sample ($n < 30$) • Population approximately normal	$\bar{x} \pm z^* \frac{s}{\sqrt{n}}$ $\bar{x} \pm t^* \frac{s}{\sqrt{n}}, df = n - 1$
• Population proportion: p • Estimator: \hat{p}	• SRS • Large population size relative to sample size • $n\hat{p} \geq 5, n(1 - \hat{p}) \geq 5$ (or $n\hat{p} \geq 10, n(1 - \hat{p}) \geq 10$)	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$
• Difference of population means: $\mu_1 - \mu_2$ • Estimator: $\bar{x}_1 - \bar{x}_2$	• Independent SRSs • Large samples ($n_1 \geq 30$ and $n_2 \geq 30$) or normal populations • Independent SRSs • Approximately normal populations	$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $df = \min\{n_1 - 1, n_2 - 1\}$ or $df = n_1 + n_2 - 2$ ($n_1 = n_2$) or $df =$ (software) ($n_1 \neq n_2$)
• Difference of population proportions: $p_1 - p_2$ • Estimator: $\hat{p}_1 - \hat{p}_2$	• SRSs from independent populations • Large population sizes relative to sample sizes • $n_1\hat{p}_1 \geq 5, n_1(1 - \hat{p}_1) \geq 5$ $n_2\hat{p}_2 \geq 5, n_2(1 - \hat{p}_2) \geq 5$	$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$

Special note concerning the degrees of freedom for the sampling distribution of the difference of two means: There are several cases: (1) population variances are assumed to be equal; (2) variances are unequal, sample sizes are the same; (3) variances are unequal, sample sizes are unequal.

In any situation, a **conservative**, and usually acceptable, approach is to choose $df = \min\{n_1 - 1, n_2 - 1\}$. This is “conservative” in the sense that it will give a smaller number of degrees of freedom than other methods, which means that there will be more area in the tails.

- (1) If the population variances are equal, we can “pool” our estimates of the population standard deviation. In practice, this is rarely done because the statistical test for equal variances is quite weak. However, if we can make that assumption, then $df = n_1 + n_2 - 2$, and the standard error becomes

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}, \text{ where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

You will never be required to use this method, although you should know when it is allowed.

- (2) If the sample sizes are the same, then we can use $df = n_1 + n_2 - 2$.
 (3) If the sample sizes are unequal, and you are not using the conservative method, then it can be done by calculator or computer, which will compute the degrees of freedom as follows:

$$df = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{1}{n_1 - 1} \frac{s_1^2}{n_1} + \frac{1}{n_2 - 1} \frac{s_2^2}{n_2}}.$$

You probably don't want to do computation such as this by hand, but you could!

example: An airline is interested in determining the average number of unoccupied seats for all of its flights. It selects an SRS of 100 flights and determines that the average number of unoccupied seats for the sample is 12.5 seats with a sample standard deviation of 3.9 seats. Construct a 95% confidence interval for the true number of unoccupied seats for all flights.

solution: The problem states that the sample is an SRS and we have a large sample, so we are justified in constructing a large sample (z) confidence interval. For a 95% confidence interval, $z^* = 1.96$. We have:

$$12.5 \pm (1.96) \left(\frac{3.9}{\sqrt{100}} \right) = \langle 11.7, 13.3 \rangle.$$

example: Interpret the confidence interval from the previous example in the context of the problem.

solution: We are 95% confident that the true mean number of unoccupied seats is between 11.7 and 13.3 seats. (Remember that we are not making any probability statement about the particular interval we have constructed. Either the true mean is in the interval or it isn't)

For large sample confidence intervals utilizing z procedures, it is probably worth memorizing the critical values of z for the most common C levels of .90, .95, and .99. They are

C level	z^*
.90	1.645
.95	1.96
.99	2.576

example: Brittany thinks she has a bad penny because, after 150 flips she counted 88 heads. Find a 99% confidence interval for the true proportion of heads. Do you think the coin is biased?

solution: First we need to check to see if using a z interval is justified.

$$\hat{p} = \frac{88}{150} = .587, n\hat{p} = 150(.587) = 88.1, n(1 - \hat{p}) = 150(.413) = 62.$$

Because $n\hat{p}$ and $n(1 - \hat{p})$ are both greater than or equal to 5, we can construct a 99% z interval:

$$\begin{aligned} .587 \pm 2.576 \sqrt{\frac{(.587)(.413)}{150}} &= .587 \pm 2.576(.04) \\ &= .587 \pm .103 = \langle .484, .69 \rangle . \end{aligned}$$

We are 99% confident the true proportion of heads for this coin is between .484 and .69. If the coin were fair, we would expect, on average, 50% heads. Because .50 is in the interval, it is a likely population value for this coin. Brittany may well have a fair coin.



Calculator Tip: Your calculator will be perfectly happy to generate a confidence interval for you. The example above could have been done by *STAT TESTS 1-PropZInt* . . . All you have to do is enter x (the count), n (the sample size), and the C level. Each of the confidence intervals we do can be done on the TI-83.



Exam Tip: you must be careful not to just give answers directly from your calculator without supporting arguments and justifications, as in the solution to the example. You must communicate your process to the reader—answers alone, even if correct, will usually not receive full credit. The best compromise: do the problem as shown above and use the calculator to check your answer.

example: The following data were collected as part of a study. Construct a 90% confidence interval for the true difference between the means ($\mu_1 - \mu_2$). Does it seem likely that, despite the sample differences, that there is not real difference between the population means? The samples were SRSs from independent, approximately normal, populations.

Population	n	\bar{x}	s
1	20	9.87	4.7
2	18	7.13	4.2

solution: The relatively small values of n tells us that we need to use a 2-sample t interval. The conditions necessary for using this interval are given in the problem: SRSs from independent, approximately normal, populations. Using the “conservative” method of choosing the degrees of freedom:

$$df = \min\{n_1 - 1, n_2 - 1\} = \min\{19, 17\} = 17 \rightarrow t^* = 1.740.$$

$$\begin{aligned} (9.87 - 7.13) \pm 1.740 \sqrt{\frac{4.7^2}{20} + \frac{4.2^2}{18}} &= 2.74 \pm 1.740(1.444) \\ &= \langle .227, 5.25 \rangle. \end{aligned}$$

We are 90% confident that the true difference between the means lies in the interval from .227 to 5.25. If the true difference between the means is zero, we would expect to find 0 in the interval. Because it isn't, this interval provides evidence that there might be a real difference between the means.

If you do the same problems on your calculator (*STAT TESTS 2-SampTInt* . . .), you get $\langle .302, 5.178 \rangle$ with $df = 35.999$. This interval is narrower, highlighting the conservative nature of using $df = \min\{n_1 - 1, n_2 - 1\}$. Also, note the calculator calculates the number of degrees of freedom using df

$$= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1} \right) + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2} \right)}.$$

example: Construct a 95% confidence interval for $p_1 - p_2$ given that $n_1 = 180$, $n_2 = 250$, $\hat{p}_1 = .31$, $\hat{p}_2 = .25$. Assume that these are data from SRSs from independent populations.

solution: $180(.31) = 55.8$, $180(1 - .31) = 124.2$, $250(.25) = 62.5$, and $250(.75) = 187.5$ are all greater than or equal to 10 so, with what is

given in the problem, we have the conditions needed to construct a two-proportion z interval.

$$\begin{aligned} (.31 - .25) \pm 1.96 \sqrt{\frac{.31(1 - .31)}{180} + \frac{.25(1 - .25)}{250}} \\ = .06 \pm 1.96(.044) = \langle -.026, .146 \rangle \end{aligned}$$

SAMPLE SIZE

It is always desirable to select as large a sample as possible when doing research because larger samples are less variable than small samples. However, it is often expensive or difficult to draw larger samples so that we try to find the optimum sample size: large enough to accomplish our goals; small enough that we can afford it or manage it. We will look at techniques in this section for selecting sample sizes in the case of a large sample test for a single population mean and for a single population proportion.

Sample Size for Estimating a Population Mean (Large Sample)

The large sample confidence interval for a population mean is given by $\bar{x} \pm z^* \frac{s}{\sqrt{n}}$.

The margin of error is $z^* \frac{s}{\sqrt{n}}$. Let M be the desired maximum margin of error. Then,

$$M = z^* \frac{s}{\sqrt{n}}$$

Solving for n :

$$n = \left(\frac{z^* s}{M} \right)^2$$

An obvious problem here is that we cannot know s until after we collect our sample data. So we must have some basis for making an estimate of the population standard deviation. One common basis would be some historical knowledge of the type of data we are examining.

example: A machine for inflating tires, when properly calibrated, inflates tires to 32 pounds, but it is known that the machine varies with a standard deviation of about .8 lbs. How large a sample is needed in order to be 99% confident that the mean inflation pressure is within a margin of error of $B = .10$ lbs.?

solution:

$$n = \left(\frac{2.576(.8)}{.10} \right)^2 = 424.49.$$

Because the sample size needed must be an integer, 424 might not be enough. You would need a sample of 425 tires. In general, you should round *up* to the next integer when choosing a sample size by this method.

Sample Size for Estimating a Population Proportion

The confidence interval for a population proportion is given by:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

The margin of error is

$$z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

Let M be the desired maximum margin of error. Then,

$$M = z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

Solving for n ,

$$n = \left(\frac{z^*}{M} \right)^2 \hat{p}(1 - \hat{p}).$$

But we do not have a value of \hat{p} until we collect data, so we need a way to estimate \hat{p} . Let $P = \text{estimated value of } \hat{p}$. Then,

$$n = \left(\frac{z^*}{M} \right)^2 P(1 - P).$$

There are two ways to choose a value of P :

1. Use a previous determined value of \hat{p} . That is, you may already have an idea, based on historical data, of about what value \hat{p} should be close to.

2. Use $P = .5$. A result from calculus tells us that the expression

$$\left(\frac{z^*}{M} \right)^2 P(1 - P)$$

achieves its maximum value when $P = .5$. Thus, n will be at its maximum if $P = .5$. If $P = .5$, the formula for n can be expressed as

$$n = \left(\frac{z^*}{2M} \right)^2.$$

It is in your interest to choose the *smallest* value of n that will match your goals, so any value of $P < .5$ would be preferable *if* you have some justification for it.

example: Historically, about 60% of a company's products are purchased by people who have purchased products from the company before. The company is preparing to introduce a new product and want to generate a 95% confidence interval for the proportion of their current customers that will purchase the new product. They want to be accurate within 3%. How many customers do they need to sample?

solution: Based on historical data, choose $P = .6$. Then

$$n = \left(\frac{1.96}{.03} \right)^2 (.6)(.4) = 1024.4.$$

They need to sample 1025 customers. Had they not had the historical data, they would have had to use $P = .5$. If

$$P = .5, n = \left(\frac{1.96}{2(.03)} \right)^2 = 1067.1, \text{ or } 1068 \text{ customers.}$$

By using $P = .6$, they were able to sample 43 less customers.

STATISTICAL SIGNIFICANCE AND P VALUE

Statistical Significance

In the first two sections of this chapter, we used confidence intervals to make estimates about population values. In one of the examples, we went further and stated that because 0 was not in the confidence interval, 0 was not a likely population value from which to have drawn the sample that generated the interval. Now, of course, 0 could be the true population

value and our interval just happened to miss it. As we progress through techniques of inference (making predictions about a population from data), we often are interested in sample values that do not seem likely.

A finding or an observation is said to be **statistically significant** if it is unlikely to have occurred by chance. That is, if we *expect* to get a certain sample result and don't, it could be because of sampling variability (in repeated sampling from the same population, we will get different sample results even though the population value is fixed), or it could be because the sample came from a different population that we thought. If the result is far enough from expected that we think something other than chance is operating, then the result is statistically significant.

example: Todd claims that he can throw a football 50 yards. If he throws the ball 50 times and averages 49.5 yards, we have no reason to doubt his claim. If he only averages 30 yards, the finding is *statistically significant* in that he is unlikely to have a sample average this extreme if his claim was true.

In the above example, most people would agree that 49.5 was consistent with Todd's claim (that is, it was a likely average if the true value is 50) and that 30 is inconsistent with the claim (it is *statistically significant*). It's a bit more complicated to decide where between 30 and 49.5 the cutoff is between "likely" and "unlikely."

There are some general agreements about how unlikely a finding needs to be in order to be significant. Typical levels, as with confidence intervals, are .1, .05, and .01. The Greek letter α is called the **significance level**. If a finding has a lower probability of occurring than the significance level, then the finding is statistically significant.

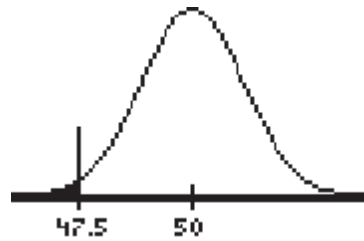
example: The school statistics teacher determined that the probability that Todd would only average 30 yards per throw if he really could throw 50 yards, is .002. If his value is so low that it seems unlikely to have occurred by chance, and so we say that the finding is significant. It is lower than any of the commonly accepted significance levels.

***P* Value**

We said that a finding is *statistically significant*, or *significant*, if it is unlikely to have occurred by chance. *P* value is what tells us just how unlikely a finding actually is. The ***P* value** is the probability of getting a finding (statistic) as extreme, or more extreme, as we obtained by chance alone. This requires that we have some expectation about what we ought to get. In other words, the *P* value is the probability of getting a finding at least as far removed from expected as we got. A decision about significance can then be made by comparing the obtained *P* value with a stated value of α .

example: Suppose it turns out that Todd's 50 throws are approximately normally distributed with mean 47.5 yards and standard deviation 8 yards. His claim is that he can average 50 yards per throw. What is the probability of getting a finding this far below expected by chance alone (that is, what is the P value) if his true average is 50 yards (assume the population standard deviation is 8 yards)? Is this finding *significant* at $\alpha = .05$? At $\alpha = .01$?

solution: We are assuming the population is normally distributed with mean 50 and standard deviation 8. The situation is pictured below:



$$z_{47.5} = \frac{47.5 - 50}{\frac{8}{\sqrt{50}}} = -2.21 \rightarrow P(\bar{X} < 47.5) = .014.$$

This is the P value: it's the probability of getting a finding as far below 50 as we did by chance alone. This finding is significant at the .05 level but not (quite) at the .01 level.

THE HYPOTHESIS-TESTING PROCEDURE

So far we have used confidence intervals to estimate the value of a population parameter (μ , p , $\mu_1 - \mu_2$, $p_1 - p_2$). In the coming chapters, we test whether the parameter has a particular value or not. We might ask if $p_1 - p_2 = 0$ or if $\mu = 3$, for example. That is, we will test the hypothesis that, say, $p_1 - p_2 = 0$. In the hypothesis-testing procedure, a researcher does not look for evidence to support this hypothesis, but instead looks evidence against the hypothesis. The process looks like this.

- **State hypotheses.** The first hypothesis, the **null hypothesis**, is the hypothesis we are actually testing. The null hypothesis usually states that there is no bias or that there is no distinction between groups. It is symbolized by H_0 . An example of a typical null hypothesis would be $H_0: \mu_1 - \mu_2 = 0$ or $H_0: \mu_1 = \mu_2$. This is the statement that μ_1 and μ_2 are the same, that populations 1 and 2 have the same mean.

The second hypothesis, the **alternative hypothesis**, is the theory that the researcher want to confirm by rejecting the *null hypothesis*. The

alternative hypothesis is symbolized by H_A . There are three forms for the alternative hypothesis: \neq , $>$, or $<$. That is, if the null is $H_0: \mu_1 - \mu_2 = 0$, then H_A could be:

$H_A: \mu_1 - \mu_2 \neq 0$ (this is called a **two-sided alternative**)

$H_A: \mu_1 - \mu_2 > 0$ (this is a **one-sided alternative**)

$H_A: \mu_1 - \mu_2 < 0$ (also a **one-sided alternative**)

[In the case of the one-sided alternative $H_A: \mu_1 - \mu_2 > 0$, the null hypothesis is sometimes written: $H_0: \mu_1 - \mu_2 \leq 0$]

- Identify which test statistic (so far, that's z or t) you intend to use and show that the conditions for its use are satisfied. We identified the conditions, or assumptions for the use of confidence interval in the first two sections of this chapter. We will identify the conditions needed to do hypothesis testing in the following chapters. For the most part, they are similar to those you have already studied.

If you are going to state a significance level, α , it can be done here.

- Compute the value of the test statistic and the P value.
- Using the value of the test statistic and/or the P value, give a conclusion in the context of the problem.



Exam Tip: The four steps above have been incorporated into AP Exam scoring for any question involving a hypothesis test. Note that the third item (compute the value of the test statistic and the P value), the mechanics in the problem, is only one part of the problem. All four steps must be present in order to receive a 4 (“complete response”) on the problem.

If you stated a significance level, the conclusion can be based on a comparison of the P value with α . If you didn't state a significance level, you can argue your conclusion based on the value of the P value alone: if it is small, you have evidence against the null; if it is not small, you do not have evidence against the null. Many statisticians will argue that you are better off to argue directly from the P value and not use a significance level. One reason for this is the arbitrariness of the P value. That is, if $\alpha = .05$, you would reject the null for a P value of .04999 but not for a P value of .05001 when, in reality, there is no practical difference between them.

The conclusion can be (1) that we reject H_0 (because of a sufficiently small P value) or (2) that we do not reject H_0 (because the P value is too large). We do not accept the null: we either reject it or fail to reject it. If we reject H_0 , we can say that we accept H_A or, preferably, that we have evidence in favor of H_A .

There are two approaches to making a decision to reject H_0 : (1) the *P value approach* (the P value is less than α) or (2) the *rejection region approach* (the value of the test statistic is larger than the critical value of the test statistic).

example: Consider one last time Todd and his claim that he can throw a ball 50 yards. His average toss, based on 50 throws, was 47.5 yards, and we assumed the population standard deviation was the same as the sample standard deviation, 8 years. A test of the hypothesis that Todd can throw the ball 50 yards on average against that alternative that he can't throw that far might look something like the following (we will fill in many of the details, especially those in the third part of the process, in the following chapters):

- Let μ be the true average distance Todd can throw a football.
 $H_0: \mu = 50$ (or $H_0: \mu \leq 50$, because the alternative is one-sided)
 $H_A: \mu < 50$
- We will use a large sample z test. We assume the 50 throws is a SRS of all his throws and the central limit theorem tells us that the sampling distribution of \bar{x} is approximately normal. We will use a significance level of $\alpha = .05$.
- In section 9.4, we determined that the P value for this situation is .014.
- Because the P value $< \alpha$ (.014 $<$.05), we can reject H_0 . We have good evidence that the true mean distance Todd can throw a football is actually less than 50 yards.

TYPE-I AND TYPE-II ERRORS AND THE POWER OF A TEST

When we do a hypothesis test as described in the previous section, we never really know if we have made the correct decision or not. We can try to minimize our chances of being wrong, but it is always a trade-off between how sure we want to be and how vague our decision must be. If we are given a hypothesis, it may be true or it may be false. We can decide to reject the hypothesis or not to reject it. This leads to four possible outcomes:

		Decision	
		Do not Reject	Reject
Hypothesis	True		
	False		

Two of the cells in the table are errors and two are not. Filling those in, we have

		Decision	
		Do not Reject	Reject
H y p o t h e s i s	True	OK	Error
	False	Error	OK

The errors have names that are rather unspectacular: If the (null) hypothesis is true, and we mistakenly reject it, it is a **type-I error**. If the hypothesis is false, and we mistakenly fail to reject it, it is a **type-II error**. We note that the probability of a type-I error is equal to α , the significance level (this is because a P value $< \alpha$ causes us to reject H_0 . If H_0 is true, and we still decide to reject it, we have made a type-I error). We call probability of a type-II error β . Filling in the table with this information, we have

		Decision	
		Do not Reject	Reject
H y p o t h e s i s	True	OK	Type-I Error $P(\text{Type-I}) = \alpha$
	False	Type-II Error $P(\text{type-II}) = \beta$	OK

The cell in the lower right-hand corner is important. An honest person does not want to foist a false hypothesis on the public and hopes that a study would lead to a correct decision to reject it. The probability of correctly rejecting a false hypothesis (in favor of the alternative) is called the **power of the test**. The power of the test equals $1 - \beta$. Finally, then, our decision table looks like this:

		Decision	
		Do not Reject	Reject
H y p o t h e s i s	True	OK	Type-I Error $P(\text{type-I}) = \alpha$
	False	Type-II Error $P(\text{type-II}) = \beta$	OK [power = $1 - \beta$]



Exam Tip: You will not need to know how to actually calculate P (type-II error) or the power of the test on the AP Exam. You *will* need to understand the concept of each.

example: Sticky Fingers is arrested for shoplifting. The judge, in her instructions to the jury, says that Sticky is innocent until proved guilty. That is, the jury's hypothesis is that Sticky is innocent. What risk is involved in a type-I and a type-II error?

solution: Our hypothesis is that Sticky is innocent. A type-I error involves mistakenly rejecting a true hypothesis. In this case, Sticky *is* innocent, but because we reject innocence, he is found guilty. The risk in a type-I error is that Sticky goes to jail for a crime he didn't commit.

A type-II error involves failing to reject a false hypothesis. If the hypothesis is false, then Sticky is guilty, but because we think he's innocent, we acquit him. The risk in type-II error is that Sticky goes free even though he is guilty.

In life we often have to choose between possible errors. In the example above, the choice was between sending an innocent person to jail (a type-I error) or setting a criminal free (a type-II error). Which of these is the most serious error is not a statistical question—it's a social one.

We can decrease the chance of type-I error by adjusting α . By making α very small, we could virtually ensure that we would never mistakenly reject a true hypothesis. However, this would result in a large type-II error because we are making it hard to reject the null under any circumstance, even when it is false.

We can reduce the probability of making a type-II error and, at the same time, increase the power of the test in the following ways:

- Increase the sample size.
- Decrease the standard deviation.
- Increase the significance level (α).
- State an alternative hypothesis that is farther away from the null.

example: A package delivery company claims that they are on time 90% of the time. Some of their clients aren't so sure, thinking that there are often delays in delivery beyond the time promised. The company states that they will change their delivery procedures if they are wrong in their claim. Suppose that, in fact, there are more delays than claimed by the company. Which of the following is equal to the power of the test?

- (a) The probability that the company will not change its delivery procedures.
- (b) The P value $\geq \alpha$.

- (c) The probability that the clients are wrong.
- (d) The probability that the company will change their delivery procedures.
- (e) The probability that the company will fail to reject H_0 .

solution: The power of the test is the probability of rejecting a false null hypothesis in favor of an alternative. In this case, the hypothesis that the company is on time 90% of the time is false. If we correctly reject this hypothesis, the company will change its delivery procedures. Hence, (d) is the correct answer.

RAPID REVIEW

1. True–False. A 95% confidence interval for a population proportion is given as $\langle .37, .52 \rangle$. This means that the probability is .95 that this interval contains the true proportion.

Answer: False. We are 95% confident that the true proportion is in this interval. The probability is .95 that the process used to generate this interval will capture the true proportion.

2. The hypothesis that the Giants would win the World Series in 2002 was held by many of their fans. What type of error has been made by a fan that refuses to accept the fact that the Giants actually lost the series to the Angels?

Answer: The hypothesis is false but the fan has failed to reject it. That is a type-II error.

3. What is the critical value of t for a 99% confidence interval based on a sample of size 26?

Answer: From the table of t distribution critical values, $t^* = 2.787$ with 25 df .

4. What is the critical value of z for a 98% confidence interval based on a large sample?

Answer: This time we have to use the *table of standard normal probabilities*. If $C = 0.98$, 0.98 of the area lies between z^* and $-z^*$. So, because of the symmetry of the distribution, 0.01 lies above z^* , which is the same as saying that 0.99 lies to the left of z^* . The nearest table entry to 0.99 is 0.9901, which corresponds to $z^* = 2.33$.

5. A hypothesis test is conducted with $\alpha = 0.01$. The P value is determined to be .037. Because P value $> \alpha$, are we justified in rejecting the null hypothesis?

Answer: No. We could only reject the null if the P value were less than the significance level. It is small probabilities that provide evidence against the null.

6. Mary comes running into your office and excitedly tells you that she got a statistically significant finding from the data on her most recent research project. What is she talking about?

Answer: Mary means that the finding she got had such a small probability of occurring by chance that she has concluded it probably wasn't just chance variation but a real difference from expected.

7. You want to create a 95% confidence interval for a population proportion with a margin or error of no more than 0.01. How large a sample do you need?

Answer: Because there is no indication in the problem that we know about what to expect for the population proportion, we will use $P = .5$. Then,

$$n = \left(\frac{1.96}{2(.05)} \right)^2 = 384.16.$$

You would need a minimum of 385 subjects for your sample.

8. Which of the following statements is correct?
- I. The t distribution has more area in its tails than the z distribution (normal).
 - II. You would always use a z procedure rather than a t procedure if you have a sample of size 30 or greater.
 - III. When constructing a 2-sample t interval, the "conservative" method of choosing degrees of freedom will result in a wider confidence interval than other methods.

Answer:

- I is correct. A t distribution because it must estimate the population standard deviation, has more variability than the normal distribution.
- II is not *necessarily* correct. In most cases, you may use, because of the central limit theorem, z procedures when you sample size is large, but there are distributions so different from normal that you might actually need even larger samples to justify the use of z procedures. Generally, it would be OK to use large-sample procedures if $n > 30$, but not always.
- III is correct. The conservative method ($df = \min\{n_1 - 1, n_2 - 1\}$) will give a larger value of t^* , which, in turn, will create a larger margin of error, which will result in a wider confidence interval than other methods for a given C level.

3 PRACTICE PROBLEMS**Multiple Choice**

1. You are going to create a 95% confidence interval for a population proportion and want the margin of error to be no more than 0.05. Historical data indicate that the population proportion has remained constant at about 0.7. What is the minimum size random sample you need to construct this interval?
 - a. 385
 - b. 322
 - c. 274
 - d. 275
 - e. 323

2. Which of the following will increase the power of a test?
 - a. Increase n .
 - b. Increase α .
 - c. Reduce the amount of variability in the sample.
 - d. Consider an alternative hypothesis further from the null.
 - e. All of these will increase the power of the test.

3. Under a null hypothesis, a sample value yields a P value of .015. Which of the following statements are true?
 - I. This finding is statistically significant at the .05 level of significance.
 - II. This finding is statistically significant at the .01 level of significance.
 - III. The probability of getting a sample value as extreme as this one by chance alone if the null hypothesis is true is .015.
 - a. I and III only
 - b. I only
 - c. III only
 - d. II and III only
 - e. I, II, and III.

4. You are going to construct a 90% t confidence interval for a population mean based on a sample of size 16. What is the critical value of t (t^*) you will use in constructing this interval?
 - a. 1.341
 - b. 1.753
 - c. 1.746
 - d. 2.131
 - e. 1.337

5. A 95% confidence interval for the difference between two population proportions is found to be $\langle 0.07, 0.19 \rangle$. Which of the following statements are true?
- I. It is unlikely that the two populations have the same proportions.
 - II. We are 95% confident that the true difference between the population proportions is between 0.07 and 0.19.
 - III. The probability is 0.95 that the true difference between the population proportions is between 0.07 and 0.19
- a. I only
 - b. II only
 - c. I and II only
 - d. I and III only
 - e. II and III only

Free Response

1. You attend a large university with approximately 15,000 students. You want to construct a 90% confidence interval estimate, within 5%, the proportion of students who favor outlawing mimes. How large a sample do you need?
2. The local farmers association in Cass County wants to estimate the mean number of bushels of corn produced per acre in the county. A random sample of 13 1-acre plots produced the following results (in number of bushels per acre): 98, 103, 95, 99, 92, 106, 101, 91, 99, 101, 97, 95, 98. Construct a 95% confidence interval for the mean number of bushels per acre in the entire county. The local association has been advertising that the mean yield per acre is 100 bushels. Do you think they are justified in this claim?
3. Two groups of 40 randomly selected students were selected to be part of a study on drop-out rates. One group was enrolled in a counseling program designed to give them skill needed to succeed in school and the other group was received no special counseling. Fifteen of the students who received counseling dropped out of school, and 23 of the students who did not receive counseling dropped out. Construct a 90% confidence interval for the true difference between the drop-out rates of the two groups. Interpret your answer in the context of the problem.
4. A hotel chain claims that the average stay for its business clients is 5 days with a standard deviation of 2.3 days. The hotel believes that the true stay may actually be fewer than 5 days. A study conducted by the hotel of 100 randomly selected clients yields a mean of 4.55 days with a standard deviation of 3.1 days. What is the probability of get-

ting a finding as extreme, or more extreme, as 4.55 if the true mean is really 5 days? That is, what is the P value of this finding?

5. One researcher wants to construct a 99% confidence interval as part of a study. A colleague says such a high level isn't necessary and that a 95% confidence level will suffice. In what ways will these intervals differ?
6. Which of the following will increase the *power* of a test?
 - (a) Increase the sample size.
 - (b) Decrease α .
 - (c) Increase α .
 - (d) Have an alternative hypothesis that is further from the null.
 - (e) Reduce the variability of the data.
7. A 99% confidence interval for a population mean is to be constructed. A sample of size 20 will be used for the study. Assuming that the population from which the sample is drawn is approximately normal, what is the upper critical value needed to construct the interval?
8. A university is worried that they might not have sufficient housing for their students for the next academic year. It's very expensive to build additional housing, so they are operating under the assumption (hypothesis) that the housing they have is sufficient, and they will spend the money to build additional housing only if they are convinced it is necessary (that is, they reject their hypothesis).
 - (a) For the university's assumption, what is the risk involved in making a type-I error?
 - (b) For the university's assumption, what is the risk involved in making a type-II error?
9. A flu vaccine is being tested for effectiveness. Three hundred fifty randomly selected people are given that vaccine and observed to see if they develop the flu during the flu season. At the end of the season, 55 of the 350 did get the flu. Construct a 95% confidence interval for the true proportion of people who will get the flu despite getting the vaccine.
10. A research study gives a 95% confidence interval for the proportion of subjects helped by a new anti-inflammatory drug is $\langle 0.56, 0.65 \rangle$.
 - (a) Interpret this interval in the context of the problem.
 - (b) What is the meaning of "95%" confidence interval as stated in the problem?
11. A study was conducted to see if attitudes toward travel have changed over the past year. In the prior year, 25% of American families took at

least one vacation away from home. In a random sample of 100 families this year, 29 families took a vacation away from home. What is the P value of getting a finding this different from expected?

Note: $s_{\hat{p}}$ is computed somewhat differently for a hypothesis test about a population proportion than $s_{\hat{p}}$ for constructing a confidence interval to estimate a population proportion. Specifically, for a confidence interval,

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

and, for a hypothesis test,

$$s_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}},$$

where p_0 is the hypothesized value of p in $H_0: p = p_0$. We do more with this in the next chapter, but you should use

$$s_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

for this problem.

12. A study done to determine if male and female 10th graders differ in performance in mathematics was conducted. Twenty-three randomly selected males and 26 randomly selected females were each given a 50-question multiple-choice test as part of the study. The scores were approximately normally distributed. The results of the study were as follows:

	Males	Females
Sample size	23	26
Mean	40.3	39.2
Std. deviation	8.3	7.6

Construct a 99% confidence interval for the true difference between the mean score for males and the mean score for females. Does the interval suggest that there is a difference between the true means for males and females?

13. Under $H_0: \mu = 35$, $H_A: \mu > 35$, a decision rule is decided upon that rejects H_0 for $\bar{x} > 36.5$. For the sample, $s_{\bar{x}} = .99$. If, in reality, $\mu = 38$, what is the power of the test?

14. You want estimate the proportion of Californians who want outlaw cigarette smoking in all public places. Generally speaking, by how much must you increase the sample size to cut the margin or error in half?
15. The Mathematics Department wants to estimate within five students, with 95% confidence, how many students will enroll in Statistics next year. They plan to ask a sample of eligible students whether or not they plan to enroll in Statistics. Over the past 5 years, the course has had between 19 and 79 students enrolled. How many students should they sample? (Note: assuming a reasonably symmetric distribution, we can estimate the standard deviation by $\text{Range}/4$).
16. A hypothesis test is conducted with $\alpha = 0.05$ to determine the true difference between the proportion of male and female students enrolling in statistics ($H_0: p_1 - p_2 = 0$). The p value of $\hat{p}_1 - \hat{p}_2$ is determined to be .03. Is this finding *statistically significant*? Explain what is meant by a statistically significant finding in the context of the problem.
17. Based on the 2000 census, the population of the United States was about 281.4 million people, and the population of Nevada was about 2 million. We are interested in generating a 95% confidence interval, with a margin of error of 3%, to estimate the proportion of people who will vote in the next presidential election. How much larger sample will we need to generate this interval for the United States than for the state of Nevada?
18. Professor Olsen has taught statistics for 41 years and has kept the scores of every test he has ever given. Every test has been worth 100 points. He is interested in the average test score over the years. He doesn't want to put all of the scores (there are thousands of them) into a computer to figure out the exact average so he asks his daughter, Anna, to randomly select 50 of the tests and use those to come up with an estimate of the population average. Anna has been studying statistics at college and decides to create a 98% confidence interval for the true average test score. The mean test score for the 50 random selected tests she selects is 73.5 with a standard deviation of 7.1. What does she tell her father?
19. A certain type of pen is claimed to operate for a mean of 190 hours. A random sample of 49 pens is tested, and the mean operating time is found to be 188 hours with a standard deviation of 6 hours.
 - (a) Construct a 95% confidence interval for the true mean operating time of this type of pen. Does the company's claim seem justified?
 - (b) Describe the steps involved in conducting a hypothesis test, at the 0.05 level of significance, that the true mean differs from 190 hours. Do not actually carry out the complete test, but do state the null and alternative hypotheses.

20. A young researcher thinks there is a difference between the mean ages at which males and females win Oscars for best actor or actress. The student found the mean age for all best actor winners and all best actress winners and constructed a 95% confidence interval for the mean difference between their ages. Is this an appropriate use of a confidence interval? Why or why not?

3 CUMULATIVE REVIEW PROBLEMS

- Use a normal approximation to the binomial to determine the probability of getting between 470 and 530 heads in 1000 flips of a fair coin.
- A survey of the number of televisions per household found the following probability distribution:

Televisions	Probability
0	.03
1	.37
2	.46
3	.10
4	.04

What is the mean number of television sets per household?

- A bag of marbles contains four red marbles and five blue marbles. A marble is drawn, its color is observed, and it is returned to the bag.
 - What is the probability that the first red marble is drawn on trial 3?
 - What is the average wait until a red marble is drawn?
- A study is conducted on which of two competing weight-loss programs is the most effective. Random samples of 50 people from each program are evaluated for losing and maintaining weight-loss over a 1-year period. The average number of pounds lost per person over the year is used as a basis for comparison.
 - Why is this an observational study and not an experiment?
 - Describe an experiment that could be used to compare the two programs. Assume that you have available 100 overweight volunteers who are not presently in any program.
- The correlation between the first and second statistics tests in a class is 0.78.
 - Interpret this value.
 - What proportion of the variation in the scores on the second test can be explained by the scores on the first test?

3 SOLUTIONS TO PRACTICE PROBLEMS

Multiple Choice

1. The correct answer is (e).

$$P = .7, M = .05, z^* = 1.96 \text{ (for a .95 } z \text{ interval)} \rightarrow$$

$$n = \left(\frac{z^*}{M} \right)^2 (P)(1 - P) = \left(\frac{1.96}{.05} \right)^2 (.7)(.3) = 322.7.$$

You need a sample of 323.

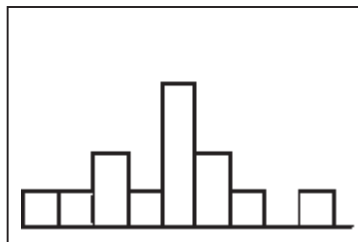
2. The correct answer is (e).
3. The correct answer is (a). It is not significant at the .01 level because .015 is greater than .01.
4. The correct answer is (b). $n = 16 \rightarrow df = 16 - 1 = 15$. Using a table of t distribution critical values, look in the row for 15 degrees of freedom and the column with 0.05 at the top (or 90% at the bottom).
5. The correct answer is (c). Because 0 is not in the interval $\langle .07, .19 \rangle$, it is unlikely to be the true difference between the proportions. III is just plain wrong! We cannot make a probability statement about an interval we have already constructed. All we can say is that the process used to generate this interval has a .95 chance of producing an interval that does contain the true population proportion.

Free Response

1. $C = .90 \rightarrow z^* = 1.645, M = .03. n = \left(\frac{1.645}{2(.05)} \right)^2 = 270.6.$

You would need to survey at least 271 students.

2. The population standard deviation is unknown, and the sample size is small (13), so we need to use a t procedure. The problem tells us that the sample is random. A histogram of the data shows no significant departure from normality, so the condition of normality is OK:



Now, $\bar{x} = 98.1$, $s = 4.21$, $df = 13 - 1 = 12 \rightarrow t^* = 2.179$. The 95% confidence interval is

$$98.1 \pm 2.179 \frac{4.21}{\sqrt{13}} = (95.56, 100.64).$$

Because 100 is contained in this interval, we do not have strong evidence against the association's claim that the mean number of bushels per acre is 100, even though the sample mean is only 98.1.

3. This is a two-proportion situation. We are told that the groups were randomly selected, but we need to check that the samples are sufficiently large:

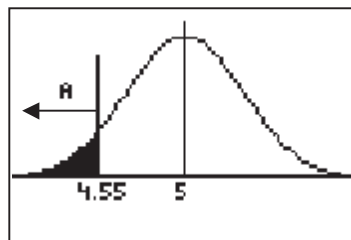
$$\begin{aligned}\hat{p}_1 &= \frac{15}{40} = .375, \hat{p}_2 = \frac{23}{40} = .575, \\ n_1\hat{p}_1 &= 40(.375) = 15, n_1(1 - \hat{p}_1) = 40(1 - .375) = 25, \\ n_2\hat{p}_2 &= 40(.575) = 23, n_2(1 - \hat{p}_2) = 40(1 - .575) = 17.\end{aligned}$$

Because all values are greater than or equal to 5, we are justified in constructing a *two-proportion z interval*. For a 90% z confidence interval, $z^* = 1.645$. Thus,

$$\begin{aligned}(.575 - .375) \pm 1.645 \sqrt{\frac{(.375)(1 - .375)}{40} + \frac{(.575)(1 - .575)}{40}} \\ = <.02, .38>.\end{aligned}$$

We are 90% confident that the true difference between dropout rates is between .02 and .38. Because 0 is not in this interval, we have evidence that the counseling program was effective at reducing the number of dropouts.

4. In this problem, $H_0: \mu = 5$ and $H_A: \mu < 5$, so we are only interested in the area to the left of our finding of $\bar{x} = 4.55$. We are interested in the area shaded in the graph:



Because $n = 100$, we use a z procedure.

$$z = \frac{4.55 - 5}{\frac{3.1}{\sqrt{100}}} = -1.45 \rightarrow A = .0735.$$

Thus, P value = .0735.

5. We will be more confident that the 99% confidence interval contains the population value being estimated, but the confidence interval will be wider than then 95% confidence interval.
6. All of these except (b) will increase the power of the test by making it easier to reject H_0 . Increasing the sample size will reduce the standard error, which makes the test statistic larger, which makes it easier to reject the null. Increasing α will also increase the rejection region. A more distant alternative makes it easier to detect an unlikely finding. Reducing the variability of the data will increase the rejection region.
7. t procedures are appropriate because the population is approximately normal. Thus, $n = 20 \rightarrow df = 19 \rightarrow t^* = 2.861$ for $C = .99$.
8. (a) A type-I error is made when we mistakenly reject a true null hypothesis. In this situation, that means that we would mistakenly reject the true hypothesis that the available housing is sufficient. The risk would be that a lot of money would be spent on building additional housing when it wasn't necessary.
(b) A type-II error is made when we mistakenly fail to reject a false hypothesis. In this situation that means the we would fail to reject the false hypothesis that the available housing is sufficient. The risk is that the university would have insufficient housing for its students.

$$9. \hat{p} = \frac{55}{350} = .157, n\hat{p} = 350(.157) = 54.95 \geq 5, n(1 - \hat{p}) \\ = 350(1 - .157) = 295.05 \geq 5.$$

The conditions are satisfied to construct a *one-sample* z interval.

$$.157 \pm 1.96 \sqrt{\frac{.157(1 - .157)}{350}} = \langle .119, .195 \rangle.$$

10. (a) We are 95% confident that the true proportion of subjects helped by a new anti-inflammatory drug is $\langle 0.56, 0.65 \rangle$
(b) The process used to construct this interval will capture the true population proportion, on average, 95 times out of 100.

11. We have $H_0: p = .25$, $H_A: p \neq .25$, $\hat{p} = .29$.

Because the hypothesis is two-sided, we are concerned about the probability of being in *either* tail of the curve even though the finding was larger than expected.

$$z_{0.29} = \frac{0.29 - 0.25}{\sqrt{\frac{(0.25)(1 - 0.25)}{100}}} = 0.92 \rightarrow \text{upper tail area}$$

$$= 1 - 0.8212 = 0.1788$$

$$P \text{ value} = 2(0.1788) = 0.3576.$$

12. The problem states that the samples were randomly selected and that the scores were approximately normally distributed, so we can construct a two-sample t interval. $df = \min\{23-1, 26-2\} = 22 \rightarrow t^* = 2.819$.

$$(40.3 - 39.2) \pm 2.819 \sqrt{\frac{8.3^2}{23} + \frac{7.6^2}{26}} = \langle -5.34, 7.54 \rangle.$$

0 is a likely value for the true difference between the means because it is in the interval. Hence, we do not have evidence that there is a difference between the true means for males and females.

[Note: Using *STAT TESTS 2-SampTInt* . . . on the TI-83, we get an interval of $\langle -5.04, 7.24 \rangle$, $df = 44.968$.]

13. The power of the test is our ability to reject the null hypothesis. In this case, we reject the null if $\bar{x} > 36.5$ when $\mu = 38$. We are given $s_{\bar{x}} = 0.99$. Thus

$$\text{Power} = P\left\{z > \frac{36.5 - 38}{.99} = -1.52\right\} = 1 - .0643 = .9357.$$

If the true mean is really 38, we are almost certain to reject the false null hypothesis.

14. For a given margin of error:

$$n = \left(\frac{z^*}{2M}\right)^2.$$

To reduce the margin of error by a factor of 0.5, we have

$$n^{**} = \left(\frac{z^*}{2(M/2)}\right)^2 = \left(\frac{2z^*}{2M}\right)^2 = 4\left(\frac{z^*}{2M}\right)^2 = 4n.$$

We need to quadruple the sample size to reduce the margin of error by a factor of $\frac{1}{2}$.

$$15. \sigma \approx \frac{\text{range}}{4} = \frac{79 - 19}{4} = 15.$$

Hence,

$$n = \left(\frac{1.96(15)}{5} \right)^2 = 34.57.$$

The department should ask at least 35 students about their intentions.

16. The finding is *statistically significant* because it is less than the significance level. In this situation, it is unlikely that we would have obtained a value of $\hat{p}_1 - \hat{p}_2$ as different from 0 as we got by chance alone if, in fact, $\hat{p}_1 - \hat{p}_2 = 0$.
17. Trick question! The sample size needed for a 95% confidence interval (or any C -level interval for that matter) is not a function of population size. The sample size is given by

$$n = \left(\frac{z^*}{M} \right)^2 P(1 - P).$$

n is a function of z^* (which is determined by the C level of the interval), M (the desired margin of error), and P (the estimated value of \hat{p}).

18. Because $n = 50$, we can use a large-sample confidence interval. For $n = 50$, $z^* = 2.33$ (that's the upper critical z value if we put 1% in each tail).

$$73.5 \pm 2.33 \left(\frac{7.1}{\sqrt{50}} \right) = < 71.16, 75.84 >.$$

Anna tells her father that he can be 98% confident that the true average test score for his 41 years of teaching is between 71.16 and 75.84.

19. The problems states that we have a SRS. With $n = 49$, we are justified in using z procedures because the central limit theorem tells us that the sampling distribution of \bar{x} is approximately normal.

$$(a) 188 \pm 1.96 \left(\frac{6}{\sqrt{49}} \right) = < 186.32, 189.68 >.$$

190 is not in this interval, so it is not a likely population value from which this sample might have been drawn. There is some doubt as to the company's claim.

- (b) Let μ = the true mean operating time of the company's pens.
- $H_0: \mu = 190$
 - $H_A: \mu \neq 190$
(the wording of the questions tells us H_A is two-sided)
 - We will use a 1 sample z test. Justify the conditions needed to use this procedure.
 - Determine the test statistic (z) and use it to identify the P value.
 - Compare the P value with α . Give a conclusion in the context of the problem.

20. It is not appropriate because confidence intervals use sample data to make estimates about unknown population values. In this case, the actual difference in the ages of actors and actresses is known, and the true difference can be calculated.

3 SOLUTIONS TO CUMULATIVE REVIEW PROBLEMS

1. Let X = the number of heads. Then X has $B(1000, 0.5)$ because the coin is fair. This binomial can be approximated by a normal distribution with mean = $1000(.5) = 500$ and standard deviation

$$s = \sqrt{1000(.5)(.5)} = 15.81. P(470 < X < 530)$$

$$= P\left(\frac{470 - 500}{15.81} < z < \frac{530 - 500}{15.81}\right) = P(-1.90 < z < 1.90) = .94.$$

2. $\mu_x = 0(0.3) + 1(.37) + 2(.46) + 3(.10) + 4(.04) = 1.75.$

3. (a) $P(\text{draw a red marble}) = 4/9.$

$$G(3) = \frac{4}{9} \left(1 - \frac{4}{9}\right)^{3-1} = \frac{4}{9} \left(\frac{5}{9}\right)^2 = .137$$

(b) Average wait

$$= \frac{1}{p} = \frac{1}{4/9} = \frac{9}{4} = 2.25.$$

4. (a) It is an observational study because the researchers are simply observing the results between two different groups. To do an experiment, the researcher must manipulate the variable of interest (different weight-loss programs) in order to compare their effects.

- (b) Inform the volunteers that they are going to be enrolled in a weight-loss program and their progress monitored. As they enroll, randomly assign them to one of two groups, say Group A and Group B (without the subjects knowledge that there are really two different programs). Group A gets one program and Group B the other. After a period of time, compare the average number of pounds lost for the two programs.
5. (a) There is a moderate to strong positive linear relationship between the scores on the first and second tests.
- (b) $0.78^2 = 0.61$

Chapter 10

Inference for Means and Proportions



We began our study of inference in the previous chapter by using confidence intervals to estimate population values. We estimated a single population mean and a single population proportion as well as the difference between two population means and between two population proportions. We discussed large sample (z) and small sample (t) procedures for estimating a population mean. In this chapter, we build on those techniques to evaluate claims about population values.

SIGNIFICANCE TESTING

Before we actually work our way through inference for means and proportions, we need to review the hypothesis testing procedure and to understand how questions involving inference will be scored on the AP Exam. In the previous chapter, we identified the steps in the hypothesis testing procedure as follows:

- I. **State hypotheses.** The first hypothesis, the **null hypothesis**, is the hypothesis we are actually testing. The null hypothesis usually states that there is no bias or that there is no distinction between groups. It is symbolized by H_0 .

The second hypothesis, the **alternative hypothesis**, is the theory that the researcher wants to confirm by rejecting the null hypothesis. The alternative hypothesis is symbolized by H_A . There are three forms for the alternative hypothesis: \neq , $>$, or $<$. That is, if the null is $H_0: \mu_1 - \mu_2 = 0$, then H_A could be:

$H_A: \mu_1 - \mu_2 \neq 0$ (this is called a **two-sided alternative**)

$H_A: \mu_1 - \mu_2 > 0$ (this is a **one-sided alternative**)

$H_A: \mu_1 - \mu_2 < 0$ (also a **one-sided alternative**)

(In the case of the one-sided alternative $H_A: \mu_1 - \mu_2 > 0$, the null hypothesis is sometimes written: $H_0: \mu_1 - \mu_2 \leq 0$)

- II. **Identify which test statistic (so far, that's z or t) you intend to use and show that the conditions for its use are satisfied.** If you are going to state a significance level, α , it can be done here.
- III. **Compute the value of the test statistic and the P value.**
- IV. Using the value of the test statistic and/or the P value, **give a conclusion in the context of the problem.**

If you stated a significance level, the conclusion can be based on a comparison of the P value with α . If you didn't state a significance level, you can argue your conclusion based on the value of the P value alone: if it is small, you have evidence against the null; if it is not small, you do not have evidence against the null.

The conclusion can be (1) that we reject H_0 (because of a sufficiently small P value) or (2) that we do not reject H_0 (because the P value is too large). We do not accept the null: we either reject it or fail to reject it. If we reject H_0 , we can say that we accept H_A or, preferably, that we have evidence in favor of H_A .

There are two approaches to making a decision to reject H_0 : (1) the P value approach (the P value is less than α) or (2) the *rejection region approach* (the value of the test statistic is larger than the critical value of the test statistic).

Significance testing involves making a decision about whether or not a finding is statistically significant. That is, is the finding sufficiently unlikely so as to provide good evidence for rejecting the null hypothesis in favor of the alternative? The four steps in the hypothesis testing process outlined above are the four steps that are required on the AP Exam when doing inference problems. In brief, every test of a hypothesis should have the following four steps:

- I. State the null and alternative hypotheses in the context of the problem.
- II. Identify the appropriate test and check that the conditions for its use are met.
- III. Do the correct mechanics, including the value of the test statistic and the P value (or rejection region).
- IV. State a correct conclusion in the context of the problem.



Exam Tip: You are not required to number the four steps on the exam, but it is a good idea to do so—then you are sure you done everything required. Note that the correct mechanics are only worth about 25% of the problem.

Large Sample versus Small Sample

In this chapter, we explore inference for means and proportions. When we deal with means, we may use, depending on the conditions, either t procedures or z procedures. With proportions, assuming the proper conditions are met, we deal only with large samples—that is, with z procedures.

We can use z procedures (large sample) for a sample mean or for the difference between sample means, when

- (a) We know the population standard deviation. We almost never will, so this is rarely a consideration.
- (b) The population from which the sample is drawn is known to be normal. In this case, the sampling distribution will also be normal, regardless of sample size.
- (c) The sample size is large enough to justify that the sampling distribution is approximately normally distributed based on the central limit theorem. In the case of inference for a single mean, a common rule of thumb is to consider the sample as “large enough” when $n > 30$.

We can use t procedures for a sample mean or for the difference between sample means, when

- (a) The sample is a random sample from the population.
- (b) The sample size is large ($n > 30$) **OR** the population from which the sample is drawn is approximately normal, or at least does not depart dramatically from normal.
- (c) The population standard deviation is unknown.

When using a t procedure on a small sample, it is important to check in step II of the hypothesis test procedure, that the data could plausibly have come from a normal distribution. A stemplot, boxplot, etc., can be used to show there are no outliers or extreme skewness in the data. t procedures are **robust** against these assumptions, which basically means that the procedures still work reasonably well even with some violation of the conditions. Some texts use the following guidelines for sample size when deciding whether or not to use t procedures:

- $n < 15$. Use t procedures if the data are close to normal (no outliers or skewness).
- $n > 15$. Use t procedures unless there are outliers or marked skewness.
- $n > 40$. Use t procedures for any distribution.

For the two-sample case discussed later, these guidelines can still be used if you replace n with $n_1 - n_2$.

Using Confidence Intervals for Two-Sided Alternatives

Consider a two-sided significance test at, say, $\alpha = .05$ and a confidence interval with $C = .95$. A sample statistic that would result in a significant finding at the .05 level would also generate a 95% confidence interval that would not contain the hypothesized value. Confidence intervals for two-sided hypothesis tests could then be used in place of generating a test statistic and finding a P value. If the sample value generates a C -level confidence interval that does not contain the hypothesized value of the parameter, then a significance test based on the same sample value would reject the null hypothesis at $\alpha = 1 - C$.

You should *not* use confidence intervals for hypothesis tests involving one-sided alternative hypotheses. For the purposes of this course, confidence intervals are considered to be two-sided.

INFERENCE FOR A SINGLE POPULATION MEAN

In step II of the hypothesis testing procedure, we need to identify the test to be used and justify the conditions needed. This involves calculating a **test statistic**. All test statistics have the following form:

$$\text{Test Statistic} = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error}}$$

When doing inference for a single mean, the estimator is \bar{x} , the hypothesized value is μ_0 in the null hypothesis $H_0: \mu = \mu_0$, and the standard error is the estimate of the standard deviation of \bar{x} , which is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \quad (df = n - 1, \text{ if you are using a } t \text{ procedure}).$$

This can be summarized in the following table.

Hypothesis	Estimator	Standard Error	Conditions	Test Statistic
• Null hypothesis $H_0: \mu = \mu_0$	• Estimator: \bar{x}	• Standard error: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$	• SRS • Large sample ($n \geq 30$) or normal population	$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$
			• SRS • Large sample OR population approximately normal	$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$, $df = n - 1$

Note: Paired samples (dependent samples) are a special case of one-sample statistics ($H_0: \mu_d = 0$.)

example: A study was done to determine if 12- to 15-year-old girls who want to be engineers differ in IQ from the average. The mean IQ of all girls in this age range is assumed to be about 100, based on prior research. A random sample of 49 girls are selected who state that they want to be engineers and their IQ is measured. The mean IQ of the girls in the sample is 104.5 with a sample standard deviation of 15. Does this finding provide evidence, at the .05 level of significance, that the mean IQ of 12- to 15-year-old girls who want to be engineers differs from the average?

solution 1 (test statistic approach): The solution to this problem will be put into a table to emphasize the format required when writing out solutions on the AP Exam. The table is *not* required.

<p>I. Let μ = the true mean IQ of girls who want to be engineers</p> <p>$H_0: \mu = 100$ $H_A: \mu \neq 100$</p> <p><i>(The alternative is two-sided because the problems wants to know if the mean IQ “differs” from 100. It would have been one-sided if it had asked whether the mean IQ of girls who want to be engineers is higher than average.)</i></p> <p>II. We will use a one-sample z test at $\alpha = .05$</p> <p>Conditions:</p> <ul style="list-style-type: none"> • The problem states that we have a random sample • The large sample size ($n = 49$) allows us to use either a t test or a z test. <p>III. $z = \frac{104.5 - 100}{15/\sqrt{49}} = 2.10 \rightarrow P \text{ value} = 2(1 - .9821) = .0358$</p> <p>IV. Because $P < \alpha$, we reject H_0. We have strong evidence that the true mean IQ for girls who want to be engineers differs from the mean IQ of all girls in this age range.</p>
--

Notes on the above solution:

- Had the alternative hypothesis been one-sided, the P value would have been $1 - .9821 = .0179$. We multiplied by 2 in step III because we needed to consider the area in **both** tails.
- The problem told us that the significance level was .05. Had it not mentioned a significance level, we could have arbitrarily chosen one,

or we could have argued a conclusion based only on the derived P value without a significance level.

- The linkage between the P value and the significance level must be made explicit in part IV. Some sort of statement, such as “Because $P < \alpha \dots$ ” or, if no significance level of stated, “Because the P value is low \dots ” will indicate that your conclusion is based on the P value determined in step III.

solution 2 (confidence interval approach):

- I. Let μ = the true mean IQ of girls who want to be engineers
 $H_0: \mu = 100$
 $H_A: \mu \neq 100$
- II. We will use a 95% confidence interval at $C = .95$.
 Conditions:
- The problem states that we have a random sample
 - The large sample size ($n = 49$) allows us to construct a z interval (although it is certainly acceptable, and perhaps preferable, to use a t interval).
- III. $\bar{x} = 104.5, z^* = 1.96$
 $104.5 \pm 1.96 \left(\frac{15}{\sqrt{49}} \right) = <100.3, 108.7>$
 We are 95% confident that the true population mean is in the interval $<100.3, 108.7>$
- IV. Because 100 is not in the 95% confidence interval for μ , we reject H_0 . We have strong evidence that the true mean IQ for girls who want to be engineers differs from the mean IQ of all girls in this age range.

example: A company president believes that there are more absences on Monday than on other days of the week. The company has 45 workers. The following table gives the number of worker absences on Mondays and Wednesdays for an 8-week period. Do the data provide evidence that there are more absences on Mondays?

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
Monday	5	9	2	3	2	6	4	1
Wednesday	2	5	4	0	3	1	2	0

solution: Because the data are paired on a weekly basis, the data we use for this problem are the difference between the days of the week for each of the 8 weeks. Adding a row to the table that gives the differences (absences on Monday minus absences on Wednesday), we have:

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
Monday	5	9	2	3	2	6	4	1
Wednesday	2	5	4	0	3	1	2	1
Difference	3	4	-2	3	-1	5	2	1

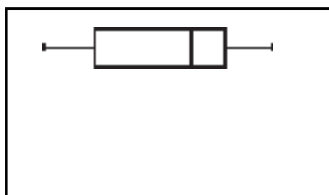
I. Let μ_d = the true mean difference between number of absences on Monday and absences on Wednesday.

$$H_0: \mu_d = 0$$

$$H_A: \mu_d > 0$$

II. We will use a *one-sample t* test for the difference scores

Conditions:



A boxplot of the data shows no significant departures from normality (no outliers or severe skewness). Also, we assume a random sample of days.

$$\text{III. } \bar{x} = 1.75, s = 2.49, t = \frac{1.75 - 0}{2.49 / \sqrt{8}} = 1.99, df = 8 - 1 = 7$$

$$\rightarrow .025 < P < .05 \text{ (from the table)}$$

[Using the TI-83, $P = .043$]

IV. The P value is small. This provides us with evidence that there are more absences on Mondays than on Wednesdays.

INFERENCE FOR THE DIFFERENCE BETWEEN TWO POPULATION MEANS

The two-sample case for the difference in the means of two independent samples is somewhat more complicated than the one-sample case. The hypothesis testing logic is identical, however, so the differences are in the mechanics needed to do the problems, not in the process. For hypotheses about the differences between two means, the procedures are summarized in the following table.

Hypothesis	Estimator	Standard Error	Conditions	Test Statistic
<ul style="list-style-type: none"> Null hypothesis: $H_0: \mu_1 - \mu_2 = 0$ OR $H_0: \mu_1 = \mu_2$ Estimator: $\bar{x}_1 - \bar{x}_2$ 			Large sample $(n_1 \geq 30 \text{ and } n_2 \geq 30)$ OR $(n_1 + n_2 > 40)$	$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
	Standard error: $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$		σ_1^2, σ_2^2 , two independent random samples from approximately normal populations OR Large samples $(n_1 \geq 30 \text{ and } n_2 \geq 30)$	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $df = \min\{n_1 - 1, n_2 - 1\},$ $df = n_1 + n_2 - 2 \text{ (} n_1 = n_2 \text{),}$ $df = \text{(software) } (n_1 \neq n_2)$
OR (if pooled):			σ_1^2, σ_2^2 , two independent random samples from approximately normal populations	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ $df = n_1 + n_2 - 2$
		$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$		

Notes on the above table:

- Generally speaking, you should not assume the population variances are equal. The bottom line in the table gives the conditions and test statistic if you can make the necessary assumption. The problem is that it is very difficult to justify the assumption.
- The middle row for conditions and test statistic give the typical t procedure approach. The first way of computing degrees of freedom is the “conservative” method introduced in Chapter 9. The “software” method is based on the complicated formula given in Section 9.2.

example: A statistics teacher, Mr. Srednih, gave a quiz to his 8:00 AM class and to his 9:00 AM class. There were 50 points possible on the quiz. The data for the two classes was as follows.

	n	\bar{x}	s
Group 1 (8:00 AM)	34	40.2	9.57
Group 2 (9:00 AM)	31	44.9	4.86

Before the quiz, some members of the 9:00 AM class had been bragging that later classes do better in statistics. Considering these two classes as random samples from the populations of 8:00 AM and 9:00 AM classes, do these data provide evidence at the .01 level of significance that 9:00 AM classes are better than 8:00 AM classes?

solution 1 (large sample)

- I. Let μ_1 = the true mean score for the 8:00 AM class
 Let μ_2 = the true mean score for the 9:00 AM class
 $H_0: \mu_1 - \mu_2 = 0$
 $H_A: \mu_1 - \mu_2 < 0$
- II. We use a *two-sample z* test at $\alpha = .01$. We assume these samples are random samples from independent populations. *z* procedures are justified because both sample sizes are larger than 30.
- III.
$$z = \frac{40.2 - 44.9}{\sqrt{\frac{9.57^2}{34} + \frac{4.86^2}{31}}} = -2.53 \rightarrow P \text{ value} = .006$$
- IV. Because $P < .01$, we reject the null hypothesis. We have good evidence that the true mean for the 9:00 AM class really is higher than the mean for the 8:00 AM class. It seems that the 9:00 AM class does have bragging rights.

solution 2 (*t* procedures):

- I. Let μ_1 = the true mean score for the 8:00 AM class
 Let μ_2 = the true mean score for the 9:00 AM class
 $H_0: \mu_1 - \mu_2 = 0$
 $H_A: \mu_1 - \mu_2 < 0$
- II. We will use a *two-sample t* test at $\alpha = .01$. We assume these samples are random samples from independent populations. *t* procedures are justified because both sample sizes are larger than 30.
- (Note: if the sample sizes were not large enough, we would need to know that the samples were drawn from populations that are approximately normally distributed.)

$$\text{III. (i) } t = \frac{40.2 - 44.9}{\sqrt{\frac{9.57^2}{34} + \frac{4.86^2}{31}}} = -2.53, df = \min\{34 - 1, 31 - 1\} = 30 \rightarrow .005 < P < .01$$

(*P* value from table; the TI-83 gives $tcdf(-100, -2.53, 30) = .008$)

(ii) Solution given by *STAT TESTS 2SampTTest* . . . :
 $t = -2.53, df = 49.92, P$ value = .007.

IV. Because $P < .01$, we reject the null hypothesis. We have good evidence that the true mean for the 9:00 AM class really is higher than the mean for the 8:00 AM class. It seems that the 9:00 AM class does have bragging rights.

INFERENCE FOR A SINGLE POPULATION PROPORTION

We are usually more interested in estimating a population proportion with a confidence interval than we are in testing that a population proportion has a particular value. However, significance testing techniques for a particular population proportion exist and follow a similar pattern to those of the previous two sections. The main difference is that the only test statistic is z . The logic is based on using a normal approximation to the binomial as discussed in Chapter 8 (see page 177).

• Hypothesis	• Estimator	• Standard Error	Conditions	Test Statistic
• Null hypothesis $H_0: p = p_0$	• Estimator: $\hat{p} = \frac{X}{n}$, where X is the count of successes.	• Standard error: $s_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}}$	• SRS • $np_0 \geq 5, n(1 - p_0) \geq 5$ (or $np_0 \geq 10, n(1 - p_0) \geq 10$)	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$

Notes on the table on page 242:

- Note that the standard error for a hypothesis test of a single population proportion is different for a confidence interval for a population proportion. The standard error for a confidence interval

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

is a function of \hat{p} , the sample proportion; whereas, the standard error for a significance test

$$s_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

is a function of the hypothesized value of p .

- Like the conditions for the use of a z interval, we require that the np_0 and $n(1 - p_0)$ be large enough to justify the normal approximation. As with determining the standard error, we use the hypothesized value of p rather than the sample value. The two different conditions given are simply the conditions that different basic statistics texts impose on the z test. You can use either 5 or 10.

example: Consider a screening test for prostate cancer that its maker claims will detect the cancer in 85% of the men that actually have the disease. One hundred seventy-five men who have been previously diagnosed with prostate cancer are given the screening test, and 141 of the men are identified as having the disease. Does this finding provide evidence that the screening test detects the cancer at a rate different from the 85% rate claimed by its manufacturer?

solution:

- I. Let p = the true proportion of men with prostate cancer that test positive.
 $H_0: p = .85$
 $H_A: p \neq .85$
- II. We want to use a z test. $175(.85) = 148.75 > 5$ and $175(1 - .85) = 26.25 > 5$, so the conditions are satisfied to use this test (*the conditions are met whether we use 5 or 10*).

$$\text{III. } \hat{p} = \frac{141}{175} = .806$$

$$z = \frac{.806 - .85}{\sqrt{\frac{.85(1 - .85)}{175}}} = 1.64 \rightarrow P \text{ value} = .10$$

[On the TI-83: STAT TESTS 1-PropZTest . . .]

- IV. Because P is reasonably large, we do not have sufficient evidence to reject the null hypothesis. The evidence is insufficient to challenge the companies claim that the test is 85% effective.

example: Maria has a quarter that she suspects is out of balance. In fact, she thinks it turns up heads more often than it would if it were fair. She has lots of time on her hands, so she decides to flip the coin 300 times and count the number of heads. There are 165 heads in the 300 flips. Does this provide evidence at the .05 level that the coin is biased in favor of heads? At the .01 level? Use a rejection region approach rather than computing a P value in your solution.

solution:

- I. Let p = the true proportion of heads if this is a fair quarter.

$$H_0: p = .50 \text{ (or } H_0: P \leq .50)$$

$$H_A: p > .50$$

- II. • We want to use a z test. $300(.50) = 150 > 5$ and $300(1 - .50) = 150 > 5$, so the conditions are satisfied to use this test.
• The upper critical z value for a one-sided significance test at .05 is $z^* = 1.645$. The upper critical z value for a one-sided test at .01 is $z^* = 2.33$.

$$\text{III. } \hat{p} = \frac{165}{300} = .55, z = \frac{.55 - .50}{\sqrt{\frac{.50(1 - .50)}{300}}} = 1.73$$

- IV. • At .05: Because $1.73 > 1.645$ (that is, $z > z^*$), we have grounds to reject the null and state that we have good evidence that the coin is biased in favor of heads.
• At .01: Because $1.73 < 2.33$, we do not have sufficient grounds to reject the null hypothesis that the coin is biased in favor of heads. At the .01 level, the coin could be fair and still yield the results we obtained.

INFERENCE FOR THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS

The logic behind inference for two population proportions is the same as for the others we have studied. As with the one-sample case, there are some differences between z intervals and z tests in terms of the computation of the standard error. The following table gives the essential details.

• Hypothesis	• Estimator	• Standard Error	Conditions	Test Statistic
<ul style="list-style-type: none"> Null hypothesis H_0: $p_1 - p_2 = 0$ (or H_0: $p_1 = p_2$) 	<ul style="list-style-type: none"> Estimator: $\hat{p}_1 - \hat{p}_2$ 	<ul style="list-style-type: none"> Standard error: 	<ul style="list-style-type: none"> SRSs from independent populations $n_1\hat{p}_1 \geq 5, n_1(1 - \hat{p}_1) \geq 5$ $n_2\hat{p}_2 \geq 5, n_2(1 - \hat{p}_2) \geq 5$ (note: some texts use 10 rather than 5) 	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ <p>where</p> $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$
	<p>where $\hat{p}_1 = \frac{X_1}{n_1}$,</p> $\hat{p}_2 = \frac{X_2}{n_2}$	$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ <p>where $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$</p>		

example: Two versions of a new vaccine are developed and a study is conducted to determine if they differ in their effectiveness. The results from the study are given in the following table.

	Vaccine A	Vaccine B
Infected	102	95
Not infected	123	190
Total	225	285

Does this study provide statistical evidence at the .01 level, that the vaccines differ in their effectiveness?

solution:

I. Let p_1 = the population proportion infected after receiving Vaccine A.

Let p_2 = the population proportion infected after receiving Vaccine B.

$$H_0: p_1 - p_2 = 0$$

$$H_A: p_1 - p_2 \neq 0$$

II. • We are going to use a *two-sample z test*.

$$\hat{p}_1 = \frac{102}{225} = .453, \hat{p}_2 = \frac{95}{285} = .333,$$

$$n_1\hat{p}_1 = 225(.453) = 102, n_1(1 - \hat{p}_1) = 225(.547) = 123,$$

$$n_2\hat{p}_2 = 285(.333) = 95, n_2(1 - \hat{p}_2) = 225(.667) = 190$$

All values are larger than 5, so we can use the *two-sample z test*.

(Note: When the values are given in table form, as in this problem, the values needed are the table entries! Being aware of this could save you some work.)

• The upper critical value for a two-sided z test at .01 is $z^* = 2.576$.

$$\text{III. } \bar{p} = \frac{102 + 95}{225 + 285} = .386$$

$$z = \frac{.453 - .333}{\sqrt{.386(1 - .386)\left(\frac{1}{225} + \frac{1}{285}\right)}} = 2.76 \rightarrow P \text{ value } .006$$

(Note: It was not strictly necessary to actually compute the P value because we had stated the upper critical value of z in Part II.)

IV. Because $P < .01$ (or since $z > 2.576$), we have grounds to reject the null hypothesis. We have strong evidence that the two vaccines differ in their effectiveness. Although this was two-sided test, we note that Vaccine A was less effective than Vaccine B.

RAPID REVIEW

1. A researcher reports that a finding of $\bar{x} = 3.1$ is significant at the .05 level of significance. What is the meaning of this statement?

Answer: Under the assumption that the null hypothesis is true, the probability of getting a value of \bar{x} at least as extreme as the one obtained is less than .05.

2. Let $\mu_1 =$ the mean score on a test of agility using a new training method and let $\mu_2 =$ the mean score on the test using the traditional method. Consider a test of $H_0: \mu_1 - \mu_2 = 0$. A large sample significance test finds $P = .04$. What conclusion, in the context of the problem, do you report if
- $\alpha = .05$?
 - $\alpha = .01$?

Answer:

- Because the P value is less than $.05$, we reject the null hypothesis. We have evidence that there is a non-zero difference between the traditional and new training methods.
 - Because the P value is greater than $.01$, we do not have sufficient evidence to reject the null hypothesis. We do not have strong support for the hypothesis that the training methods differ in effectiveness.
3. True–False: In a hypothesis test concerning a single mean, we can use either z procedures or t procedures as long as the sample size is at least 20.

Answer: False. With a sample size of only 20, we can not use z procedures unless we know that the population from which the sample was drawn is normal. We can use t procedures if the data do not have outliers or severe skewness; that is, if the population from which the sample was drawn is approximately normal.

4. We are going to do a two-sided significance test for a population proportion. The null hypothesis is $H_0: p = .3$. The simple random sample of 225 subjects yields $\hat{p} = .35$. What is the standard error involved in this procedure if
- you are constructing a confidence interval for the true population proportion?
 - you are doing a significance test for the null hypothesis?

Answer:

- (a) For a confidence interval, you use the value of \hat{p} in the standard

$$\text{error. Hence, } s_{\hat{p}} = \sqrt{\frac{(.35)(1 - .35)}{225}} = .0318.$$

- (b) For a significance test, you use the hypothesized value of p . Hence,

$$s_p = \sqrt{\frac{(.3)(1 - .3)}{225}} = .03055.$$

5. For the following data,
- justify the use of a *two-proportion z test* for $H_0: p_1 - p_2 = 0$?
 - what is the value of the test statistic for $H_0: p_1 - p_2 = 0$?
 - what is the P value of the test statistic?

	n	x	\hat{p}
Group 1	40	12	.3
Group 2	35	14	.4

Answer:

- (a) $n_1\hat{p}_1 = 40(.3) = 12$, $n_1(1 - \hat{p}_1) = 40(.7) = 28$, $n_2\hat{p}_2 = 35(.4) = 14$, $n_2(1 - \hat{p}_2) = 35(.6) = 21$. Because all values are greater than or equal to 5, we can use a *two-proportion z test*.

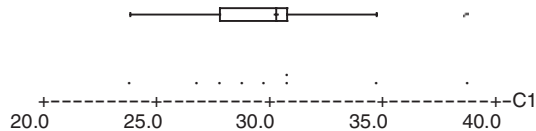
$$(b) \hat{p} = \frac{12 + 14}{40 + 35} = .35$$

$$z = \frac{.3 - .4}{\sqrt{.35(1 - .35)\left(\frac{1}{40} + \frac{1}{35}\right)}} = -.91$$

- (c) $z = -.91 \rightarrow P \text{ value} = 2(.18) = .36$

6. You want to conduct a *one-sample* significance test (*t test*) for a population mean. Your random sample of size 10 yields the following data: 26, 27, 34, 29, 38, 30, 28, 30, 30, 23. Should you proceed with your test? Explain

Answer: A boxplot of the data shows that the 38 is an outlier. Further, the dotplot of the data casts doubt on the approximate normality of the population from which this sample was drawn. Hence, you should *not* use a *t test* on these data.



7. Although it may be difficult to justify, there are conditions under which you can *pool* your estimate of the population standard deviation when doing a *2-sample* test for the difference between population means. When is this procedure justified? Why is it difficult to justify?

Answer: This procedure is justified when you can assume that the population variances (or population standard deviations) are equal. This is hard to justify because of the lack of a strong statistical test for the equality of population variances.

3 PRACTICE PROBLEMS

Multiple Choice

- A school district claims that the average teacher in the district earns \$48,000 per year. The teacher's organization argues that the average salary is less. A random sample of 25 teachers yields a mean salary of \$47,500 with a sample standard deviation of \$2000. Assuming that the distribution of all teacher's salaries is approximately normally distributed, what is the value of the t test statistic and the P value for a test of the hypothesis $H_0: \mu = 48,000$ against $H_A: \mu < 48,000$.

 - $t = 1.25, .10 < P < .15$
 - $t = -1.25, .20 < P < .30$
 - $t = 1.25, .20 < P < .30$
 - $t = -1.25, .10 < P < .15$
 - $t = -1.25, P > .25$
- Which of the following conditions are necessary to justify the use of z procedures in a significance test about a population proportion?

 - The samples must be drawn from a normal population.
 - The population must be much larger (10–20 times) than the sample
 - $np_0 \geq 5$ and $n(1 - p_0) \geq 5$
 - I only
 - I and II only
 - II and III only
 - III only
 - I, II, and III
- A minister claims that more than 70% of the adult population attends a religious service at least once a month. The null and alternative hypotheses you would use to test this claim would be:

 - $H_0: p \leq .7, H_A: p > .7$
 - $H_0: \mu = .7, H_A: \mu > .7$
 - $H_0: p = .7, H_A: p .7$
 - $H_0: p = .7, H_A: p > .7$
 - $H_0: p \geq .7, H_A: p < .7$
- A t test for the difference between the means to two populations is to be conducted. The samples, of sizes 12 and 15, are considered to be random samples from independent, approximately normally distributed, populations. Which of the following statements is true?

 - If we can assume the population variances are equal, the number is degrees of freedom is 25.

- II. An appropriate conservative estimate of the number of degree of freedom is 11.
 - III. The P value for the test statistic in this situation will be larger for 11 degrees of freedom than for 25 degrees of freedom.
 - a. I only
 - b. II only
 - c. III only
 - d. I and II only
 - e. I, II, and III
5. When is it OK to use a confidence interval instead of computing a P value in a hypothesis test?
- a. In any significance test.
 - b. In any hypothesis test with a two-sided alternative hypothesis.
 - c. Only when the hypothesized value of the parameter is *not* in the confidence interval.
 - d. Only when you are conducting a hypothesis test with a one-sided alternative.
 - e. Only when doing a test for a single population mean or a single population proportion.

Free Response

1. Which of the following is *not* a required step for a significance test?
 - a. State null and alternative hypotheses in the context of the problem.
 - b. Identify the test to be used and justify the conditions for using it.
 - c. State the significance level for which you will decide to reject the null hypothesis.
 - d. Compute the value of the test statistic and, if needed, the P value.
 - e. State a correct conclusion in the context of the problem.
2. Twenty-six pairs of identical twins are enrolled in a study to determine the impact of training on ability to memorize a string of letters. Two programs (A and B) are being studied. One member of each pair is randomly assigned to one of the two groups and the other twin goes into the other group. Each group undergoes the appropriate training program, and then the scores for pair of twins is compared. The means and standard deviations for groups A and B are determined as well as the mean and standard deviation for the difference between each twins score. Is this study a *one-sample* or *two-sample* situation, and how many degrees of freedom are involved in determining the t value?
3. Which of the following statements are correct?
 - I. A confidence interval can be used instead of a test statistic in any hypothesis test involving means or proportions.
 - II. A confidence interval can be used instead of a test statistic in a two-sided hypothesis test involving means or proportions.

- III. The standard error for constructing a confidence interval for a population proportion and the standard error for a significance test for a population proportion are the same.
- IV. The standard error for constructing a confidence interval for a population mean and the standard error for a significance test for a population mean are the same.
4. The average math SAT score at Hormone High School over the years is 520. The mathematics faculty believes that this year's class of seniors is the best the school has ever had in mathematics. One hundred seventy-five seniors take the exam and achieve an average score of 531 with a sample standard deviation of 96. Does this performance provide good statistical evidence that this year's class is, in fact, superior?
5. Under which of the following conditions are we justified in using z procedures?
- We know the population standard deviation.
 - Sample size is at least 30.
 - The conditions are met for doing inference for proportions.
 - I only
 - II only
 - I and II only
 - II and III only
 - I, II, and III
6. An avid reader, Booker Worm, claims that he reads books that average more than 375 pages in length. A random sample of 13 books on his shelf had the following number of pages: 595, 353, 434, 382, 420, 225, 408, 422, 315, 502, 503, 384, 420. Do the data support Booker's claim? Test at the .05 level of significance.
7. The statistics teacher, Dr. Tukey, gave a 50-point quiz to his class of 10 students and they didn't do very well, at least by Dr. Tukey's standards. Rather than continuing to the next chapter, he spent some time reviewing the material and then gave another quiz. The quizzes were comparable in length and difficulty. The results of the two quizzes were as follows.

Student	1	2	3	4	5	6	7	8	9	10
Quiz 1	42	38	34	37	36	26	44	32	38	31
Quiz 2	45	40	36	38	34	28	44	35	42	30

Do the data indicate that the review was successful, at the .05 level, of improving the performance of the students on this material? Give good statistical evidence for your conclusion.

8. The new reality TV show, "I Want to Marry a Statistician," has been showing on Monday evenings, and ratings show that it has

been watched by 55% of the viewing audience each week. The producers are moving the show to Wednesday night but are concerned that such a move might reduce the percentage of the viewing public watching the show. After the show has been moved, a random sample of 500 people who are watching television on Wednesday night are surveyed and asked what show they are watching. Two hundred fifty-five respond that they are watching “I Want to Marry a Statistician.” Does this finding provide evidence at the .01 level of significance that the percentage of the viewing public watching “I Want to Marry a Statistician” has declined?

9. Which of the following best describes what we mean when say that *t* procedures are robust?
- The *t* procedures work well with almost any distribution.
 - Its numerical value is not affected by outliers.
 - The *t* procedures will still work reasonably well even if the assumption of normality is violated.
 - t* procedures can be used as long as the sample size is at least 40.
 - t* procedures are as accurate as *z* procedures.
10. A company uses two different models, call them model A and model B, of a machine to produce electronic locks for hotels. The company has several hundred of each machine in use in its various factories. The machines are not perfect, and the company would like to phase out of service the one that produces the most defects in the locks. A random sample of 13 model A machines and 11 model B machines are tested and the data for the average number of defects per week are given in the following table.

	<i>n</i>	\bar{x}	<i>s</i>
Model A	13	11.5	2.3
Model B	11	13.1	2.9

Dotplots of the data indicate that there are no outliers or strong skewness in the data and that there are no strong departures from normal. Do these data provide statistically convincing evidence that the two machines differ in terms of the number of defects produced?

11. Take another look at the preceding problem. Suppose there were 30 of each model machine that were sampled. How might this have changed the hypothesis test you performed in that question? Answer both in terms of using the same test you used in Question 10 and in terms of a different test.
12. The directors of a large metropolitan airport claim that security procedures are 98% accurate in detecting banned metal objects that

passenger may try to carry onto a plane. The local agency charged with enforcing security thinks the security procedures are not as good as claimed. A study of 250 passengers showed that screeners missed nine banned carry-on items. What is the P value for this test and what conclusion would you draw based on it?

13. A group of 175 married couples are enrolled in a study to see if women have a stronger reaction to videos that contain violent material. At the conclusion of the study, each couple is given a questionnaire designed to measure the intensity of their reaction. Higher values indicate a stronger reaction. The means and standard deviations for all men, all women, and the differences between husbands and wives are as follows:

	\bar{x}	s
Men	8.56	1.42
Women	8.97	1.84
Difference (Husband–Wife)	–.38	1.77

Do the data give strong statistical evidence that wives have a stronger reaction to violence in videos than do their husbands?

14. An election is bitterly contested between two rivals. In a poll of 750 potential voters taken 4 weeks before the election, 420 indicated a preference for candidate Grumpy over candidate Dopey. Two weeks later, a new poll of 900 randomly selected potential voters found 465 who plan to vote for Grumpy. Dopey immediately began advertising that support for Grumpy was slipping dramatically and that he was going to win the election. Statistically speaking (say at the .05 level), how happy should Dopey be (i.e., how sure is he that support for Grumpy has dropped)?
15. Consider once again a problem given earlier in this problem set:

The new reality TV show, “I Want to Marry a Statistician,” has been showing on Monday evenings, and ratings show that it has been watched by 55% of the viewing audience each week. The producers are moving the show to Wednesday night but are concerned that such a move might reduce the percentage of the viewing public watching the show. After the show has been moved, a random sample of 500 people who are watching television on Wednesday night are surveyed and asked what show they are watching. Two hundred fifty-five respond that they are watching “I Want to Marry a Statistician.” Does this finding provide evidence at the .01 level of significance that the percentage of the viewing public watching “I Want to Marry a Statistician” has declined?

Suppose that the producers had no feeling for how the move to Wednesday night might affect their ratings but are interested in finding out. Redo the hypothesis test described in the box but use the two-sided alternative. If appropriate, use a confidence interval rather than a significance test for your statistical evidence.

16. A *two-sample* study for the difference between two population means will utilize *t procedures* and is to be done at the .05 level of significance. The sample sizes are 23 and 27. What is the upper critical *t score* (t^*) for the rejection region if
- the alternative hypothesis is one-sided, and the conservative method is used to determine the degrees of freedom?
 - the alternative hypothesis is two-sided and the conservative method is used to determine the degrees of freedom?
 - the alternative hypothesis is one-sided and you assume that the population variances are equal?
 - the alternative hypothesis is two-sided, and you assume that the population variances are equal?

3 CUMULATIVE REVIEW PROBLEMS

- How large a sample is needed to estimate a population proportion within 2.5% at the 99% level of confidence if
 - you have no reasonable estimate of the population proportion?
 - you have data that shows the population proportion should be about .7?
- Let X be a binomial random variable with $n = 250$ and $p = .6$. Use a normal approximation to the binomial to approximate $P(X > 160)$.
- Write the mathematical expression you would use to evaluate $P(X > 2)$ for a binomial random variable X that has $B(5, .3)$ (that is, X is a binomial random variable equal to the number of successes out of 5 trials of an event that occurs with probability of success $p = .3$). Do not evaluate.
- An individual is picked at random from a group of 55 office workers. Thirty of the workers are female, and 25 are male. Six of the women are administrators. Given that the individual picked is female, what is the probability she is an administrator?
- A random sample of 25 cigarettes of a certain brand were tested for nicotine content, and the mean was found to be 1.85 mg with a standard deviation of .75 mg. Find a 90% confidence interval for the mean number of mg in this type of cigarette. Assume that the amount of nicotine in cigarettes is approximately normally distributed. Interpret the interval in the context of the problem.

3 SOLUTIONS TO PRACTICE PROBLEMS

Multiple Choice

1. The correct answer is (d).

$$t = \frac{47500 - 48000}{2000/\sqrt{25}} = -1.25 \rightarrow .10 < P < .15$$

for the one-sided alternative. The calculator answer is $P = .112$. Had the alternative been two-sided, the correct answer would have been (b).

2. The correct answer is (c). If the sample size conditions are met, it is not necessary that the samples be drawn from a normal population.
3. The correct answer is (a). Often you will see the null written as $H_0: p = .7$ rather than $H_0: p \leq .7$. Either is correct.
4. The correct answer is (e). If we can assume that the variances are equal, then $df = n_1 + n_2 - 2 = 12 + 15 - 2 = 25$. A conservative estimate for the number of degrees of freedom is $df = \min\{n_1 - 1, n_2 - 1\} = \min\{12 - 1, 25 - 1\} = 11$. For a given test statistic, the greater the number of degrees of freedom, the less the P value.
5. The correct answer is (b). In this course, we consider confidence interval to all be two-sided. An two-sided α -level significance test will reject the null whenever the hypothesized value of the parameter is not contained in the $C = 1 - \alpha$ level confidence interval.

Free Response

1. (c) is not one of the required steps. You can state a significance level that you will later compare with the P value, but it is not required. You can simply argue the strength of the evidence against the null hypothesis based on the P value alone: low values of P provide evidence against the null.

Notice the wording in choice (d). It says compute the P value “if needed.” That’s because a rejection region approach only requires the value of the test statistic to be compared to a critical value of the test statistic.

2. This is a paired study because the scores for each pair of twins is compared. Hence, it is a *1-sample* situation, and there are 26 pieces of data to be analyzed, which are the 26 difference scores between the twins. Hence, $df = 26 - 1 = 25$.

3. • I is not correct. A confidence interval, at least in this course, cannot be used in any hypothesis test—only two-sided tests.
- II is correct. A confidence interval constructed from a random sample that does not contain the hypothesized value of the parameter can be considered significant evidence against the null hypothesis.
- III is not correct. The standard error for a confidence interval is based on the data

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

whereas, the standard error for a significance test is based on the hypothesized population value

$$s_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

- IV is correct.

4. I. Let μ = the population mean score for all students taking the exam this year

$$H_0: \mu = 520$$

$$H_A: \mu > 520$$

- II. We want to use a *one-sample z test*. We consider the 175 students to be a random sample of students taking the exam. The sample size is large, so *z procedures* are justified.

III. $\bar{x} = 531, s = 96, s_{\bar{x}} = \frac{96}{\sqrt{175}} = 7.26$

(because of the large sample size, we can consider s to be a good estimate of the population standard deviation.)

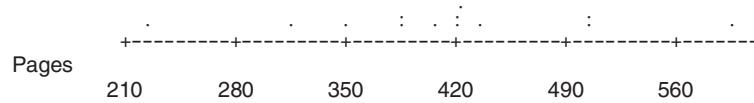
$$z = \frac{531 - 520}{\frac{96}{\sqrt{175}}} = 1.52 \text{ ? } P \text{ value} = .064$$

[On the TI-83: STAT TESTS *z Test* . . .]

- IV. The P value of .064 is reasonably low but is not low enough to provide strong evidence that the current class is superior in math ability as measure by the SAT.

5. e

6. I. Let μ = the true average number of pages in the books Booker reads.
 $H_0: \mu \leq 375$
 $H_A: \mu > 375$
- II. We are going to use a *one-sample t test* test at $\alpha = .05$. The problem states that the sample is a random sample. A dot plot of the data shows good symmetry and no significant departures from normality (although the data do spread out quite a bit from the mean, there are no outliers):



The conditions for the *t test* are met.

- III. $n = 13, \bar{x} = 412.5, s = 91.35$
 $t = \frac{412.5 - 375}{91.35 / \sqrt{13}} = 1.48, df = 12? \quad .05 < P < .10$
 (from the table) ($P = .082$ from $tcdf(1.48, 100, 12)$ or from *STAT TESTS t test . . .*)
- IV. Because $P > .05$, we cannot reject H_0 . We do not have strong evidence to back up Booker's claim that the books he reads actually average more than 375 pages in length.

7. The data are paired by individual students, so we need to test the difference scores for the students rather than the means for each quiz. The differences are given in the following table.

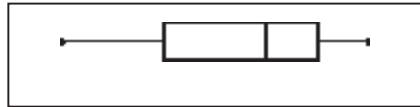
Student	1	2	3	4	5	6	7	8	9	10
Quiz 1	42	38	34	37	36	26	44	32	38	31
Quiz 2	45	40	36	38	34	28	44	35	42	30
Difference (Q2-Q1)	3	2	2	1	-2	2	0	3	4	-1

- I. Let μ_d = the mean of the differences between the scores of students on Quiz 2 and Quiz 1.

$$H_0: \mu_d = 0$$

$$H_A: \mu_d > 0$$

- II. This is a matched *pairs t test*. That is, it is a *one-sample t test* for a *population mean*. We assume that these are random samples from the populations of all students who took both quizzes. The significance level is $\alpha = .05$.



A boxplot of the difference scores shows no significant departures from normality, so the conditions to use the *one-sample t test* are satisfied.

- III. $n = 10$, $\bar{x}_d = 1.4$, $s_d = 1.90$

$$t = \frac{1.4 - 0}{1.90 / \sqrt{10}} = 2.33, df = 9? \quad .02 < P < .025$$

(from the table; $P = .022$ from *STAT TESTS t Test . . .*)

- IV. Because $P < .05$, we reject the null. The data provide evidence at the .05 level that the review was successful at improving student performance on the material.

8. I. Let p = the true proportion of Wednesday night television viewers who are watching “I Want to Marry a Statistician.”

$$H_0: p = .55$$

$$H_A: p < .55$$

- II. We want to use a *one-proportion z test* at $\alpha = .01$. $500(.55) = 275 > 5$ and $500(1 - .55) = 225 > 5$. Thus, the conditions needed for this test have been met.

III. $\hat{p} = \frac{255}{500} = .51$

$$z = \frac{.51 - .55}{\sqrt{\frac{.55(1 - .55)}{500}}} = -1.80 \rightarrow P \text{ value} = .036$$

[On the TI-83: *STAT TESTS 1-PropZTest . . .*]

IV. Because $P > .01$, we do not have sufficient evidence to reject the null hypothesis. The evidence is insufficient to conclude at the .01 level that the proportion of viewers has dropped when the program was moved to a different night.

9. (c) is the most correct response. (a) is incorrect because *t procedures* do not work well with small samples that come from non-normal populations. (b) is false because the numerical value of *t* is, like *z*, affected by outliers. *t procedures* are generally OK to use for samples of size 40 or larger, but this is not what is meant by *robust* [so (d) is incorrect]. (e) is not correct because the *t distribution* is more variable than the standard normal. It becomes closer to *z* as sample size increases but are “as accurate” only in the limiting case.

10. I. Let μ_1 = the true mean number of defects produced by machine A.
Let μ_2 = the true mean number of defects produced by machine B.
- $H_0: \mu_1 - \mu_2 = 0$
 $H_A: \mu_1 - \mu_2 \neq 0$
- II. We use a *two-sample t test* for the difference between means. The conditions for this procedure are given in the problem: both samples are simple random samples from independent, approximately normal populations.
- III. $df = \min \{13 - 1, 11 - 1\} = 10$.
- $$t = \frac{11.5 - 13.1}{\sqrt{\frac{2.3^2}{13} + \frac{2.9^2}{11}}} = -1.48 \rightarrow .10 < P < .20 \text{ (from table)}$$
- (When read directly from the table, $t = -1.48$ with 10 df is between tail probabilities of .05 and .10. However, those are one-sided values and must be doubled for the two-sided alternative.)
- [From the TI-83, *STAT TESTS 2-SampTTest* . . . , $P = .16$, $df = 18.99$.]
- IV. The *P* value is too large to be strong evidence against the null hypothesis that there is no difference between the machines. We do not have strong evidence that the types of machines actually differ in the number of defects produced.

11. • Using a *two-sample t test*. Steps I and II would not change. Step III would change to

$$t = \frac{11.5 - 13.1}{\sqrt{\frac{2.3^2}{30} + \frac{2.9^2}{30}}} = -2.37 \Rightarrow .02 < P < .04$$

based on either $df = \min\{30 - 1, 30 - 1\} = 29$, or $df = 30 + 30 - 2 = 58$ (which could be used because of the equal sample sizes). Step IV would probably arrive at a different conclusion—reject the null because the P value is small.

- Step I does not change. Because the samples are large (both = 30), we could declare the use of a *two-sample z test* in step II. In step III, the z score would be the same as the t computed earlier. Thus, $z = -2.37 \rightarrow P = .018$. Step IV would agree with the t test with the n's = 30 (Reject).

12. $H_0: p = .98, H_A: p < .98, \hat{p} = \frac{241}{250} = .964$.

$$z = \frac{.964 - .98}{\sqrt{\frac{(.98)(.02)}{250}}} = -1.81 \rightarrow P \text{ value} = .035.$$

This P value is quite low and provides evidence against the null and in favor of the alternative that security procedures actually detect less than the claimed percentage of banned objects.

13. I. Let μ_d = the mean of the differences between the scores of husbands and wives

$$H_0: \mu_d = 0$$

$$H_A: \mu_d < 0$$

- II. This is a *matched pairs z test*. That is, it is a *one-sample z test for a population mean*. The data were collected on pairs of husbands and wives, which means the data are paired. We assume that this is a random sample of married couples. It is quite a large sample ($n = 175$), so we are justified in using *z procedures*.

III. $z = \frac{-.38 - 0}{1.77 / \sqrt{175}} = -2.84 \rightarrow P \text{ value} = .002$

- IV. Because P is very small, we reject H_0 . The data provide strong evidence that women have a stronger reaction to violence in videos than do men.

14. I. Let p_1 = the true proportion of voters who plan to vote for Grumpy 4 weeks before the election.
Let p_2 = the true proportion of voters who plan to vote for Grumpy 2 weeks before the election.

$$H_0: p_1 - p_2 = 0$$

$$H_A: p_1 - p_2 > 0$$

- II. We will use a *two-sample z test for the difference between two population proportions*. Both samples are random samples of the voting populations at the time.

$$\hat{p}_1 = \frac{420}{750} = .56, \hat{p}_2 = \frac{465}{900} = .517$$

$$\text{Also, } n_1\hat{p}_1 = 750(.56) = 420, n_1(1 - \hat{p}_1) \\ = 750(.44) = 330,$$

$$n_2\hat{p}_2 = 900(.517) = 465.3, n_2(1 - \hat{p}_2) \\ = 900(.483) = 434.7$$

All combinations are larger than 5, so the conditions for the *two-proportion z test* are met.

$$\text{III. } \hat{p} = \frac{420 + 465}{750 + 900} = \frac{885}{1650} = .54$$

$$z = \frac{.56 - .517}{\sqrt{.54(.46)\left(\frac{1}{750} + \frac{1}{900}\right)}} = 1.75 \rightarrow P \text{ value} = .04$$

[From the TI-83, *STAT TESTS 2-PropZTest* . . . , $P = .039$.]

- IV. Because $P < .05$, we can reject the null hypothesis. Candidate Grumpy does seem to have cause for celebration—there is strong evidence that support for candidate Dopey is dropping.

15. I. Let p = the true proportion of Wednesday night television viewers who are watching “I Want to Marry a Statistician”

$$H_0: p = .55$$

$$H_A: p \neq .55$$

- II. We want to use a 99% confidence interval for the true proportion of people watching the show after the move to Wednesday night.

$$\hat{p} = \frac{255}{500} = .51$$

$500(.51) = 255 > 5$ and $500(1-.51) = 245 > 5$. Thus, the conditions needed for this test have been met.

(Remember: the conditions for using a confidence interval depend on \hat{p} ; whereas, a significance test depends on p_0 , the hypothesized value of P .)

$$\text{III. } \hat{p} = \frac{255}{500} = .51, C = .99 \rightarrow z^* = 1.96$$

$$.51 \pm 2.576 \sqrt{\frac{.51(1-.51)}{500}} = \langle .45, .57 \rangle$$

[On the TI-83: STAT TESTS 1-PropZInt . . .]

IV. Because the hypothesized value of p (.55) is contained in the 99% confidence interval generated, we do not have sufficient evidence to reject the null hypothesis. The evidence is insufficient to conclude at the .01 level that the proportion of viewers has dropped when the program was moved to a different night.

16. (a) $df = \min\{23 - 1, 27 - 1\} = 22 \rightarrow t^* = 1.717$
 (b) $df = 22 \rightarrow t^* = 2.074$
 (c) $df = 23 + 27 - 2 = 48 \rightarrow t^* = 1.684$ (round down to 40 degrees of freedom in the table)
 (d) $df = 48 \rightarrow t^* = 2.021$

3 SOLUTIONS TO CUMULATIVE REVIEW PROBLEMS

$$1. \text{ a. } n = \frac{\left(\frac{2.576}{2(.025)}\right)^2 - 2654.3}{.} = 2655$$

$$\text{b. } \left(\frac{2.576}{.025}\right)^2 (.7)(1-.7) = 2229.6 \text{ Round up to } n = 2230$$

$$2. \mu_X = 250(.6) = 150, \sigma_X = \sqrt{250(.6)(.4)} = 7.75$$

$$P(x > 80) = P\left(z > \frac{160 - 150}{7.75} - 1.29\right) = .099$$

[The exact binomial given by the TI-83 is $1 - \text{binomcdf}(250, .6, 160) = .087$. If you are familiar with using a continuity correction for the normal approximation, $\text{normalcdf}(160.5, 1000, 150, 7.75) = .088$, which is much closer to the exact binomial.]

$$3. \binom{5}{3}(.3)^3(.7)^2 + \binom{5}{4}(.3)^4(.7)^1 + \binom{5}{5}(.3)^5(.7)^0$$

4. $P(\text{the worker is an administrator} \mid \text{the worker is female})$

$$= \frac{P(A \text{ and } F)}{P(F)} = \frac{\cancel{6}/\cancel{55}}{\cancel{30}/\cancel{55}} = \frac{6}{30} = .2.$$

5. A confidence interval is justified because we are dealing with a random sample from an approximately normally distributed population. $df = 25 - 1 = 24 \rightarrow t^* = 1.711$.

$$1.85 \pm 1.711 \left(\frac{.75}{\sqrt{25}} \right) = (1.59, 2.11).$$

We are 90% confidence that the true mean number of mg per cigarette for this type of cigarette is between 1.59 mg and 2.11 mg.

Chapter 11

Inference for Regression

2 SIMPLE LINEAR REGRESSION

When we studied data analysis earlier in this text, we distinguished between *statistics* and *parameters*. *Statistics* are measurements or values that describe samples, and *parameters* are measurements that describe populations. We have also seen that statistics can be used to estimate parameters. Thus, we have used \bar{x} to estimate the population mean μ , s to estimate the population standard deviation σ , etc. In Chapter 5, we introduced the *least-squares regression line* ($\hat{y} = a + bx$), which was based on a set of bivariate ordered pairs. \hat{y} is actually a *statistic* because it is based on sample data. In this chapter, we study the *parameter*, μ_y , that is estimated by \hat{y} .

Before we look at the model for linear regression, let's consider an example to remind us of what we did in Chapter 5:

example: The following data are pulse rates and heights for a group of 10 female statistics students:

Height	70	60	70	63	59	55	64	64	72	66
Pulse	78	70	65	62	63	68	76	58	73	53

- What is the least-squares regression line for predicting pulse rate from height.
- What is the correlation coefficient between height and pulse rate.
- What is the predicted pulse rate of a 67" tall student?
- Interpret the slope of the regression line in the context of the problem.

solution:

- $Pulse\ rate = 47.17 + .302(Height)$ [done on the TI-83: Height in L1, Pulse in L2, STAT CALC LinReg(a + bx) L1,L2,Y1]
- $r = .21$
- $Pulse\ rate = 47.17 + .302(67) = 67.4$ [On the Ti-83: Y1(67) = 67.42]
- For each increase in height of one inch, the pulse rate is predicted to increase by 0.302 beats per minute (or: the pulse rate will increase, on average, by 0.302 beats per minute)

To do inference for regression, we use $\hat{y} = a + bx$. Similar to what we have done with other statistics used for inference, we use a and b as estimators of population parameters α and β , the intercept and slope of the population regression line. The conditions necessary for doing inference for regression are:

- For each given value of x , the values of the response variable y are independent and normally distributed.
- For each given value of x , the standard deviation, σ , of y is the same.
- The mean response of the y values for the fixed values of x are linearly related by the equation $\mu_y = \alpha + \beta x$.

Remember that a *residual* was the error involved when making a prediction from a regression equation ($residual = actual\ value\ of\ y - predicted\ value\ of\ y = y_i - \hat{y}_i$). Not surprisingly, the standard error of the predictions is a function of the squared residuals.

$$s = \sqrt{\frac{SS_{RES}}{n - 2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

s is an estimator of σ , the standard deviation of the residuals. Thus, there are actually three parameters to worry about in regression: α , β , and σ , which are estimated by a , b , and s , respectively.

The final statistic we need to do inference for regression is the standard error of the slope of the regression line:

$$s_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

In summary, inference for regression depends upon estimating $\mu_y = \alpha + \beta x$ with $\hat{y} = a + bx$. For each x , the response values of y are independent and follow a normal distribution, each distribution having the same standard deviation. Inference for regression depends on the following statistics:

- a , the estimate of the y intercept, α , of μ_y
- b , the estimate of the slope, β , of μ_y

- s , the standard error of the residuals
- s_b , the standard error of the slope of the regression line

In the section that follows, we explore inference for the slope of a regression line in terms of a significance test and a confidence interval for the slope.

INFERENCE FOR THE SLOPE OF A REGRESSION LINE

In doing inference for the slope of a regression line, we do a significance test of the hypothesis: $H_0: \beta = \beta_0$ (usually just $H_0: \beta = 0$) and construct confidence intervals for the slope of a regression line. Our interest is the extent to which a least-squares regression line is a good model for the data. That is, the significance test is a test of a linear model for the data.

We note that in theory we could test whether the slope of the regression line is equal to any specific value. However, the usual test is whether the slope of the regression line is zero or not. If the slope of the line is zero, then there is no linear relationship between the x and y variables (remember:

$$b = r \frac{s_y}{s_x}; \text{ if } r = 0 \rightarrow b = 0).$$

The alternative hypothesis is often two sided (that the slope of the regression line is not zero). We can, and often will, do a one-sided test if we believed that the data were positively or negatively related.

Significance Test for the Slope of a Regression Line

The basic details of a significance test for the slope of a regression line are given in the following table:

• Hypothesis • Estimator • Standard Error	Conditions	Test Statistic
<ul style="list-style-type: none"> • Null hypothesis $H_0: \beta = \beta_0$ (<i>most often: $H_0: \beta = 0$</i>) • Estimator: b (<i>from: $\hat{y} = a + bx$</i>) 	<ul style="list-style-type: none"> • For each given value of x, the values of the response variable y are independent and normally distributed. 	$t = \frac{b - \beta_0}{s_b}$ $= \frac{b}{s_b} \text{ (if } \beta_0 = 0),$ $df = n - 2$

- Standard error of the residuals:
- For each given value of x , the standard deviation of y is the same.
- The mean response of the y values for the fixed values of x are linearly related by the equation $\mu_y = \alpha + \beta x$

$$s = \sqrt{\frac{SS_{RES}}{n - 2}}$$

$$= \sqrt{\frac{\sum (y_i - \bar{y}_i)^2}{n - 2}}$$

- Standard error of the slope:

$$s_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

example: The data in the following table give the top 15 states in terms of per pupil expenditure in 1985 and the average teacher salary in the state for that year.

State/Salary	Per Pupil Expenditure
MN 27360	3982
CO 25892	4042
OR 25788	4123
PA 25853	4168
WI 26525	4247
MD 27186	4349
DE 24624	4517
MA 26800	4642
RI 29470	4669
CT 26610	4888
DC 33990	5020
WY 27224	5440
NJ 27170	5536
NY 30678	5710
AK 41480	8349

Test the hypothesis, at the .01 level, that there is no straight-line relationship between per pupil expenditure and teacher salary. Assume that the conditions necessary for linear regression have been satisfied.

solution:

- I. Let β the true slope of the regression line.
- $H_0: \beta = 0$
 $H_A: \beta \neq 0$

II. We will use the t test for the slope of the regression line. The problem states that the conditions necessary for linear regression are satisfied.

III. The regression equation is
Salary = 12027 + 3.34 PPE ($s = 2281$, $s_b = .5536$)

$$t = \frac{3.34 - 0}{.5536} = 6.04, df = 15 - 2$$

$$= 13 \rightarrow P \text{ value} = .00004$$

[The LSRL, s , t , and P were obtained directly from the TI-83. The value of s_b can be found in computer output. Minitab would report the P value as .0000. We look at this more closely in the next section.]

IV. Because $P < \alpha$, we reject H_0 . We have evidence that the true slope of the regression line is not zero. We conclude that there seems to be a strong linear relationship between amount of per pupil expenditure and teacher salary.

A significance test that the slope of a regression line equals zero is closely related to a test that there is no correlation between the variables. That is, if ρ is the population correlation coefficient, then the test statistic for $H_0: \beta = 0$ is equal to the test statistic for $H_0: \rho = 0$. You aren't required to know it for the AP Exam, but the t test statistic for $H_0: \beta = 0$, where r is the sample correlation coefficient, is

$$t = r \sqrt{\frac{n-2}{1-r^2}}, df = n-2.$$

Because this and the test for a non-zero slope are equivalent, we have

$$r \sqrt{\frac{n-2}{1-r^2}} = \frac{b}{s_b}.$$

Confidence Interval for the Slope of a Regression Line

In addition to doing hypothesis tests on $H_0: \beta = \beta_0$, we can construct a confidence interval for the true slope of the regression line. The details follow:

• Parameter • Estimator	Conditions	Formula
<ul style="list-style-type: none"> • Population slope: β • Estimator: b (from: $\hat{y} = a + bx$) • Standard error of the residuals: $s = \sqrt{\frac{SS_{RES}}{n - 2}}$ $= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$ <ul style="list-style-type: none"> • Standard error of the slope: $s_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$	<ul style="list-style-type: none"> • For each given value of x, the values of the response variable y are independent and normally distributed. • For each given value of x, the standard deviation of y is the same. • The mean response of the y values for the fixed values of x are linearly related by the equation $\mu_y = \alpha + \beta x$ 	$b \pm t^* s_b, df = n - 2$ (where t^* is the upper critical value of t for a C -level confidence interval)

example: Consider once again the earlier example on predicting teacher salary from per pupil expenditure. Construct a 95% confidence interval for the slope of the population regression line.

solution: When we were doing a test of $H_0: \beta = 0$ for that problem, we noted that $\text{Salary} = 12027 + 3.34 \text{ PPE}$ ($s = 2281, s_b = .5536$). The slope of the regression line for the 15 points is $b = 3.34$. For $C = .95, df = 15 - 2 = 13 \rightarrow t^* = 2.160$ (remember, the confidence interval is two-sided). Thus, the 95% confidence interval for the true slope of the regression line is $3.34 \pm 2.160(.5536) = \langle 2.14, 4.54 \rangle$.

We are 95% confident that the true slope of the regression line is between 2.14 and 4.54. Note that because 0 is *not* in this interval, this finding is consistent with our earlier rejection of the hypothesis that the slope equals 0.

INFERENCE FOR REGRESSION USING TECHNOLOGY

If you had to do them from the raw data, the computations involved in doing inference for the slope of a regression line would be daunting. For example, t

$$s_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Fortunately, you probably will never have to do this by hand but can rely on computer output you are given or you will be able to use your calculator to do the computations.

Consider the following data that were gathered by counting the number of cricket chirps in 15 seconds and noting the temperature.

Number of Chirps	22	27	35	15	28	30	39	23	25	18	35	29
Temperature (F)	64	68	78	60	72	76	82	66	70	62	80	74

We want to use technology to test the hypothesis that the slope of the regression line is 0 and to construct a confidence interval for the true slope of the regression line.

First let us look at the Minitab regression output for this data.

The regression equation is				
Temp = 44.0 + 0.993 Number				
Predictor	Coef	St Dev	t ratio	P
Constant	44.013	1.827	24.09	.000
Number	0.99340	0.06523	15.23	.000
s = 1.538		R-sq = 95.9%		R-sq(adj) = 95.5%

You should be able to read most of this table, but you are not responsible for all of it. You see the following table entries:

- The regression equation, $Temp = 44.0 + 0.993 \text{ Number}$, is the LSRL for predicting temperature from the number of cricket chirps.
- Under “Predictor” are the y intercept and explanatory variable of the regression equation, called “Constant,” and “Number.”
- Under “Coef” are the values of the “Constant” (the intercept is 44.013) and the slope of the regression line ($b = .99340$).
- For the purposes of this course, we are not concerned with the “St Dev,” “t ratio,” or “P” for “Constant.”
- “St Dev” of “Number” is the standard error of the slope (what we have called s_b , which equals

$$\frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- “t ratio” is the value of the t test statistic

$$t = \frac{b}{s_b}, df = n - 2;$$

and “ P ” is the P value associated with the test statistic assuming a two-sided test (if you were doing a *one*-sided test, you would need to divide the given P value by 2).

- “ s ” is the standard error of the residuals

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

- “R-sq” is the coefficient of determination (r^2). You don’t need to worry about “R-sq(adj).”

All of the mechanics needed to do a t test for the slope of a regression line are contained in this printout. You need only to quote the appropriate values in your write up. Thus, for the problem given above, we see that $t = 15.23 \rightarrow P$ value = .000.



Exam Tip: You may be given a problem that has both the raw data and the computer print out. If so, there is no advantage to doing the computations all over again because it has already been done for you.

A C -level confidence interval for the slope is $b \pm t^* s_b$ based on $n - 2$ degrees of freedom. A 99% confidence interval for the slope in this situation ($df = 10 \rightarrow t^* = 3.055$) is: $44.0 \pm 3.055(.06523) = \langle 43.8, 44.2 \rangle$.

You can also do a t test for linear regression on the TI-83, but it is a bit trickier to generate a confidence interval because the calculator doesn’t return s_b (there is a work-around—more about that in a bit)

To use the calculator to do the regression, enter the data in, say, $L1$ and $L2$. Then go to **STAT TESTS LinRegTTest**. . . . Enter the data as requested (response variable in the $Ylist$:). Assuming the alternative is two sided ($H_A: \beta \neq 0$), choose β & $\rho \neq 0$. Then **Calculate**. You will get two screens of data:

```
LinRegTTest
y=a+bx
β≠0 and ρ≠0
t=15.22949379
F=3.0213567E-8
df=10
↓a=44.01259748
```

```
LinRegTTest
y=a+bx
β≠0 and ρ≠0
↑b=.9934013197
s=1.537610858
r²=.9586670079
r=.9791154211
```

This contains all of the information in the computer print out except s_b . It does give the number of degrees of freedom, which the computer does not, as well as greater accuracy. Note that the calculator lumps together the test for both the slope (β) and the correlation coefficient (ρ) because, as we noted earlier, the *test statistics* are the same for both.

If you *had* to do a confidence interval using the calculator, you first need to determine s_b . Because you know that

$$t = \frac{b}{s_b},$$

it follows algebraically that

$$s_b = \frac{b}{t}.$$

Thus, for this problem,

$$s_b = \frac{.9934}{15.2295} = .0652,$$

which agrees with the standard error of the slope (“St Dev” of “Number”) given in the computer printout.

RAPID REVIEW

1. The regression equation for predicting grade point average from number of hours studied is determined to be $\text{GPA} = 1.95 + .05(\text{Hours})$. Interpret the slope of the regression line.

Answer: For each additional hour studied, the GPA is predicted to increase by .05.

2. Which of the following is *not* a necessary condition for doing inference for the slope of a regression line?
 - a. For each given value of the independent variable, the response variable is normally distributed.
 - b. The values of the predictor and response variables are independent.
 - c. For each given value of the independent variable, the distribution of the response variable has the same standard deviation.
 - d. The mean response values lie on a line.

Answer: (b) is not a condition for doing inference for the slope of a regression line. In fact, we are trying to find out the degree to which they are not independent.

3. True–False: Significance tests for the slope of a regression line are always based on the hypothesis $H_0: \beta = 0$ versus the alternative $H_A: \beta \neq 0$.

Answer: False. While the stated null and alternative may be the usual hypotheses in a test about the slope of the regression line, it is possible to test that the slope has some particular non-zero value and the alternative can be one-sided.

4. Consider the following Minitab print out:

The regression equation is				
$y = 282 + 0.634 x$				
Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	282.459	3.928	71.91	.000
x	0.63383	0.07039	9.00	.000
$s = 9.282$	$R\text{-sq} = 81.0\%$	$R\text{-sq(adj)} = 80.0\%$		

- What is the slope of the regression line?
- What is the standard error of the residuals?
- What is the standard error of the slope?
- Do the data strongly indicate a linear relationship between x and y ?

Answer:

- .634
 - 9.282
 - .07039
 - Yes, the t test statistic = 9.00 \rightarrow P value = .000. That is, the probability is close to zero of getting a slope of .634 if, in fact, the true slope was zero.
5. A t test for the slope of the regression line is to be conducted at the .02 level of significance based on 18 datapoints. What is the upper critical t value for this test (i.e., what is t^*)?

Answer: Note that there are $18 - 2 = 16$ degrees of freedom. Accordingly, $t^* = 2.583$. If you mistakenly got $t^* = 2.567$, you used $n - 1$ degrees of freedom.

6. In the printout from Problem 4, we were given the regression equation $y = 282 + 0.634 x$. The t test for $H_0: \beta = 0$ yielded a P value of .0000. What is the conclusion you would arrive at based on these data?

Answer: Because P is very small, we have evidence to reject the null hypothesis that the slope of the regression line is 0. We have strong evidence of a linear relationship between x and y .

7. Suppose the computer output for regression reports $P = .036$. What is the P value for $H_A: \beta > 0$ (assuming the test was in the correct direction for the data)?

Answer: .018. Computer output for regression assumes the alternative is two-sided ($H_A: \beta \neq 0$). Hence the P value reported assumes the finding could have been in either tail of the t distribution. The correct P value for the one-sided test is one-half of this value.

3 PRACTICE PROBLEMS

Multiple Choice

1. Which of the following statements are true?
 - I. In the computer output for regression, s is the estimator of σ , the standard deviation of the residuals.
 - II. The t test statistic for the $H_0: \beta_1 = 0$ has the same value as the t test statistic for $H_0: \rho = 0$.
 - III. The t test for the slope of a regression line is always two-sided ($H_A: \beta_1 \neq 0$)
 - a. I only
 - b. II only
 - c. III only
 - d. I and II only
 - e. I and III only

Use the following output in answering Questions 2 and 3:

A study attempted to establish a linear relationship between IQ score and musical aptitude. The following table is a partial printout of the regression analysis and is based on a sample of 20 individuals.

The regression equation is MusApt = -22.3 + 0.493 IQ				
Predictor	Coef	St Dev	t ratio	P
Constant	-22.26	12.94	-1.72	.102
IQ	0.4925	0.1215		
$s = 6.143$	R-sq = 47.7%		R-sq(adj) = 44.8%	

2. The value of the t test statistic for $H_0: \beta_1 = 0$ is
 - a. 4.05
 - b. -1.72
 - c. .4925
 - d. 6.143
 - e. .0802
3. A 99% confidence interval for the slope of the regression line is
 - a. $.4925 \pm 2.878(6.143)$
 - b. $.4925 \pm 2.861(.1215)$
 - c. $.4925 \pm 2.861(6.143)$
 - d. $.4925 \pm 2.845(.1215)$
 - e. $.4925 \pm 2.878(.1215)$

4. Which of the following conditions is *not* needed to do inference for the slope of a regression line?
- For a given value of x , each response y is independent.
 - The populations from which the x and y variables are sampled must be approximately normally distributed.
 - For each value of x , the response variable y varies normally.
 - For each value of x , the response variable y varies with the same standard deviation.
 - The mean response of y has a linear relationship with x .
5. The two screens shown below were taken from a TI-83 *LinReg TTest*. . . . What is the standard error of the slope of the regression line (s_b)?

```
LinRegTTest
y=a+bx
b≠0 and ρ≠0
t=4.177610769
p=.0058289601
df=6
↓a=-10923.26676
■
```

```
LinRegTTest
y=a+bx
b≠0 and ρ≠0
↑b=9225.164765
s=17033.52924
r²=.7441629906
r=.8626488223
■
```

- 17033.53
- 6953.91
- 2206.98
- 9225.16 ± 17033.53
- 3115.84

Free Response

- 1.–5. The following table gives the ages in months of a sample of children and their mean height (in inches) at that age.

Age	18	19	20	21	22	23	24	25	26	27	28
Height	30.0	30.7	30.7	30.8	31.0	31.4	31.5	31.9	32.0	32.6	32.9

- Find the correlation coefficient and the least-squares regression line for predicting height (in inches) from age (in months).
- Draw a scatterplot of the data and the LSRL on the plot. Does the line appear to be a good model for the data?

3. Construct a residual plot for the data. Does the line still appear to be a good model for the data?
4. Use your LSRL to predict the height of a child of 35 months. How confident should you be in this prediction?
5. Interpret the slope of the regression line found in Problem #1 in terms of the context of the problem.
6. In 2002, there were 23 states in which more than 50% of high school graduates took the SAT test. The following printout gives the regression analysis for predicting SAT Math from SAT Verbal from these 23 states.

The regression equation is				
<input style="width: 100%; height: 15px;" type="text"/>				
Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	185.77	71.45	2.60	.017
Verbal	0.6419	0.1420	4.52	.000
s = 7.457		R-sq = 49.3%		R-sq(adj) = 46.9%

- a. What is the equation of the least-squares regression line for predicting Math SAT score from Verbal SAT score?
 - b. What is the standard error of the slope of the regression line?
 - c. Assuming that the conditions are met for doing inference for regression, what is the hypothesis being tested in this analysis, what test statistic is used in the analysis, and what conclusion would you make concerning the hypothesis?
7. For the regression analysis of Problem 6
 - a. construct and interpret a 95% confidence interval for the true slope of the regression line.
 - b. explain what is meant by “95% confidence interval” in the context of the problem.
 8. It has been argued that the average score on the SAT tests drop as more students take the test (nationally, about 46% of graduating students took the SAT). The following data are the Minitab output for predicting SAT Math score from the percentage taking the test (PCT) for each of the 50 states. Assuming that the conditions for doing inference for regression are met, test the hypothesis that scores decline as proportion of students taking the test rises. That is, test to determine if the slope of the regression line is negative. Test at the .01 level of significance.

The regression equation is
 SAT Math = 574 – 99.5 PCT

Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	574.179	4.123	139.25	.000
PCT	-99.516	8.832	-11.27	.000

$s = 17.45$ R-sq = 72.6% R-sq(adj) = 72.0%

9. Some bored researchers got the idea that they could predict a person's pulse rate from his or her height (earlier studies had shown a very weak linear relationship between pulse rate and weight). They collected data on 20 college-age women. The following table is part of the Minitab output of their findings.

The regression equation is

Pulse = Height

Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	52.00	37.24	1.40	.180
Height	0.2647	0.5687	<input type="text"/>	<input type="text"/>

$s = 10.25$ R-sq = 1.2% R-sq(adj) = 0.0%

- Determine the *t* ratio and the *P* value for the test.
 - Construct a 99% confidence interval for the slope of the regression line.
 - Do you think there is a predictive linear relationship between height and pulse rate? Explain.
 - Suppose the researcher was hoping to show that there was a positive linear relationship between pulse rate and height. Are the *t* ratio and *P* value the same as in Part (a)? If not, what are they?
10. The following table gives the number of manatees killed by powerboats along the Florida coast in the years 1977 to 1990, along with the number of powerboat registrations (in thousands) during those years:

Year	Powerboat Registrations	Manatees Killed
1977	447	13
1978	460	21
1979	481	24
1980	498	16
1981	513	24

1982	512	20
1983	526	15
1984	559	34
1985	585	33
1986	614	33
1987	645	39
1988	675	43
1989	711	50
1990	719	47

Use your calculator, or a computer statistics package, to

- test the hypothesis that there is a positive linear relationship between the number of powerboat registrations and the number of manatees killed by powerboats. Assume that the conditions needed to do inference for regression have been met.
- Use a residual plot to assess the appropriateness of the model.
- construct a 90% confidence interval for the true slope of the regression line.

3 CUMULATIVE REVIEW PROBLEMS

- You are testing the hypothesis $H_0: p = .6$. You sample 75 people as part of your study and calculate that $\hat{p} = .7$.
 - What is $s_{\hat{p}}$ for a significance test for p ?
 - What is $s_{\hat{p}}$ for a confidence interval for p ?
- A manufacturer of light bulbs claims a mean life of 1500 hours. A mean of 1450 hours would represent a significant departure from this claim. Suppose, in fact, the mean life of bulbs is only 1450 hours. In this context, what is meant by the power of the test?
- Complete the following table by filling in the shape of the sampling distribution of \bar{x} for each situation.

Situation	Shape of Sampling Distribution
<ul style="list-style-type: none"> Shape of parent population: normal Sample size: $n = 8$ 	
<ul style="list-style-type: none"> Shape of parent population: normal Sample size: $n = 35$ 	
<ul style="list-style-type: none"> Shape of parent population: strongly skewed to the left. Sample size: $n = 8$ 	

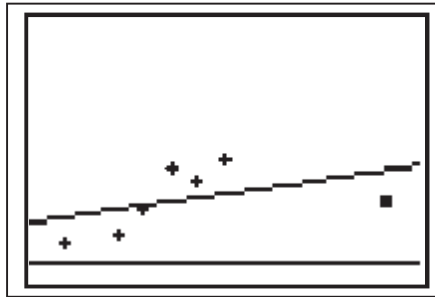
- | |
|---|
| <ul style="list-style-type: none"> • Shape of parent population: strongly skewed to the left • Sample size: $n = 35$ |
| <ul style="list-style-type: none"> • Shape of parent population: unknown • Sample size: $n = 8$ |
| <ul style="list-style-type: none"> • Shape of parent population: unknown • Sample size: $n = 50$ |

4. The following is (most of) a probability distribution for a discrete random variable.

X	2	6	7	9
$p(x)$.15	.25		.40

Find mean and standard deviation of this distribution.

5. Consider the following scatterplot and regression line.



- Would you describe the point marked with a box as an outlier, influential point, neither, or both?
- What would be the effect on the correlation coefficient of removing the box-point?
- What would be the effect on the slope of the regression line of removing the box-point?

3 SOLUTIONS TO PRACTICE PROBLEMS

Multiple Choice

1. The correct answer is (d). II is true since it can be shown that

$$t = \frac{b}{s_b} = r \sqrt{\frac{n-2}{1-r^2}}. \text{ III is not true since, although we often use the}$$

alternative $H_A: \beta_1 \neq 0$, we can certainly test a null with an alternative that states that there is a positive or a negative association between the variables.

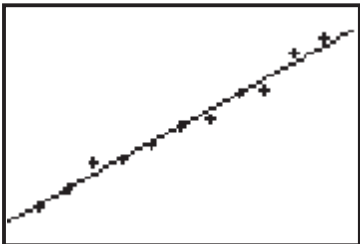
- The correct answer is (a). $t = \frac{b}{s_b} = \frac{.4925}{.1215} = 4.05$
- The correct answer is (e). For $n = 20$, $df = 20 - 2 = 18 \rightarrow t^* = 2.878$ for $C = .99$ level confidence interval.
- The correct answer is (b). All of the rest are necessary conditions.
- The correct answer is (c). The TI-83 does not give the standard error of the slope directly. However,

$$t = \frac{b}{s_b} \rightarrow s_b = \frac{b}{t} = \frac{9225.16}{4.18} = 2206.98.$$

Free Response

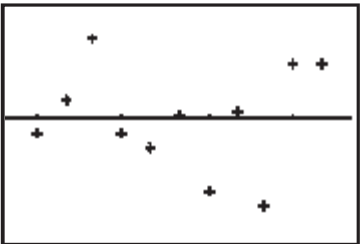
- $r = .9817$, $\text{Height} = 25.41 + .261(\text{Age})$

[Assuming that you have put the age data in L1 and the Height data in L2, remember that this can be done on the TI-83 as follows: STAT CALC LinReg L1,L2,Y1]

- 

The line does appear to be a good model for the data.

[After the regression equation was calculated and the LSRL stored in Y1, this was constructed using STAT PLOT by drawing a scatterplot with the Xlist: L1 and Ylist:L2]

- 

The residual pattern seems quite random. A line still appears to be a good model for the data.

[This was drawn by STAT PLOT scatterplot with Xlist:L1 and Ylist:RESID—remember that the list of residuals is saved in a list names RESID each time you do a regression]

- $\text{Height} = 25.41 + .26(35) = 34.51$ [$Y1(35) = 34.65$]. You probably shouldn't be too confident in this prediction. Thirty-five is well outside of the data on which the LSRL was constructed and, even

though the line appears to be a good fit for the data, there is reason to believe that a linear pattern is going to continue indefinitely (if it did, a 25-year-old would have a height of $25.07 + .27(12 \cdot 25) = 106.1''$, or 8.8 feet).

5. The slope of the regression line is .261. This means that for each increase in age of 1 year, the height is predicted to increase by .27 inches. You could also say that, for each increase in age of one year, that the height will increase on average by .27 inches.
6. a. $Math = 185.77 + .6419(Verbal)$
 b. $s_b = .1420$
 c. • $H_0: \beta = 0$ and $H_0: \rho = 0$ (the true slope of the regression line is 0, or the correlation between the variable is 0)
 $H_A: \beta \neq 0$
 • The test statistic is

$$t = \frac{b}{s_b},$$

with $df = 23 - 2 = 21$.

- $t = 4.52 \rightarrow P = .000$. The probability is extremely small that the true slope of the regression line is equal to 0. We have strong evidence that there is a linear relationship between scores on SAT verbal and SAT Math.
7. a. $df = 23 - 2 = 21 \rightarrow t^* = 2.080$. $.6419 \neq 2.080(.1420) = \langle .35, .94 \rangle$. We are 95% confident that the true slope of the regression line lies in the interval $\langle .35, .94 \rangle$.
 b. The procedure used to generate the confidence interval would produce intervals that contain the true slope of the regression line, on average, .95 of the time.

8.

<p>I. Let β – the true slope of the regression line for predicting SAT Math score from the percentage of graduating seniors taking the test.</p> <p>$H_0: \beta = 0$ $H_A: \beta < 0$</p> <p>II. We use a linear regression t test with $\alpha = .01$. The problem states that the conditions for doing inference for regression are met.</p>

III. We see from the printout that

$$t = \frac{b}{s_b} = \frac{-99.516}{8.832} = -11.27$$

based on $50 - 2 = 48$ degrees of freedom. The P value is 0.000. (Note: the P value in the printout is for a two-sided test. However, because the P value for a one-sided test would only be half as large, it is still 0.000)

IV. Because $P < .01$, we reject the null hypothesis. We have very strong evidence that there is a negative linear relationship between the proportion of students taking SAT math and the average score on the test.

9. a. $t = \frac{b}{s_b} = \frac{.2647}{.5687} = .47, df = 20 - 2 = 18 \rightarrow P \text{ value} = .644$

b. $df = 18 \rightarrow t^* = 2.878; .2647 \pm 2.878(.5687) = \langle -1.37, 1.90 \rangle.$

c. No. The P value is very large giving no grounds to reject the null hypothesis that the slope of the regression line is 0. Furthermore, the correlation coefficient is only $r = \sqrt{1.2} = 1.1$, which is very close to 0. Finally, the confidence interval constructed in Part (b) contains the value 0 as a likely value of the slope of the population regression line.

d. The t ratio would still be .47. The P value, however, would be half of the .644, or .322 because the computer output assumes a two-sided test. This is a lower P value but is still much too large to infer any significant linear relationship between pulse rate and height.

10. a.

I. Let β the true slope of the regression line for predicting the number of manatees killed by powerboats from the number of powerboat registrations.

$$H_0: \beta = 0$$

$$H_A: \beta > 0$$

II. We use a t test for the slope of the regression line. The problem tells us that the conditions necessary to do inference for regression have been met.

III. We do this problem on the TI-83 as well as Minitab.

- On the TI-83: Enter the number of powerboat registrations in $L1$ and the number of manatees killed in $L2$.

Then go to *STAT TESTS LinRegTTest...* and set it up as shown below.

```
LinRegTTest
Xlist:L1
Ylist:L2
Freq:1
B & P:≠0 <0
RegEQ:Y1
Calculate
```

After “Calculate,” we have the following two screens.

```
LinRegTTest
y=a+bx
B>0 and P>0
t=9.675470539
p=2.5545306E-7
df=12
↓a=-41.43043895
```

```
LinRegTTest
y=a+bx
B>0 and P>0
↑b=.1248616923
s=4.276387771
r²=.8863794853
r=.9414772888
```

The *Minitab* output for this problem is:

The regression equation is
 $\text{Man} = -41.4 + 0.125 \text{ PB Reg}$

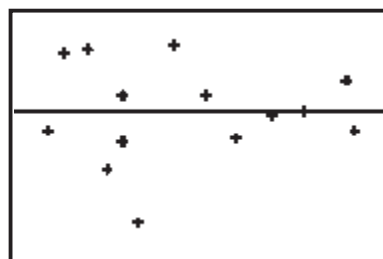
Predictor	Coef	St Dev	t ratio	P
Constant	-41.430	7.412	-5.59	.000
PB Reg	0.12486	0.01290	9.68	.000

s = 4.276 R-sq = 88.6% R-sq(adj) = 87.7%

IV. Because the *P* value is very small, we reject the null. We have very strong evidence of a positive linear relationship between the number of powerboat registrations and the number of manatees killed by powerboats.

b. From the TI-83, using the residuals generated when we did the linear regression above, we have:

```
Plot1 Plot2 Plot3
Off
Type: [ ] [ ] [ ]
Xlist:L1
Ylist:RESID
Mark: [ ] [ ] [ ]
```



There appears to be no pattern in the residual plot that would cause us to doubt the appropriateness of the model. A line does seem to be a good model for the data.

- c. (i) Using the *TI-83* results. $df = 12 \rightarrow t^* = 1.782$. We need to determine s_b . We have

$$t \frac{b}{s_b} \rightarrow s_b \frac{b}{t} \frac{.125}{9.68} = .013.$$

The confidence interval is: $.125 \pm 1.782(.013) = \langle .10, .15 \rangle$.

- (ii) Directly from the *Minitab* output: $.125 \pm 1.782(.013) = \langle .10, .15 \rangle$.

3 SOLUTIONS TO CUMULATIVE REVIEW PROBLEMS

1. a. $\sqrt{\frac{(.6)(.4)}{75}} = .057$

b. $\sqrt{\frac{(.7)(.3)}{75}} = .053$

2. The *power of the test* is the probability of correctly rejecting a false hypothesis against a particular alternative. In other words, the *power* of this test is the probability of rejecting the claim that the true mean is 1500 hours against the alternative that the true mean is only 1450 hours.

3.

Situation	Shape of Sampling Distribution
<ul style="list-style-type: none"> • Shape of parent population: normal • Sample size: $n = 8$ 	Normal
<ul style="list-style-type: none"> • Shape of parent population: normal • Sample size: $n = 35$ 	Normal
<ul style="list-style-type: none"> • Shape of parent population: strongly skewed to the left. • Sample size: $n = 8$ 	Skewed somewhat to the left
<ul style="list-style-type: none"> • Shape of parent population: strongly skewed to the left • Sample size: $n = 35$ 	Approximately normal (central limit theorem)
<ul style="list-style-type: none"> • Shape of parent population: unknown • Sample size: $n = 8$ 	Similar to parent population
<ul style="list-style-type: none"> • Shape of parent population: unknown • Sample size: $n = 50$ 	Approximately normal (central limit theorem)

$$4. P(7) = 1 - (.15 + .25 + .40) = .20.$$

$$\mu_x = 2(.15) + 6(.25) + 7(.20) + 9(.40) = 6.8$$

$$\sigma_x = \sqrt{(2 - 6.8)^2(.15) + (6 - 6.8)^2(.25) + (7 - 6.8)^2(.20) + (9 - 6.8)^2(.40)} = 2.36$$

[Remember that this can be done by putting the X values in L1, the p(x) values in L2, and doing STAT CALC 1-Var Stats L1,L2.]

5. a. The point is both an outlier and an influential point. It's an outlier because it is removed from the general pattern of the data. It is an influential observation because it is an outlier in the x direction and its removal would have an impact on the slope of the regression line.
- b. Removing the point would increase the correlation coefficient. That is, the remaining data are better modeled by a line without the box-point than with it.
- c. Removing the point would make the slope of the regression line more positive than it is already.

Chapter 12

Inference for Categorical Data

Chi-square

2 CHI-SQUARE GOODNESS-OF-FIT TEST

The following are the approximate percentages for the different blood types among white Americans: A: 40%; B: 11%; AB: 4%; O: 45%. A sample of 1000 black Americans yielded the following blood type data: A: 270; B: 200; AB: 40; O: 490. Does this sample indicate that the distribution of blood types among black Americans differs from that of white Americans or could the sample values simply be due to sampling variation? This is the kind of question we can answer with the **chi-square (χ^2) goodness-of-fit test**.

To answer this question, we need to compare the **observed values** in the sample with the **expected values** we would get *if* the sample of black Americans really had the same distribution of blood types as white Americans. These values we need for this are summarized in the following table.

Blood Type	Proportion of Population	Expected Values	Observed Values
A	.40	$(.4)(1000) = 400$	270
B	.11	$(.11)(1000) = 110$	200
AB	.04	$(.04)(1000) = 40$	40
O	.45	$(.45)(1000) = 450$	490

It appears that the numbers vary noticeably for types A and B, but not as much for types AB and O. The table can be condensed as follows.

Blood Type	Observed Values	Expected Values
A	270	400
B	200	110
AB	40	40
O	490	450

The chi-square statistic (χ^2) calculates the squared difference between the observed and expected values relative to the expected value for each category. The χ^2 statistic is computed as follows:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \sum \frac{(O - E)^2}{E}$$

The chi-square distribution is based on the number of degrees of freedom that equals, for the goodness-of-fit test, the number of categories minus 1 ($df = n - 1$). The χ^2 statistic follows approximately a unique chi-square distribution, assuming a random sample and a large enough sample, for each different number of degrees of freedom. The probability that a sample has a χ^2 value as large as it does can be read from a table of χ^2 critical values, or determined from a calculator. There is a χ^2 table in the back of this book, and you will be supplied a table on the AP exam. We will demonstrate both the use of tables and the calculator in the examples and problems that follow.

A hypothesis test for χ^2 goodness-of-fit follows the, by now, familiar pattern. The essential parts of the test are summarized in the following table.

Hypotheses	Conditions	Test Statistic
<ul style="list-style-type: none"> Null hypothesis: H_0: $p_1 =$ population proportion for category 1 $p_2 =$ population proportion for category 2 . . . $p_n =$ population proportion for category n Alternative hypothesis: H_A: at least one of the proportions in H_0 is not equal to the claimed population proportion 	<ul style="list-style-type: none"> Observations are based on a random sample The number of each expected count is at least 5. <i>(Some texts use the following condition for expected counts: at least 80% of the counts are greater than 5 and none is less than 1.)</i> 	$\chi^2 = \sum \frac{(O - E)^2}{E}$ $df = n - 1$

The hypothesis testing procedure follows the usual pattern. Let's do the above illustration in detail.

example: The following are the approximate percentages for the different blood types among white Americans: A: 40%; B: 11%; AB: 4%; O: 45%. A random sample of 1000 black Americans yielded the following blood type data: A: 270; B: 200; AB: 40; O: 490. Does this sample indicate that the distribution of blood types among black Americans differs from that of white Americans?

solution:

- I. Let p_A = proportion with type A blood; p_B = proportion with type B blood; p_{AB} = proportion with type AB blood; p_O = proportion with type O blood
 $H_0: p_A = .40, p_B = .11, p_{AB} = .04, p_O = .45$
 H_A : the actual population proportions are different from those in H_0

- II. We will use the χ^2 goodness-of-fit test. The problem states that the sample is a random sample. The expected values are: type A: 400; type B: 110; type AB: 40; type O: 450. Each of these is greater than 5. The conditions for the test are satisfied.

- III. The data are summarized in the table:

Blood Type	Observed Values	Expected Values
A	270	400
B	200	110
AB	40	40
O	490	450

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(270 - 400)^2}{400} + \frac{(200 - 110)^2}{110} + \frac{(40 - 40)^2}{40} + \frac{(490 - 450)^2}{450} = 119.44,$$

$$df = 4 - 1 = 3.$$

From the χ^2 table, we see that 119.44 is larger than any value for 3 df . Hence, P value $< .001$.

[The TI-83 gives a P value of 1.02×10^{-25} !!!—more about how to do this coming up.]

- IV. Because $P < .001$, we reject the null and conclude that not all the proportions are as hypothesized in the null. We have very strong evidence that the proportions of the various blood types among black Americans differs from the proportions among white Americans.

III.

Face Value	Observed	Expected
1	19	16
2	15	16
3	10	16
4	14	16
5	17	16
6	21	16

$\chi^2 = 4.75$ (calculator result), $df = 6 - 1 = 5 \rightarrow P$ value $> .25$ (table) or P value = .45 (calculator).

[Remember: to get χ^2 on the calculator, put the Observed in L1, the Expected in L2, let $L3 = (L1 - L2)^2/L2$, then LIST MATH $\text{sum}(L3)$ will be χ^2 . The corresponding probability is then found by $\text{DISTR } \chi^2 \text{cdf}(4.75, 100, 5)$.]

IV. Because the P value is $> .25$, we have no evidence to reject the null. We cannot conclude that the calculator is failing to simulate a fair die.

INFERENCE FOR TWO-WAY TABLES

Two-Way Tables (Contingency Tables) Defined

A **two-way table**, or **contingency table**, for bivariate categorical data is simply a rectangular array of cells. Each cell contains the frequencies for the joint values of the row and column variables. If the row variable has r values, then there will be r rows of data in the table. If the column variable has c values, then there will be c columns of data in the table. Thus, there are $r \times c$ cells in the table. The **marginal totals** are the sums of the observations for each row and each column.

example: A class of 36 students is polled concerning their political party preference. The results are presented in the following *two-way* table.

Political Party Preference

		Democrat	Republican	Independent	
Gender	Male	11	7	2	20
	Female	7	8	1	16
		18	15	3	36

The values of the row variable (Gender) are “Male” and “Female.” The values of the column variable (Political Party Preference) are “Democrat,” “Republican,” and “Independent.” There are $r=2$ rows and $c=3$ columns. We refer to this as a 2×3 table (the number of rows always comes first). The **dimension** of a two-way table is $r \times c$. The row marginal totals are 20 and 16; the column marginal totals are 18, 15, and 3. Note that the sum of the row and column marginal totals must both add to the total number in the sample.

In the example above, we had one population of 36 students and two categorical variables (gender and party preference). In this type of situation, we are interested in whether or not the variables are independent in the population. That is, does knowledge of one variable provide you with information about the other variable. Another study might have drawn a simple random sample of 20 males from, say, the senior class and another simple random sample of 16 females. Now we have two populations rather than one, but only one categorical variable. Now we might ask if the proportions of Democrats, Republicans, and Independents in each population are the same. Either way we do it, we end up with the same contingency table given in the example. We will look at how these differences play out in the next couple of sections.

Chi-square Test for Independence

A random sample of 400 residents of large western city are polled to determine their attitudes concerning the affirmative action admissions policy of the local university. The residents are classified according to ethnicity (white, black, Asian) and whether or not they favor the affirmative action policy. The results are presented in the following table.

Attitude Toward Affirmative Action

		Favor	Do Not Favor	
Ethnicity	White	130	120	250
	Black	75	35	110
	Asian	28	12	40
		233	167	400

We are interested in whether or not, in this population of 400 citizens, ethnicity and attitude toward affirmative action are related. That is, does knowledge of a person's ethnicity give us information about that person's attitude toward affirmative action? Another way of asking this is, "are the variables independent in the population." As part of a hypothesis test, the null hypothesis is that the two variables are independent, and the alternative is that they are not: H_0 : the variables are independent in the population vs. H_A : the variables are not independent in the population. Alternatively, we could say H_0 : the variables are not related in the population vs. H_A : the variables are related in the population.

The test statistic for the independence hypothesis is the same chi-square statistic we saw for the goodness-of-fit test

$$\left(\chi^2 = \sum \frac{(O - E)^2}{E} \right)$$

For a two-way table, the number of *degrees of freedom* is calculated as $(\text{number of rows} - 1)(\text{number of columns} - 1) = (r - 1)(c - 1)$. As with the goodness-of-fit test, we require that we are dealing with a random

sample and that the number of expected values in each cell be at least 5 (or: there are no empty cells and at least 80% of the cells have more than 5 expected values).

Calculation of the expected values for chi-square can be labor intensive, but is *usually* done by technology. However, you should know how expected values are arrived at.

example (calculation of expected value): suppose we are testing for independence of the variables in the illustration above. For the two-way table with the given marginal values, find the expected value for the cell marked “****.”

		Favor	Do Not Favor	
Ethnicity	White			250
	Black		****	110
	Asian			40
		233	167	400

solution: The probability that a sample respondent is black is $110/400$. The probability that a respondent does not favor the policy is $167/400$. Because we are assuming the variables are independent, the probability that a black respondent does not favor the proposal [i.e., $P(\text{****})$] is $(110/400)(167/400)$. The expected value of “****” is the probability of being in that cell times the total number of respondents. Hence, the expected value in the cell “****” is $(110/400)(167/400)(400) = 45.925$.



Calculator Tip: The easiest way to obtain the expected values is to use your calculator. To do this, let’s use the data from the previous example:

130	120
75	35
28	12

Go to *MATRIX EDIT* [A], enter these data in a 3×2 matrix. Then go to *STAT TESTS* χ^2 Test. . . . Enter [A] for Observed (you can do this by *MATRIX NAMES* [A]) and *MATRIX NAMES* [B] for *Expected*.

Then choose *Calculate*. The calculator will return the results of the χ^2 test. In the process, the calculator created and stored the set of expected values in Matrix [B]. To see this matrix, enter *MATRIX NAMES* [B]. You should get the following matrix of expected values:

145.625	104.475
64.075	45.925
23.3	16.7

Note that the entry in the 2nd row and 2nd column (45.925) agrees with what we calculated in the example above.

The hypothesis testing procedure for testing independence in two-way tables can be summarized as follows.

Hypotheses	Conditions	Test Statistic
<ul style="list-style-type: none"> Null hypothesis: H_0: The row and column variables are independent (or: they are not related) Alternative hypothesis: H_A: The row and column variables are not independent (or: they are related) 	<ul style="list-style-type: none"> Observations are based on a random sample The number of each expected count is at least 5. (Some texts use the following condition: all expected counts are greater than 1, and at least 80% of the expected counts are greater than 5.) 	$\chi^2 = \sum \frac{(O - E)^2}{E}$ $df = (r - 1)(c - 1)$

example: A study of 150 cities wanted to determine if crime rate is related to outdoor temperature. The results of the study are summarized in the following table:

Crime Rate

	Below	Normal	Above
Temperature Below	12	8	5
Normal	35	41	24
Above	4	7	14

Do these data provide evidence, at the .02 level of significance, that the crime rate is related to the temperature at the time of the crime?

solution:

I.	H_0 : The crime rate is independent of temperature H_A : The crime rate is not independent of temperature												
II.	We will use a chi-square test for independence. A matrix of expected values is found to be: <table style="display: inline-table; vertical-align: middle;"> <tr> <td></td> <td>8.5</td> <td>9.33</td> <td>7.17</td> </tr> <tr> <td>34</td> <td>37.33</td> <td>28.67</td> <td></td> </tr> <tr> <td>8.5</td> <td>9.33</td> <td>7.17</td> <td></td> </tr> </table> Because all expected values are greater than 5, we can use the chi-square test.		8.5	9.33	7.17	34	37.33	28.67		8.5	9.33	7.17	
	8.5	9.33	7.17										
34	37.33	28.67											
8.5	9.33	7.17											
III.	$\chi^2 = \sum \frac{(O - E)^2}{E} = 12.92, df = (3 - 1)(3 - 1) = 4 \rightarrow$ $.01 < P \text{ value} < .02 \text{ (table). [} P \text{ value} = .012 \text{ (calculator).]}$												

IV. Because $P < .02$, we reject H_0 . We have strong evidence that the number of crimes committed is related to the temperature at the time of the crime.

Chi-square Test for Homogeneity of Proportions

In the previous section, we tested for the independence of two categorical variables measured on a single population. In this section we again use the chi-square statistic but will investigate whether or not the values of a single categorical variable are proportional among two or more populations.

In the previous section, we considered a situation in which a sample of 36 students was selected and were then categorized according to gender and political party preference. We then asked if gender and party preference are independent in the population. Now suppose instead that we had selected a random sample of 20 males from the population of males in the school and another, independent, random sample of 16 females from the population of females in the school. Within each sample we classify the students as Democrat, Republican, or Independent. The results are presented in the following table, which you should notice is *exactly* the same table we presented earlier when gender was a category.

Political Party Preference

		Democrat	Republican	Independent	
Gender	Male	11	7	2	20
	Female	7	8	1	16
		18	15	3	36

Because “Male” and “Female” are now considered separate populations, we do not ask if gender and political party preference are independent in the population of students. We ask instead if the proportions of Democrats, Republicans, and Independents are the same within the populations of Males and Females. This is the test for **homogeneity of proportions** (or *homogeneity of populations*). Let the proportion of Male Democrats be p_1 ; the proportion of Female Democrats be p_2 ; the proportion of Male Republicans be p_3 ; the proportion of Female Republicans be p_4 ; the proportion of Independent Males be p_5 ; and the proportion of Independent Females be p_6 . Our null and alternative hypotheses are then

$$H_0: p_1 = p_2, p_3 = p_4, p_5 = p_6$$

H_A : Not all of the proportions stated in the null hypothesis are true.

It works just as well, and might be a bit easier, to state the hypotheses as follows.

H_0 : The proportions of Democrats, Republicans, and Independents are the same among Male and Female students

H_A : Not all of the proportions stated in the null hypothesis are equal.

For a given two-way table, expected values are the same under a hypothesis of homogeneity as independence but are calculated in a different manner. Let’s look again at the example we considered earlier.

example: Under the assumption that the samples of white, black, and Asian populations are independent samples from larger populations, calculate the expected value of the cell marked with “****.”

		Favor	Do Not Favor	
Ethnicity	White			250
	Black		****	110
	Asian			40
		233	167	400

solution: Under the assumption of homogeneity, we expect the proportions of white, black, and Asian who do not favor the proposal to be the same. Because 167/400 of all respondents “Do not favor” the proposal, we would expect $167/400(110) = 45.925$ blacks to not favor the proposal. This is the same value we obtained for that cell earlier under the assumption of independence. What differs is the method of calculating the value based on the type of test being done.

example: A university dean suspects that there is a difference between tenured and nontenured professors concerning a proposed salary increase. She randomly selects 20 nontenured instructors and 25 tenured staff to see if there is a difference. She gets the following results.

	Favor Plan	Do Not Favor Plan
Tenured	15	10
Nontenured	8	12

Do these data provide good statistical evidence that tenured and nontenured faculty differ in their attitudes toward the proposed salary increase?

solution:

- I. Let p_1 = the proportion of tenured faculty that favor the plan and let p_2 = the proportion of nontenured faculty that favor the plan.
 $H_0: p_1 = p_2$
 $H_A: p_1 \neq p_2$

- II. We will use a chi-square test for homogeneity of proportions. The samples of tenured and nontenured instructors are given as random. We determine that the expected values are given by the matrix $\begin{bmatrix} 12.78 & 12.22 \\ 10.22 & 9.78 \end{bmatrix}$. Because all expected values are greater than 5, the conditions for the test are met.

- III. $\chi^2 = 1.78$, $df = 1 \rightarrow .15 < P \text{ value} < .20$ (from the χ^2 table).
[from the calculator, $P = .18$.]
- IV. The P value is not small enough to reject the null hypothesis. These data does not provide strong statistical evidence that tenured and nontenured faculty differ in their attitudes toward the proposed salary plan.

χ^2 versus z^2

The example just completed could have been done as a two-sample for proportions where p_1 and p_2 are defined the same way as in the example (that is, the proportions of tenured and nontenured staff that favor the new plan). Then

$$\hat{p}_1 = \frac{15}{25}$$

and

$$\hat{p}_2 = \frac{8}{20}.$$

Computation of the z -test statistics for the 2-proportion z test yields $z = 1.333$. Now, $z^2 = 1.78$. Because the chi-square test and the 2-proportion z test are testing the same thing, it should come as little surprise that z^2 equals the obtained value of χ^2 in the example. For a 2×2 table, the χ^2 test statistic and the value of z^2 obtained for the same data in a 2-proportion z test are the same.

RAPID REVIEW

- A study yields a chi-square statistic value of 20 ($\chi^2 = 20$). What is the P value of the test if
 - the study was a goodness-of-fit test with $n = 12$?
 - the study was a test of independence between two categorical variables, the row variable with 3 values and the column variable with 4 values?

Answer:

- $n = 12 \rightarrow df = 12 - 1 = 11 \rightarrow .025 < P < .05$
[using the calculator: $\chi^2 \text{ cdf}(20, 1000, 11) = .045$.]
- $r = 3, c = 4 \rightarrow df = (3 - 1)(4 - 1) = 6 \rightarrow .0025 < P < .005$
[using the calculator: $\chi^2 \text{ cdf}(20, 1000, 6) = .0028$.]

- 2–4. The following data were collected while conducting a chi-square test for independence:

Preference

	Brand A	Brand B	Brand C
Male	16	22	15
Female	18 (X)	30	28

2. What null hypothesis is being tested?

Answer: H_0 : Gender and preference are independent in the population.

3. What is the expected value of the cell marked with the X?

Answer: Identifying the marginals on the table we have

16	22	15	53
18 (X)	30	28	76
34	52	43	129

The expected value is

$$\left(\frac{76}{129}\right)\left(\frac{34}{129}\right)(129) = 20.03.$$

4. How many degrees of freedom are involved in the test?

Answer: $df = (2 - 1)(3 - 1) = 2$

5. The null hypothesis for a chi-square goodness-of-fit test is given as:

$H_0: p_1 = .2, p_2 = .3, p_3 = .4, p_4 = .1$. Which of the following is an appropriate alternative hypothesis?

- $H_A: p_1 \neq .2, p_2 \neq .3, p_3 \neq .4, p_4 \neq .1$
- $H_A: p_1 = p_2 = p_3 = p_4$.
- H_A : Not all of the p 's stated in H_0 are correct.
- $H_A: p_1 \neq p_2 \neq p_3 \neq p_4$.

Answer: c

3 PRACTICE PROBLEMS

Multiple Choice

1. Find the expected value of the cell marked with the “***” in the following 3×2 table (the bold face values are the marginal totals):

observation	observation	19
observation	***	31
observation	observation	27
45	32	77

- 74.60
 - 18.12
 - 12.88
 - 19.65
 - 18.70
2. A χ^2 goodness-of-fit test is performed on a random sample of 360 individuals to see if the number of birthdays each month is proportional to the number of days in the month. χ^2 is determined to be 23.5. The P value for this test is
- $.001 < P < .005$
 - $.02 < P < .025$
 - $.025 < P < .05$
 - $.01 < P < .02$
 - $.05 < P < .10$
3. Two random samples, one of high school teachers and one of college teachers, are selected and each sample is asked about their job satisfaction. Which of the following are appropriate null and alternative hypotheses for this situation?
- H_0 : The proportion of each level of job satisfaction is the same for high school teachers and college teachers.
 H_A : The proportions of teachers highly satisfied with their jobs is higher for college teachers.
 - H_0 : Teaching level and job satisfaction are independent.
 H_A : Teaching level and job satisfaction are not independent.
 - H_0 : Teaching level and job satisfaction are related.
 H_A : Teaching level and job satisfaction are not related.
 - H_0 : The proportion of each level of job satisfaction is the same for high school teachers and college teachers.
 H_A : Not all of the proportions of each level of job satisfaction are the same for high school teachers and college teachers.
 - H_0 : Teaching level and job satisfaction are independent.
 H_A : Teaching level and job satisfaction are not related.
4. A group separated into men and women are asked their preference toward certain types of television shows. The following table gives the results.

	Program Type A	Program Type B
Men	5	20
Women	3	12

Which of the following statements are true?

- I. The variables gender and program preference are independent
 - II. For these data, $\chi^2 = 0$.
 - III. The variables gender and program preference are related.
- a. I only
 - b. I and II only
 - c. II only
 - d. III only
 - e. II and III only

5. For the following two-way table, compute the value of χ^2 .

	C	D
A	15	25
B	10	30

- a. 2.63
- b. 1.22
- c. 1.89
- d. 2.04
- e. 1.45

Free Response

1. An AP Statistics students noted that the probability distribution for a binomial random variable with $n = 4$ and $p = .3$ is approximately given by:


The student decides to test the randBin function on her calculator by putting 500 values into a list from this function ($randBin(4, .3, 500) \rightarrow L1$) and counting the number of each outcome. She obtained

n	P
0	.24
1	.41
2	.27
3	.08
4	.01

n	Observed
0	110
1	190
2	160
3	36
4	4

Do these data indicate that the “randBin” function on the calculator is failing to correctly generate the correct quantities of 0, 1, 2, and 3 from this distribution?

2. A chi-square test for the homogeneity of proportions is conducted on three populations and one categorical variable that has four values. Computation of the chi-square statistic yields $\chi^2 = 17.2$. Is this finding significant and the .01 level of significance?
3. Which of the following best described the difference between a test for independence and a test for homogeneity of proportions?
 - a. There is no difference because they both produce the same value of the chi-square test statistic.
 - b. A test for independence has one population and two categorical variables, whereas a test for homogeneity of proportions has more than one population and only one categorical variable.
 - c. A test for homogeneity of proportions has one population and two categorical variables; whereas, a test for independence has more than one population and only one categorical variable.
 - d. A test for independence uses count data when calculating chi-square and a test for homogeneity uses percentages or proportions when calculating chi-square.
4. Compute the expected value for the cell that contains the frog. You are given the marginal distribution.

	D	E	F	G	Total
A					94
B					96
C					119
Total	74	69	128	38	309

5. Restaurants in two parts of a major city were compared on customer satisfaction to see if location influences customer satisfaction. A random sample of 38 patrons from the Big Steak Restaurant in the eastern part of town and another random sample of 36 patrons from the Big Steak Restaurant on the western side of town were interviewed for the study. The restaurants are under the same management, and the researcher established that they are virtually identical in terms of décor and service. The results are presented in the following table.

Patron's Ratings of Restaurants

	Excellent	Good	Fair	Poor
Eastern	10	12	11	5
Western	6	15	7	8

Do these data provide good evidence that location influences customer satisfaction?

6. A chi-square test for goodness-of-fit is done on a variable with 15 categories. What is the minimum value of χ^2 necessary to reject the null hypothesis at the .02 level of significance?
7. The number of defects from a manufacturing process by day of the week are as follows:

	Monday	Tuesday	Wednesday	Thursday	Friday
Number:	36	23	26	25	40

The manufacturer is concerned that the number of defects is greater on Monday and Friday. Test, at the .05 level of significance, the claim that the proportion of defects is the same each day of the week.

8. A study was done on opinions concerning the legalization of marijuana at Mile High College. One hundred fifty-seven respondents were randomly selected from a large pool of faculty, students, and parents at the college. Respondents were given a choice of favoring the legalization of marijuana, opposing the legalization of marijuana, or favoring making marijuana a legal but controlled substance. The results of the survey were as follows.

	Favor Legalization	Oppose Legalization	Favor Legalization with Control
Students	17	9	6
Faculty	33	40	27
Parents	5	8	12

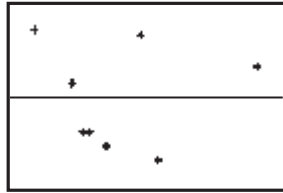
Do these data support, at the .05 level, the contention that the type of respondent (student, faculty, or parent) is related to the opinion toward legalization? Is this a test of independence or a test of homogeneity of proportions?

3 CUMULATIVE REVIEW PROBLEMS

Use the computer output given below to answer Questions 1 and 2.

The regression equation is				
$y = -136 + 3.98 X$				
Predictor	Coef	St Dev	<i>t</i> ratio	<i>P</i>
Constant	-136.10	42.47	-3.20	.024
X	3.9795	0.6529	6.09	.002
s = 8.434	R-sq = 88.1%	R-sq(adj) = 85.8%		

- Based on the computer output above, what conclusion would you draw concerning $H_0: \beta = 0$? Justify your answer.
- Use the computer output above to construct a 99% confidence interval for the slope of the regression line ($n = 8$). Interpret your interval.
- If you roll two dice, the probability that you roll a sum of 10 is approximately .083.
 - What is the probability that you first roll a ten on the 10th roll?
 - What is the average wait until you first roll a 10?
- An experiment is conducted by taking measurements of a personality trait on identical twins and then comparing the results for each set of twins. Would the analysis of the results assume two independent populations or proceed as though there were a single population? Explain.
- The lengths and widths of a certain type of fish were measured and the least-squares regression line for predicting width from length was found to be: $width = -.826 + .193 (length)$. The graph that follows is a residual plot for the data:



- The fish whose width is 3.8 had a length of 25.5. What is the residual for this point?
- Based on the residual plot, does a line seem like a good model for the data?

3 SOLUTIONS TO PRACTICE PROBLEMS

Multiple Choice

- The correct answer is (c). The expected value for that cell can be found as follows: $32/77(31) = 12.88$.
- The correct answer is (d). Because $n = 12$, $df = 12 - 1 = 11$. Reading from the χ^2 table, we have $.01 < P < .02$.
- The correct answer is (d). Because we have independent random samples of teachers from each of the two levels, this is a test of homogeneity of proportions.
- The correct answer is (b). The expected values for the cells are exactly equal to the observed values (e.g., for the 1st row, 1st column, $Exp = (25/40)(8) = 5$), so χ^2 must equal 0 \rightarrow the variables are independent, not related.

5. The correct answer is (e). The expected values for this two-way table are given by the matrix: $\begin{bmatrix} 12.5 & 27.5 \\ 12.5 & 27.5 \end{bmatrix}$.

Then,

$$\chi^2 = \frac{(15 - 12.5)^2}{12.5} + \frac{(25 - 27.5)^2}{27.5} + \frac{(10 - 12.5)^2}{12.5} + \frac{(30 - 27.5)^2}{27.5} = 1.45$$

Free Response

1. I. Let p_1 = the proportion of 0s, p_2 = the proportion of 1s, p_3 = the proportion of 2s, and p_4 = the proportion of 3s, p_5 = the proportion of 4s
 $H_0: p_1 = .24; p_2 = .41; p_3 = .27; p_4 = .07; p_5 = .01$
 H_A : Not all of the proportions stated in H_0 are correct.
- II. We will use a chi-square goodness-of-fit test. The expected values for 500 trials are
- | n | Expected | Because all expected values are at least 5, the conditions for the test have been met. |
|-----|--------------------|--|
| 0 | $(.24)(500) = 120$ | |
| 1 | $(.41)(500) = 205$ | |
| 2 | $(.27)(500) = 135$ | |
| 3 | $(.08)(500) = 40$ | |
| 4 | $(.01)(500) = 5$ | |
- III. $\chi^2 = \sum \frac{(O - E)^2}{E} = 7.16, df = 4 \rightarrow .10 < P \text{ value} < .15$
 [on the calculator, $\chi^2 \text{cdf}(7.16, 1000, 4) = .128$]
- IV. The P value is greater than .10, which is too high to reject the null. We do not have strong evidence that the calculator is not performing as it should.
2. For a 3×4 two-way table, $df = (3 - 1)(4 - 1) = 6 \rightarrow .005 < P \text{ value} < .01$. The finding is significant at the .01 level of significance. [$\chi^2 \text{cdf}(17.2, 1000, 6) = .009$]
3. b

4. The expected value of the cell with the frog is $128/309(96) = 39.77$.
5. Because of the manner in which the test was set up, this is a test of homogeneity of proportions.

- I. H_0 : the proportions of patrons from each side of town that rate the restaurant Excellent, Good, Fair, and Poor are same for eastern and western.
 H_A : Not all the proportions are the same.
- II. We will use the chi-square test for homogeneity of proportions. Calculation of expected values yields the following results:

8.22	13.86	9.24	6.68
7.78	13.14	8.76	6.32

 Because all expected values are at least 5, the conditions for the test are met.
- III. $\chi^2 = \sum \frac{(O - E)^2}{E} = 2.86$, $df = (2 - 1)(4 - 1) = 3 \rightarrow P \text{ value} > .25$ (from table.) [P value = .41 from calculator.]
- IV. The P value is large. We cannot reject the null. These data do not provide us with evidence that location influences customer satisfaction.

6. If $n = 15$, then $df = 15 - 1 = 14$. In the table we find the entry in the column for tail probability of .02 and the row with 14 degrees of freedom. That value is 26.87. Any value of χ^2 larger than 26.87 will yield a P value less than .02.

7.

- I. H_0 : $p_1 = p_2 = p_3 = p_4 = p_5$ (the proportion of defects is the same each day of the week.)
 H_A : At least one proportion is not equal to the others (the proportion of defects is not the same each day of the week.)
- II. We will use a chi-square goodness-of-fit test. The number of expected defects is the same for each day of the week. Because there were a total of 150 defects during the week, we would expect, if the null is true, to have 30 defects each day. Because the number of expected defects is greater than 5 for each day, the conditions for the test are met.
- III. $\chi^2 = \sum \frac{(O - E)^2}{E} = 7.533$, $df = 4 \rightarrow .10 < P \text{ value} < .15$ (tables.) [P value = .11 on the calculator.]

IV. The P value is too large to reject the null. We do not have strong evidence that there are more defects produced on Monday and Friday than on other days of the week.

8. Because we have a single population from which we drew our sample and we are asking if two variables are related *within* that population, this is a chi-square test for independence.

I. H_0 : Type of respondent and opinion toward the legalization of marijuana are independent.
 H_A : Type of respondent and opinion toward the legalization of marijuana are not independent.

II. We will do a χ^2 test of independence. The expected values are given by

11.21	11.62	9.17
35.03	36.31	28.66
18.76	9.08	7.16

Because all expected values are greater than 5, the conditions for the test have been met.

III. $\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(17 - 11.21)^2}{11.21} + \dots + \frac{(12 - 7.16)^2}{7.16} = 10.27$,
 $df = (3 - 1)(3 - 1) = 4 \rightarrow .025 < P \text{ value} < .05$ (table.)
 [From the calculator, $P \text{ value} = \chi^2 cdf(10.27, 1000, 4) = .036$.]

IV. Because $P < .05$, reject H_0 . We have evidence that the type of respondent is related to opinion concerning the legalization of marijuana.

3 SOLUTIONS TO CUMULATIVE REVIEW PROBLEMS

1. The t test statistic for the slope of the regression line (under $H_0: \beta = 0$) is 6.09. This translates into a P value of .002, as shown in the printout. This low P allows us to reject the null hypothesis. We conclude that we have strong evidence that there is a linear relationship between X and Y .
2. For $C = .99$ (a 99% confidence interval) at $df = n - 2 = 6$, $t^* = 3.707$. The required interval is: $3.9795 \pm 3.707(.6529) = <1.56, 6.40>$. We are 99% confident that the true slope of the population regression line is between 1.56 and 6.40.
3.
 - a. This is a geometric distribution problem. The probability that the first success occurs on the 10th roll is $G(10) = (.083)(1 - .083)^9 = .038$. [On the calculator, $geometpdf(.083, 10) = .038$ – difference due to rounding.]

- b. $1/.083 = 12.05$ On average it will take about 12 rolls before you get a 10. (Actually, it's *exactly* 12 rolls on average because the exact probability of rolling a 10 is $1/12$, not .083.)
4. This is an example of a *matched pairs design*. The identical twins form a block for comparison of the treatment. Statistics are based on the differences between the scores of the pairs of twins. Hence, the situation is univariate—one population of values, each constructed as a difference of the matched pairs.
5. a. $Residual = actual - predicted = 3.8 - [-.826 + .193(25.5)] = 3.8 - 4.01 = -.295$.
- b. There does not seem to be any consistent pattern in the residual plot, indicating that a line is probably a good model for the data.