

Statistics



[http://www.textbooksfree.org/
Excel-Statistics-Book.htm](http://www.textbooksfree.org/Excel-Statistics-Book.htm)

Walter Antoniotti
21st Century Learning Products

Third Edition

ISBN 1-929850-01-8

Copyright © 2001 by 21st Century Learning Products

All rights reserved.

QUICK NOTES is a registered trademark of 21st Century Learning Products

Fred and Lulu have been provided by Corel Draw and Image Club Graphics (403-262-8008) respectively.

21st Century Learning Products
227 Baboosic Lake Road
Merrimack, NH 03054
603-424-4665
800-253-6595
antonw@ix.netcom.com
www.businessbookmall.com

Dedication

This book is dedicated to the many teachers who spend countless hours developing class handouts to meet the learning styles and ability levels particular to their students. I have been privileged to learn from many such teachers, the most pertinent of whom is the late Dr. Paul R. Gawthrop, my Marietta College Statistics teacher. **Quick Notes Statistics** was modeled after his Statistics course outline.

A Very Special Thank You

To Professors Carl T. Brezovec, Normand A. Dion, William H. Jack, Jr., Candace B. McKinniss, Robert F. Wiesenauer, and P. Teresa Farnum of Franklin Pierce College, Rindge, New Hampshire, whose suggestions and encouragement improved the book and made the project more enjoyable. To my Franklin Pierce College Division of Professional Studies Statistics students who enhanced the development of **The Quick Notes Learning System** for Statistics. To Jill Moon, graduate statistics student at George Mason University, Washington, DC, who extensively reviewed an early draft of the book. To Professor William R. Benoit, Chair of the Business Department, Plymouth State College, Plymouth, New Hampshire, for his invaluable suggestions.

About The Author

Walter Antoniotti began teaching Statistics over 30 years ago for Daniel Webster College, Nashua, New Hampshire, where he became an Associate Professor of Business Administration and Chairperson of the Department of Aviation Management. During the past 21 years, as Director and then Dean of Continuing Education for Franklin Pierce College, Rindge, New Hampshire, Walter helped build one of New England's most successful Continuing Education Programs. Today, as Franklin Pierce College's Special Assistant for Professional Studies Program Development, Walter enjoys teaching, writing, and investigating areas of interest to himself and the College. Walter Antoniotti has a Bachelor of Science degree in Business Administration from Marietta College, Marietta, Ohio, and a Masters of Business Administration degree from Northeastern University, Boston, Massachusetts.

THE QUICK NOTES PHILOSOPHY

The Theory of Optimum Amounts

There exists for every CONCEPT to be learned,
an optimum amount of explanatory material.
There exists for every TOPIC to be learned,
an optimum number of concepts to be integrated.
There exists for every SUBJECT to be learned,
an optimum number of topics to be mastered.

By limiting explanatory material to optimum
amounts, **Quick Notes** maximizes learning.

The Theory of Optimum Placement

There exists for every CONCEPT to be learned and
integrated into a TOPIC of concern, a unique place-
ment of elements that will maximize learning.

By placing related elements on the same page
or facing pages, **Quick Notes** maximizes learning.

The Optimum Relationship Between Content and Process

Education is the learning of content and process.
Content is the what of learning—it's the arithmetic
of *mathematics* and the *grammar of communication*.
Process is the application of content—it's the problem-
solving of *mathematics* and the writing of *communication*.
Learning begins with content and expands to process.

By making the learning of content easier,
Quick Notes makes the learning of process easier.

Education Requires Sacrifice and Discipline

Sacrifice and discipline, which are required to do schoolwork
and homework, are essential parts of the educational process.
Applying the **Quick Notes Philosophy** will make this sacrifice
and discipline less frustrating, but it will not make education fun.
If schoolwork and homework were supposed to be fun, they would
be called schoolfun and homefun.

By learning to sacrifice and exhibit discipline while going to school,
a young person begins the process of becoming an adult.

The World of Multiple Intelligence

Howard Garner's Theory of Multiple Intelligence defines these eight kinds of human intelligence.

1. Mathematical-logical (problem solving, fix or repair, program)
2. Spatial (dance, sports, driving a bus)
3. Bodily-kinesthetic (acting, mime, sports)
4. Musical-rhythmic (composing, playing music, clapping)
5. Verbal-linguistic (reading, using words, public speaking, storytelling)
6. Interpersonal (social skills, reading other people, working in a group)
7. Intrapersonal (introspection, self-assessment, goal making, vision, planning)
8. Naturalist (able to distinguish among, classify, and use environmental features)

Mathematical-logical and Verbal intelligence represent **core intelligence**. Skills related to core intelligence are emphasized by traditional schools. People with above average ability in any of the eight areas of intelligence, have **special intelligence**. The world of work rewards people who develop skills associated with their special intelligence, provided they meet minimum skill requirements associated with core intelligence.

Determining Appropriate Education for a World of Multiple Intelligence

Determining educational requirements begins by matching a person's special intelligence with careers that reward this intelligence. Careers have many levels of competition. Choosing one's appropriate level requires honest analysis of intelligence, motivation, and personal needs. For example, the health industry requires doctors and nurses, hospital directors and floor supervisors, x-ray technicians and physical therapists. Career success will be enhanced by choosing an appropriate level of competition, one in which core and special intelligence requirements are reasonably satisfied. Once the competitive level is set, the appropriate education, considering minimum core intelligence and special intelligence requirements, can be determined.

Success at any level will be enhanced by improving skills related to non-core and non-special intelligence. A person might not like going to the office picnic or talking to potential customers, but developing these skills is important to economic success.

The dynamic nature of business may cause skill requirements for a particular career level to change. In addition, people often want to compete at a higher level. As a result, an individual may frequently have to compare their core and special intelligence with new skill requirements. Once this analysis is completed, choosing an education appropriate for the enhancement of these skills may begin.

Developing Special Skills is Important

Once minimum core intelligence skill requirements have been satisfied for a given career level, economic and academic returns from education will be maximized by developing special intelligence skills. People who ignore the process of determining appropriate education for a world of multiple intelligence may receive little return from their education.

Bureau of the Census 1992 data indicates that approximately 25% of the bachelor degree holders earn less than the median high school graduate and approximately 20% of the high school graduates earn more than the median college graduate. Percentages vary depending upon age, gender, and other demographic characteristics.

National Survey of Adult Literacy tests measuring Prose, Document (understanding forms), and Quantitative skills conducted by the Department of Education in 1992 reported that 15 to 20% of four-year college graduates have skill levels below median high school graduates.

Using The Quick Notes Learning System

Quick Notes

explain basic statistics principles with clear, concise, two-page outlines. The beginning of each outline contains basic definitions, theories, and concepts. The nature of statistics is explained at the beginning of chapter 1. See page 162 for a complete review of areas covered by **Quick Notes Statistics**.

Chapter 1 Statistics Is About Using Data in Decision Making

- I. The nature of statistics
 - A. Many disciplines use statistics.
 1. Business and Economics
 2. Natural and Social Sciences
 3. Physical Sciences
 4. Education
 5. Politics
 - B. Basic definitions
 1. **Population**: totality under study such as the students attending a school
 2. **Sample**: subset of a population such as the students in one class of a school
 3. **Parameter**: a characteristic of a population such as the average age of students attending a school
 4. **Statistic**: a characteristic of a sample such as the average age of students in a class of a school
 - C. **Statistics** is the science of collecting, organizing, presenting, analyzing, and interpreting numerical data in relation to the decision-making process.
 1. **Descriptive statistics** summarizes numerical data using numbers and graphs. The grades of students in a class can be summarized with averages and line graphs.
 2. **Inferential statistics** uses sample statistics to estimate population parameters. The average age of students in a class can be used to estimate the average age of students attending a school.

Linda's Video Showcase

A continuous example of how Linda Smith calculates statistics and uses them when making business decisions for Linda's Video Showcase is an integral part of **Quick Notes Statistics**. Videotape rentals will be analyzed to learn about the relationship between sales revenue and advertising expenditures. Customer satisfaction will be measured, as will the effectiveness of her sales team. See page 164 for a complete review of the topics she will explore.

Chapter 2 Summarizing Data

- I. Linda's Video Showcase
 - A. Upon graduating from college, Linda Smith opened **Linda's Video Showcase**, a retail business specializing in videotape rentals.
 - B. Linda will use descriptive statistics to analyze this daily video rentals data set.
 1. **76, 88, 53, 66, 97, 73, 64, 82, 77, 57, 93, 85, 70, 76, 68**
 2. Linda's first step was to make a list of data by order of magnitude called an **array**. She also calculated a range (high number minus the low number) for the data.
Array: 53, 57, 64, 66, 68, 70, 73, 76, 76, 77, 82, 85, 88, 93, 97
Range: High - Low = 97 - 53 = 44
- II. Frequency distributions
 - A. A **frequency distribution** divides data into numerical groupings and depicts the number of observations occurring within each grouping. Academic grades are often summarized with a frequency distribution with each of the five grades representing a group. A grade of B is usually between 79 and 90. The first three columns of the chart at the bottom of this page are a frequency distribution of the above rental data.

Practice Sets Provide Reinforcement

Each Quick Notes chapter is followed by a Practice Set of similar design. If you have trouble answering a Practice Set problem, just turn back two pages and look at the same page location for the appropriate Quick Notes demonstration problem.

Darin's Music Emporium

Practice Sets deal with how Darin Jones calculates and uses statistics when managing Darin's Music Emporium. Then he buys Future Horizons Corporation. It requires he study product quality control and other issues of concern to manufacturing companies.

Quick Questions

follow Practice Sets and review definitions and other important concepts.

Reviews and Tests

The first 4 parts of Quick Notes end with a formula review and a test. Part V is a unique review of Quick Notes Statistics.

Complete Solutions

to Practice Sets, Quick Questions, and tests have been provided to help with difficult concepts.

Practice Set 2 Summarizing Data

I. Darin's Music Emporium

- A. Upon graduating from college, Darin Jones opened **Darin's Music Emporium**. The company sells music-related hardware and software. We will use descriptive statistics to analyze company sales data.
- B. Darin recently collected the following Walkman CD Recorder sales data.

Units sold per day: 17, 22, 17, 8, 12, 15, 14, 16, 21, 16

1. Make an array and calculate the range of this data.
 2. Calculate an appropriate class width for this data.
- II. **Make a 5-class frequency distribution using stated class limits for the first class of 5-9 sales units. Those using statistics software should try other class limits with their software and print the one with the most symmetrical distribution.**

Quick Questions 2 Summarizing Data

- I. Place the number of the appropriate formula or phrase next to the item it describes.

- A. Mutually-exclusive events _____
- B. Relative frequency _____
- C. Class midpoint _____
- D. Approximate class width _____
- E. All-inclusive events _____
- F. Ogive _____

1.	$\frac{\text{range}}{\text{\# of classes}}$
2.	do not contain the same outcome
3.	$\frac{X_1 + X_2}{2}$
4.	$\frac{\text{class frequency}}{\text{total frequencies}}$
5.	cumulative frequency distribution
6.	a place for every outcome

Probability Formula Review

I. Types and characteristics of probability

A. Types of probability

1. Classical: $P(A) = \frac{A}{N}$

2. Empirical: $P(A) = \frac{A}{n}$

Probability Test

- I. Average hours worked by manufacturing workers is normally distributed with a mean of 41 hours and a standard deviation of .5 hours. Graph and solve the following problems.

A. $P(41.0 \text{ hours} \leq x < 42.5 \text{ hours})$

Introducing Fred Look Ahead and Lulu Review

I'm Fred Look Ahead.

Use me as a reminder to look over
the main points of a learning unit before
reading it in detail. Looking around
first will make learning easier.



I'm Lulu Review.

I'm here to remind you
to review once in a while.
So jump on board, and
we will review together.



Message to Quick Users

For All Users

1. Quick Notes summarize difficult concepts. Most students review them a number of times.
2. Complete solutions to all Practice Sets and Quick Questions are provided in Part VI. Reading these answers is a great way to review basic concepts, especially when studying for a test!
3. Chapters 25 to 27 review important concepts and are designed to tie everything together. Relevant sections of these chapters should be reviewed after completing each part of Quick Notes.

For People Not Using Statistics Software

1. Information provided on page 22 of chapter 5 and in all of chapter 6 has been provided in chapters 3 and 4 and may be skipped. **Warning!** This information may be required by those taking a college statistics course. Check your syllabus to see if "grouped measures" are required.
2. Quick answers may differ slightly from your answers because of rounding. When answers differ, compare your procedures with those of the appropriate Quick Notes demonstration problem and check your math.
3. **Ignore Data Sets For People Using Statistics Software.**

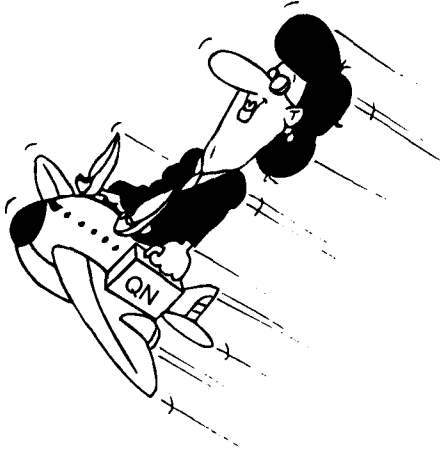
For People Using Quick Notes Data Files and Statistics Software Directions

1. Data files, practice set instructions, and computer generated answers are available for popular statistics programs. If purchased, they are on the disk affixed to the back cover. Set your word processor to Rich Text Format and load the file "compdir" for directions on how to use your software with Quick Notes Statistics™.
2. Information provided on page 22 of chapter 5 and in all of chapter 6 has been provided elsewhere and may be skipped. **Warning!** This information may be required for those taking a college statistics course. Check your syllabus to see if "grouped measures" are required. Grouped calculations will differ from ungrouped calculations.

Free Study Aids to help with Statistics, Excel, Accounting, Economics, Management, and Mathematics are available at www.businessbookmall.com.

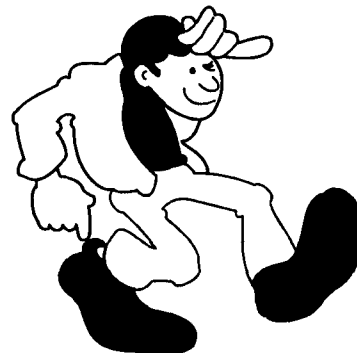
Table of Contents

Chapter	Part I Descriptive Statistics	Page
1	Statistics Is About Using Data in Decision Making	2
2	Summarizing Data	4
3	Measuring Central Tendency of Ungrouped Data	10
4	Measuring Dispersion of Ungrouped Data	16
5	Measuring Central Tendency of Grouped Data	22
6	Measuring Dispersion of Grouped Data	28
	Descriptive Statistics Formula Review and Test	34
Part II Probability, The Basis for Inferential Statistics		
7	Understanding Probability	40
8	Probability Part II Multiplication Rules	46
9	Discrete Probability Distributions	52
10	Continuous Normal Probability Distributions	58
11	Sampling and the Sampling Distribution of the Means	66
12	Sampling Distributions Part II	70
	Probability Formula Review and Test	76
Part III Inferential Statistics		
13	Large Sample Hypothesis Testing	84
14	Large Sample Hypothesis Testing Part II	88
15	Hypothesis Testing of Population Proportions	94
16	Small Sample Hypothesis Testing Using Student's t Test	98
17	Statistical Quality Control	102
18	Analysis of Variance	108
19	Two-Factor Analysis of Variance	114
20	Nonparametric Hypothesis Testing of Nominal Data	120
21	Nonparametric Hypothesis Testing of Ordinal Data Part I	126
22	Nonparametric Hypothesis Testing of Ordinal Data Part II	132
	Inferential Statistics Executive Summary, Formula Review, and Test	135



Part IV Correlation and Regression		
23	Correlation Analysis	146
24	Simple Linear Regression Analysis	152
	Correlation and Regression Formula Review and Test	158
Part V Cumulative Review		
25	Taxonomy of Statistics	162
26	Taxonomy of Parametric Statistics	163
27	Problem Review	164
Part VI The Professor's Answer Book		
	Appendix I Complete Solutions to Practice Sets	PS 5
	Appendix II Complete Solutions to Quick Questions	QQ 3
	Appendix III Complete Solutions to Tests	T 35
Part VII Statistical Tables		ST 1
Part VIII Index		I 1

Let's get to work before Lulu gets us into trouble!



Chapter 1 Statistics Is About Using Data in Decision Making

Remember to look at the key points of a learning unit before studying them in detail. Here you will see that this unit covers definitions related to the nature of statistics, the nature of measurement, and the collection of data.

I. The nature of statistics

A. Many disciplines use statistics.

1. Business and Economics
2. Natural and Social Sciences
3. Physical Sciences
4. Education
5. Politics

B. Basic definitions

1. **Population:** totality under study such as the students attending a school
2. **Sample:** subset of a population such as the students in one class of a school
3. **Parameter:** a characteristic of a population such as the average age of students attending a school
4. **Statistic:** a characteristic of a sample such as the average age of students in a class of a school

C. **Statistics** is the science of collecting, organizing, presenting, analyzing, and interpreting numerical data in relation to the decision making process.

1. **Descriptive statistics** summarizes numerical data using numbers and graphs. The grades of students in a class can be summarized with averages and line graphs.
2. **Inferential statistics** uses sample statistics to estimate population parameters. The average age of students in a class can be used to estimate the average age of students attending a school.



II. The nature of measurement

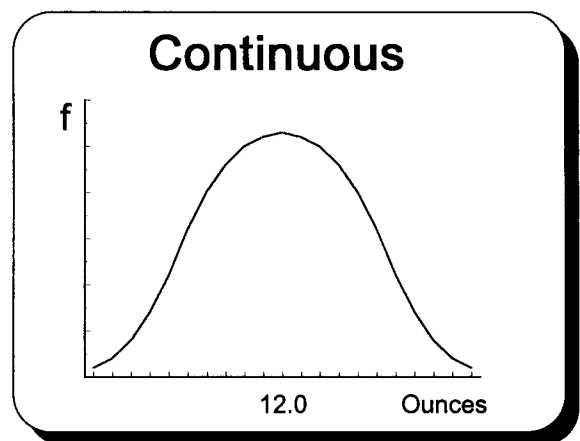
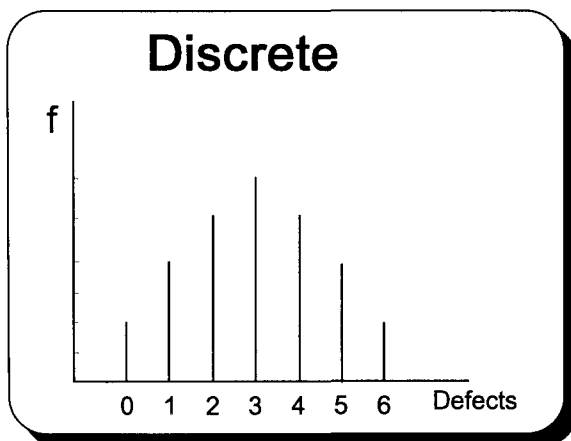
A. **Variable:** an activity subject to variation, e.g., grades on a statistics test and how someone feels

B. Quantitative versus qualitative variables

1. **Quantitative variable:** expressed numerically, e.g., a grade of 85 and a body temperature of 101 degrees
2. **Qualitative variable:** not expressed numerically, e.g., a grade of B and someone feeling poorly

C. Discrete versus continuous variables

1. **Discrete:** only finite values, such as the countable numbers, can exist on the x-axis, e.g., defects in a tire and the number correct on a true or false statistics exam
2. **Continuous:** measurement may assume any value associated with an uninterrupted scale, e.g., a bottle may contain 12.02 ounces of liquid refreshment and a person may weigh 175.25 pounds



D. The x-axis, as shown here, represents 1 of 4 measurement scales important to our study of statistics.

E. The y-axis often measures how often an x-axis measurement has occurred. This is called frequency (f).

F. Measurement scales (levels) determine data's exactness

1. **Nominal** scaled data is the weakest, providing the least information. Data can only be put into groups called categories and be counted. No order or scale exists. Examples include the number of shoppers who buy or do not buy when going into a store and the number of parts that pass or do not pass inspection.
2. **Ordinal** scaled data can be arranged in order. An example would be the number of customers who think a product is poor, average, or good. While good is better than average, no attempt is made to quantify such differences into measurable intervals.
3. **Interval** scaled data allows for the quantification of difference. Fahrenheit and Celsius thermometers have interval scales. These scales have equal intervals. But, their measure of zero is arbitrary because zero degrees does not measure the absence of heat. Such arbitrary starting points place restrictions on the math operations that can be done with interval scaled data. For example, the use of proportions is not appropriate.
4. **Ratio scaled data** has an inherent starting point. Temperature measured on a Kelvin scale is ratio scaled data because zero represents the absence of heat. Total variable costs are ratio scaled data because costs are zero when production is zero. Total costs, because of fixed costs, are interval scaled data.

III. Collecting data

A. Primary versus secondary sources of data

1. **Primary source data** is published by the original collector (data collected by the Bureau of the Census).
2. **Secondary source data** is published by a noncollector (Bureau of the Census data printed in a newspaper).

B. Methods of gathering data

1. **Observation**
2. **Personal interview**
3. **Telephone interview**
4. **Self-administration** is when a form (questionnaire) is completed by the respondent (individual, company, etc.).
5. **Registration** is when the respondent is responsible for bringing the desired information to a prescribed location (registering a car).

C. Data gathering alternatives

1. A **survey** is the collecting of information concerning existing material.
 - a. A **census** contains information from an entire population.
 - b. A **sample** contains information from part of a population.
 - 1) **Sampling error** occurs because a sample is taken rather than a census. The primary cause of sampling error is the sample is not representative of the population.
 - 2) **Nonsampling error**, which occurs with any survey, exists because of poor collection techniques. Because a sample is smaller than a census, more effort may be put into eliminating nonsampling error. This means that limited funds may make a sample more accurate than a census.
2. An **experiment** is a process for generating and measuring data.

Quick Questions 1 Statistics Is About Using Data in Decision Making

Place the number of the appropriate description next to the item it describes.

- | | |
|---------------------------------|--|
| A. Statistic _____ | 1. Subset of a population |
| B. Parameter _____ | 2. Expressed numerically |
| C. Population _____ | 3. The use of sample statistics to estimate population parameters |
| D. Discrete _____ | 4. Characteristic of a sample |
| E. Quantitative variable _____ | 5. Only finite values can exist on the x-axis |
| F. Secondary source data _____ | 6. Published by the original collector |
| G. Sample _____ | 7. Measurement may assume any value associated with an uninterrupted scale |
| H. Inferential statistics _____ | 8. Published by a noncollector |
| I. Continuous _____ | 9. Characteristic of a population |
| J. Primary source data _____ | 10. Totality under study |

See page QQ 3 of Appendix II for Complete Solutions to Quick Questions.

Chapter 2 Summarizing Data

I. Linda's Video Showcase

- A. Upon graduating from college, Linda Smith opened **Linda's Video Showcase**, a retail business specializing in videotape rentals.
- B. Linda will use descriptive statistics to analyze this daily video rentals data set.

1. 76, 88, 53, 66, 97, 73, 64, 82, 77, 57, 93, 85, 70, 76, 68

2. Linda's first step was to make a list of data by order of magnitude called an **array**. She also calculated a range (high number minus the low number) for the data.

Array: 53, 57, 64, 66, 68, 70, 73, 76, 76, 77, 82, 85, 88, 93, 97

Range: High - Low = 97 - 53 = 44

II. Frequency distributions

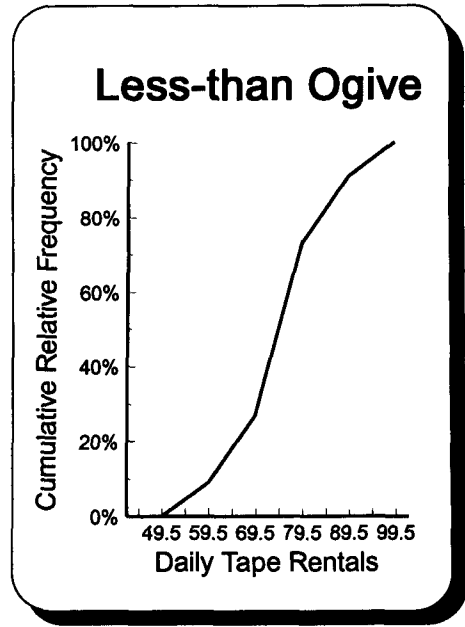
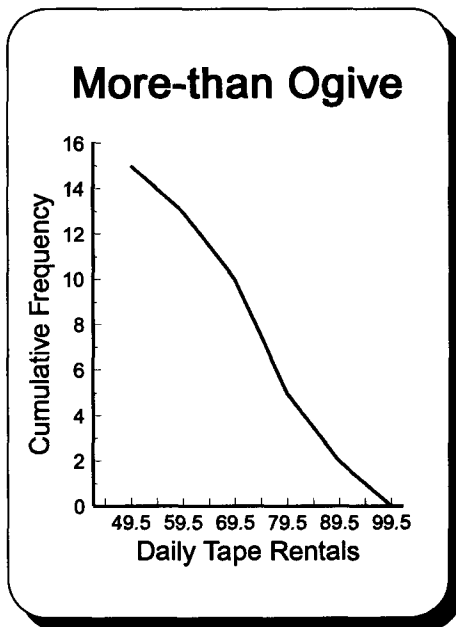
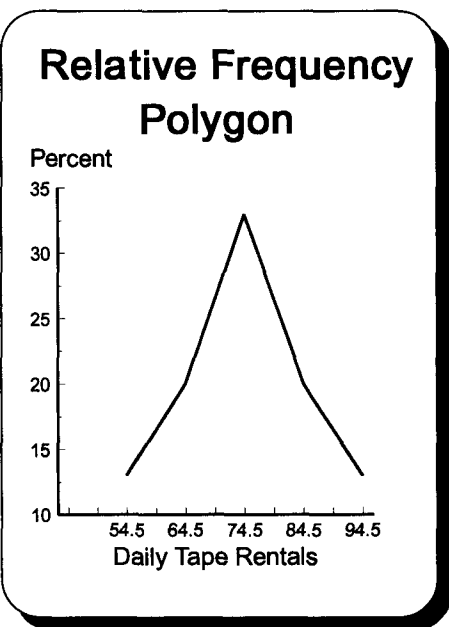
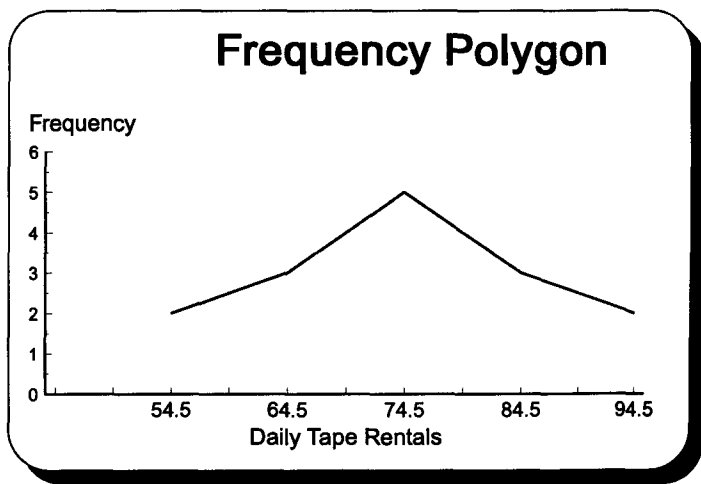
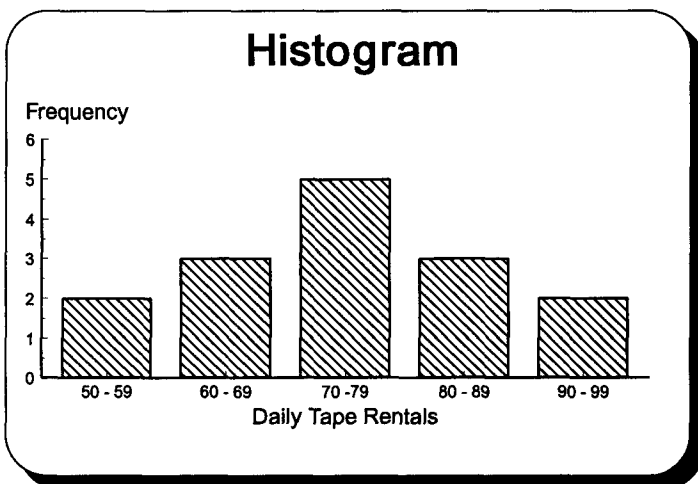
- A. A **frequency distribution** divides data into numerical groupings and depicts the number of observations occurring within each grouping. Academic grades are often summarized with a frequency distribution with each of the five grades representing a group. A grade of B is usually between 79 and 90. The first three columns of the chart at the bottom of this page are a frequency distribution of the above rental data.
- B. A grouping is called a **class**.
1. **Class limits** state the extremes of a class. Their difference is called the **class width**.
 2. Classes must be **mutually exclusive** in that a piece of data (outcome) may belong to only one class.
 3. Classes must be **all-inclusive (collectively exhaustive)** in that there must be a class for every outcome.
- C. Data is often summarized with 5 to 15 classes.
1. A class width should be easily divisible, i.e., 5, 10, 50, 100, 500, etc.
 2. This formula is used with a number such as five to determine an approximate class width.
 3. Data that is naturally clustered should be so clustered in the distribution. If possible, all classes should be of equal size and contain at least one outcome.
- $$\frac{\text{range}}{\# \text{ of classes}} = \frac{44}{5} = 8.8$$
- D. Rounded class limits are called **stated class limits**.
1. For example, a first class with stated class limits of 50-59 would have **real class limits** of 49.5-59.5.
 2. Outcomes equal to the upper real limit belong to the next higher class. That is, the outcome 59.5 would belong to the second class.
- E. A **tally** is a vertical line used to count class outcomes.
1. The total outcomes of a class are its **frequency** (rate of occurrence).
 2. Frequency, expressed as a decimal, is called **relative frequency**.
- $$\text{relative frequency} = \frac{\text{class frequency}}{\text{total frequencies}}$$
- F. **Cumulative frequency** is measured by **more-than** and **less-than ogives**. Ogives summarize the cumulative number of outcomes over or under each real class limit.
- G. Frequency, relative frequency, and cumulative frequency are calculated below and graphed on the next page.

Linda's Video Showcase Daily Rentals Beginning 1/2/98						
Stated Class Limits	Real Class Limits	Tally	Frequency (f)	Relative Frequency $f \div n$	Cumulative Frequency	
					More-than	Less-than
50 - 59	49.5 - 59.5		2	0.13	49.5 is 15	49.5 is 0
60 - 69	59.5 - 69.5		3	0.20	59.5 is 13	59.5 is 2
70 - 79	69.5 - 79.5		5	0.34	69.5 is 10	69.5 is 5
80 - 89	79.5 - 89.5		3	0.20	79.5 is 5	79.5 is 10
90 - 99	89.5 - 99.5		2	0.13	89.5 is 2	89.5 is 13
Totals			n = 15	1.00	99.5 is 0	99.5 is 15

H. Graphing frequency distributions

1. A **histogram** is a vertical bar chart depicting a frequency distribution. The x-axis is for the variable being measured and the y-axis is for the frequency.
2. A **frequency polygon** (a many-sided figure) is a line graph depicting a frequency distribution.
 - a. Each frequency is depicted at the midpoint of the class it represents.
 - b. The midpoint is the stated or real class limits added together and divided by two. Both yield the same answer.
3. A **relative frequency polygon** is similar to a frequency polygon except it has the relative frequency of each class on the y-axis.
4. **Cumulative frequency distributions** (Ogives) measure the accumulation of frequencies above and below each real class limit.
 - a. A **more-than cumulative frequency distribution** begins with the number of frequencies that are above the real lower limit of the lowest class. The answer is equal to total frequency. It is located near the top of the y-axis above the lower real class limit. Each successive class limit is associated with a smaller and smaller number of frequencies being above the successively higher class limits. The final value on the y-axis will be zero because none of the outcomes can be higher than the upper limit of the upper class. Cumulative frequency distributions can also be constructed on a relative basis with the cumulative frequency percentage graphed on the y-axis.
 - b. A **less-than cumulative frequency distribution** is the complement of the more-than frequency distribution. Its y-axis value at origin is zero, and at the upper class limit, y will be equal to total frequency.

$$\frac{X_1 + X_2}{2} = \frac{50 + 59}{2} = 54.5$$



Note the y-axis scale.

Practice Set 2 Summarizing Data

See pages PS 6 and PS 7 of Appendix I for complete solutions to this Practice Set.

- I. Darin's Music Emporium
- A. Upon graduating from college, Darin Jones opened **Darin's Music Emporium**. The company sells music-related hardware and software. We will use descriptive statistics to analyze company sales data.
- B. Darin recently collected the following Walkman CD Recorder sales data.

Units sold per day: 17, 22, 17, 8, 12, 15, 14, 16, 21, 29, 16

1. Make an array and calculate the range of this data.

 2. Calculate an appropriate class width for this data.
- II. Make a 5-class frequency distribution using stated class limits for the first class of 5-9 sales units. Those using statistics software should try other class limits with their software and print the one with the most symmetrical distribution.

Darin's Music Emporium Walkman Sales Data						
Stated Class Limits						
5 - 9						

- A. Draw or print a histogram.

Note: The x-axis may be labeled with the lower stated or real class limits, the class midpoints, or each class range.

B. Draw or print a frequency polygon.

C. Draw or print a less-than cumulative relative frequency polygon (Ogive) and a relative frequency polygon.

Note: A less-than cumulative relative frequency distribution can be used to estimate the percentiles defined in chapter 3.

See pages PS 6 and PS 7 of Appendix I for complete solutions to this Practice Set.

Quick Questions 2 Summarizing Data

I. Place the number of the appropriate formula or phrase next to the item it describes.

- A. Mutually-exclusive events _____
- B. Relative frequency _____
- C. Class midpoint _____
- D. Approximate class width _____
- E. All-inclusive events (collectively exhaustive) _____
- F. Ogive _____

1.	$\frac{\text{range}}{\text{\# of classes}}$
2.	do not contain the same outcome
3.	$\frac{X_1 + X_2}{2}$
4.	$\frac{\text{class frequency}}{\text{total frequencies}}$
5.	cumulative frequency distribution
6.	a place for every outcome

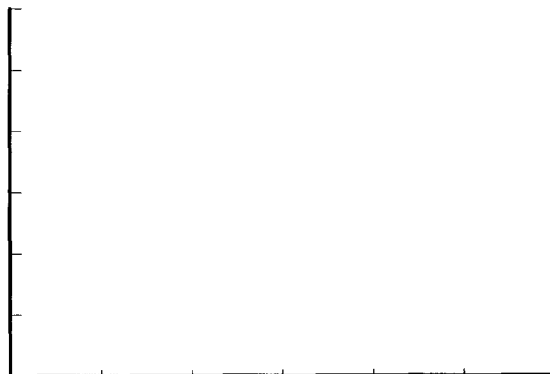
II. Complete the following using this data.

Data: 38, 48, 27, 14, 31, 23, 46, 38, 54, 26, 44, 33, 17, 34, 6, 37

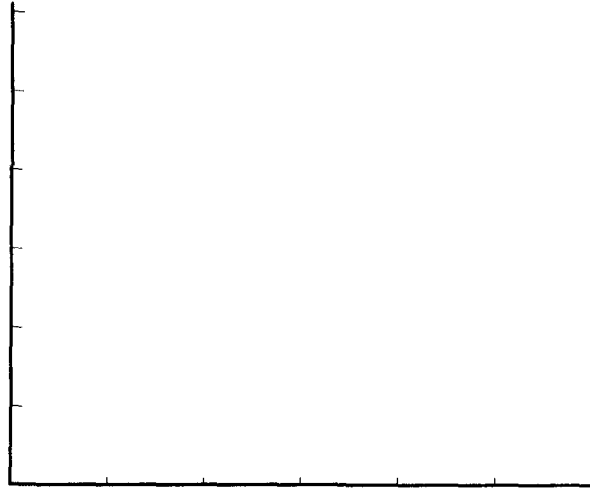
- A. Array
- B. Range
- C. Approximate class width
- D. Complete this chart. People using statistics software should print a frequency distribution, relative frequency distribution, and less-than cumulative frequency distribution.

Stated Class Limits	Real Class Limits	Tally	Frequency (f)	Relative Frequency	Cumulative Frequency	
					More-than	Less-than
5 - 14						

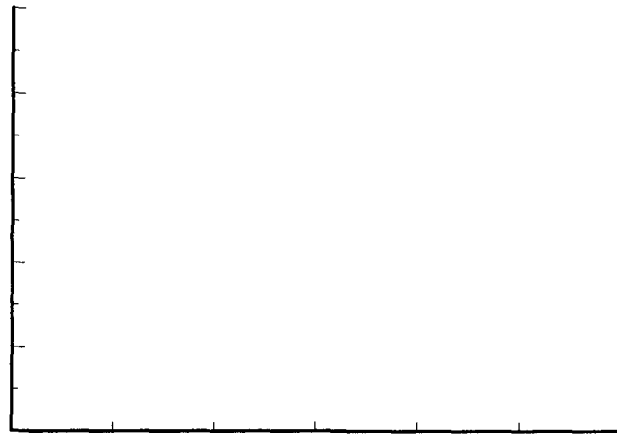
E. Frequency polygon



F. Histogram



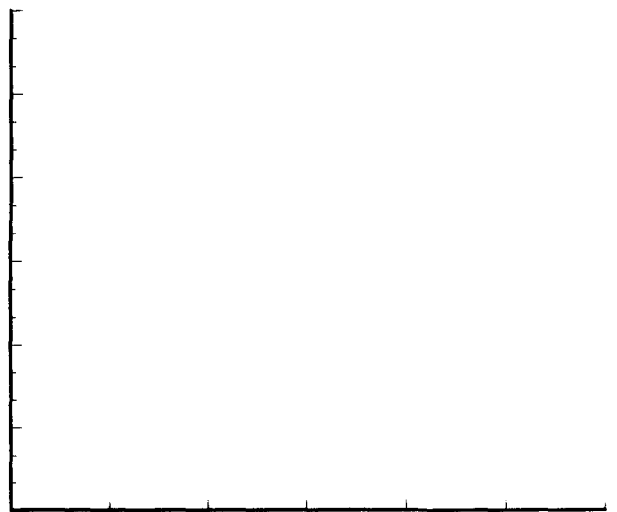
G. Relative frequency polygon



H. More-than cumulative frequency polygon



I. Less-than cumulative frequency polygon

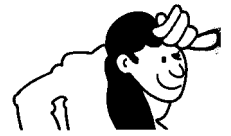


Chapter 3 Measuring Central Tendency of Ungrouped Data

I. Introduction

- Central tendency describes the middle of data. It represents a typical value.
- Measures of central tendency are called averages.
- The **arithmetic mean** is the most common average. It is used to measure grades, success in sports, business success, and many other interesting subjects.
- Population parameters are represented by Greek capital letters.
- Sample statistics are represented by Arabic lowercase letters.

Don't forget
to look ahead



II. The mean

A. The sample mean (\bar{x})

- Linda is interested in how many self-help videotapes she rented last year. If substantial, she will stock a larger variety of tapes. To make an estimate, she counted last week's self-help tape rentals and recorded the following sample data. The data is a sample because she only included part of last year's data.
- Daily self-help tape rentals were: 3, 7, 7, 4, 1, 8, 5.

$\bar{x} = \frac{\sum x}{n}$ where \bar{x} , read x bar, is the sample mean. x is the variable being measured.

Σ is the Greek capital letter sigma. It is the symbol for addition. n is the sample size.

$$\bar{x} = \frac{\sum x}{n} = \frac{3+7+7+4+1+8+5}{7} = \frac{35}{7} = 5$$

B. The population mean (μ)

- Had Linda used all of last year's data, this population mean formula would have been used.
- μ is the Greek capital letter for M and it is read Mu.
- N is the population size.

$$\mu = \frac{\sum x}{N}$$

C. A weighted mean (\bar{x}_w)

- When a data set has a number of duplicate values, a weighted mean is often calculated.
- Each variable occurring more than once is assigned a variable name consisting of capital x with a subscript and a weight (W) with a similar subscript.
- Linda's Video Showcase rents tapes for \$2, \$3, and \$4. The weighted mean of receipts per tape for a day of 36, 18, and 6 respective rentals is calculated as follows:

$$\bar{x}_w = \frac{W_1X_1 + W_2X_2 + W_3X_3 + \dots + W_nX_n}{W_1 + W_2 + W_3 + \dots + W_n} = \frac{\sum (W_x X_x)}{\sum w_x}$$

$$\bar{x}_w = \frac{(36)(\$2) + (18)(\$3) + (6)(\$4)}{36 + 18 + 6} = \frac{\$72 + \$54 + \$24}{60} = \frac{\$150}{60} = \$2.50$$

Note: W_1 refers to how often X_1 happens.

D. The sum of the deviations around a mean equals zero.

- $\Sigma(x - \mu) = 0$
- The mean of 1, 3, and 8 is 4.
- The sum of the deviations around the mean would be calculated as follows:

$$\begin{aligned} \Sigma(x - \mu) &= (1 - 4) + (3 - 4) + (8 - 4) \\ &= (-3) + (-1) + (4) = 0 \end{aligned}$$

- The primary disadvantage of using the mean as a measure of central tendency concerns it being severely affected by a few values at either extreme. Using the data at the top of this page as an example, the mean is small because a big snowstorm resulted in a day with only 1 rental.

III. The median

- A. The median is the middle number of data arranged into an array.
- B. The median as a measure of central tendency
 - 1. The median may be thought of as the geometric middle while the mean is the arithmetic middle.
 - 2. The geometric nature of the median results in it not being influenced by a few large numbers at either extreme.
- C. Determining the median
 - 1. Arrange the data into an array.
 - 2. Determine the median's position using this expression. $\frac{n}{2} + .5$
 - 3. Count this number of spaces from either extreme to find the median. An even number for n will result in the location being halfway between two numbers. Add the numbers and divide by 2 to determine the median.
- D. Example
 - 1. Linda Smith wants to calculate last week's median number of self-help rentals.
 - 2. Daily self-help rentals from page 10 were 3, 7, 7, 4, 1, 8, and 5.

Array: 1, 3, 4, 5, 7, 7, 8

$$\frac{n}{2} + .5 = \frac{7}{2} + .5 = 4 \rightarrow 5$$

The arrow means go to the array. Counting from either direction, the fourth number is 5.

IV. The mode

- A. The mode is the value occurring most often.
- B. It was 7 for self-help tape rentals.
- C. Some data sets have no modes while others have two (**bimodal**) or more (**multimodal**) modes.
- D. For many data sets, the mode is not a good representation of the data's middle value. As a result, it is the least used measure of central tendency. However, knowing the value that occurred most often is often of interest.

V. Measures of position

- A. These measures locate interesting points along data arranged into an array.
- B. The median is an example.
- C. **Quartiles** separate data into quarters.
 - 1. Q_1 separates the first and second quarters.
 - 2. Q_2 , the median, separates the second and third quarters.
 - 3. Q_3 separates the third and fourth quarters.

Quartile	Location	Finding the quartiles for the above data	Analysis
Q_1	$\frac{n}{4} + .5$	$\frac{7}{4} + .5 = 1.75 + .5 = 2.25 \rightarrow 3.25$	Note: 3.25 is .25 of the distance between 3, the second number, and 4, the third number.
Q_2	$\frac{n}{2} + .5$	$\frac{7}{2} + .5 = 3.5 + .5 = 4 \rightarrow 5$	This data is not symmetrical. It is a coincidence that the mean and median are equal.
Q_3	$\frac{3n}{4} + .5$	$\frac{21}{4} + .5 = 5.25 + .5 = 5.75 \rightarrow 7$	Note: 7 is .75 of the distance between 7 and 7.

D. Interquartile range

- 1. The interquartile range is the difference between Q_3 and Q_1 .

$$Q_3 - Q_1 = 7 - 3.25 = 3.75$$

- E. **Deciles** separate data into tenths. The 3rd decile would be calculated as follows:

$$\frac{xn}{10} + .5 = \frac{3(7)}{10} + .5 = 2.6 \rightarrow 3.6$$

F. Percentiles

- 1. Percentiles separate data into 100 parts.
- 2. Let x equal the percentile of interest.
- 3. The location of the x percentile would be stated as follows:
- 4. The 90th percentile of daily self-help rentals would be

$$\frac{xn}{100} + .5 = \frac{90(7)}{100} + .5 = \frac{630}{100} + .5 = 6.8 \rightarrow 7.8$$

$$\frac{xn}{100} + .5$$

Note: Computer software may use different formulas to locate the position of data. As a result, their answers for measures of position may differ slightly from these answers.

Practice Set 3 Measuring Central Tendency of Ungrouped Data

- I. Darin Jones wants to know more about the sales of Walkman CD recorders/players described on page 6. Calculate the sample mean using this Walkman sales data from the last Practice Set. State the formula for the population mean.

Array of daily Walkman sales: 8, 12, 14, 15, 16, 16, 17, 17, 21, 22, 29

A. Sample mean

Having trouble with these problems?
Please look back 2 pages to about the
same page location in Quick Notes to
see how Linda solved a similar problem.

B. Population mean formula

- II. Darin sells three different Walkman CD recorders; one for \$149, one for \$159, and a third for \$169. Of the 187 machines sold during this eleven-day period; 43 were the least expensive, 90 were moderately priced, and 54 were the expensive model. Calculate the weighted mean sales price for these machines.



- III. Using the data from question I, prove that the sum of the deviations from a mean is _____.

IV. The median number of Walkman units sold is _____ .

V. The mode for this data is _____ .

VI. This data can be described as _____ .

VII. Calculate the following measures of position. Those using computer software should use a less-than cumulative relative frequency distribution to answer these questions.

A. Q_1 _____

B. Q_3 _____

C. Interquartile range _____

D. 6th decile _____

E. 95th percentile _____

Quick Questions 3 Measuring Central Tendency of Ungrouped Data

I. Write the number of the appropriate formula next to the item it describes.

A. Sample mean _____

B. Population mean _____

C. Location of the median _____

D. Location of Q_1 _____

E. Weighted mean _____

F. Location of Q_3 _____

1. $\frac{3n}{4} + .5$	4. $\frac{n}{2} + .5$
2. $\frac{\sum X}{N}$	5. $\frac{\sum (W_x X_x)}{\sum w_x}$
3. $\frac{\sum x}{n}$	6. $\frac{n}{4} + .5$

II. List and calculate the 3 measures of central tendency.

Data: 5, 7, 3, 8, 6, 10, 9, 8

A. _____

B. _____

C. _____

III. What is the primary disadvantage of the mean as a measure of central tendency?

IV. Using this data, prove that the sum of the deviations around an arithmetic mean is _____ .

Data: 3, 7, 5

V. Calculate a weighted mean of parking tickets costing \$25, \$35, and \$45 with corresponding weights of 10, 20, and 10 respectively. Why must the answer be \$35?

VI. Calculate the following for the question II data.

A. Q_1

B. Q_3

C. Interquartile range

D. 2nd decile

E. 85th percentile

Chapter 4 Measuring Dispersion of Ungrouped Data

I. Introduction

- A. Dispersion refers to the spread of data, its variability.
- B. Dispersion is important because it determines the reliability of central tendency measurements.
- C. Comparing the dispersion of different data sets may be revealing. Two students might have the same grade point average with one having all B's and the other having half A's and half C's.
- D. This page will explore population parameters. Where sample statistic formulas differ, calculations will be done on the next page.
- E. The sample data for self-help rentals presented in chapter 3 will be used here as population data.

3, 7, 7, 4, 1, 8, 5 and $\mu = 5$

II. Range

- A. The range is the highest value (H) minus the lowest value (L).
- B. $H - L = 8 - 1 = 7$
- C. While easy to calculate, the range is severely affected by unusual circumstances. In this case, a snowstorm caused Linda to close early limiting that day's rentals to one unit.

III. Population average deviation (AD)

- A. The **average deviation** is the mean of the absolute values of the deviations from the mean.

$$AD = \frac{\sum |x - \mu|}{N} = \frac{14}{7} = 2$$

Note: N is population size.

- B. Using the absolute value of the deviations is necessary because the sum of the deviations is zero.
- C. The average deviation is a quick way to measure dispersion. The soon to be explained variance and standard deviation are more valuable measures.

Self-Help Rentals			
x	μ	$x - \mu$	$ x - \mu $
3	5	-2	2
7	5	2	2
7	5	2	2
4	5	-1	1
1	5	-4	4
8	5	3	3
5	5	0	0
Totals		0	14

IV. Population variance (σ^2) and standard deviation (σ)

- A. The variance solves the problem of the sum of the variations from a mean being zero by squaring the differences.
- B. The **variance** is the average of the squared deviations of the data from their mean.
- C. The resulting measure is similar to the averaged deviation although it is larger because the variation was squared.
- D. This problem is solved with the **standard deviation** which is the square root of the variance.
- E. The **population variance**

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

$$= \frac{38}{7} = 5.4$$

Alternative Formula

$$\sigma^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N} \right)^2$$

$$= \frac{213}{7} - \left(\frac{35}{7} \right)^2$$

$$= 30.4 - 25 = 5.4$$

Self-Help Rentals				
x	μ	$x - \mu$	$(x - \mu)^2$	x^2
3	5	-2	4	9
7	5	2	4	49
7	5	2	4	49
4	5	-1	1	16
1	5	-4	16	1
8	5	3	9	64
5	5	0	0	25
Total = 35			Total = 38	Total = 213

F. Population standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{5.4} = 2.3$$

V. The sample variance (s^2) and standard deviation (s)

A. Sample variance

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

Alternate formula

$$S^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

B. Note that N has been replaced by n-1 in the denominator.

C. In chapter 11, the sample variance will be used to predict the population variance. If one is not subtracted from n when calculating the sample standard deviation, it will be bias (not representative of σ).

D. Sample standard deviation (Assume data on page 16 is sample data.)

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

or

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{213 - \frac{(35)^2}{7}}{7-1}} = \sqrt{\frac{213-175}{6}} = 2.5$$

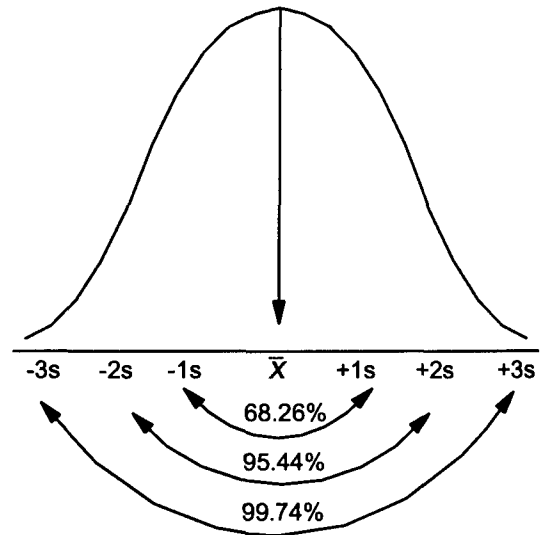
VI. Using the standard deviation as a measure of variability

A. The empirical rule is used for normal, bell-shaped data.

- For symmetrical or bell-shaped data, 68.26% of the item will be within one standard deviation of the mean, 95.44% will be within two standard deviations of the mean, and 99.74% will be within three standard deviations of the mean. If $\mu = 500$ and $\sigma = 100$, then 95.44% of the population will be between 300 and 700.

$$500 \pm 2(100)$$

$$500 \pm 200 \rightarrow 300 \leftrightarrow 700$$



- Students would like a small standard deviation around a test mean of 95 so everyone receives a grade of A.
- B. Chebyshev's rule is used for nonsymmetrical distributions.
- Russian mathematician P. Chebyshev developed a method to estimate the minimum proportion of items that are within a designated number of standard deviations from the mean for nonsymmetrical distributions with means greater than 1. As with the empirical rule, the estimate works for both samples and populations.
 - The proportion of items within K standard deviations of the mean is at least 1 minus 1 over K squared provided K is a constant greater than 1.
 - The proportion of the data falling within 2 standard deviations of a mean is calculated as follows:

$$1 - \frac{1}{K^2}$$

$$1 - \frac{1}{K^2} = 1 - \frac{1}{(2)^2} = 1 - \frac{1}{4} = \frac{3}{4}$$

VII. Coefficient of variation (CV)

- Comparing the variability of data sets of differing magnitudes is accomplished using the coefficient of variation.
- Department A with \$40 million in sales will have a much larger standard deviation than Department B which has only \$3 million in sales. Suppose Department A's σ was \$4 million and Department B's σ was \$400,000.
- The coefficient of variation, which expresses the standard deviation as a percent of the mean, reveals which department has the largest relative sales variability. $CV = \frac{\sigma}{\bar{x}}(100)$ for sample data.

For Department A

$$C.V. = \frac{\sigma}{\bar{\mu}}(100) = \frac{\$4,000,000}{\$40,000,000}(100) = 10\%$$

For Department B

$$C.V. = \frac{\sigma}{\bar{\mu}}(100) = \frac{\$400,000}{\$3,000,000}(100) = 13.3\%$$

Note: Department A had less sales dollar variability even though it had a larger standard deviation.

Practice Set 4

Measuring Dispersion of Ungrouped Data

- I. Darin is concerned about Walkman sales variability. First calculate the range for Walkman sales and then the average deviation, the standard deviation, and the variance.

Array of daily Walkman sales: 8, 12, 14, 15, 16, 16, 17, 17, 21, 22, 29

Sample mean: 17

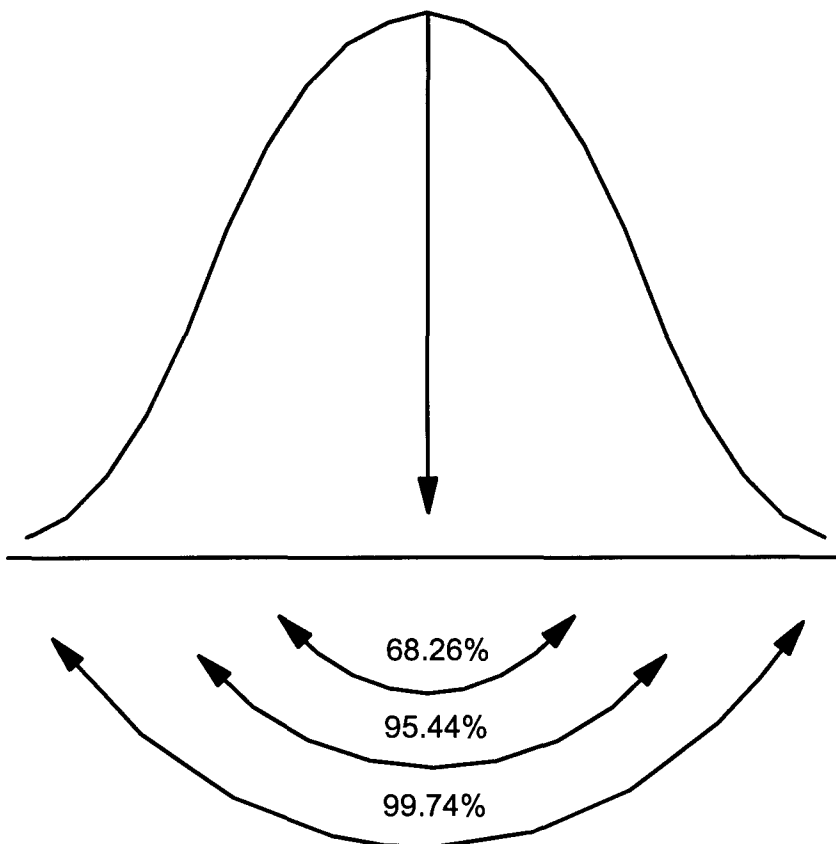
A. Range

B. Sample average deviation

C. Sample variance

D. Sample standard deviation

- II. Label this graph depicting the empirical rule.



III. Last year's mean weekly Walkman sales were 16 and the standard deviation was 4. Use the empirical rule to determine a range for Walkman sales for one, two, and three sample standard deviations from the mean.

A. One Standard Deviation

B. Two Standard Deviations

C. Three Standard Deviations

IV. Use Chebyshev's rule to determine a range for Walkman sales being within two sample standard deviations of the mean.

V. Darin read in a trade publication that the average Walkman sales and standard deviation for a store his size and type are 18 and 3 respectively. Using the sample data from page 18, are Darin's Walkman sales more or less variable than those of his industry? Use the standard deviation calculated in problem I.

Quick Questions 4 Measuring Dispersion of Ungrouped Data

I. Place the number of the appropriate formula next to the parameter or statistic it describes.

- A. Population average deviation _____
- B. Population variance _____
- C. Population standard deviation _____
- D. Alternative population variance _____
- E. Alternative population standard deviation _____
- F. Chebyshev's rule _____
- G. Sample variance _____
- H. Sample standard deviation _____
- I. Alternative sample variance _____
- J. Alternative sample standard deviation _____

1. $\frac{\sum x - \mu }{N}$	6. $1 - \frac{1}{k^2}$
2. $\frac{\sum (x - \mu)^2}{N}$	7. $\frac{\sum (x - \bar{x})^2}{n - 1}$
3. $\sqrt{\frac{\sum (x - \mu)^2}{N}}$	8. $\sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$
4. $\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2$	9. $\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}$
5. $\sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$	10. $\sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}}$

II. Calculate the following statistics using this sample data.

Data: 5, 7, 3, 8, 6, 10, 9, 8

$\bar{x} = 7$

- A. Variance (use alternative formula)

B. Standard deviation

C. Average deviation

III. Use Chebyshev's rule to calculate the percentage of question II outcomes that will be within 3 standard deviations of the mean. Was this prediction correct?

IV. A data set of grades is normally distributed and has a mean of 84 and a standard deviation of 4. Calculate a range of grades that will include the middle 95.44% of the data set.

Chapter 5 Measuring Central Tendency of Grouped Data

I. Introduction

- A. When actual data is unavailable or of an unmanageable volume, it may be necessary to determine parameters and statistics using a frequency distribution.
- B. Important symbols:



Symbol	Definition	Symbol	Definition
\bar{X}	the sample mean	fx	frequency times the class midpoint
X	the midpoint of a class	$\sum fx$	summation of fx
f	the frequency of a class	n	total frequency



II. The grouped sample mean

$$\bar{X} = \frac{\sum fx}{n}$$

- A. Linda needs to estimate this year's tape rentals for a bank loan application. She will use the page 4 tape rentals summarized with a frequency distribution to estimate average daily rentals for the year.
- B. Linda must calculate each class midpoint and then multiply it by the class frequency.

The midpoint formula is

$$X = \frac{X_1 + X_2}{2}$$

For class one

$$X = \frac{50 + 59}{2} = 54.5$$

$$\bar{X} = \frac{\sum fx}{n} = \frac{1,117.5}{15} = 74.5$$

Stated Class Limits	Frequency (f)	x	fx
50 - 59	2.0	54.5	109.0
60 - 69	3.0	64.5	193.5
70 - 79	5.0	74.5	372.5
80 - 89	3.0	84.5	253.5
90 - 99	<u>2.0</u>	94.5	<u>189.0</u>
Totals	n = 15.0		$\sum fx = 1,117.5$

Estimated yearly tape rentals would be $(52)(7)(74.5) = 27,118$.

III. The grouped median

- A. The median is the middle number.

$$L + \frac{\frac{n}{2} - CF_b}{f}(i)$$

Symbols	Definitions
L	lower real limit of the median's class
CF_B	cumulative frequency before the median's frequency
i	class interval (width)

- B. Use $\frac{n}{2}$ to determine the location of the middle frequency.

$$\frac{n}{2} = \frac{15}{2} = 7.5$$

- C. Beginning at the top of the frequency distribution and counting down the frequency column reveals that the 7.5 frequency is located in the third class from the top (or bottom for that matter). The lower real limit of the median's class is 69.5 and the class is 10 wide.

Class Limits	Frequency
50 - 59	2
60 - 69	3
70 - 79	5
80 - 89	3
90 - 99	<u>2</u>
	15

lower limit →

→ Used 2 here

→ Used 3 here

→ Need 2.5 from here to get to 7.5

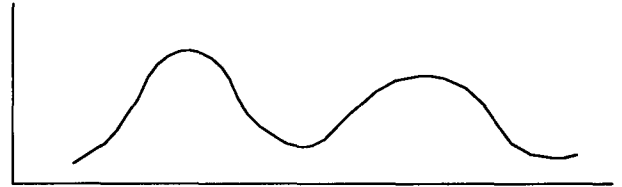
→ Out of 5

$$\begin{aligned}
 &L + \frac{\frac{n}{2} - CF_b}{f}(i) \\
 &= 69.5 + \frac{\frac{15}{2} - 5}{5}(10) \\
 &= 69.5 + 5 = 74.5
 \end{aligned}$$

IV. The grouped mode

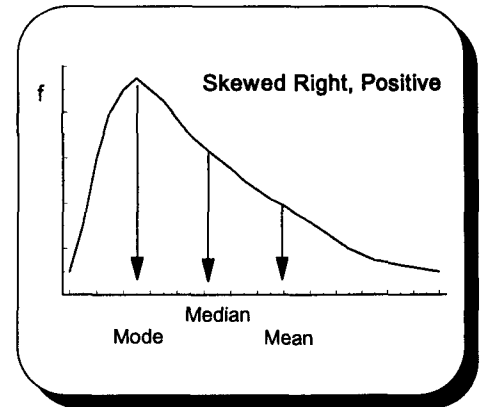
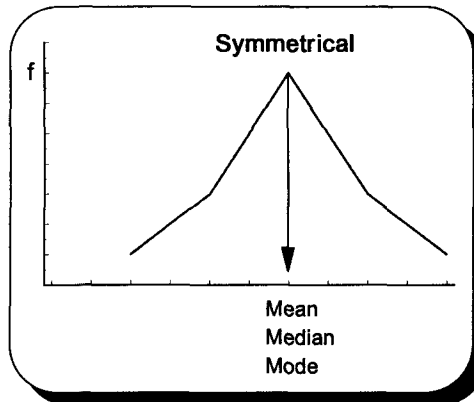
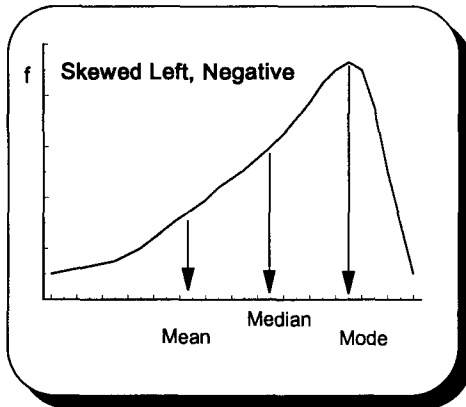
- The grouped mode is the midpoint of the class with the highest frequency.
- For the page 22 distribution, the mode is 74.5.
- The mean, median, and mode are all 74.5 because this distribution is symmetrical (normal).
- Frequency distributions with two peaks are said to be **bimodal**. More than two is **multimodal**.

A Bimodal Distribution



V. Nonsymmetrical distributions

- Frequency distributions that are not symmetrical are said to be **skewed**.
 - With negatively skewed data, the mean is the smallest of the three measures of central tendency.
 - With positively skewed data, the mean is the largest of the three measures of central tendency.



B. Measuring skewness

- The degree to which a distribution (curve) is skewed is measured by **Pearson's coefficient of skewness**.
- The measure applies to both sample and population data.
- When data is positively skewed, the mean is larger than the median, and the measure is positive.
- When data is negatively skewed, the mean is smaller than the median, and the measure is negative.
- An increase in skewness increases the difference between the mean and the median. This causes an increase in the coefficient of skewness.
- Normal distributions have a zero coefficient of skewness.

$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

$$= \frac{3(74.5 - 74.5)}{12.5} = 0$$

Note: The standard deviation of 12.5 was taken from page 28.

- For highly skewed distributions, the median measures central tendency better than the mean because it is not as influenced by extreme values.
 - Income is skewed right (positive) by a few people making a large amount of money.
 - Comparing the mean and median salaries of these unionized workers yields interesting results.

\$14,000
 \$15,000
 → **\$16,000**
 \$17,000
 \$28,000

$$\mu = \frac{\sum x}{N} = \frac{\$90,000}{5} = \$18,000 \quad \text{The median is } \$16,000.$$

- In situations like this, a union would use the median salary to make the average look low. Management would use the mean salary to make the average look high.

Note: Suppose the top salary of \$28,000 was increased to \$48,000. The mean would increase from \$18,000 to \$22,000, but the median would remain unchanged.

Practice Set 5 Measuring Central Tendency of Grouped Data

- I. Label the top row of this Walkman sales data chart and calculate the following measures of central tendency.

Array of Walkman sales from page 6

8, 12, 14, 15, 16, 16, 17, 17, 21, 22, 29

Darin's Music Emporium Walkman Sales Data			
5 - 9	1	7	7
10 - 14	2	12	24
15 - 19	5	17	85
20 - 24	2	22	44
25 - 29	<u>1</u>	27	<u>27</u>
	11		187

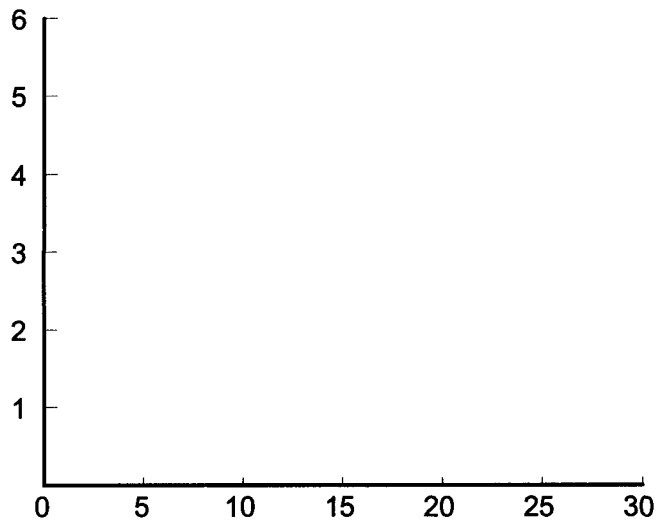
A. Grouped mean

B. Grouped median

C. Grouped mode

II. Do your answers to question 1 differ from those calculated on pages 12 and 13? Is the difference large? Could the difference be large?

III. Draw a frequency polygon of page 24 data and locate the mean, median, and mode.



IV. Using the mean of 17 and the median of 17 calculated on page 24, and the sample standard deviation of 5.5 to be calculated on page 30, calculate Pearson's coefficient of skewness.

Quick Questions 5 Measuring Central Tendency of Grouped Data

I. Place the number of the appropriate formula next to the item it describes.

- A. Grouped sample mean _____
- B. Location of the grouped median _____
- C. Grouped median _____
- D. Class midpoint _____

1.	$\frac{X_1+X_2}{2}$
2.	$\frac{\sum fx}{n}$
3.	$L + \frac{\frac{n}{2}-CF_b}{f}(i)$
4.	$\frac{n}{2}$

II. Fill in the middle column and then use this frequency distribution to answer the following questions. Information needed to do this problem was presented on page 4.

Stated Class Limits	x	Frequency (f)
10 - 14		2
15 - 19		3
20 - 24		5

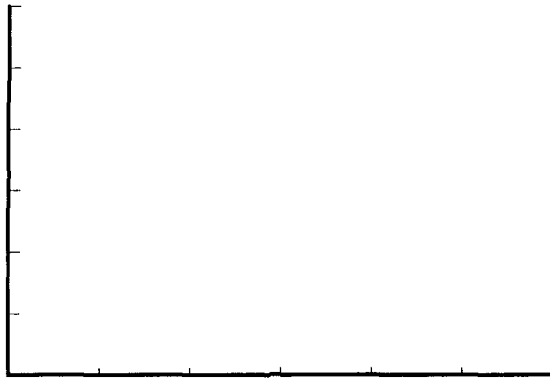
- A. The first class has real class limits of _____ and _____.
- B. The first class has stated class limits of _____ and _____.
- C. The class width is _____.
- D. The midpoint of the first class is _____.
- E. The range using real class limits is from _____ to _____.

III. Calculate the following statistics using this frequency distribution of exam grades.

Stated Class Limits	x	Frequency (f)	
50 - 59	54.5	1	
60 - 69	64.5	3	
70 - 79	74.5	5	
80 - 89	84.5	7	
90 - 99	94.5	2	

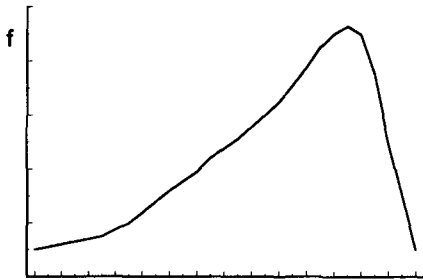
- A. Mean
- B. Median
- C. Mode

IV. Draw a frequency polygon for the question III data and locate the mean, median, and mode.

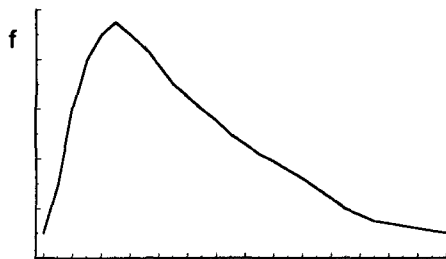


V. Show the approximate location of the mean, median, and mode on the x-axis of these frequency distributions.

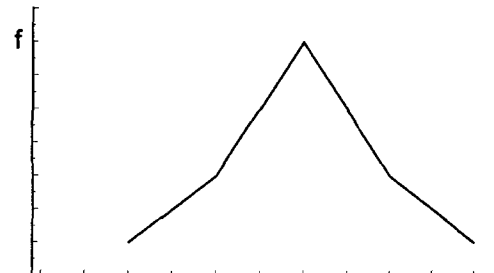
Curve #1



Curve #2



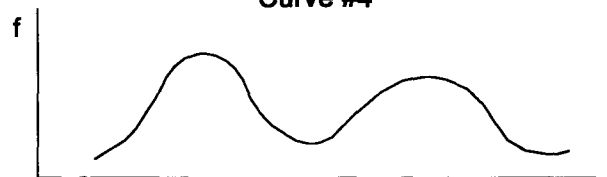
Curve #3



VI. Answer these questions using Curves #1 to #4.

- A. Curve #1 is skewed to the _____.
- B. Curve #3 is not skewed and is said to be _____.
- C. Curve #4 is _____.

Curve #4



- D. You represent a union and data indicates member salaries are distributed similar to Curve #1. Which measure of central tendency would you use when talking to local media about current wage negotiations? _____
- E. Using the mean of 77.8 and the median of 79.5 from page 26, calculate Pearson's coefficient of skewness. The standard deviation, which will be calculated on page 32, is 14.2.

Chapter 6 Measuring Dispersion of Grouped Data

I. Introduction

A. Daily tape rentals summarized in chapter 2 will be analyzed.

B.

Daily Rentals Beginning 1/2/98					
Stated Class Limits	Frequency (f)	x	fx	x ²	fx ²
50 - 59	2.00	54.50	109.00	2,970.25	5,940.50
60 - 69	3.00	64.50	193.50	4,160.25	12,480.75
70 - 79	5.00	74.50	372.50	5,550.25	27,751.25
80 - 89	3.00	84.50	253.50	7,140.25	21,420.75
90 - 99	2.00	94.50	189.00	8,930.25	17,860.50
Totals	n = 15.00		1,117.50		85,453.75

II. Range

A. Range = H - L

B. Use the real class limits for H and L

$$H - L = 99.5 - 49.5 = 50$$

III. Sample standard deviation

A. Ungrouped data

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

B. Grouped data

$$\begin{aligned}
 S &= \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}} \\
 &= \sqrt{\frac{85,453.75 - \frac{(1,117.5)^2}{15}}{15-1}} = \sqrt{\frac{85,453.75 - 83,253.75}{14}} \\
 &= \sqrt{\frac{2,200}{14}} = \sqrt{157.143} = 12.5
 \end{aligned}$$

IV. Variance

$$S^2 = (S)^2 = (12.5)^2 = 156.25 \approx 156.3$$

V. Measures of position

A. Measures of position locate interesting points along data arranged into an array.

B. **Quartiles** separate data into quarters.

1. Q_1 separates the first and second quarters.
2. Q_2 , the median, separates the second and third quarters.
3. Q_3 separates the third and fourth quarters.

$$Q_1 = L + \frac{\frac{n}{4} - CF_b}{f}(i)$$

$$Q_2 = L + \frac{\frac{n}{2} - CF_b}{f}(i)$$

$$Q_3 = L + \frac{\frac{3n}{4} - CF_b}{f}(i)$$

Symbols	Definitions
L	lower real limit of the measure's class
CF_b	cumulative frequency before the measure's frequency
i	class interval (width)

4. The first and third quartiles

- a. The location of the median is $\frac{n}{2}$, the first quartile's location is $\frac{n}{4}$, and the third quartile's location is $\frac{3n}{4}$.
- b. Sample size divided by four equals $15/4 = 3.75$. Counting down the frequency distribution on the previous page reveals that the first quartile is near the middle of the second class.
- c. $\frac{3n}{4} = \frac{3 \times 15}{4} = \frac{45}{4} = 11.25$ Counting down reveals the third quartile is in the fourth class.

$$Q_1 = L + \frac{\frac{n}{4} - CF_b}{f}(i)$$

$$= 59.5 + \frac{\frac{15}{4} - 2}{3}(10)$$

$$= 59.5 + \frac{3.75 - 2}{3}(10)$$

$$Q_1 = 59.5 + 5.8 = 65.3$$

From page 23

 $Q_2 = 74.5$

$$Q_3 = L + \frac{\frac{3n}{4} - CF_b}{f}(i)$$

$$= 79.5 + \frac{\frac{45}{4} - 10}{3}(10)$$

$$= 79.5 + \frac{11.25 - 10}{3}(10)$$

$$Q_3 = 79.5 + 4.2 = 83.7$$

50.0 65.3 74.5 83.7 99.0

first quarter second quarter third quarter fourth quarter

C. Interquartile range

- 1. The interquartile range is the difference between Q_3 and Q_1 .
- 2. $Q_3 - Q_1 = 83.7 - 65.3 = 18.4$

D. Percentiles

- 1. Percentiles separate data into 100 parts.
- 2. Let x equal the percentile of interest.
- 3. Here, the 90th percentile of daily rentals beginning 1/2/98 is of interest.
- 4. The location of the 90th percentile is found using this expression.

$$\frac{xn}{100}$$

$$\frac{xn}{100} = \frac{90(15)}{100} = 13.5$$

Counting down the frequencies reveals the 90th percentile is in the bottom class.

$$P_x = L + \frac{\frac{xn}{100} - CF_b}{f}(i)$$

$$P_x = L + \frac{\frac{xn}{100} - CF_b}{f}(i)$$

$$P_{90} = 89.5 + \frac{\frac{90(15)}{100} - 13}{2}(10)$$

$$= 89.5 + \frac{13.5 - 13}{2}(10)$$

$$= 92.0$$

VI. Kurtosis describes the peak of a curve.

A **platykurtic** curve is flat, items are evenly distributed.

A **mesokurtic** curve is not flat or peaked.

A **leptokurtic** curve is thin, items are concentrated in the middle.

Practice Set 6 Measuring Dispersion of Grouped Data

I. Label this chart of the page 24 frequency distribution and calculate the following measurements.

Darin's Music Emporium Walkman Sales Data					
5 - 9	1	7	7	49	49
10 - 14	2	12	24	144	288
15 - 19	5	17	85	289	1,445
20 - 24	2	22	44	484	968
25 - 29	1	27	<u>27</u>	729	<u>729</u>
	11		187		3,479

A. Range

B. Sample variance

C. Sample standard deviation

D. Quartiles

First

Second

Third

E. Interquartile range

F. 80th percentile

II. Locate the three quartile measures calculated above on this number line. Label the four quartiles.



Quick Questions 6 Measuring Dispersion of Grouped Data

I. Place the number of the appropriate formula next to the item it describes.

- A. Grouped sample standard deviation _____
- B. First quartile _____
- C. Median (second quartile) _____
- D. Third quartile _____
- E. Interquartile range _____
- F. Percentile _____

1. $L + \frac{\frac{n}{2} - CF_b}{f}(i)$	4. $L + \frac{\frac{3n}{4} - CF_b}{f}(i)$
2. $\sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}}$	5. $L + \frac{\frac{xn}{100} - CF_b}{f}(i)$
3. $L + \frac{\frac{n}{4} - CF_b}{f}(i)$	6. $Q_3 - Q_1$

II. Complete the first row of this table and calculate the following measurements.

Stated Class Limits	Frequency (f)				
40 - 49	1	44.5	44.5	1,980.25	1,980.25
50 - 59	2	54.5	109.0	2,970.25	5,940.50
60 - 69	3	64.5	193.5	4,160.25	12,480.75
70 - 79	5	74.5	372.5	5,550.25	27,751.25
80 - 89	3	84.5	253.5	7,140.25	21,420.75
90 - 99	2	94.5	189.0	8,930.25	17,860.50
Totals	16	417.0	1,162.0	30,731.50	87,434.00

A. Range

B. Sample variance

C. Sample standard deviation

D. Calculate the following:

First Quartile

Median

Third Quartile

E. Locate the three quartiles and the four quarters on this figure.



F. Calculate the interquartile range.

G. Calculate the 95th percentile.

It's time to review for the Part I Quiz beginning on page 35. I know it's a pain in the neck, but Fred will be upset if you don't do well. Begin with the formula review on page 34. Then look at the relevant sections of pages 162, 164, and 166.



Descriptive Statistics Formula Review

Ungrouped Measures

1. Population mean $\mu = \frac{\sum x}{N}$

2. Sample mean $\bar{X} = \frac{\sum x}{n}$

3. First quartile $\frac{n}{4} + .5$

4. Median $\frac{n}{2} + .5$

5. Third quartile $\frac{3n}{4} + .5$

6. Interquartile range $Q_3 - Q_1$

7. x percentiles $\frac{xn}{100} + .5$

8. x deciles $\frac{xn}{10} + .5$

9. Weighted mean $\frac{\sum (W_x X_x)}{\sum w_x}$

10. Average deviation $\frac{\sum |x - \mu|}{N}$

11. Population standard deviation $\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$

12. Sample standard deviation $S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$

13. Population variance $\sigma^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2$

14. Sample variance $S^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$

15. Coefficient of variation $C.V. = \frac{\sigma}{\mu}(100)$

16. Range $H - L$

17. Chebyshev's rule $1 - \frac{1}{k^2}$

18. Pearson's coefficient of skewness $\frac{3(\bar{x} - md.)}{S}$

Grouped Measures

19. Approximate class width $\frac{\text{range}}{\# \text{ of classes}}$

20. Class midpoint $\frac{X_1 + X_2}{2}$

21. Population mean $\mu = \frac{\sum fx}{N}$

22. Sample mean $\bar{X} = \frac{\sum fx}{n}$

23. Location of the median $\frac{n}{2}$

24. Median $L + \frac{\frac{n}{2} - CF_b}{f}(i)$

25. Population standard deviation $\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}}$

26. Sample standard deviation $S = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}}$

27. Relative frequency $\frac{\text{class frequency}}{\text{total frequencies}}$

28. Sample variance $S^2 = \frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}$

Descriptive Statistics Test

I. Place the number of the appropriate definition next to the item it describes.

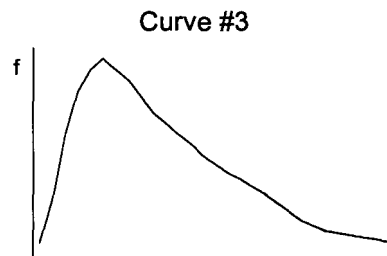
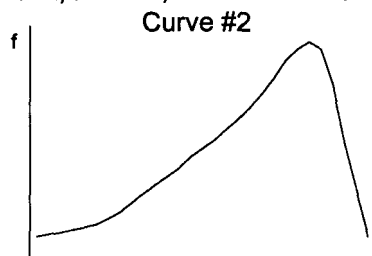
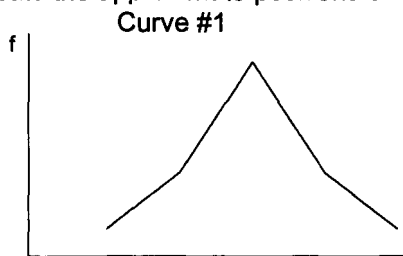
- | | |
|---------------------------------|---|
| A. Statistic _____ | 1. A place for every outcome |
| B. Parameter _____ | 2. Do not contain the same outcome |
| C. All-inclusive _____ | 3. The use of sample statistics to draw conclusions concerning the population |
| D. Discrete _____ | 4. A numerical characteristic of a sample |
| E. Mutually exclusive _____ | 5. Only finite values can exist on the x-axis |
| F. Zero _____ | 6. Published by the original collector |
| G. Continuous _____ | 7. Severely affected by a few extreme values |
| H. Inferential statistics _____ | 8. Measurement may assume any value associated with an uninterrupted scale |
| I. Arithmetic mean _____ | 9. A numerical characteristic of a population |
| J. Primary data _____ | 10. Sum of the deviations around a mean |

II. Answer questions A - E using the information in this chart.

Stated Class Limits	x	Frequency (f)
10 - 24	17	2
25 - 39	32	3
40 - 54	47	5

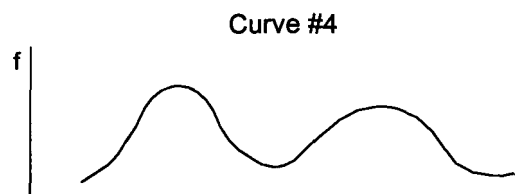
- A. The second class has real class limits of _____ and _____.
- B. The first class has stated class limits of _____ and _____.
- C. The class width is _____.
- D. The midpoint of the third class is _____.
- E. The range using real class limits is from _____ to _____.

III. Locate the approximate positions of the mean, median, and mode on these graphs.



IV. Answer questions A - E using Curves #1 to #4.

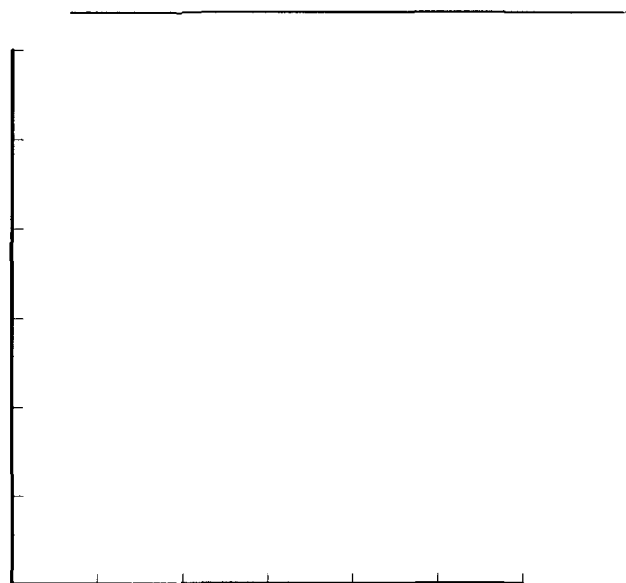
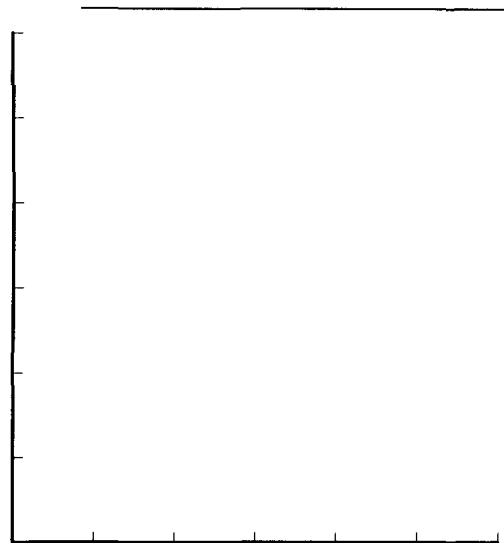
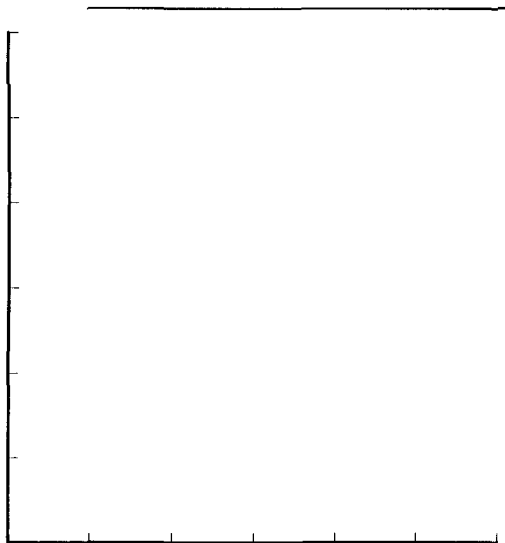
- A. Curve #1 is not skewed and is said to be _____.
- B. Curve #2 is skewed to the _____.
- C. Curve #3 is skewed to the _____.
- D. Curve #4 is _____.
- E. A curve with more than two peaks is _____.



- V. Using the following frequency distribution, construct and completely label a frequency polygon, histogram, and less-than ogive.

Linda's Video Showcase Daily Rental Figures	
Stated Class Limits	Frequency (f)
50 - 59	2
60 - 69	4
70 - 79	5
80 - 89	1

For People Using Statistics Software	
Data Set:	62, 66, 74, 58, 78, 71, 64, 84, 76, 53, 68, 75



VI. Use this sample data when calculating the following statistics. Those not using statistics software may want to use the page 39 formulas.

Data: 4, 6, 3, 7, 6, 8, 17, 5

A. Mean

B. Median

C. Mode

D. Variance

E. Standard deviation

F. Use Chebyshev's rule to calculate the minimum proportion of items that will be within 3 standard deviations of the mean.

G. T F Chebyshev's rule only applies to normally distributed data. (true or false)

H. Calculate Pearson's coefficient of skewness.

VII. Label this chart. Calculate the following sample statistics being sure to state the symbol and formula for each measure. Formulas are given on page 39. **This problem is only for people not using statistics software.**

Stated Class Limits	Frequency (f)				
40 - 49	1	44.5	44.5	1,980.25	1,980.25
50 - 59	2	54.5	109.0	2,970.25	5,940.50
60 - 69	3	64.5	193.5	4,160.25	12,480.75
70 - 79	5	74.5	372.5	5,550.25	27,751.25
80 - 89	5	84.5	422.5	7,140.25	35,701.25
90 - 99	2	94.5	189.0	8,930.25	17,860.50
Totals	18	417.0	1,331.0	30,731.50	101,714.50

A. Standard deviation

B. Variance

C. Median

D. 85th percentile

VIII. Place the number of each formula next to the appropriate description of its function.

Ungrouped Measures

1. Population mean _____
2. Sample mean _____
3. Median _____
4. First quartile _____
5. Third quartile _____
6. x percentile _____
7. Interquartile range _____
8. Population average deviation _____
9. Population standard deviation _____
10. Sample standard deviation _____
11. Population variance _____
12. Sample variance _____
13. Chebyshev's rule _____
14. Coefficient of variation _____
15. Weighted mean _____
16. Pearson's coefficient of skewness _____

Ungrouped Formulas

1. $\frac{\sum x}{n}$	2. $\frac{n}{2} + .5$
3. $\frac{\sum (W_x X_x)}{\sum W_x}$	4. $\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2$
5. $\sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2}$	6. $\frac{\sum x - \mu }{N}$
7. $\frac{\sum x}{N}$	8. $\frac{3n}{4} + .5$
9. $Q_3 - Q_1$	10. $\frac{\sigma}{\mu}(100)$
11. $\frac{xn}{100} + .5$	12. $1 - \frac{1}{k^2}$
13. $\frac{n}{4} + .5$	14. $\sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$
15. $\frac{3(\bar{x} - Md.)}{S}$	16. $\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$

Grouped Measures

1. Approximate class width _____
2. Class midpoint _____
3. Population mean _____
4. Sample mean _____
5. Location of the median _____
6. Median _____
7. Range _____
8. Sample standard deviation _____
9. Sample variance _____
10. Relative frequency _____

Grouped Formulas

1. $\frac{n}{2}$	2. $\frac{\text{class frequency}}{\text{total frequencies}}$
3. $\frac{\sum fx}{N}$	4. $\frac{\text{range}}{\# \text{ of classes}}$
5. $L + \frac{\frac{n}{2} - CF_b}{f}(i)$	6. $\frac{X_1 + X_2}{2}$
7. $\sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}}$	8. $\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}$
9. $\frac{\sum fx}{n}$	10. $H - L$

See page T 35 of Appendix III for complete solutions to this test.

Chapter 7 Understanding Probability

I. Introduction

- A. **Probability**, the likelihood of something happening, deals with uncertainty.
- B. Probability is the basis for inferential statistics.
- C. **Inferential statistics** involves estimating population parameters using sample statistics.

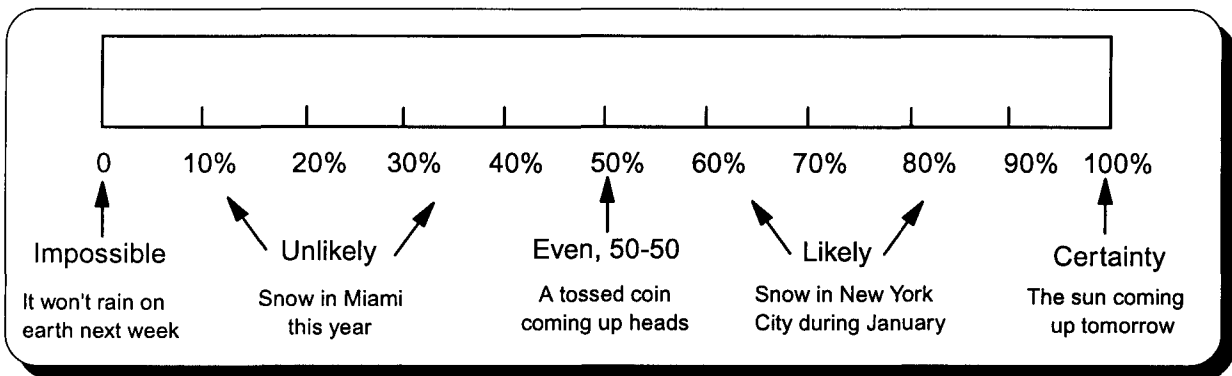
II. Basic data

- A. Linda Smith wants to understand the relationship between monthly advertising expenditures and monthly sales revenue. A recent study (experiment) revealed the following monthly data in thousands of dollars.

Advertising (000)	5	2	7	6	10	4	6	5	3	8
Sales (000)	50	25	80	50	90	30	60	60	40	80

III. Understanding probability

- A. The data for measuring probability comes from an experiment.
- B. An **experiment** is a repeatable process resulting in measurements (collecting this advertising and sales data).
- C. An **outcome** is a measurement from an experiment (a month's sales).
- D. An **event** is one or more outcomes (10 months of sales).
- E. A **simple event** cannot be divided (a month's sales).
- F. A **compound event** is a collection of simple events (10 months of sales).
- G. Events that do not share common outcomes are **mutually exclusive** (total sales and total advertising).
- H. Events that contain all the outcomes of an experiment are **all-inclusive (collectively exhaustive)**.
- I. Probability may be expressed as a fraction, decimal, or as a percentage.
- J. A **sample space** contains all the outcomes of an experiment.
- K.



IV. Types of probability

A. Classical probability

1. An experiment is not required to determine an outcome (rate of occurrence) because it is known. For example, the probability of getting a head when flipping a fair coin is known to be one-half.
2. Each simple event has an equal chance of happening.
3. The probability of event A; where $P(A)$ is the probability of event A, A is the number of times A occurred, and N is the total number of possible outcomes, is represented by this formula.

$$P(A) = \frac{\text{number of times A occurs}}{\text{total number of possible outcomes}} = \frac{A}{N}$$

4. For example, this is the probability of drawing a queen out of a 52-card deck containing 4 queens.

$$P(Q) = \frac{Q}{N} = \frac{4}{52} = \frac{1}{13}$$

B. Relative probability

1. Relative probability requires an experiment to measure outcomes.
2. Relative probability is called **empirical probability** because it is verifiable by experimentation.
3. For example, a personnel manager's survey (experiment) revealed 50 out of 1,000 adults are part-time college students.
4. With relative probability, the population is constantly changing, so the denominator may be thought of as a sample.

$$P(A) = \frac{\text{observations of A}}{\text{sample size}} = \frac{A}{n}$$

$$P(C) = \frac{C}{n} = \frac{50}{1,000} = \frac{5}{100} = .05 = 5\%$$

C. Subjective probability

1. Classical and empirical probability are objective because they are based upon long-run observations of repeatable events.
2. Some events occur only once. Probability statements concerning these events are based upon personal beliefs and are called subjective probability.
3. Example: With the economy coming out of a recession, Linda feels there is an 80% chance this year's sales will be up 10%. This is subjective probability because this recession has never happened before.

V. Probability rules

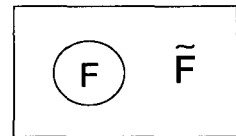
A. Introduction

1. $0 \leq P(A) \leq 1$ is a range for all probability statements. It means that probability can't be negative or greater than one.
2. The **complement** of an event is everything from the sample space that is not the event.
 - a. If F stands for female then \bar{F} , read not F, would be the symbol for male.
 - b. $P(\bar{F}) = 1 - P(F)$
 - c. If 45% of Linda's customers are female, the probability of \bar{F} (male) would be calculated as follows:

$$P(\bar{F}) = 1 - P(F) = 1 - .45 = .55 = 55\%$$

d. **Venn diagrams** are drawings of probability statements.

- 1) A rectangle represents the sample space (everything that can happen).
- 2) A circle represents an event.



3. The page 40 advertising and sales data can each be divided into 2 events.
 - a. Advertising will now be months of less than or equal to \$5,000 and months of greater than \$5,000.
 - b. Sales will now be months of less than or equal to \$50,000 and months of greater than \$50,000.

Sales	Less than or equal to \$50,000 (≤ 50)	Greater than \$50,000 (> 50)	Totals
Advertising			
Less than or equal to \$5,000 (≤ 5)	4	1	5
Greater than \$5,000 (> 5)	1	4	5
Totals	5	5	10

B. Addition rule for adding two events

1. Addition is used to determine the probability of A or B. It is the **union** of two events.

2. **General rule for addition is** $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.

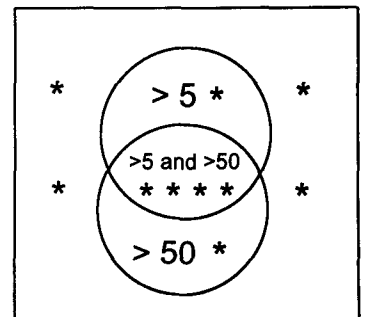
a. $P(A \text{ and } B)$ is called the **intersection** or **joint probability** because it represents how some outcomes overlap and are common to both events.

$$P(> 5 \text{ or } > 50)$$

$$P(> 5) + P(> 50) - P(> 5 \text{ and } > 50)$$

$$\frac{5}{10} + \frac{5}{10} - \frac{4}{10} = \frac{6}{10} = \frac{3}{5}$$

Please locate the 6 out of 10 outcomes in the above table and in the Venn diagram to the right.



3. **Special rule for addition**

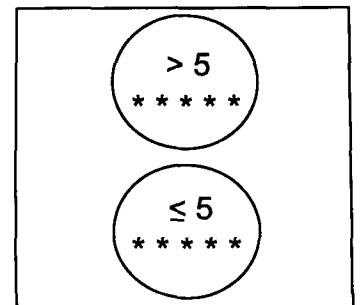
a. When the two events being combined do not contain common outcomes, there isn't an intersection. These events are **mutually exclusive** because they cannot happen at the same time. When adding mutually exclusive events, there isn't an intersection to subtract.

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(\leq 5 \text{ or } > 5) = P(\leq 5) + P(> 5)$$

$$= \frac{5}{10} + \frac{5}{10}$$

$$= 1 = 100\%$$



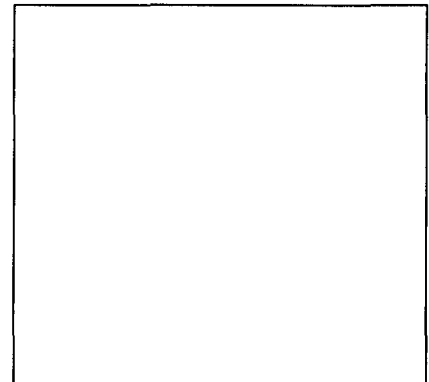
Practice Set 7 Understanding Probability

I. Darin collected the following information concerning customer age and making a sale. Please complete this chart.

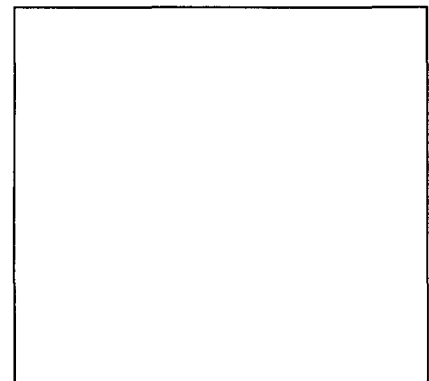
Customer Age and Making A Sale			
Customer Age	Less than or equal to 20	Over 20	Totals
Making A Sale			
No		8	
Yes	24		36
Totals	40		

II. Solve the following problems using the data from question 1. Be sure to use a formula and draw a Venn diagram.

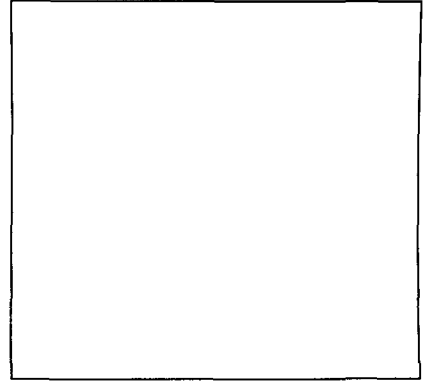
A. The probability of making a sale.



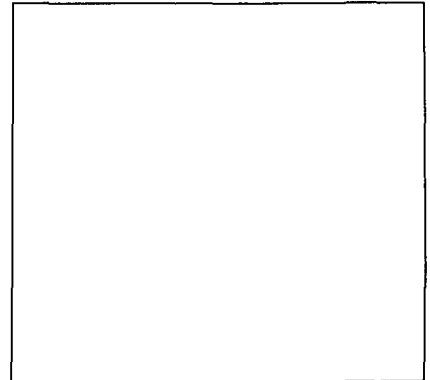
B. The probability of a customer being over 20.



C. The probability of making a sale or a customer being less than or equal to 20.



D. The probability of making a sale or not making a sale.



E. State the addition rule used to answer question C. What condition is necessary to apply this rule?

F. State the addition rule used to answer question D. What condition is necessary to apply this rule?

Quick Questions 7 Understanding Probability

I. List the three types of probability.

II. Place the letter of the appropriate definition, formula, or expression next to the concept it defines.

1. Probability		A. Each outcome has a known, equal chance of happening
2. Inferential statistics		B. Combines two or more simple events
3. Experiment		C. $1 - P(A)$
4. Outcome		D. Mutually exclusive
5. Event		E. The likelihood of something happening
6. Compound event		F. Cannot be divided
7. Simple event		G. $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
8. Probability of A's complement		H. Empirical probability
9. A range for the probability of A		I. $P(A \text{ or } B) = P(A) + P(B)$
10. When A does not intersect B		J. Estimating population parameters using sample statistics
11. General rule of addition		K. Measurements resulting from an experiment
12. The complement of A		L. \bar{A}
13. Another name for relative probability		M. A process resulting in one or more measurements
14. Special rule of addition		N. $0 \leq P(A) \leq 1$
15. Classical probability		O. Collection of outcomes

III. Identify these probability situations by placing in the space provided a C for Classical, E for Empirical, or S for Subjective.

1. Flipping a coin	
2. Drawing a red card from a deck of cards	
3. The chance of drivers stopping at a stop sign in the city of Boston	
4. Mary earning a grade of B or higher in Statistics I next term	
5. Darin Jones having a 10% increase in sales next year	
6. Salesperson A making a sale	
7. Drawing a red ball from a container of 3 red balls and 4 blue balls	
8. An advertising campaign increasing this December's sales	
9. School being called off in January because of inclement weather	
10. School being called off next Tuesday because of inclement weather	

- IV. The following data concerns the buying habits of people entering a retail store in relation to their gender. Please complete the chart.

Customer Buying Habits and Gender			
Customer Gender	Male	Female	Totals
Making a Sale			
Yes	42		56
No		6	
Totals	60		

- V. Using the above data, draw a Venn diagram and determine, using a formula, the probability of each of these events.

A. The probability of making a sale is _____.

B. The probability of a customer being female is _____.

C. The probability of making a sale or a customer being male is _____.

D. The probability of making a sale or not making a sale is _____.

E. State the rule used to answer questions C and D.
What condition is necessary to apply each rule?

Chapter 8 Probability Part II Multiplication Rules

I. Special rule of multiplication

- A. Events A and B are **independent** when event A happening does not affect the probability of event B happening.
 B. The intersection of two events represents how often they happen together.
 1. This probability of this intersection is called **joint probability**.
 2. When two events are independent, their joint probability is the product of their individual probabilities.

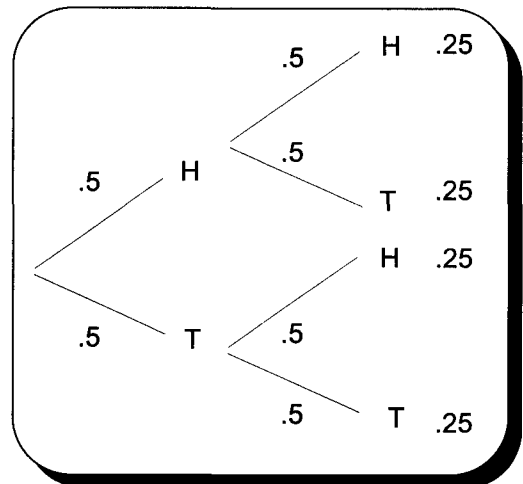
$$P(A \text{ and } B) = P(A) \times P(B)$$

- C. Flipping a coin results in independent events because the outcome of the first flip (event #1) does not affect the outcome of the second flip (event #2).

$$P(H \text{ and } H) = P(H) \times P(H) = (.5)(.5) = .25$$

- D. Special multiplication problems can also be solved with a contingency table and a tree diagram.

Toss 2	Toss 1	P(H) = .5	P(T) = .5	Totals
P(H) = .5		.25	.25	0.50
P(T) = .5		.25	.25	0.50
Totals		.50	.50	1.00



- The probability of a third head is still .5 and the probability of three heads in a row is $(.5)(.5)(.5) = .125$.
- The P(A) is referred to as **marginal probability** because its probability is located in the margins of a contingency table.
- For independent events, joint probability $P(A \text{ and } B)$ is the product of marginal probability (see the highlighted boxes of the contingency table).

II. General rule of multiplication

- A. Events A and B are **dependent** when event A happening has an affect on the probability of event B happening.
 B. Rather than simply multiplying $P(A) \times P(B)$, the P(B) is adjusted for the effect of A having happened.
 C. The idea of event A happening first and affecting B is known as **conditional probability**. A conditional probability statement would be written $P(B | A)$ and read the probability of B given A.
 D. When events are dependent, A affects B and the general rule of multiplication is appropriate.

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

Note that $P(B | A)$ is weighted by the probability of A happening.

- E. Suppose Linda wants to determine the probability of advertising expenditures being greater than \$5,000 and sales revenue being greater than \$50,000.

$$\begin{aligned}
 &P(A > \$5,000 \text{ and } S > \$50,000) \\
 &= P(A > \$5,000) P(S > \$50,000 | A > \$5,000) \\
 &= \frac{5}{10} \times \frac{4}{5} = \frac{20}{50} = .4 = 40\%
 \end{aligned}$$

Note: This answer can be read directly from this table. $\frac{4}{10} = .4$

	Sales Revenue		Totals
Advertising Expenditures	Less than or equal to \$50,000 (≤ 50)	Greater than \$50,000 (> 50)	
Less than or equal to \$5,000 (≤ 5)	4	1	5
Greater than \$5,000 (> 5)	1	4	5
Totals	5	5	10

Note: The general probability rules for addition and multiplication work for all cases. The special rule for addition may be used when events are mutually exclusive. The special rule for multiplication may be used when events are independent.

III. Bayes' theorem

A. Bayes' theorem is used to find the probability of conditional events.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A) \times P(B|A)}{P(A) \times P(B|A) + P(\bar{A}) \times P(B|\bar{A})}$$

B. Logic of Bayes' theorem

1. The condition is that B has occurred. The denominator contains the situations when this happens (the whole).
 - a. Therefore, it contains B happening with A plus B happening with \bar{A} .
 - b. B happening with A is weighted by how often A occurs.
 - c. B happening with \bar{A} is weighted by how often \bar{A} occurs.
2. The numerator is the part of the denominator that is of concern. In this case it is A happening with B.

C. Linda wants to determine the probability of sales being over \$50,000 when she spends over \$5,000 on advertising. First Bayes' theorem is written using symbols more representative of the problem. Second, substitute and solve.

$$P(>\$50 | >\$5) = \frac{P(>50 \text{ and } >5)}{P(>5)}$$

$$= \frac{P(>50) \times P(>5 | >50)}{P(>50) \times P(>5 | >50) + P(>50 | \bar{>5}) \times P(>5 | \bar{>50})} = \frac{\frac{5}{10} \times \frac{4}{5}}{\frac{5}{10} \times \frac{4}{5} + \frac{5}{10} \times \frac{1}{5}} = \frac{\frac{20}{50}}{\frac{20}{50} + \frac{5}{50}} = \frac{20}{25} = .8 = 80\%$$

IV. Joint and conditional probability may easily be read from a contingency table converted to decimals.

Monthly Advertising and Sales					
	Sales	Less than or equal to \$50,000	Greater than \$50,000	Totals	The page 46 answer to sales over \$50,000 and advertising over \$5,000 of 40% can be read directly from this chart.
Advertising	Less than or equal to \$5,000	0.40	0.10	0.50	
	Greater than \$5,000	0.10	0.40	0.50	The answer to the conditional statement above can be read off the chart as .4 divided by .5 or 80%.
	Totals	0.50	0.50	1.00	

Advertising and sales are dependent so the special rule for multiplication does not apply. Note that joint probability is not the product of marginal probability.

V. Counting relevant outcomes

- A. As problems become more complex, counting total outcomes and outcomes of interest will also be more complex.
- B. **The counting rule:** If one event can happen M ways and a second event can happen N ways, then the two events can happen in sequence (M)(N) ways. Linda wants to visit her 3 competitors, each of whom have 2 stores. There are (3)(2) = 6 stores she can visit. The total counting for three events would be (M)(N)(O).
- C. **The factorial rule** involves arranging N available items.
 1. Linda can visit the 6 stores of her competitors using 6! alternative routes.
 2. $N! = 6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ alternative routes
 3. When she begins, she has 6 alternatives. Having been to a store, she then has 5 alternatives, then 4, etc.
- D. **The permutation rule** involves arranging R of N available items.
 1. Order is important as a, b, c and c, a, b are different and each is counted as an outcome.
 2. Here is how many ways Linda could arrange 4 of 7 posters as a window display. $N = 7$ and $R = 4$

$${}_N P_R = \frac{N!}{(N-R)!} = \frac{\text{Totality}}{\text{What is not of interest}} = {}_7 P_4 = \frac{7!}{(7-4)!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = 7 \times 6 \times 5 \times 4 = 840$$

- E. **The combination rule** involves choosing (not arranging) R of N available items. Because items are not being arranged, order is not important. Items abc and cba are the same and are not counted twice.
 1. Just hanging (not arranging) 4 of 7 posters has fewer possibilities because order doesn't count.

$${}_N C_R = \frac{N!}{(N-R)!(R!)}$$

$${}_7 C_4 = \frac{7!}{(7-4)!4!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 4 \times 3 \times 2 \times 1} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 35$$

2. The use of R! in the denominator eliminates the multiple counting of items of interest.

Practice Set 8 Probability Part II Multiplication Rules

- I. Below is the data Darin Jones collected concerning sales to customers of different ages. (see page 42) Convert Table 1 to decimals and place the information into Table 2.

Analysis of Sales By Age of Customer (Table 1)				Decimals (Table 2)		
Customer Age Sale	Less than or equal to 20	Over 20	Totals	Less than or equal to 20	Over 20	Totals
No	16	8	24			
Yes	24	12	36			
Totals	40	20	60			

- II. Use a formula to calculate the probability of these events and check your answers using Table 2.

A. The probability of a customer being over 20 years old is _____.

B. The probability of a customer being over 20 years old and not making a sale is _____.

C. The probability of a customer being less than or equal to 20 years old and over 20 years old is _____.

D. Was the special rule of multiplication applicable to question B? Why or why not? Could the special rule of multiplication be used by Linda with the page 46 advertising data? Why or why not?

III. Use Bayes' theorem to calculate the probability of making a sale given a customer is less than or equal to 20 years of age.

IV. Recalculate your answer to question III using Table 2 on page 48.

V. Use Linda's page 46 advertising data to calculate the possibility of having monthly advertising over \$5,000 and monthly sales over \$50,000.

VI. Answer these questions about 5 posters Darin has to advertise a new CD recorder/player. Be sure to show all formulas.

A. How many ways can he arrange these posters in a horizontal line across a wall?

B. How many ways can he arrange only 3 posters? Arrange implies that order counts. AB is not the same as BA and that both should be counted.

C. How many ways can he just hang them? (order doesn't count)

Just Cruis'in



Quick Questions 8 Probability Part II Multiplication Rules

I. Place the letter of the appropriate definition or formula next to the concept it defines.

1. General rule for multiplication		A. $P(A \text{ and } B) = P(A) \times P(B)$
2. Independent events		B. Marginal probability
3. Special rule for multiplication		C. $P(A \text{ and } B)$
4. $P(A)$		D. Event A does not affect the probability of event B
5. Counting rule		E. $P(A \text{ and } B) = P(A) \times P(B A)$
6. Combination rule		F. $P(A) \times P(B A) + P(\bar{A}) \times P(B \bar{A})$
7. Joint probability		G. $(M)(N)$
8. Denominator of Bayes' theorem		H. N items can be arranged $N!$ ways
9. Factorial rule		I. $\frac{N!}{(N-R)!}$
10. Permutation rule		J. $\frac{N!}{(N-R)!(R!)}$

Note that G represents how two sets of items can be ordered and H, I, and J represent how one set of items can be ordered.

II. Complete this chart concerning the number of hours students studied for a test and their exam grades.

Hours studying	Less than 4	Greater than or equal to 4	Total
Test score			
Less than 85		2	10
Greater than or equal to 85	2		
Totals		10	

III. Use a formula and the data in question II to answer the following questions.

A. The probability of earning a grade less than 85.

B. The probability of someone studying 4 or more hours and earning a grade of 85 or higher.

C. Was the special rule of multiplication applicable to question B? Why or why not?

D. Use Bayes' theorem to calculate the probability of someone scoring 85 or higher if they studied 4 or more hours.

E. Prove your answer to question D using the chart on page 50.

IV. How many stores will a salesperson visit if they must visit 3 locations in each of 4 cities?

V. An advertising manager has 6 advertisements of equal size to place horizontally across a magazine page.

A. How many ways can the 6 ads be arranged?

B. How many ways can 4 of the 6 ads be arranged if order counts?

C. How many ways can 4 of the 6 ads be arranged if order does not count and a, b, c, d and d, c, b, a are considered the same arrangement?

Chapter 9 Discrete Probability Distributions

I. Understanding probability distributions

- A. A **random variable** measures a numerical event, the value of which, is determined by chance.
- B. The experimental outcomes described in chapter 8 are random variables. Examples include flipping a coin and customer buying habits based upon gender.
- C. **Random variables** are either discrete or continuous.
 - 1. **Discrete:** Only finite values, such as the countable numbers, can exist on the x-axis. Examples include tire defects and the number correct on a true or false exam.
 - 2. **Continuous:** Measurement may assume any value associated with an uninterrupted scale. Examples include the exact weight of a one-pound box of cookies and the average length of computer parts.
- D. A **probability distribution** lists all the probability values associated with a random variable (x).
- E. Example: In chapter 3, Linda found that 36, 18, and 6 tapes were rented for \$2, \$3, and \$4 respectively.
 - 1. The amount received is a discrete random variable with possible values (outcomes) of \$2, \$3, and \$4.
 - 2. Below is the probability distribution associated with tape rental fees.

Discrete Probability Distribution					
Rental Fees (x)	Number of Tapes Rented	Probability P(x)	[x • P(x)]	x ²	[x ² • P(x)]
\$2.00	36	36/60 = .60	\$1.20	4	\$2.40
3.00	18	18/60 = .30	0.90	9	2.70
4.00	6	6/60 = .10	0.40	16	1.60
	60	1.0	\$2.50		\$6.70

Note: This distribution is similar to a frequency distribution with P(x) replacing f.

F. The mean and variance of a discrete probability distribution

- 1. Random variable parameter calculations are similar to grouped data parameter calculations. However, division is not necessary for random variable calculations because the observations total 1.0 (100%).
- 2. The mean of random variable x is called the expected value of x or E(x).
- 3. The variance of x is V(x).

$$E(x) = \sum [x \cdot P(x)] = \$2.50$$

See chart calculations

$$V(x) = [\sum x^2 \cdot P(x)] - [E(x)]^2$$

$$= \$6.70 - (\$2.50)^2$$

$$= \$6.70 - \$6.25 = \$.45$$

Note: These formulas may be written using Greek letters with μ for E(x) and σ^2 for V(x).

II. The binomial probability distribution

- A. Binomial experiments have the following characteristics.
 - 1. The experiment consists of a fixed number of trials. Two mutually-exclusive outcomes result from each trial.
 - 2. Defined as success and failure, each set of outcomes can be counted and represent an independent event.
 - 3. The probability of success and the probability of failure must be constant with $P(F) = 1 - P(S)$.
- B. Binomial experiments include flipping a coin, counting product defects, and marketing response rates.
- C. Determining the binomial distribution requires calculating $P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$ where:

n is number of trials	x is number of successes	p is probability of success	q, the probability of failure, is 1 - p
-----------------------	--------------------------	-----------------------------	---

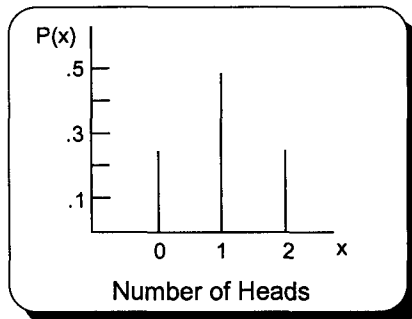
- 1. The page 46 coin flipping experiment, solved with a contingency table and a decision tree, is a binomial experiment. The probability of having exactly one head with two tosses is calculated below.
- 2. n = 2, x = 1 (head), p = .5, q = .5 **Note:** 0! = 1, x⁰ = 1, and x¹ = x

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$P(1) = \frac{2!}{1!(2-1)!} (.5^1 \cdot .5^{2-1})$$

$$= \frac{2 \times 1}{1(1)} (.5^1 \cdot .5^{2-1})$$

$$= 2 \times .5 \times .5 = .5$$



The Binomial Probability Distribution for n = 2 and p = .5	
# of Heads (x)	P(x)
0	.25
1	.50
2	.25
Total	1.00

D. Binomial tables

1. Extensive tables have been developed to solve binomial experiments. See Table 1 page ST 1.
2. Below is a table for a two trial ($n = 2$) experiment and some relevant probabilities.
3. Note the distribution for the page 46 coin problem is under the .5 column.
4. If the probability of a defective part is .05, then getting 2 out of 2 defects would be .0025 or .25%.

Probability of x successful outcomes given the following probability values (p) and trials (n)												Note: A binomial table has 2 defining characteristics, n and p. Note: For this table, n = 2.
X	0.0500	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.9500	
0	0.9025	0.81	0.64	0.49	0.36	0.25	0.16	0.09	0.04	0.01	0.0025	
1	0.0950	0.18	0.32	0.42	0.48	0.50	0.48	0.42	0.32	0.18	0.0950	
2	0.0025	0.01	0.04	0.09	0.16	0.25	0.36	0.49	0.64	0.81	0.9025	

E. The shape of binomial distributions

1. Distributions are symmetrical when $P(x) = .5$. High or low probabilities have highly skewed distributions.
2. When the $p(x) \neq .5$, the distribution is skewed and a larger n will result in a more symmetrical distribution.

III. The Poisson distribution

- A. A Poisson distribution is similar to a binomial distribution except the $P(x)$ must be small. A Poisson distribution is defined by only 1 characteristic, its mean. The distribution is highly skewed to the right.
- B. Events related to time, such as customers arriving per 5-minute periods, often follow a Poisson distribution.
- C. The mean is needed when using a Poisson distribution.

$$\mu = E(x) = \sum [x \cdot P(x)] \text{ (see page 52)}$$

- D. A Poisson distribution may be determined with a formula or looked up in a table.

1. Calls per 15 minute period to Linda's repair facility follow a Poisson distribution with $\mu = 1.0$. What is the probability of exactly three service calls being received in a randomly selected 15-minute period?
- 2.

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

$$P(3) = \frac{(1^3)2.7183^{-1}}{3!} = \frac{(1)(0.3679)}{6} = 0.0613$$

See table below

IV. The Poisson approximation of the binomial probability distribution

- A. A Poisson distribution is often used to approximate a binomial distribution for problems such as errors on a typed page, circuit board defects, and customers bouncing checks at Linda's Video Showcase.
- B. This is done to save the time and money necessary to solve extensive binomial experiments.
- C. These two distributions have similar skewness provided the number of trials is large ($n \geq 30$) and the probability of occurrence (p) is small (either np or $nq < 5$).
- D. The mean for a Poisson approximation of a binomial is $\mu = np$ (n = trials and p is the probability of an event).
- E. Recent observations revealed that 4 of 40 items purchased by customers are returned. This is a binomial problem with a sample mean of $P(s) = 4/40 = .10$. Determining the entire distribution by using the binomial formula 40 times would be a tremendous task.
 1. Using the Poisson approximation with a sample mean of .10 for μ is much easier. Its use is appropriate as $n \geq 30$ and $np < 5$ ($40 \times .1 = 4$). The above formula yields a probability of 0 returns equal to .904837.
 2. Using a Poisson distribution table to solve this problem only requires locating the appropriate outcome (value of x) under the appropriate mean. $\mu = .1$ and $x = 0 \rightarrow .9048$ (see Table 2 page ST 2)

x	Probability of x outcomes given the following population means									
	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	0.3679
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647	0.1839
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494	0.0613
4		0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111	0.0153
5				0.0001	0.0002	0.0004	0.0007	0.0012	0.0020	0.0031
6							0.0001	0.0002	0.0003	0.0005
7										0.0001

Practice Set 9 Discrete Probability Distributions

- I. Darin sells three different Walkman CD recorders; one for \$149, one for \$159, and a third for \$169. Of the 187 machines sold during a recent period, 43 were the least expensive, 90 were moderately priced, and 54 were the expensive model.

A. Calculate the expected price of Walkman sales.

B. Compare this answer to the page 12 weighted mean sales value of Walkman sales.

C. In theory, what is the difference between a weighted mean of variable x and the expected value of x ?

- II. When waiting on a customer, Darin's salespeople make a sale 60% of the time (see page 42). Use the binomial formula or your statistics software to calculate the probability of making exactly 3 sales to 5 customers.

- III. Using the appropriate table or your statistics software, complete the binomial distribution described by question II.

Special Note

- I. Variables that may follow a binomial probability distribution
 - A. Probability of an employee contributing to the company pension plan
 - B. Probability of collecting an overdue accounts receivable
 - C. Probability of receiving a positive response to a marketing campaign
 - D. Probability of a part being defective
- II. Variables that may follow a Poisson probability distribution
 - A. Number of defects on a 300 foot roll of aluminum
 - B. Errors on a typed page
 - C. Customers arriving at a drive up window within a 5 minute period
 - D. Number of rare disease cases per 1,000,000 people

IV. Using the answer to question III or statistics software, answer the following questions.

- A. $P(x = 4)$ is _____ B. $P(x > 2)$ is _____ C. $P(x < 3)$ is _____ D. $P(x \leq 4)$ is _____

V. Darin is interested in how busy his complaint department is during any 20-minute period. Data shows the expected number of calls is highly skewed with an average of only 1.0 calls per 20-minute period.

- A. Assuming a Poisson probability distribution and using a formula or statistics software, is the probability of zero calls being received in a 20-minute period over or under 50%?

B. Using a table or statistics software, complete and draw this distribution.

C. What is the probability that at least 3 calls are received in a 20-minute period?

VI. Darin is interested in the number of customers who bounce a check. Last year only .2% of the 1,000 checks deposited from customers did not clear. This year Darin expects 1,500 customers will pay by check and with the economy being about the same, the same percent of checks should bounce.

A. Can the Poisson approximation of the binomial be used to solve this problem? Why?

B. What is the expected number of bounced checks for this year?

C. What is the probability that no one will bounce a check this year?

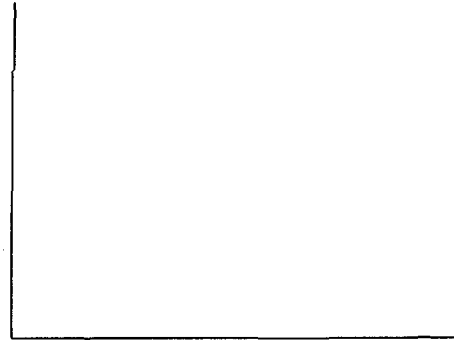
D. What is the probability that at least 2 checks will bounce?

E. What would you think if 5 checks had bounced by the end of May?

III. Five percent of the parts coming off an assembly line are defective.

A. Using the binomial formula or your statistics software, calculate the probability of exactly 2 out of 5 parts being defective.

B. Determine the distribution of defective parts using a table in the back of this book. Graph the distribution.



IV. A bank found that the average number of cars waiting during the noon hour at a drive-up window follows a Poisson distribution with a mean of 2 cars. Make a chart of this distribution using a Poisson distribution table. Graph the distribution and answer these questions concerning the probability of cars waiting at the drive-up window.

A.



B. No cars waiting

C. Two cars waiting

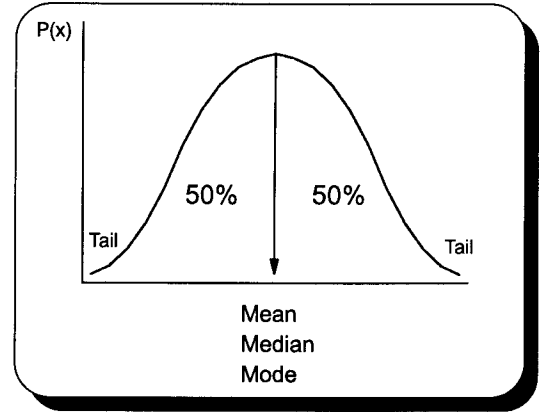
D. At least three cars waiting

E. Not as many as 3 cars waiting

Chapter 10 Continuous Normal Probability Distributions

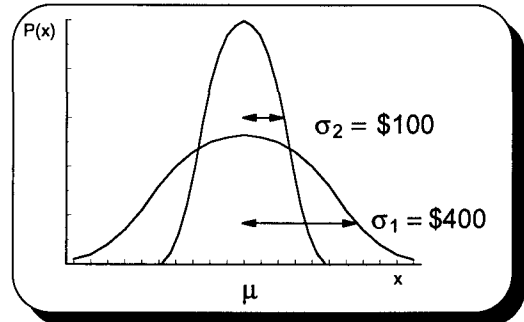
I. Characteristics of normal probability distributions

- A. **Continuous:** An infinite number of x-axis values may exist.
- B. **Symmetrical around the mean:** Each half (50%) of the curve is a mirror image of the other half.
- C. **Bell-shaped:** There is 1 peak above the mean, median, and mode.
- D. **Asymptotic:** The tail of the distribution approaches, but never touches, the x-axis.
- E. Variables dealing with size (income, weight, and intelligence) are often normally distributed.



II. The standard deviation

- A. 100% of the outcomes are under the normal curve.
- B. One standard deviation spans approximately 34% of the outcomes in each direction from the mean.
- C. Curves that are flat, more spread out, have a larger standard deviation.
 - 1. Two of Linda's video stores may each have average weekly sales of \$1,800, but sales variability may differ from store 1 to store 2.
 - 2. As shown here, store 1 has a $\sigma = \$400$. This represents more variability in weekly sales than exists at store 2, with a $\sigma = \$100$.



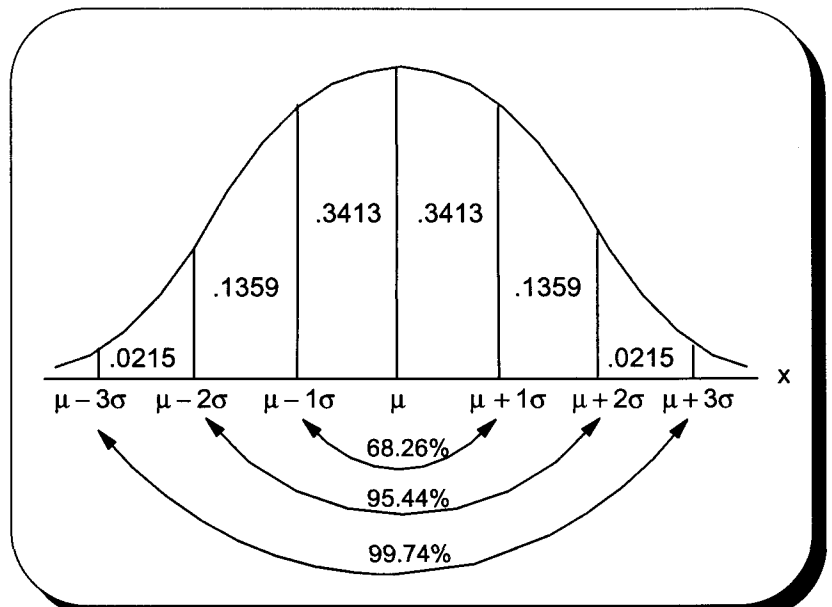
III. The standard normal distribution

- A. Continuous probability distributions may have an infinite number of means, each with an infinite number of standard deviations. This makes it impossible to have a table for each possible mean and standard deviation.
- B. The standard normal probability distribution avoids this problem by having a mean (μ) of 0 and a standard deviation (σ) of 1. The standard normal distribution is kind of a generic distribution.
 - 1. Distances from the mean are measured in standard deviations.
 - 2. This distance (or number of standard deviations) is called z.
 - 3. The probability distribution associated with all possible z values is the standard normal probability distribution.
 - 4. This distribution is presented as a z table.
 - 5. Chapter 4 used the empirical rule to measure the percent of data within a set number of standard deviations of the mean. The standard normal probability distribution is the mathematical basis for the empirical rule.

IV. In the above example, store 1 had mean weekly sales of \$1,800 and a standard deviation of \$400. The following example reviews how the empirical rule is used to determine the sales range for 1, 2, and 3 standard deviations. Note how a set probability is associated with each number of standard deviations.

Given: $\mu = \$1,800$ and $\sigma = \$400$

<p>A. $\mu \pm 1\sigma$ $\\$1,800 \pm 1(\\$400)$ $\\$1,800 \pm \\400 68.26% are between \$1,400 and \$2,200</p>
<p>B. $\mu \pm 2\sigma$ $\\$1,800 \pm 2(\\$400)$ $\\$1,800 \pm \\800 95.44% are between \$1,000 and \$2,600</p>
<p>C. $\mu \pm 3\sigma$ $\\$1,800 \pm 3(\\$400)$ $\\$1,800 \pm \\$1,200$ 99.74% are between \$600 and \$3,000</p>



V. Using a z table to answer questions about probability given a range for the random variable

A. A z table shows the probability (area under curve) associated with a number of standard deviations (z) from the mean. Z standard deviations carried to 1 decimal place are shown in the first column of the table. Probability for a z is in the second column. Carrying z to 2 decimal places requires using other table columns.

B.
$$Z = \frac{x - \mu}{\sigma}$$

Partial z Table			
z	0.00	0.01	0.02
0.5	0.1915	0.1950	0.1985
1.0	0.3413	0.3438	0.3461
1.5	0.4332	0.4345	0.4357
2.0	0.4772	0.4778	0.4783

See Table 3, page ST 3, for a complete z table

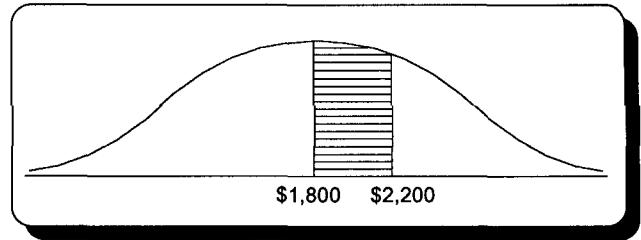
C. Continuing the page 58 example. What is the probability of store 1 having sales between \$1,800 and \$2,200? Calculate z and determine the probability using a z table.

Given: $\mu = \$1,800$ and $\sigma = \$400$

$$Z = \frac{x - \mu}{\sigma}$$

$$= \frac{\$2,200 - \$1,800}{\$400} = \frac{\$400}{\$400} = 1 \rightarrow .3413$$

34.13% are between \$1,800 and \$2,200



D. What is the probability of store 1 having sales between \$1,200 and \$2,004?

$$Z = \frac{x - \mu}{\sigma}$$

$$= \frac{\$1,200 - \$1,800}{\$400}$$

$$= \frac{-\$600}{\$400}$$

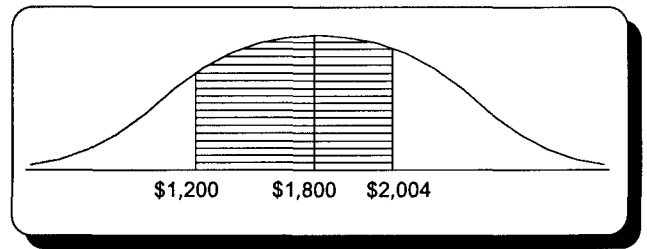
$$= -1.5 \rightarrow .4332$$

$$Z = \frac{x - \mu}{\sigma}$$

$$= \frac{\$2,004 - \$1,800}{\$400}$$

$$= \frac{\$204}{\$400}$$

$$= .51 \rightarrow .1950$$



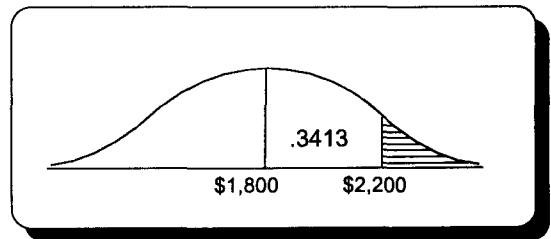
Note: Because a normal curve is symmetrical around the mean, z values below the mean are negative and represent the same probability as their positive counterparts.

43.32% are between \$1,200 and \$1,800
 19.50% are between \$1,800 and \$2,004
 62.82% are between \$1,200 and \$2,004

E. What is the probability of store 1 having sales over \$2,200?

- From question C, we know that 34.13% are between \$1,800 and \$2,200 (1σ).
- 50% are greater than \$1,800. Therefore,

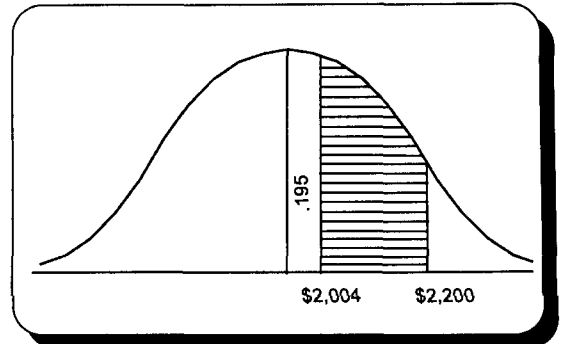
15.87% or (50% - 34.13%) are greater than \$2,200.



F. What is the probability of store 1 having sales between \$2,004 and \$2,200?

- From question C, we know 34.13% are between \$1,800 and \$2,200.
- From question D, we know 19.5% are between \$1,800 and \$2,004. Therefore,

14.63% (34.13% - 19.50%) are between \$2,004 and \$2,200.



VI. Using a z table to answer questions about a range for the random variable given a probability

- A. Determine the problem's relevant probability. Locate the row and column of this probability in the body of a z table. The one decimal place z value for this probability is in the first column of this row. The second decimal place is directly above in the first row of the table.
- B. The range is then found using this expression. $\mu \pm z\sigma$

Partial z Table			
z	0.00	0.01	0.02
0.5	0.1915	0.1950	0.1985
1.0	0.3413	0.3438	0.3461
1.5	0.4332	0.4345	0.4357
2.0	0.4772	0.4778	0.4783
2.5	0.4938	0.4940	0.4941

- C. Monthly sales at Linda's stores are normally distributed with a mean of \$55,000 and a standard deviation of \$15,000.
1. Using symmetrical limits around the mean, 95.44% of her monthly sales are between x_1 and x_2 .

Given:
 $\mu = \$55,000$
 $\sigma = \$15,000$

$\frac{.9544}{2} = .4772 \rightarrow Z = 2$

$\mu \pm z\sigma$
 $\$55,000 \pm 2(\$15,000)$
 Range is \$25,000 to \$85,000.

2. Find the first and third quartiles.

From page ST 3, .25 $\rightarrow z = .67$

$\mu \pm z\sigma$
 $\$55,000 \pm .67(\$15,000)$
 $\$55,000 \pm \$10,050$
 Range is \$44,950 to \$65,050.

3. Find the top decile.

$50\% - 10\% = 40\% \rightarrow z = 1.28$

$\mu \pm z\sigma$
 $\$55,000 + 1.28(\$15,000)$
 $\$55,000 + \$19,200$
 Sales must be above \$74,200.

4. Find the second decile from the bottom.

a. The lower limit of x is associated with 40% and z for 40% is 1.28.

$\mu \pm z\sigma$
 $\$55,000 - 1.28(\$15,000)$
 $\$55,000 - \$19,200$
 Lower limit is \$35,800.

b. The upper limit of x is associated with 30% and z for 30% is .84.

$\mu \pm z\sigma$
 $\$55,000 - .84(\$15,000)$
 $\$55,000 - \$12,600$
 Upper limit is \$42,400.

The second decile has sales between \$35,800 and \$42,400.

VII. The normal approximation to a binomial distribution

- A. In chapter 9, a Poisson distribution was used to approximate a binomial distribution when $n \geq 30$ and either $np < 5$ or $nq < 5$.
- B. Here we learn how the standard normal distribution may be used to approximate a binomial distribution when $n \geq 30$ and both np and nq are ≥ 5 .
- C. Linda wants to know how many sales will result from calls to 40 previous customers. Past history indicates 25% of these customers will make an additional purchase. Calculate the probability of making at least 12 sales.

- Using the binomial would require solving $P_{\text{binomial}}(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$ many times.
- The normal approximation to the binomial distribution is appropriate to solve this problem.
 - $n \geq 30$ as $n = 40$
 - np and nq are ≥ 5 as $np = 40 \times .25 = 10$ and $nq = 40 \times .75 = 30$

- The mean and standard deviation would be calculated as follows:

$$\mu = np = (40)(.25) = 10$$

$$\sigma = \sqrt{npq} = \sqrt{(40)(.25)(.75)} = 2.7386$$

Note: Mean and standard deviation formulas for a binomial distribution were given on page 56.

- The continuity correction factor
 - Because a discrete event (12 sales) has to be considered a continuous interval when using the continuous normal probability distribution, the number 12 must be expressed as the interval of 11.5 to 12.5. This is done to ensure that the entire area under a normal curve is included in the analysis.
 - Linda's question includes 12 so the lower limit of 11.5 is appropriate. Had she excluded 12 with $p(x) > 12$, then 12.5 would have been the value of x .
- Calculate $P(x \geq 12)$ using the normal approximate of the binomial.

$$Z = \frac{x-\mu}{\sigma} = \frac{11.5-10}{2.7386} = \frac{1.5}{2.7386} = .55 \rightarrow .2088 \text{ and } 50\% - 20.88\% = 29.12\%$$

VIII. Summary of problem types

- A. Finding the probability given a range for the random variable

The problems on page 59 describe situations where the **range is known** and the probability for that range must be determined.

- Calculate z using this formula. $Z = \frac{x-\mu}{\sigma}$
- Find z in the margins of the z table.
- Look in the body of the table to find the probability.
- Continue until the problem is solved.

- B. Finding a range for the random variable given a probability

The problems on page 60 describe situations where the **probability is known**, and the range for that probability must be determined.

- Find the probability in the body of the z table.
- Look to the margins of the table to find z .
- Find the range for x using this formula. $\mu \pm z\sigma$
- Continue until the problem is solved.

Practice Set 10 Continuous Normal Probability Distributions

- I. Sales commissions paid by Darin's Music Emporium are normally distributed with a mean of \$25,000 and a standard deviation of \$5,000. Solve the following being sure to draw a graph of each distribution.

A. $P(\$15,000 \leq x < \$25,000)$

Note: This question is read "What is the probability that \$15,000 is less than or equal to x which is less than \$25,000?"

B. $P(\$20,000 \leq x < \$30,000)$

C. $P(x \geq \$37,550)$

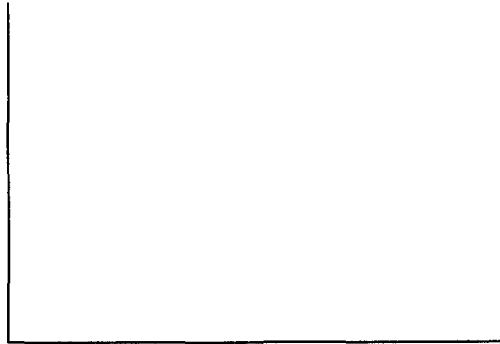
D. $P(\$27,500 \leq x < \$32,500)$

II. The number of customers returning merchandise to Darin's Music Emporium is normally distributed with a mean of 6.3 per week and a standard deviation of 1.5. Given the following probabilities, calculate the appropriate value or values for x .

- A. Half of the time, returns will be above _____
- B. Ninety percent of the time, returns will be below _____

C. Find the interquartile range for returns to Darin's Music Emporium.

D. Draw a graph of the eighth decile for returns to Darin's Music Emporium.



III. A recent study indicated 5% of Darin's customers return merchandise sold for credit. What is the probability of Darin having less than 20 returns for a 500 credit sales week?

Quick Questions 10 Continuous Normal Probability Distributions

- I. The average income of 30-year-old college graduates from State University is normally distributed with a mean of \$30,000 and a standard deviation of \$4,000. Calculate the following being sure to graph each question.

A. $P(x < \$34,000)$

B. $P(x > \$38,000)$

C. $P(\$18,000 \leq x < \$19,800)$

D. $P(x > \$30,000)$

II. Grades of State University graduates are normally distributed with a mean of 3.0 and a standard deviation of .3. Calculate the following being sure to graph each question.

A. What grade point average is required to be in the top 5% of the graduating class?

B. Calculate the interquartile range.

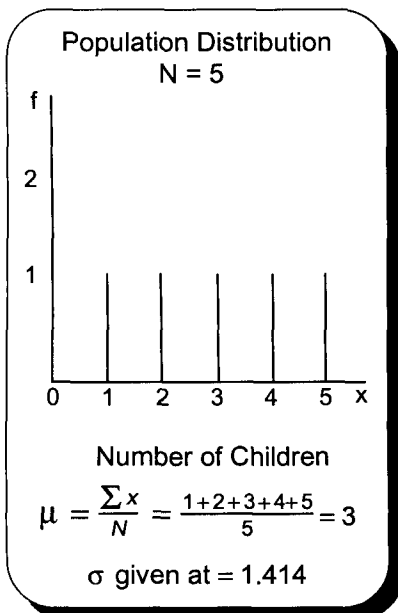
C. An eccentric alumnus left scholarship money for students in the third decile from the bottom of their class. Determine the range of the third decile.
Would a student with a 2.8 grade point average qualify for this scholarship?

D. What is the median grade point average of this class?

Chapter 11 Sampling and the Sampling Distribution of the Means

- I. Inferential statistics uses sample statistics to estimate population parameters. This chapter will explore how a sample mean (\bar{x}) is used to predict its population mean (μ).
- II. Why use sample data
 - A. The cost of a census is prohibitive.
 - B. The time required to take a census is not available.
 - C. Measuring a parameter destroys the item being tested (measuring the mean lifetime in hours of light bulbs).
 - D. A sample will yield adequate results.
- III. Probability samples
 - A. A probability sample is one in which the likelihood of an item being chosen is known.
 - B. Probability sampling methods
 1. Simple random samples
 - a. Each population member has an equal chance of being chosen.
 - b. Put an identification (name, serial number, etc.) into a hat, mix, and select.
 - c. A table of random digits or a computer program often replace the hat.
 - d. To sample 30 out of 485 students using their ID numbers from 1 to 485:
 - 1) Arbitrarily choose a starting point on a table of random digits.
 - 2) Working in some direction (horizontally, vertically, or diagonally), and using the first or last three digits, choose 30 student numbers ignoring those over 485.
 2. Systematic random samples
 - a. Use every n th item beginning at some random point on a list of population members.
 - b. This method could be biased because population members at the beginning of a list (Mr. Abbot or employee 0001) and end of a list (Ms. Zona or employee 9999) might not have an equal chance of being chosen.
 3. Stratified random samples
 - a. Divide population into homogeneous subgroups and sample each subgroup.
 - b. This type of sample can be more representative than a simple random sample because a small diverse section of a population might not be chosen with a simple random sample.
 - C. Sampling and nonsampling error
 1. Sampling error exists because a nonrepresentative sample was used in place of a census.
 2. Nonsampling error, which occurs with any survey, exists primarily because of poor collection techniques. High nonsampling error can make a census less accurate than a sample. Why? Limited funds and having to survey all population members cause poor collection techniques and high nonsampling error.
- IV. Sampling distribution of the means
 - A. The sampling distribution of the means consists of all the possible sample means of size n that may be drawn from a population size N . It is important. Taking one sample is really taking one out of many possible samples. The sampling distribution is the key to why accurate predictions can be made with inferential statistics.
 - B. Population members A,B,C,D, and E have 1,2,3,4, and 5 children respectively. The sampling distribution of the means, its mean and standard deviation, for a sample of 3 out of 5 has been calculated and demonstrated below.

1318	7677	9619	2786
2122	8297	1190	1379
0037	6355	4717	5184
4788	9044	5583	0292

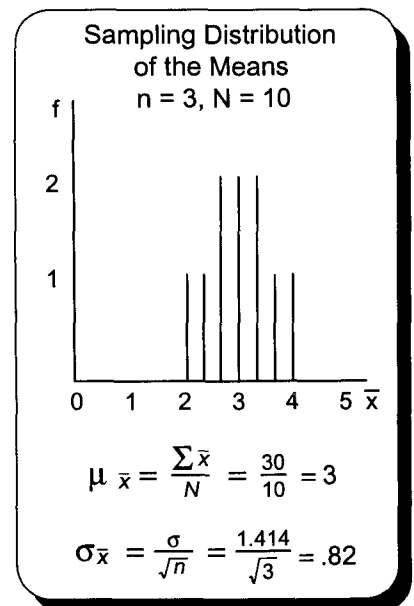


10 possible samples result from a sample of 3 out of 5

	x	\bar{x}
ABC	1,2,3	2.00
ABD	1,2,4	2.33
ABE	1,2,5	2.67
ACD	1,3,4	2.67
ACE	1,3,5	3.00
ADE	1,4,5	3.33
BCD	2,3,4	3.00
BCE	2,3,5	3.33
BDE	2,4,5	3.67
CDE	3,4,5	4.00
	$\sum \bar{x} =$	30.00

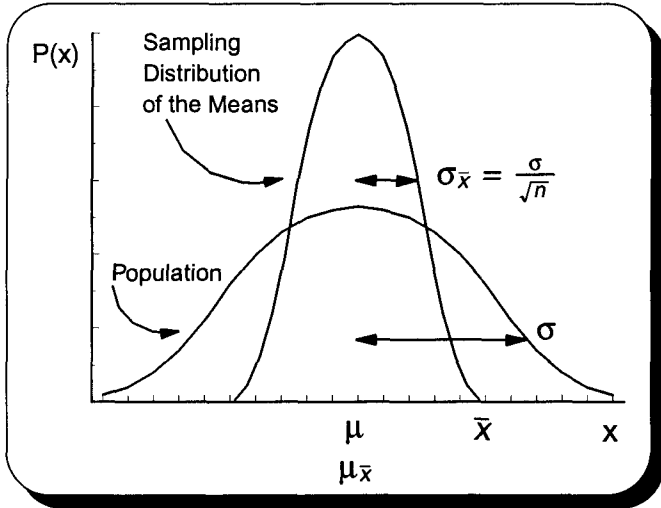
Sampling Distribution of the Means

\bar{x}	f
2.00	1
2.33	1
2.67	2
3.00	2
3.33	2
3.67	1
4.00	1



V. Central limit theorem

- A. The sampling distribution of the means will be normal whenever the population is normal.
- B. The central limit theorem also applies to skewed populations provided the sample is large ($n \geq 30$).
- C. The relationship between the parameters of a population and its sampling distribution is shown below.



Note: Because the sampling distribution is normal regardless of its population's skewness, a sampling distribution's mean can be used to make predictions about one of its sample means. Prediction procedures will be similar to those followed in chapter 10, where the population mean was used to make predictions about a value of x . In practice, the sample mean is known and used to make estimates about the sampling distribution's mean. These estimates also apply to the population mean because said means are equal. These estimates of a population mean can be very accurate because the sampling distribution's standard deviation will be smaller if the sample size is increased. Diminishing returns apply to larger samples being more accurate as the denominator of $\sigma_{\bar{x}}$ is not n but the square root of n . A sample of 49 is only slightly more accurate than a sample of 36. Why? Because the denominator is only slightly larger (7 vs. 6), and the sampling distribution's standard deviation is not proportionately smaller.

VI. Using a large sample ($n \geq 30$) to determine point and interval estimates of population parameters

A. Point estimates

- 1. A point estimate is a one-number estimate.
- 2. Important point estimates
 - a. A sample mean for its population mean
 - b. A sample standard deviation for its population standard deviation

Section B Note: When $n < 30$ and σ is unknown, the t distribution, discussed in chapter 16, must be substituted for the z distribution when making interval estimates. Many statistics software programs do all interval calculations, regardless of sample size, using the t distribution.

B. Interval estimates

- 1. An interval estimate is a range.
- 2. A range for μ , called a **confidence interval**, is determined using this expression.
- 3. The standard deviation of a sampling distribution ($\sigma_{\bar{x}}$), called the **standard error of the mean**, is very important in determining an interval estimate of a population mean.
- 4. Below are two important confidence intervals for $\mu_{\bar{x}}$ and therefore μ .

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

- a. 95 percent confidence interval

z for $.95/2 = .4750 \rightarrow 1.96$	$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
--	--

- b. 99 percent confidence interval

z for $.99/2 = .4950 \rightarrow 2.58$	$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$
--	--

Note: These interval estimates are based upon the relationship between z , the population distribution, and the sampling distribution of the means.

$$Z = \frac{x - \mu}{\sigma} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- 5. When population standard deviation is unknown, the sample standard deviation may be used as a point estimate of the population standard deviation provided the sample is large. Small samples will be examined in chapter 16.

$$\bar{x} \pm 2.58 \frac{s}{\sqrt{n}}$$

- C. **Example:** Linda took a random sample of 49 customer orders and found the mean purchase amounted to \$7.50. The population standard deviation is known to be \$.70. The 99% confidence interval for the population mean purchase has been calculated in this frame.

Note: Linda can lower the range by accepting a confidence interval of only 95% or by increasing the sample size.

Given:	$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$
$\bar{x} = \$7.50$	$\$7.50 \pm 2.58 \frac{\$.70}{\sqrt{49}}$
$\sigma = \$.70$	$\$7.50 \pm 2.58(\$.10)$
$n = 49$	$\$7.50 \pm \$.258$
z for .99 is 2.58	$\$7.24 \leftrightarrow \7.76

Practice Set 11 Sampling and the Sampling Distribution of the Means

- I. Darin's new company, Future Horizons Corporation, manufactures a component for computer chips. Darin wants to know the average weight of 1,000 recently produced components. A sample of 36 had a mean weight of 30.025 milligrams and a standard deviation of .065 milligrams. Calculate the 98% confidence interval for the population mean weight of these components.

Raw Data For People Using Statistics Software											
29.89	29.96	29.97	30.05	29.97	29.98	29.98	30.06	30.04	30.07	30.05	30.06
29.97	29.95	30.05	30.05	29.95	30.09	29.95	29.99	30.06	30.06	29.89	30.09
29.99	29.99	29.98	30.02	30.08	30.01	30.09	30.06	30.08	30.12	30.16	30.15

- II. Calculate the 95% confidence interval using problem I information.

- III. What can Darin do to make this interval smaller?

Quick Questions 11 Sampling and the Sampling Distribution of the Means

I. Place the number of the appropriate formula next to the concept it defines.

- A. The 99% confidence interval _____
- B. Standard error of the mean _____
- C. Used when the population variance is unknown and the sample is large. _____
- D. The 95% confidence interval _____
- E. The mean of the sampling distribution of the means _____

1.	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
2.	$\mu_{\bar{x}}$
3.	$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$
4.	$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
5.	$\bar{x} \pm 2.58 \frac{s}{\sqrt{n}}$

II. Answer the following true or false and fill in the blank questions.

- A. The primary cause of sampling error is poor collection techniques. T F
- B. The standard error of the mean is halved when the sample size is doubled. T F
- C. A one-number estimate of the population mean is called a _____ estimate of the mean.
- D. A range for a population parameter is called the _____.
- E. A _____ may be more accurate than a simple random sample because a small diverse section of the population might not be chosen with a simple random sample.

III. Calculate the 95% and 99% confidence intervals for the population mean given a sample of 36 resulted in a mean of 55 and a standard deviation of 18.

Data Set For People Using Statistics Software								
55	55	39	50	81	48	43	85	38
58	50	57	52	75	55	85	55	81
52	47	62	25	54	71	32	73	40
72	98	53	35	56	55	21	26	46

Chapter 12 Sampling Distributions Part II

I. Estimating the population proportion

- A. The population proportion is the average part of a population having a particular trait.
 B. It may be expressed as a fraction, decimal, or percentage.
 C. The sample proportion is $\bar{p} = \frac{x}{n}$.
 D. The population proportion is used to measure traits such as consumer attitudes toward a product, voter preference, and the proportion of parts passing inspection.
 E. Experiments described here must meet the binomial experiment conditions described on page 52 and the normal approximation of the binomial conditions described on page 61.
 F. Estimating a confidence interval for the population proportion using a large sample is explained below.

$p = \frac{\text{successes}}{\text{population size}}$

Some texts use π for the population proportion.

1. $\bar{p} \pm Z\sigma_{\bar{p}}$ where $\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$ and

1) \bar{p} is the sample proportion
 2) n , the sample size, is ≥ 30
 3) z is based upon the desired confidence interval

2. Example: Linda Smith randomly called 100 customers and found that 80 were happy with the service they received when shopping at Linda's Video Showcase. Calculate a 95% confidence interval for the population proportion. Given: $n = 100$ and z for 95% confidence is 1.96

$$\bar{p} = \frac{x}{n}$$

$$= \frac{80}{100} = .80$$

$n = 100 \geq 30$

$np = 100 \times .8 = 80 \geq 5$

$nq = 100 \times .2 = 20 \geq 5$

The normal approximation of the binomial applies.

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= \sqrt{\frac{.8(1-.8)}{100}}$$

$$= \sqrt{.0016} = .04$$

$$\bar{p} \pm Z\sigma_{\bar{p}}$$

$$.80 \pm 1.96(.04)$$

$$.80 \pm .0784$$

$$.722 \leftrightarrow .878$$

II. Finite correction factor

- A. Thus far, formulas used to calculate the **standard error of the mean** ($\sigma_{\bar{x}}$) and the **standard error of the proportion** ($\sigma_{\bar{p}}$) have been based upon infinitely large populations.
- B. If the population is finite, then the relative size of our sample has increased, and the standard error can be reduced using the finite correction factor.

$$\sqrt{\frac{N-n}{N-1}}$$

- C. The finite correction factor is used to calculate the standard error when $\frac{n}{N} \geq .05$. Smaller ratios are immaterial.

Standard Error of the Mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Standard Error of the Proportion

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \sqrt{\frac{N-n}{N-1}}$$

- D. Linda must adjust her interval calculation because her customer pool totaled 1,000.

$$\frac{n}{N} = \frac{100}{1,000} = .10 \geq .05$$

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$= .04 \sqrt{\frac{1,000-100}{1,000-1}}$$

$$= .04 \sqrt{.9009}$$

$$= .038$$

$$\bar{p} \pm Z\sigma_{\bar{p}}$$

$$.80 \pm 1.96(.038)$$

$$.80 \pm .0745$$

$$.726 \leftrightarrow .875$$

Note: Because the range is slightly smaller, the prediction may be more useful.

III. Determining sample size

- A. A small sample may give an inadequate answer (too large a confidence interval).
- B. A large sample requires excess time and money.
- C. Three factors are used to determine an appropriate sample size.
 1. The population variance (σ^2)
 - a. A large population variance means a larger sample is needed to yield acceptable results.
 - b. If the population variance is not known, it may be estimated with a small preliminary survey.
 2. The required degree of confidence (z)
 - a. A given confidence interval (90%) has a matching **degree of confidence**. In the long run, there is a 90% degree of confidence that the population parameter being measured will fall within the 90% confidence interval.
 - b. A higher degree of confidence requires a larger sample.
 3. The amount of acceptable error (E)
 - a. A study will have some logical acceptable range for the confidence interval.
 - 1) Income may be estimated to within \$500 of the mean.
 - 2) A part's size may be estimated to within .01 millimeters.
 - b. A small acceptable error requires a larger sample.
- D. Sample size determination when estimating the population mean
 1. Solving $\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$ for n gives the following sample size formula.

$$n = \left(\frac{z\sigma}{E} \right)^2$$

Note: A large degree of confidence, a large variance, and a small acceptable error all make the sample size larger.

3. Suppose Linda was unhappy with the average customer purchase range first described on page 67 and summarized below. How large a sample would be required to lower the acceptable error from \$.26 to \$.10? Assume the finite correction factor is not applicable.

Problem Review

Given:
 $\bar{x} = \$7.50$
 z for .99 is 2.58
 $\sigma = \$0.70$
 $n = 49$

$$\begin{aligned} \bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}} \\ \$7.50 \pm 2.58 \frac{\$.70}{\sqrt{49}} \\ \$7.50 \pm \$.258 \\ \$7.24 \leftrightarrow \$7.76 \end{aligned}$$

$$\begin{aligned} n &= \left(\frac{z\sigma}{E} \right)^2 \\ &= \left[\frac{(2.58)(.7)}{.1} \right]^2 \\ &= [18.06]^2 = 326.16 \rightarrow 327 \end{aligned}$$

Note: always round up

4. Check your answer by calculating the confidence interval using the new sample size. If the interval is acceptable (within \$.10), conduct your new survey with a sample of 327.
5. When determining the sample size for both mean and proportion problems, answers less than 30 should be rounded up to 30 because sample size formulas are based upon a normal population.

$$\begin{aligned} \bar{x} \pm Z \frac{\sigma}{\sqrt{n}} \\ \bar{x} \pm 2.58 \frac{\$.70}{\sqrt{327}} \\ \bar{x} \pm .09987 \\ \text{and } .09987 < .10 \end{aligned}$$

E. Sample size determination when estimating the population proportion

1.
$$n = \bar{p}(1 - \bar{p}) \left(\frac{z}{E} \right)^2$$

2. Using the problem II data from the previous page, Linda would like to lower the acceptable error associated with the 95% confidence interval for customer satisfaction from $\pm 7.45\%$ to $\pm 5\%$. What sample size is required?
3. The sample size formula must include the page 70 finite correction factor because n/N is $> .05$.
4. From these calculations, it appears that Linda can reduce the range of the confidence interval to $\pm 5\%$ by increasing the sample size to 234.
5. If \bar{p} is not known, it may be estimated with a sample of 100. Also, using \bar{p} of .5 will give the maximum appropriate sample size.

$$\begin{aligned} n &= \bar{p}(1 - \bar{p}) \left(\frac{z}{E} \right)^2 \sqrt{\frac{N-n}{N-1}} \\ &= .80(1 - .80) \left(\frac{1.96}{.05} \right)^2 (.949) \\ &= .80(.20)(39.2)^2 (.949) \\ &= 233.3 \rightarrow 234 \end{aligned}$$

Practice Set 12 Sampling Distributions Part II

- I. Darin wants to know the proportion of page 68 parts passing inspection. Fifty parts were randomly selected from a recent production run of 1,000 parts and 45 passed inspection.

Data Set For People Using Statistics Software												
P	P	P	P	P	F	P	P	P	P	P	P	P
P	P	P	P	P	P	P	P	F	P	P	F	F
P	P	P	P	P	P	P	P	P	P	P	P	
P	P	P	P	P	P	P	P	P	F	P	P	

- A. Calculate the proportion of parts passing inspection.
- B. Darin would like to use last week's data to predict a range for the proportion of future production runs passing inspection. Calculate the 95% confidence interval for the proportion of parts produced by this production process passing inspection.
- C. What assumption is Darin making when using last week's data to predict future manufacturing quality?
- D. Calculate the 99% confidence interval for the proportion of parts passing inspection.
- E. What sample size is necessary to reduce acceptable error to $\pm 5\%$?

- II. Darin is also concerned about the weight of page 68 parts. It must be possible for the mean weight of parts to be ≤ 30 mg with a 99% degree of confidence. As indicated on page 68 and reviewed below, a recent test was barely successful. Darin wants to reduce error from the current $\pm .0279$ mg to $\pm .025$ mg. What sample size is required?

Page 68 Problem Review
(see page PS 68)

Given

$$\bar{x} = 30.025 \text{ mg}$$

$$n = 36$$

$$z = 2.58$$

$$s = .065 \text{ mg}$$

$$\bar{x} \pm zS_{\bar{x}}$$

$$30.025 \pm .0279$$

$$29.997 \text{ mg} \leftrightarrow 30.053 \text{ mg}$$

Note: This range indicates the population mean could be under 30 mg.

- III. Check your answer to problem II by calculating the 99% confidence interval using a sample size of 45 and a sample standard deviation of .065. Analyze the result.

- IV. How would the solution to problem III change if the sample of 45 had been taken from a population of 500 items?

- V. Recalculate the answer to problem III using the finite correction factor.

Quick Questions 12 Sampling Distributions Part II

I. Place the number of the appropriate formula next to the item it describes.

- A. Population proportion _____
- B. Standard error of the proportion _____
- C. Confidence interval for the population proportion _____
- D. Finite correction factor _____
- E. When to use the finite correction factor _____
- F. Sample size when predicting the population mean _____
- G. Sample size when predicting the population proportion _____

1.	$\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$
2.	$\sqrt{\frac{N-n}{N-1}}$
3.	$\frac{n}{N} \geq .05$
4.	$\bar{p} \pm z\sigma_{\bar{p}}$
5.	$\frac{x}{n}$
6.	$\bar{p}(1-\bar{p})\left(\frac{z}{E}\right)^2$
7.	$\left(\frac{zS}{E}\right)^2$

II. A survey of 80 New York City voters revealed 60 planned to vote in the next election. Calculate both the 99% and 95% confidence interval for the population proportion.

- A. 99% confidence interval
- B. 95% confidence interval

Data Set For People Using Statistics Software				
Y	N	Y	Y	Y
N	Y	Y	Y	Y
N	Y	Y	Y	N
Y	N	N	Y	Y
Y	Y	Y	N	Y
Y	N	Y	Y	N
Y	Y	N	Y	N
Y	Y	Y	Y	N
Y	Y	N	Y	Y
Y	Y	Y	Y	Y
N	Y	Y	Y	Y
Y	Y	N	Y	N
Y	Y	Y	N	Y
Y	Y	Y	N	Y
Y	Y	N	Y	Y
Y	N	Y	Y	Y

C. Using the same data, calculate the 99% confidence interval assuming the results came from a city of 1,500 voters.

III. Restaurant customers leave a tip approximately 70% of the time. A 95% confidence interval for the tips proportion is desired. The answer should be correct within $\pm 5\%$. How many customers must be surveyed? Computer users set s to $\sqrt{.21} = .458$.

IV. Linda will consider opening a new video showcase in towns with average family income over \$35,000. She requires a 99% confidence interval. The estimate should be within \$1,000 of the population mean. Recently gathered data indicates the population standard deviation is \$4,000. What size sample is required?

Can you believe it's time to review again?
Begin with the Formula Review on pages 76-77.
Then look at the relevant sections of pages 162,
164, and 166.



Probability Formula Review

I. Types and characteristics of probability

A. Types of probability

1. Classical: $P(A) = \frac{A}{N}$

2. Empirical: $P(A) = \frac{A}{n}$

3. Subjective: Use empirical formula assuming past data of similar events is appropriate.

B. Probability characteristics

1. Range for probability: $0 \leq P(A) \leq 1$

2. Value of complements: $P(\bar{A}) = 1 - P(A)$

II. Probability rules

A. Addition is used to find the sum or union of 2 events.

1. General rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

2. Special rule: $P(A \text{ or } B) = P(A) + P(B)$ is used when events are mutually exclusive.

B. Multiplication is used to determine joint probability or the intersection of 2 events.

1. General rule: $P(A \text{ and } B) = P(A) \times P(B | A)$

2. Special rule: $P(A \text{ and } B) = P(A) \times P(B)$ is used when the events are independent.

Note: For independent events, the joint probability is the product of the marginal probabilities.

C. Bayes' theorem is used to find conditional probability.

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(A) \times P(B|A) + P(\bar{A}) \times P(B|\bar{A})}$$

Note: The denominator is when condition B happens. It happens with A and with \bar{A} .

III. Counting rules

A. The **counting rule of multiple events**: If one event can happen M ways and a second event can happen N ways, then the two events can happen (M)(N) ways. For 3 events, use (M)(N)(O).

B. **Factorial rule** for arranging all of the items of one event: N items can be arranged in N! ways.

C. **Permutation rule** for arranging some of the items of one event:
(order is important: a, b, c and c, a, b are different)

$${}_N P_R = \frac{N!}{(N-R)!}$$

D. **Combination rule** for choosing some of the items of one event:

(order is not important: abc and cba are the same and are not counted twice)

$${}_N C_R = \frac{N!}{(N-R)!(R!)}$$

IV. Discrete probability distributions

A. Probability distributions

1. $P(x) = [x \bullet P(x)]$ is calculated for each value of x.

2. Mean of a probability distribution: $\mu = E(x) = \sum [x \bullet P(x)]$

3. Variance of a probability distribution: $V(x) = [\sum x^2 \bullet P(x)] - [E(x)]^2$

B. Binomial distributions

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x} \quad \text{where}$$

n is number of trials	x is number of successes
p is probability of success	q, the probability of failure, is 1 - p
$\mu = np, \sigma^2 = npq \text{ and } \sigma = \sqrt{npq}$	

C. Poisson distributions

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} \quad \text{where } \mu = np$$

Poisson approximation of the binomial requires $n \geq 30$ and $np < 5$ or $nq < 5$.

V. The continuous normal probability distribution

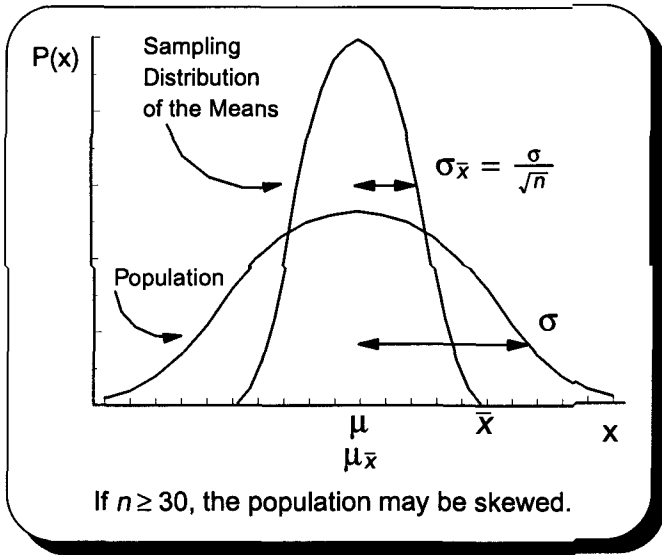
A. To find the probability of x being within a given range:

$$Z = \frac{x - \mu}{\sigma}$$

Normal approximation of the binomial requires $n \geq 30$ and both np and nq are ≥ 5 . The continuity correction factor applies.

B. To find a range for x given the probability: $\mu \pm Z\sigma$

VI. Central limit theorem



VII. Point estimates

- A. \bar{x} for μ B. s for σ C. \bar{p} for p D. $S_{\bar{x}}$ for $\sigma_{\bar{x}}$ where $S_{\bar{x}} = \frac{s}{\sqrt{n}}$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

VIII. Interval estimates when $n \geq 30$

A. For a population mean $\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$ or $\bar{x} \pm z \frac{s}{\sqrt{n}}$

B. For a population proportion $\bar{p} \pm z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$ where $\bar{p} = \frac{x}{n}$

Note: Use the finite correction factor in section VIII formulas when $n/N \geq .05$. $\sqrt{\frac{N-n}{N-1}}$

Section VIII Note: When $n < 30$ and σ is unknown, the t distribution, to be discussed in chapter 16, must be substituted for the z distribution when making interval estimates. Many statistics software programs do all interval calculations, regardless of sample size, using the t distribution.

IX. Determining sample size

A. When estimating the population mean $n = \left(\frac{z\sigma}{E}\right)^2$

B. When estimating the population proportion $n = \bar{p}(1 - \bar{p})\left(\frac{z}{E}\right)^2$

Probability Test

- I. Average hours worked by manufacturing workers is normally distributed with a mean of 41 hours and a standard deviation of .5 hours. Graph and solve the following problems.
- A. $P(41 \text{ hours} \leq x < 42.5 \text{ hours})$
 - B. $P(x < 40.345 \text{ hours})$
 - C. $P(41.75 \text{ hours} \leq x < 42 \text{ hours})$
 - D. $P(39.5 \text{ hours} \leq x < 42.5 \text{ hours})$
- II. Study time at State University is normally distributed with a mean of 15 hours per week and a standard deviation of 3 hours. Graph and solve the following problems.
- A. How many hours must a student study to be in the top 1% of the students attending State University?
 - B. Calculate the fourth decile.

III. Answer the following questions based upon this study of money spent on souvenirs at a virtual reality theme park.

Age	Money spent on souvenirs	Under \$5	\$5 and over	Totals
Under 22		5	15	20
22 and older		20	20	40
Totals		25	35	60

A. Use a formula to calculate the $P(\text{Age} < 22 \text{ or } \text{Age} \geq 22)$.

B. The events in question A are _____ and therefore, the _____ rule for _____ is applicable.

C. Use a formula to calculate the probability of someone being at least 22 years old and spending \$5 and over.

D. Question C required the _____ rule for _____ because the events are _____.

E. Use Bayes' theorem to calculate the probability of someone at least 22 years old spending \$5 or more.

F. Using the above chart, calculate the probability of someone at least 22 years old spending less than \$5.

G. Why does your answer to question F make sense?

- IV. Use a formula to calculate the probability of tossing a coin 3 times and getting exactly 3 heads. What is the probability of a head coming up on the fourth toss?
- V. Four customers have three bank branches and you will visit the manager and assistant manager at each branch. How many managers and assistant managers will you visit?
- VI. A salesperson must visit 4 of 6 stores and order is important. That is, AB and BA represent different routes. How many routes are available to the salesperson?
- VII. Redo problem VI assuming order does not count. AB and BA are the same and count as one route. Be sure to use a formula and show all work.
- VIII. How many different 3-person subcommittees can be chosen from an 8-person committee?
- IX. Three of 8 committee members must be chosen to give a speech. All 8 have very different personalities and order is important. How many different speaker arrangements are possible?

X. How many 4-place random numbers can be generated from 10 digits? Repeating digits is allowed.

XI. Six parts are to be inspected from a production process designed to have approximately 5% defective parts. Using the binomial formula, determine the probability of zero defects. Use a table to determine the probability of at least 2 defective parts. State the entire probability distribution. What is the probability of 2 defective parts?

XIII. Place the number of the appropriate item in the space provided.

- A. Standard error of the mean _____
- B. 99% confidence interval _____
- C. Standard error of the proportion _____
- D. Requires n be ≥ 30 _____
- E. Acceptable error _____

1.	$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$
2.	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
3.	$\bar{x} \pm 2.58 \frac{s}{\sqrt{n}}$
4.	E
5.	$\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

XIV. Answer the following true or false and fill in the blank questions.

- A. The standard error of the mean will be halved if the sample size is doubled. _____
- B. Sampling error exists because a nonrepresentative sample was taken in place of a census. _____
- C. A one-number estimate of the population mean is called a _____ estimate of the mean.
- D. A range for a population parameter is called the _____.
- E. A _____ may be more accurate than a simple random sample because a small diverse section of the population might not be represented in a simple random sample.

XV. A sample of 36 out of 25,000 baseball fans attending a game revealed average refreshment spending of \$7.60. The standard deviation for the population is \$2.10. Calculate the 95% confidence interval for average refreshment spending by fans attending this game.

4.50	8.00	9.00	9.00
6.95	4.90	7.00	8.05
10.00	8.00	9.50	2.00
11.00	9.00	5.00	8.00
8.05	8.50	10.00	4.80
6.00	4.90	11.00	9.00
6.50	7.00	7.00	8.00
11.00	8.00	5.00	5.75
9.10	6.00	9.10	9.00

XVI. A marketing test of chocolate flavored shaving cream revealed a favorable response from 35 of 50 test subjects. Test subjects were chosen at random from the company's 1,200 employees. Calculate the following:

- A. The 90% confidence interval for this market test.
- B. The company is unhappy with the confidence interval calculated above and would like to lower acceptable error from 11% to 5%. How large a sample must be taken?

U	F	F	F	F
F	U	F	F	U
U	F	U	F	F
U	F	F	F	U
F	U	F	F	F
U	F	F	U	F
F	F	F	F	F
U	F	F	U	U
F	F	F	F	F
F	F	F	U	U

XVII. Match each item on the right with the concept it defines.

1. Bayes' theorem _____
2. Addition rule when events are mutually exclusive _____
3. Variance of a binomial probability distribution _____
4. Factorial rule for arranging all of the items of one event _____
5. Range for probability _____
6. Multiplication rule when the events are independent _____
7. Empirical probability _____
8. Subjective probability _____
9. General rule for addition _____
10. Permutation rule _____
11. To find a range given the probability _____
12. Classical probability _____
13. Mean of a probability distribution _____
14. Value of a complement _____
15. For independent events _____
16. Binomial distribution _____
17. To find the probability given a range _____
18. Combination rule _____
19. Poisson distribution _____
20. The complement of A _____
21. Variance of a probability distribution _____
22. The counting rule for multiple events _____
23. Is calculated for each value of x when determining a probability distribution _____
24. Mean of a binomial probability distribution _____
25. General rule for multiplication _____

1.	$P(A) + P(B)$
2.	$\frac{N!}{(N-R)!}$
3.	$M \times N$
4.	$\frac{\mu^x e^{-\mu}}{x!}$
5.	$0 \leq P(A) \leq 1$
6.	$\frac{x-\mu}{\sigma}$
7.	Joint probability is the product of the marginal probabilities
8.	$\mu \pm Z\sigma$
9.	$\frac{N!}{(N-R)!(R!)}$
10.	$\frac{A}{n}$
11.	(\bar{A})
12.	$1 - P(A)$
13.	$\frac{A}{N}$
14.	$P(A) + P(B) - P(A \text{ and } B)$
15.	$P(A) \times P(B)$
16.	$N!$ ways
17.	$[\sum x^2 \cdot P(x)] - [E(x)]^2$
18.	$\frac{n!}{x!(n-x)!} p^x q^{n-x}$
19.	npq
20.	$\frac{P(A) \times P(B A)}{P(A) \times P(B A) + P(\bar{A}) \times P(B \bar{A})}$
21.	$x \cdot P(x)$
22.	$P(A) \times P(B A)$
23.	Use empirical formula assuming past data of similar events is appropriate
24.	np
25.	$\sum [x \cdot P(x)]$

Chapter 13 Large Sample Hypothesis Testing

I. Introduction

- A. Chapter 13 explores a systematic method for testing claims about the population mean using a sample mean.
- B. Large sample ($n \geq 30$) tests using z will be considered. The standard deviation (σ) may be known or unknown.
- C. Small sample ($n < 30$) t distribution tests used by most statistics software will be explored in chapter 16.
- D. Issues to be tested include
 - 1. Quality control issues such as the weight of a computer part
 - 2. Marketing research issues such as the proportion of consumers liking a new product
 - 3. Political issues such as the proportion of voters planning to vote for a political candidate

II. Definitions

- A. **The null hypothesis (H_0)** states some hypothesized value for a population parameter such as the mean.
 - 1. Read "H sub-zero," its acceptance implies **no statistical difference** between a parameter(μ) and a statistic(\bar{x}).
 - 2. Linda Smith wants to know whether the average customer purchase has decreased from last year's mean of \$7.75 because a recent sample of 49 had a mean of only \$7.50 (see page 67).
 - a. A null hypothesis might read "the average purchase has not decreased from \$7.75."
 - b. In effect, $H_0 : \mu \geq \$7.75$
 - 3. The direction of the inequality is greater than or equal to because this implies the mean has not decreased.
 - 4. H_0 is rejected if the measured difference between the hypothesized μ and \bar{x} is large and seldom happens.
- B. **The alternate "research" hypothesis (H_1)** represents the possible difference being studied.
 - 1. Read "H sub-one," it implies **there is a statistical difference**. It is the complement of the null hypothesis. $H_1 : \mu < \$7.75$
 - 2. An alternate hypothesis might read "the mean purchase is under \$7.75."

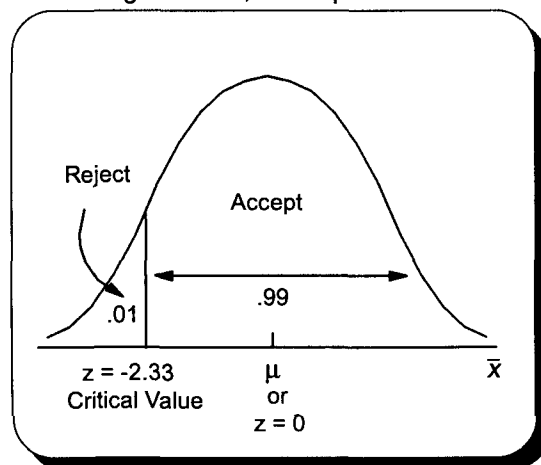
C. Level of significance

- 1. Rejection of a true null hypothesis should rarely happen.
 - a. The level of significance states the maximum probability of such an error.
 - b. A .01 significance level indicates a sample statistic at least this different from some hypothesized parameter will happen no more than 1% of the time. Therefore, the maximum error is one percent.
 - c. The significance level provides a limit for the sample statistic. Beyond this limit, H_0 is rejected.
 - d. The cost associated with making an incorrect decision determines the appropriate level of significance.
- 2. **Type I or alpha error (α)**
 - a. Alpha error equals the level of significance. It measures the risk of rejecting a true null hypothesis.
 - b. Deciding to reject the null hypothesis about the average purchase of \$7.75 creates the possibility of type I error (accepting a decrease when there is not a decrease).
 - c. Traditional alpha errors include .05 for marketing research questions and .01 for quality control questions.
- 3. **Type II or beta error (β)**, accepting a false null hypothesis, is examined on page 89.

Error Summary		
Decision	Nature's True State	
	H_0 is true	H_0 is false
Accept H_0	Correct	Type II error
Reject H_0	Type I error	Correct

D. Test statistics and their critical values

- 1. Test statistics are used to determine the validity of a null hypothesis. Examples include \bar{x} and \bar{p} .
- 2. Here, \bar{x} will be used to test a null hypothesis concerning population mean purchases described above.
- 3. **We begin by assuming the null hypothesis is true.** For the .01 level of significance, a sample mean that separates 1% of the sampling distribution's sample means from the other 99% will be the critical value.
- 4. When testing a null hypothesis related to a normal sampling distribution, the test statistic is often converted into its z value. This z value is like the **critical value** because it separates the region of acceptance from the region of rejection.
- 5. Here we have a critical value for z of -2.33 for the .01 level of significance as $.49 \rightarrow z = -2.33$. This means $\leq 1\%$ of the sample means are beyond -2.33 standard deviations from μ and result in the error of rejecting a true null hypothesis.
- 6. The alternate hypothesis points toward the region of rejection. In this **one-tail problem**, with an H_1 of $\mu < \$7.75$, the critical area is to the left because Linda is concerned that a low sample mean of \$7.50 indicates the population mean has decreased.



III. A 5-step approach to hypothesis testing

- A. State the null hypothesis and alternate hypothesis.
 1. Determine the condition (claim, concern, difference) being tested using $>$, $<$, or \neq . Call it H_1 .
 2. Determine the condition's complement using \leq , \geq , or $=$. Call it H_0 .
 3. H_0 implies no difference by containing an equality sign. It is stated first.
- B. Select the level of significance based upon acceptable type I error.
- C. Determine the relevant test statistics (\bar{x} for now, \bar{p} and others will follow).
- D. Determine the decision rule using a graph of the critical values of z .
 1. Accept the null hypothesis if the test statistic z is not beyond the critical value of z .
 2. Otherwise, reject the null hypothesis.
- E. Apply the decision rule.

Simply put, if the test statistic is extreme enough, beyond the critical value, reject the null hypothesis.

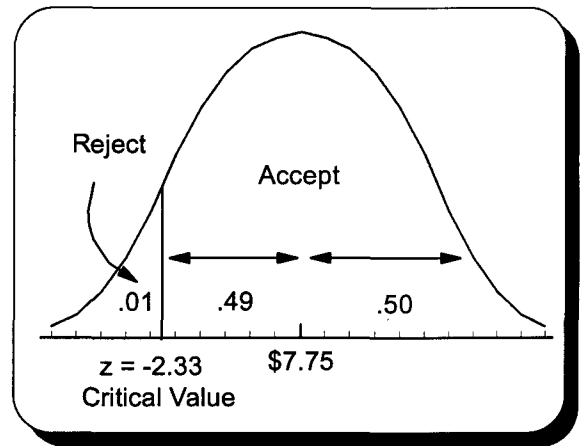
IV. One-tail testing of one sample mean

Linda Smith thinks average customer purchases could be lower than last year's \$7.75 because a sample of 49 (see page 67) had a mean of only \$7.50. The population standard deviation is \$.70. Linda wants type I error, the chance of rejecting a true null hypothesis, to be 1%.

- A. $H_0 : \mu \geq \$7.75$ and $H_1 : \mu < \$7.75$
- B. Type I error is 1%.
- C. The test statistic is \bar{x} .
- D. If z from the test statistic is beyond -2.33, reject the null hypothesis.
- E.

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\$7.50 - \$7.75}{\frac{\$.70}{\sqrt{49}}} = \frac{-\$0.25}{\frac{\$.70}{7}} = \frac{-0.25}{.10} = -2.50$$

Reject the null hypothesis because a z of -2.50 is beyond (smaller) the critical value of -2.33. A sample mean of \$7.50 happens less-than 1% of the time when $\mu \geq \$7.75$.

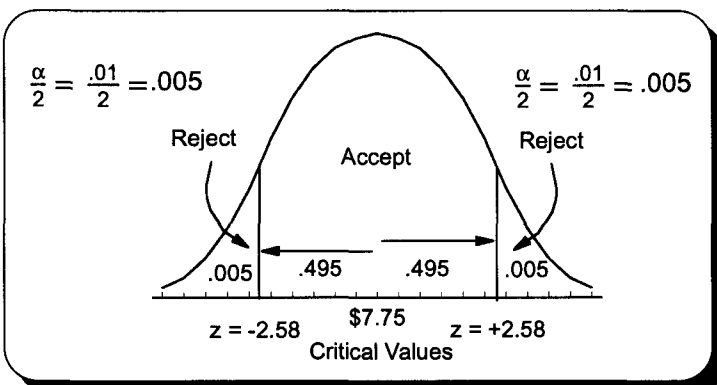


Note: If the area beyond the test statistic (the tail) is less-than the level of significance, the measured difference is significant and H_0 is also rejected. This approach, called p-value hypothesis testing, is used by most statistics software. After completing this page, statistics software users should read part II of page 88 and chapter 16.

V. Two-tail testing of one sample mean

- A. **Two-tail problems** concern any change, regardless of direction.
- B. In the problem above, Linda was not concerned about the average purchase going up. Now she is. The claim concerning the average purchase must be changed to include any difference from last year's average purchase of \$7.75. The null hypothesis and alternate hypothesis would be:

$H_0 : \mu = \$7.75$ $H_1 : \mu \neq \$7.75$
- C. For this two-tail problem, the alternate hypothesis does not state the direction of the change (difference).
- D. Using a .01 level of significance, alpha risk must be divided evenly between the 2 tails of a normal curve.



The test statistic remains ± 2.50 . The analysis to the left indicates the critical value has changed to ± 2.58 . **Accept** the null hypothesis as z of -2.50 is not beyond the critical value of -2.58. At the .01 level of significance, a sample mean of \$7.50 is not low enough to conclude the population mean is not \$7.75. Note how splitting the .01 level of significance (risk) between two tails increases the critical value. As a result, what was a significant difference is now an acceptable difference.

Practice Set 13 Large Sample Hypothesis Testing

- I. Darin Jones is very concerned that parts designed to weigh less than or equal to 30 mg may be too heavy and not pass inspection. From page 68, we know that a sample of 36 parts resulted in a sample mean of 30.025 mg and a sample standard deviation of .065 mg. Darin wants to control type I error (the probability of deciding the parts that are too heavy when they are not) to the .01 level of significance. Solve this problem using the 5-step approach to hypothesis testing. **Special Note: We know the population mean can be less than or equal to 30 mg at the .011 level of significance because the 98% confidence interval calculated for this population mean on page 68 had a lower limit of 29.999 mg.**
- II. Using problem I data and a .01 level of significance, determine whether the population mean has changed from 30 milligrams.
- III. Redo problem II using a .05 level of significance.

Quick Questions 13 Large Sample Hypothesis Testing

I. Complete the following chart and questions.

Error Summary		
Decision Concerning Null Hypothesis	Nature's True State	
	H_0 is true	H_0 is false
Accept H_0		
Reject H_0		

- A. Type I error is called _____ error.
- B. Type II error is called _____ error.
- C. When z calculated from sample data is beyond the critical value (less than for left tail problems and greater than for right tail problems), the null hypothesis is _____.

D. T F By setting the confidence level to 99%, we are trying to assure that the alternate (research) hypothesis will not be easily accepted.

II. Make these tests using the 5-step approach to hypothesis testing.

- A. A light bulb warranty states average bulb life is at least 20,000 hours. A sample of 49 bulbs had an average life of 19,000 hours. The population standard deviation is 1,400 hours. Test the warranty claim to the .01 level of significance.

For People Using Statistics Software				
Life of Light Bulbs (Thousands of Hours)				
19	17	18	19	19
20	19	21	20	22
20	19	19	21	19
19	18	19	17	19
19	19	19	16	20
19	20	17	19	18
18	18	21	17	18
20	21	18	16	21
17	19	20	22	19
20	18	20	18	

- B. Average weekly manufacturing earnings were \$480 and the standard deviation was \$72. A recent sample of 36 resulted in a mean of \$450. The standard deviation has not changed. Test to the .05 level whether average weekly earnings changed.

For People Using Statistics Software					
Weekly Manufacturing Earnings					
500	520	490	580	470	475
565	610	490	420	480	400
445	580	300	440	450	480
400	420	480	410	440	430
390	480	390	460	460	450
420	385	350	500	360	280

Chapter 14 Large Sample Hypothesis Testing Part II

I. Two-tail testing of two sample means from independent populations

A. Variables are independent when the occurrence of one variable does not affect the value of the other variable.

B. Linda is interested in whether the average customer purchase is different at two of her stores.

1. A sample of 50 from store #1 had a mean of \$7.50 and a standard deviation of \$1.00.
2. A sample of 32 from store #2 had a mean of \$7.40 and a standard deviation of \$.80.

C. The 5-step approach to hypothesis testing

1. State the null and alternate hypothesis.

a. This is a two-tail problem because the claim involves any difference in average purchase.

b. $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$

2. Since the claim is marketing oriented, the test will be at the .05 level of significance.

3. Determine the relevant test statistics.

a. \bar{x} is the relevant test statistic.

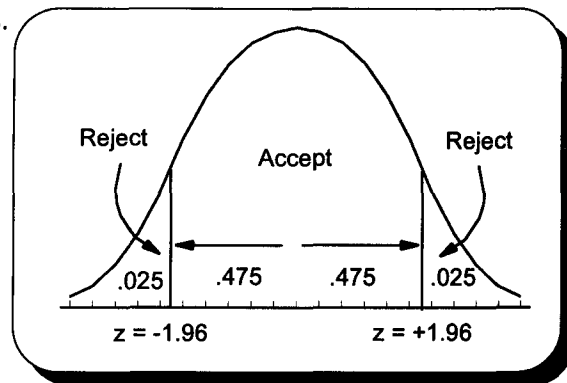
b.
$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Note: If the difference between the two sample means is large relative to their average standard errors, z for the test will be larger than the critical value of z and the null hypothesis will be rejected.

4. Determine the decision rule using a graph of the critical values.

The critical value of z for $\alpha/2 = .05/2 = .025$ is ± 1.96 .
If z from the test statistic is beyond ± 1.96 the null hypothesis will be rejected.

Note: This would be a one-tail problem if Linda wanted to know whether one store had a larger average purchase than the other store.



5. Apply the decision rule.

Store # 1	$n_1 = 50$	$\bar{x}_1 = \$7.50$	$s_1 = \$1.00$
Store # 2	$n_2 = 32$	$\bar{x}_2 = \$7.40$	$s_2 = \$.80$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{7.50 - 7.40}{\sqrt{\frac{(1.00)^2}{50} + \frac{(.80)^2}{32}}} = \frac{.10}{\sqrt{.02 + .02}} = \frac{.10}{.2} = .50$$

Accept H_0 because $.50 < 1.96$.
Sales are the same at the .05 level of significance.

II. Hypothesis testing using p-values

A. The p-value approach to hypothesis testing compares the probability associated with the test statistic's tail or tails (p) with the level of significance. P measures the significance of the test data.

1. If the p-value is smaller than the level of significance, the probability of a test statistic this extreme is unlikely (less than the level of significance), and the null hypothesis is rejected.
2. A small p-value (a tail of .003) means substantial difference and H_0 is rejected.
3. A large p-value (a tail of .30) means little difference and H_0 is easily accepted.

B. For example, a p-value analysis of the one-tail and two-tail problems on page 85, where z for the test statistic was 2.50 and the level of significance was .01, would be done as follows.

One-tail Problem

$z = 2.50 \rightarrow .4938 \rightarrow (.5000 - .4938) = .0062$
Reject H_0 because p of $.0062 < .01$.

Two-tail Problem

$z = 2.50 \rightarrow .4938 \rightarrow (.5000 - .4938) = .0062$
Because this is a two-tail problem, $\alpha/2 = .01/2 = .005$.
Accept H_0 because p of $.0062 > .005$.

III. Analyzing type II error (β)

- A. Type II error is the probability of accepting a false null hypothesis.
 B. Linda's two-tailed study concerning any change in the average purchase price from last year's \$7.75 (see page 85) will be analyzed. First we will calculate the lower critical value, an accept/reject point for this null hypothesis.

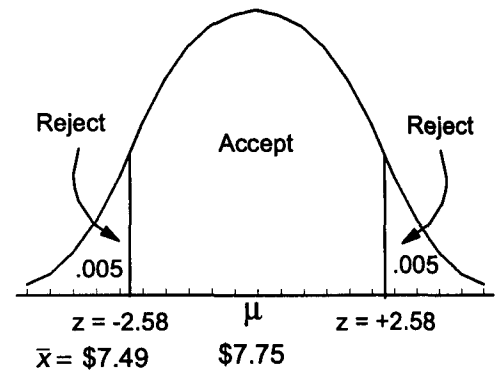
$$\alpha/2 = .01/2 = .005$$

$$.50 - .005 = .495 \rightarrow z = \pm 2.58$$

$$\mu - z(\sigma_{\bar{x}})$$

$$\$7.75 - 2.58(\$.10)$$

$$\$7.75 - \$.258 = \$7.49$$



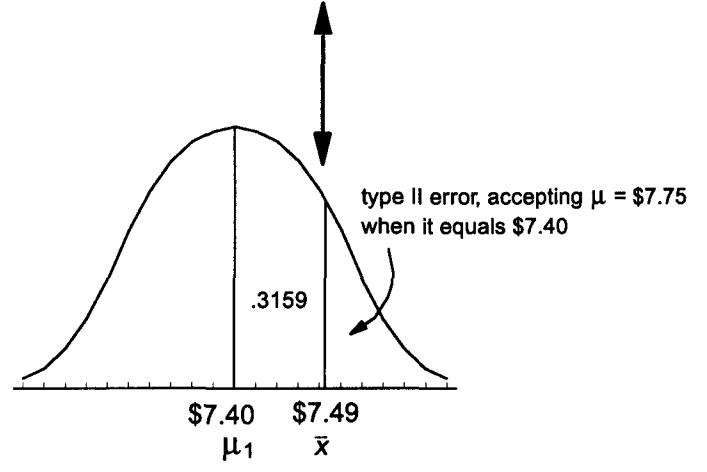
- Here, type II error exists everywhere except for $\mu = \$7.75$.
- This means the amount of type II error varies depending upon the value of the true population mean.
- We will calculate the probability of type II error for a population mean (μ_1) of \$7.40.

$$Z = \frac{\bar{x} - \mu_1}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{\$7.49 - \$7.40}{\frac{\$.70}{\sqrt{49}}} = \frac{\$.09}{\$.10} = .90 \rightarrow .3159$$

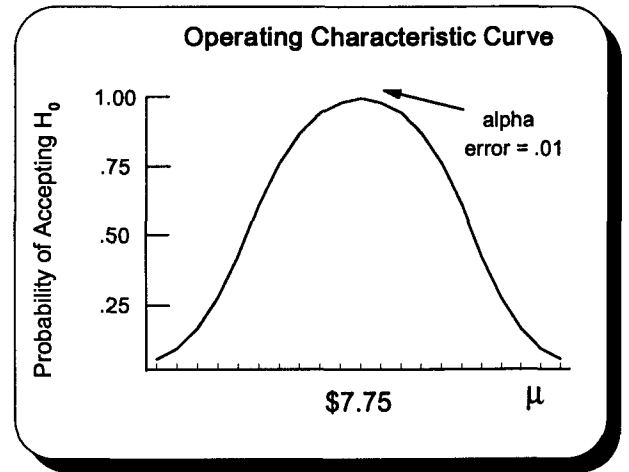
type II error is $.50 - .3159 = .1841$

When the mean is \$7.40, Linda's decision rule has a type II error of 18.41%.



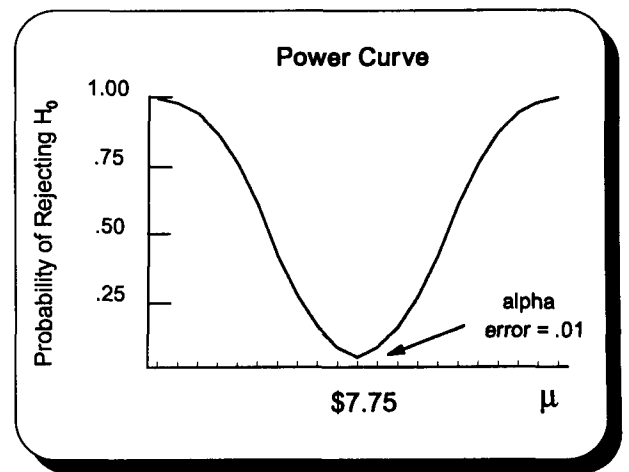
C. Operating characteristic curves

- The operating characteristic curve graphs the probability of type II error. It depicts all possible type II errors given some acceptable level of type I error. It measures accepting no change when there has been change.
- As the true population mean in the above example drops, accepting a false null hypothesis becomes less likely as the right tail area of the second graph becomes smaller. Eventually the true population mean is so small that accepting a false null hypothesis is almost impossible.
- As the true mean approaches \$7.75, the area to the right gets larger. It reaches a peak of 98+ percent just before \$7.75. Type II error does not exist for $\mu = \$7.75$ because the null hypothesis is not false.
- At a point just beyond \$7.75, beta error is still 98+ percent and it drops toward zero as the true population mean increases.



D. Power curves

- A power curve graphs the probability of not making a type II error. It measures:
 - how often you correctly reject a false null hypothesis
 - how often you accept a correct research hypothesis
- It is the complement of type II error or $1 - \text{type II error}$.
- The power curve shows accepting a change in quality, consumer attitude, and voter preference when there has been changes in these areas.
- Lowering type II error comes at the expense of increasing type I error and vice versa.



Practice Set 14 Large Sample Hypothesis Testing Part II

- I. Darin buys material for his 30-milligram parts from suppliers A and B. A sample of 30 orders placed with supplier A had a mean delivery time of 24 days and a standard deviation of 9 days. A sample of 40 orders placed with supplier B had a mean delivery time of 27 days and a standard deviation of 10 days. Using a .05 level of significance, determine whether these suppliers have different mean delivery times.

Supplier A: 10, 22, 14, 39, 37, 40, 30, 29, 30, 16, 11, 27, 32, 32, 26, 26, 29, 24, 29, 19, 10, 19, 22, 12, 17, 31, 26, 35, 11, 15,

Supplier B: 14, 37, 20, 19, 12, 18, 22, 23, 26, 21, 19, 39, 34, 27, 34, 40, 17, 41, 35, 26, 11, 42, 25, 29, 36, 17, 21, 42, 10, 37, 31, 38, 27, 38, 34, 13, 40, 22, 11, 32

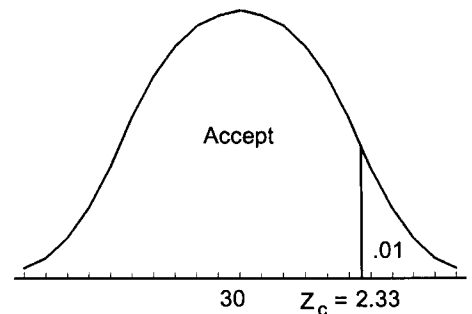
- II. Darin has decided to determine the p-value associated with the test of the 30-milligram parts conducted in problem 1 on page 86. This data was first analyzed on page 68.

Problem Review

Given: $\bar{x} = 30.025$ mg, $n = 36$, $s = .065$ mg, and $\alpha = .01$

$H_0 : \mu \leq 30.00$ mg $H_1 : \mu > 30.00$ mg

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{(30.025 - 30.000)}{\frac{.065}{\sqrt{36}}} = 2.315 < 2.33, \text{ accept } H_0$$



Note: c is for critical value.

- A. Calculate the p-value associated with this study.

B. Use this p-value to accept or reject the null hypothesis. Does your answer agree with the page 86 answer?

C. What does this p-value indicate is the strength or validity of the decision made concerning the null hypothesis?

III. Past experience indicates that the population mean weight of material containers used to make computer parts is 5,000 kilograms. The standard deviation is 28 kilograms. Type I error for a sample of 49 will be controlled to the .01 level of significance. The 99% confidence interval is 4,989.68 kilograms to 5,010.32 kilograms.

A. Calculate the type II error for a two-tail problem using each of these possible population means.

$$\mu = 4,985 \text{ kg}$$

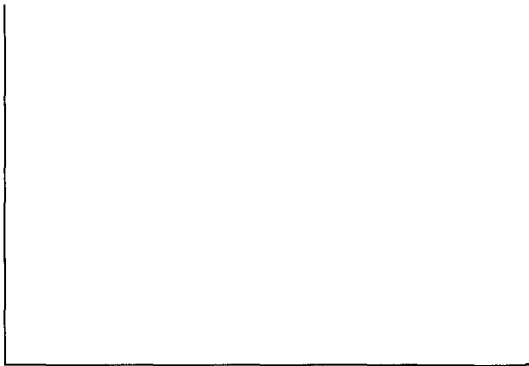
$$\mu = 4,995 \text{ kg}$$

$$\mu = 5,000 \text{ kg}$$

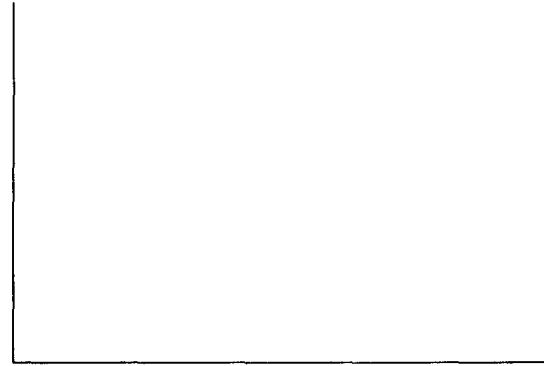
$$\mu = 5,005 \text{ kg}$$

$$\mu = 5,015 \text{ kg}$$

B. Using the data calculated in problem A, sketch and label an operating characteristic curve.



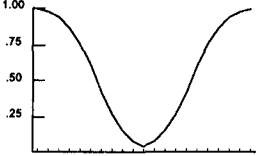
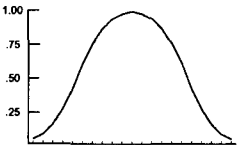
C. Using the data calculated in problem A, sketch and label a power curve.



Note: An operating characteristic curve and power curve for a one-tail problem is limited to one side of the population mean. Both look like half a normal curve stopping at the mean.

Quick Questions 14 Large Sample Hypothesis Testing Part II

I. Place the number of the description next to the item it describes.

1. Area beyond the test statistic	2. $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	3. 	4. 
-----------------------------------	---	---	--

A. Power curve _____

C. Z for testing two means _____

B. P-value _____

D. Operating characteristics curve _____

II. Ace Realty wants to determine whether the average time it takes to sell homes is different for its two offices. A sample of 40 from office #1 revealed a mean of 90 days and a standard deviation of 15 days. A sample of 50 from office #2 revealed a mean of 100 days and a standard deviation of 20 days. Use a .05 level of significance.

For People Using Statistics Software						
Days to Sell a Home						
Office #1			Office #2			
52	95	89	129	108	57	60
80	102	90	64	94	58	123
63	94	90	110	93	63	83
106	91	91	87	109	74	117
80	99	93	127	106	106	137
105	89	95	78	98	116	90
86	83	123	93	93	120	110
85	119	82	90	103	122	118
86	58	98	100	124	124	
100	75	103	84	100	110	
108	70	69	98	92	74	
90	80	107	127	106	116	
95	82		119	98	84	
90	107		93	105	110	

III. Tough Tire Company is concerned that tread life of its new all weather tire may be below the 70,000 mile warranty. A sample of 36 revealed a mean of 69,800 miles and a standard deviation of 750 miles. Using a .05 level of significance and the p-value approach, test Tough Tire's warranty claim.

For People Using Statistics Software		
Tire Mileage		
69850	71200	69700
69400	69550	70625
70150	69300	70175
70100	69950	70400
68950	68416	69150
71834	70200	70750
69904	68650	69700
69620	68850	69475
70350	70300	69300
70450	70250	68550
70200	68825	69900
68850	69725	70200

IV. The Easy Loan Company wants to determine whether the average length of car loans has increased from last year's population mean of 50 months. A sample of 49 had a mean of 53 months and a standard deviation of 14 months.

A. Test $H_0 : \mu \leq 50$ and $H_1 : \mu > 50$ at the .05 level of significance.

For People Using Statistics Software						
Length of Car Loans						
47	58	20	53	52	52	79
72	40	48	55	61	62	68
27	55	49	56	53	78	55
52	44	73	53	57	66	63
33	49	51	75	42	45	71
69	67	52	53	38	36	43
38	46	32	73	60	23	53

B. Calculate the critical value of \bar{x} .

C. Calculate type II error for $\mu = 55$ months.

D. What is the type II error for these population means?

54 months

53.31 months

50.01 months

Chapter 15 Hypothesis Testing of Population Proportions

I. Introduction

- A. The population proportion, first described on page 70, is the average part of a population having a certain characteristic.
1. The **population proportion (p)** follows a binomial probability distribution.
 2. It may be expressed as a fraction, decimal, or percentage.
 3. Important statistics

Don't forget to look ahead



Sample Proportion

$$\bar{p} = \frac{\text{\# of successes}}{\text{sample size}} = \frac{x}{n}$$

Interval Estimate for p

$$\bar{p} \pm z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

- B. Proportion tests must meet binomial experiment requirements.
1. The experiment must involve two mutually-exclusive outcomes defined as success or failure.
 2. Outcomes, which can be counted, must be independent and constant.
 3.

n is number of trials	p is probability of success	q, the probability of failure, is 1 - p
-----------------------	-----------------------------	---
- C. These proportion tests use the normal approximation of the binomial. This means both np and nq must be ≥ 5 and n must be ≥ 30 . The recommended requirement for n varies from 30-100.

II. One-tail testing of one sample proportion

- A. Linda is applying for a Flopbuster Video franchise. Flopbuster requires at least 85% of Linda's customers be happy with service at the .05 level of significance. Page 70 sample data indicated 80 of 100 customers were happy with service.
- B. Before using the normal approximation to the binomial, the appropriateness of the data must be checked.
1. Both np and nq are ≥ 5 as $(100)(.85) = 85$ and $100(.15) = 15$.
 2. The sample size of 100 is ≥ 30 .
- C. The 5-step approach to hypothesis testing
1. The null hypothesis and alternate hypothesis are $H_0 : p \geq .85$ and $H_1 : p < .85$.
 2. The level of significance will be .05 and the critical value of z is -1.645.
 3. The relevant statistic will be \bar{p} .

$$Z = \frac{\bar{p} - p}{\sigma_p} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Note: The standard error of the population proportion is based upon the hypothesized population proportion p (sometimes labeled π), and not the sample proportion.

4. Either of 2 decision rules may be used.
 - a. If z from the test statistic is beyond the critical value of z, the null hypothesis will be rejected.
 - b. If the p-value is less than the .05 level of significance, the null hypothesis will be rejected.

5. Apply the decision rule.

$$\bar{p} = \frac{x}{n} = \frac{80}{100} = .80$$

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.80 - .85}{\sqrt{\frac{.85(1-.85)}{100}}} = -1.40$$

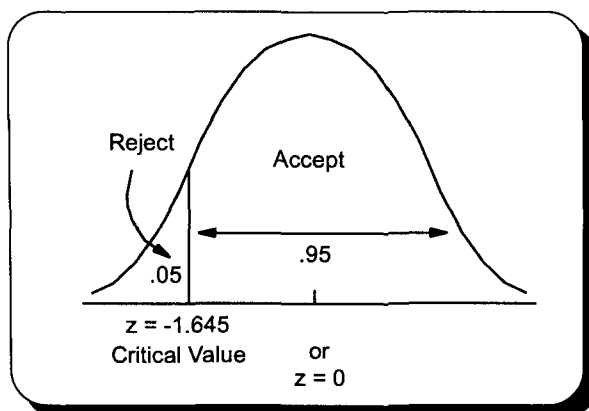
Accept H_0 because -1.40 is not beyond -1.645. Customer satisfaction is $\geq 85\%$.

The p method yields the same answer.

$$z = -1.40 \rightarrow .4192$$

$$p = .5000 - .4192 = .0808$$

Accept H_0 because $.0808 > .05$.



III. Two-tail testing of one sample proportion

- A. When any change is being measured, a two-tail problem exists.
- B. If the above problem were stated as a two-tail problem, then $H_0 : p = .85$ and $H_1 : p \neq .85$ would be appropriate.
- C. With a two-tail test, p must be doubled to $2(.0808) = .1616$. Accept H_0 because $.1616 > .05$.

IV. Two-tail testing of two sample proportions

- A. Many interesting problems involve two population proportions.
- Does consumer satisfaction differ because of gender, age, income, etc.?
 - Does machine A produce fewer defects than machine B?
 - Does taking a certain drug lower the incidence of illness?
- B. A two-tail problem
- Linda wants to know at .05 level of significance whether two of her stores have equal levels of customer satisfaction. Store #1 had 80 of 100 satisfied customers while store #2 had 45 of 50 satisfied customers.
 - The 5-step approach to hypothesis testing
 - The null hypothesis and alternate hypothesis are:
 - $H_0: p_1 = p_2$
 - $H_1: p_1 \neq p_2$
 - The level of significance will be .05 and $\alpha/2 = .05/2 = .025 \rightarrow z = \pm 1.96$.
 - The test statistic will be \bar{p} .

n_1 is sample size #1 and x_1 is successful responses from this sample.
n_2 is sample size #2 and x_2 is successful responses from this sample.
\bar{p}_1 , the sample proportion for population # 1, is $\frac{x_1}{n_1} = \frac{80}{100} = .80$.
\bar{p}_2 , the sample proportion for population # 2, is $\frac{x_2}{n_2} = \frac{45}{50} = .90$.
\bar{p}_w is the weighted or pooled estimate of the population mean.

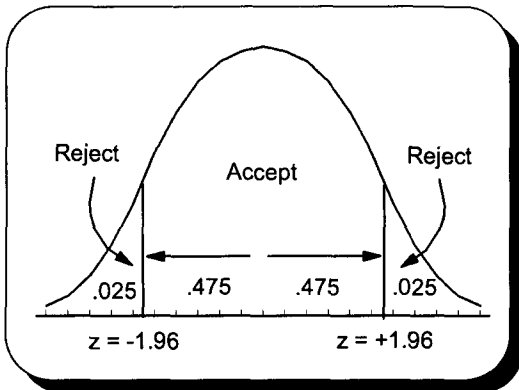
$$\bar{p}_w = \frac{\text{total successes}}{\text{total sampled}} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{\bar{p}_w(1-\bar{p}_w)}{n_1} + \frac{\bar{p}_w(1-\bar{p}_w)}{n_2}}}$$

- d. The decision rule will be, if z from the test statistic is beyond the critical value of z, the null hypothesis will be rejected.

- e. Apply the decision rule.

$$\bar{p}_w = \frac{x_1 + x_2}{n_1 + n_2} = \frac{80 + 45}{100 + 50} = .833$$



$$\begin{aligned} Z &= \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{\bar{p}_w(1-\bar{p}_w)}{n_1} + \frac{\bar{p}_w(1-\bar{p}_w)}{n_2}}} \\ &= \frac{.80 - .90}{\sqrt{\frac{.833(1-.833)}{100} + \frac{.833(1-.833)}{50}}} \\ &= -1.55 \end{aligned}$$

Accept H_0 because -1.55 is not beyond -1.96. Customer satisfaction is the same at the .05 level of significance.

The p-value method yields the same answer.

$z = -1.55 \rightarrow .4394$ and $.5000 - .4394 = .0606$ for one tail
Accept H_0 because $P = 2(.0606) = .1212$ and $.1212 > .05$.

V. One-tail testing of two sample proportions

- A. One-tail problems involve change in one direction.
- B. Doing the above problem as a one-tail problem, the question could be; does store #2 give better service?

$$H_0: p_2 \leq p_1 \text{ and } H_1: p_2 > p_1$$

1. Using z yields the following analysis.

Accept H_0 because $\alpha = .05 \rightarrow z$ of ± 1.645 and -1.55 is not beyond -1.645.

2. The p method yields the following analysis.

$z = -1.55 \rightarrow .4394$ and $p = .5000 - .4394 = .0606$
Accept H_0 because $.0606 > .05$.

Quick Questions 15 Hypothesis Testing of Population Proportions

- I. Place the number of the appropriate formula or expression next to the item it describes.
- A. When using the normal approximation to the binomial distribution,
1. np and $n(1 - p)$ must be _____
 2. n must be _____
- B. A one population test _____
- C. $\bar{p}_w =$ _____
- D. A two population test _____

1.	$\frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{\bar{p}_w(1-\bar{p}_w)}{n_1} + \frac{\bar{p}_w(1-\bar{p}_w)}{n_2}}}$
2.	≥ 30
3.	$\frac{x_1 + x_2}{n_1 + n_2}$
4.	≥ 5
5.	$\frac{\bar{p} - p}{\sigma_{\bar{p}}}$

- II. A national video publication stated long-term tape rentals average 20% of all tape rentals. A 150 customer study at *Linda's Video Showcase* revealed 24 long-term rentals. Test at the .05 level of significance whether Linda's long-term rentals are less than the national average.

For People Using Statistics Software	
Length of Video Rentals	
m	m m m m m m m m m m m
m	m m m m m m m m m m m m m m
	m m m m m m m m m m m m m
m	m m m m m m m m m m m m
m	m m m m m m m m m m m m
m	m m m m m m m m m m m m
m	m m m m m m m m m m m m m
m	m m m m m m m m m m m
m	m m m m m m m m m m m m
m	m m m m m m m m m m m

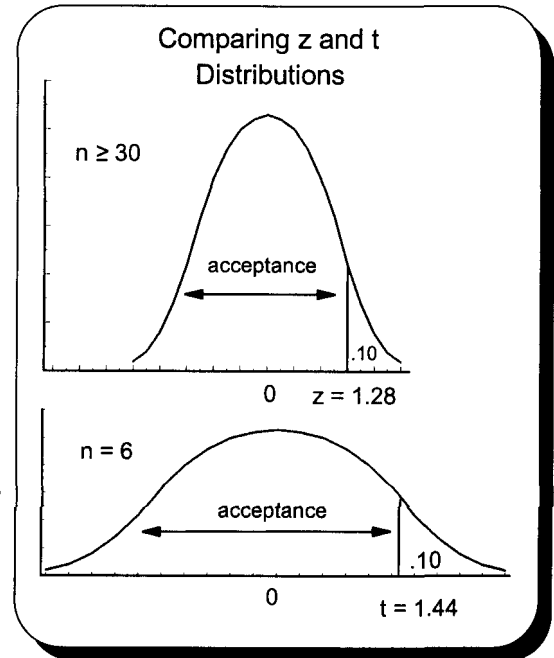
- III. Linda Smith found that 70 out of 100 customers rented 2 or more tapes at one store and 44 out of 50 rented 2 or more tapes at a second store. Test at the .05 level of significance whether there is a difference between the proportion of customers at these two stores renting 2 or more tapes.

For People Using Statistics Software	
Number of Video Rentals	
Store 1	Store 2
2 2 1 2 2 1 2 1 2 1	2 2 2 2 3
1 2 2 2 2 2 2 1 2	4 1 2 2 2
5 1 2 1 2 2 3 1 2 2	2 2 3 2 2
3 2 1 2 4 2 1 2 2 1	3 4 1 5 2
1 2 2 2 1 2 3 2 1 2	2 2 2 2 2
2 2 1 3 2 2 2 1 2 5	1 2 3 2 2
2 1 2 2 2 1 2 2 1 2	3 2 2 4 2
2 2 1 1 4 2 2 1 3 1	2 2 2 3 1
3 1 2 2 1 2 1 2 3 4	2 3 3 1 4
2 2 1 2 2 1 3 2 1 3	3 4 1 2 2

Chapter 16 Small Sample Hypothesis Testing Using Student's t Test

I. Large versus small samples

- A. The standard normal distribution (z) is appropriate for large samples ($n \geq 30$). The population may be normal or skewed.
 1. If σ is unknown, use s .
 2. For small samples, $n < 30$, z is appropriate provided the population is normal and σ is known.
- B. The student t distribution is appropriate for small samples, $n < 30$, provided the population is normal, and σ is not known (use s).
- C. Small skewed distributions will be discussed in chapter 20.



II. The t distribution's characteristics

- A. The t distribution is a family of distributions.
 1. A distribution's **degrees of freedom (df)** is determined by the number of samples involved with the distribution and the size of these samples.
 2. Degrees of freedom and level of significance determine t values.
- B. The t distribution is approximately normal and flatter than the z distribution.
- C. Values for t are larger than their corresponding z values, though the difference is negligible when n is over 29. Some statistics software use t values even when n is larger than 29.

III. One-tail testing of one sample mean using t

- A. Linda wants to know whether average tapes rented per customer has decreased from last year's mean of 2.6 tapes. A recent sample of 9 customers had a mean of 2.3 tapes and a standard deviation of .3. Test at the .01 level of significance whether average tape rentals decreased. Assume a normal distribution.
- B. The 5-step approach to hypothesis testing
 1. Here are the null hypothesis and alternate hypothesis.

$$H_0 : \mu \geq 2.6 \text{ and } H_1 : \mu < 2.6$$

2. The level of significance is .01.
3. The relevant statistic is \bar{x} .

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

4. Reject the null hypothesis when t from the test statistic is beyond t's critical value.
 - a. When testing one mean, there are $n - 1$ degrees of freedom.

$$n - 1 = 9 - 1 = 8$$

- b. The critical value of t is -2.896 for the .01 level.

5. Apply the decision rule.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{2.3 - 2.6}{\frac{.3}{\sqrt{9}}} = \frac{-.3}{\frac{.3}{3}} = -3.0$$

Reject H_0 because -3.0 is beyond -2.896.
Average tape rentals decreased.

Partial Student t Distribution

Degrees of freedom	Area from the mean to a critical value				
	0.40	0.45	0.475	0.49	0.495
df	α for a one-tail problem				
	0.10	0.05	0.025	0.01	0.005
df	α for a two-tail problem				
	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.883	2.262	2.821	3.250
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750

See page ST 4 for a more complete t table.

IV. Two-tail testing of one sample mean using t

- A. This test involves measuring change in either direction.
- B. Procedures are the same as those described in earlier chapters.

V. Two-tail testing of two sample means from independent populations

- A. Populations are independent when a sample selected from one is not related to a sample selected from the other.
- B. Examples of independent populations include production time using two different assembly procedures and industrial accidents at two plants.
- C. These tests assume the populations are approximately normal with equal variances.
 - 1. These equal variances make a weighted (pooled) point estimate the best estimate of the population σ^2 .

2. $S_w^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$ S_1^2 is the variance of sample #1 and S_2^2 is the variance of sample #2.

- D. Linda Smith wants to compare the time salespeople spend with customers at two of her stores. A sample of 6 salespeople from one store had a mean of 4.5 minutes and variance of 3. A sample of 5 from a second store had a mean of 5.1 minutes and a variance of 3.1. Linda will conduct a .05 level test to determine whether the means are the same for these normally distributed populations.

E. The 5-step approach to hypothesis testing

- 1. These are the null hypothesis and alternate hypothesis.
 - a. $H_0 : \mu_1 = \mu_2$
 - b. $H_1 : \mu_1 \neq \mu_2$
- 2. The level of significance is .05 for this two-tail test.
- 3. The relevant test statistic is \bar{x} .

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_w^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

\bar{x}_1 is 4.5, s_1^2 is 3.0, \bar{x}_2 is 5.1, and s_2^2 is 3.1.
n_1 is 6 and n_2 is 5.
S_w^2 is the weighted or pooled estimate of the population variance.
df = items tested - number of samples

- 4. Reject the null hypothesis when the test statistic is beyond the critical value.
- 5. Apply the decision rule.

$$S_w^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(6-1)3.0 + (5-1)3.1}{6 + 5 - 2} = 3.0$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_w^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$= \frac{4.5 - 5.1}{\sqrt{3.0 \left(\frac{1}{6} + \frac{1}{5} \right)}} = -.57$$

$df = n_1 + n_2 - 2 = 6 + 5 - 2 = 9 \rightarrow t = \pm 2.262$

Accept H_0 because $-.57$ is not beyond -2.262 . Average customer waiting time at these stores is the same.

VI. Two-tail testing of two sample means from dependent populations using a "paired difference test"

- A. Paired sample are used to test a change in environment. Examples include production time before and after training and accidents before and after a safety campaign. A large difference means variables are dependent.
- B. Weekly sales at three of Linda's stores, before and after a big promotion, were \$1,200, \$1,300 and \$1,400 and \$1,400, \$1,500 and \$1,500 respectively. Linda will conduct a .10 level test to determine whether the promotion increased sales at these three stores. This is a one-tail test. Any change in sales would be a two-tail test.
 - 1. Paired tests treat data sets as one sample. A large difference results in a negative measure ($\mu_d < 0$).
 - 2. The 5-step approach to hypothesis testing
 - a. The null hypothesis and alternate hypothesis are $H_0 : \mu_d \geq 0$ and $H_1 : \mu_d < 0$.
 - b. The level of significance is .10.
 - c. The relevant statistic is \bar{d} .

\bar{d} is the mean difference of paired observations.
s_d is the standard deviation of paired differences.
n is the number of paired observations.
$df = n - 1 = 3 - 1 = 2 \rightarrow t = \pm 1.886$

Store	Sales Dollars		Difference d	d ²
	Before	After		
1	1,200	1,400	-200	40,000
2	1,300	1,500	-200	40,000
3	1,400	1,500	-100	10,000
Totals			-500	90,000

$$\bar{d} = \frac{\sum d}{n} = \frac{-500}{3} = -\$166.67$$

$$S_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{90,000 - \frac{(-500)^2}{3}}{3-1}}$$

$$= 57.7$$

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{-166.67}{\frac{57.7}{\sqrt{3}}}$$

$$= -5.03$$

Reject H_0 because -5.03 is beyond -1.886 . Sales increased.

Note: With independent populations, we test the mathematical relationship between different environments. With dependent populations, we test to see if a change in environment affects population parameters.

Practice Set 16 Small Sample Hypothesis Testing Using Student's t Test

- I. Darin wants to determine whether there is a difference in the number of sick days taken by employees based upon their education. A sample of 11 high school graduates had a mean of 5 sick days per year and a standard deviation of 2.5 days. Twelve non-graduates averaged 10 sick days per year. Their standard deviation was 3.25 days. Is there a difference in sick days taken based upon education? Use the .01 level of significance.

Data Set For People Using Statistics Software
Graduates' sick days: 5, 4, 7, 2, 7, 7, 0, 3, 6, 8, 6
Non-graduates' sick days: 9, 13, 8, 6, 14, 6, 12, 16, 8, 10, 7, 11

- II. Darin conducted a training program for 5 recently-hired employees. Test at the .01 level whether the training program increased employee efficiency.

Employee	Efficiency Rating	
	Before	After
1	8	9
2	6	8
3	7	8
4	7	9
5	8	10

Quick Questions 16 Small Sample Hypothesis Testing Using Student's t Test

I. Place the number of the appropriate definition or formula next to the concept it defines.

A. Weighted or pooled estimate of the population variance _____

B. Standard deviation of the differences _____

C. t when comparing two dependent populations _____

D. t when comparing two independent populations _____

E. Used with one population _____

F. Requires the use of the t distribution _____

1. $\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$	4. $\sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$
2. $\frac{x_1 - x_2}{\sqrt{s_w^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$	5. $\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$
3. the population is approximately normal, $n \leq 30$, and the population variance isn't known	6. $\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

II. Linda is tracking the number of work days missed by employees before and after taking part in a company-sponsored lunchtime physical fitness program. Test at the .01 level of significance whether the average number of days missed went down for program participants.

Employee	A	B	C	D	E	F	G
Before	8	9	6	8	3	4	5
After	6	7	5	6	5	2	5

III. Eight men applying to State University had a sample mean and variance on college board tests of 1,050 and 2,500 respectively. The respective numbers for nine women were 1,075 and 3,600. Test at the .05 level of significance whether women did better than men on these tests.

For People Using Statistics Software	
Men	Women
1,041	1,001
1,100	1,114
1,025	1,081
1,114	1,080
1,060	1,140
950	955
1,060	1,125
1,050	1,079
	1,100

Chapter 17 Statistical Quality Control

I. Introduction

- A. Increased competition in manufacturing has made statistical quality control more important than ever.
- B. Variation in manufactured parts is natural. It must be measured and controlled to achieve high quality.
 - 1. **Random variation** is due to chance. When it is the primary cause for variation over a period of time (production run), a process is in control.
 - 2. **Assignable variation** is not random and results from an identifiable cause. When it is excessive, a process is statistically out of control. Assignable variation is usually controllable by adjusting equipment, materials, atmospheric conditions, and other environmental factors.
- C. Quality control charts
 - 1. A **control chart** measures a process value (statistic) sequentially over a period of time.
 - 2. Whether a statistic such as \bar{X} is within upper and lower limits determines whether a process is in control. These **control limits** are similar to the interval estimates examined on pages 67 and 70.
 - 3. Control charts
 - a. An \bar{x} **chart** measures whether the **mean** size, weight, temperature, etc., is getting too high or too low.
 - b. A **range chart** measures whether **variation** in size, weight, temperature, etc., is too large.
 - c. A \bar{p} **chart** measures whether the **proportion** of some attribute (good or defective parts) is appropriate.

II. The \bar{x} chart

- A. Interval estimates based on the central limits theory (chapter 11) provide the theoretical foundation for the \bar{x} chart.
 - 1. Confidence intervals for 99.74%, 95%, and 90% are common. Confidence intervals are called control limits.
 - 2. These formulas are used to determine the 3 sigma or 99.74% confidence interval for the sample mean.

$$UCL = \bar{\bar{x}} + 3 \frac{\bar{s}}{\sqrt{n}} \quad LCL = \bar{\bar{x}} - 3 \frac{\bar{s}}{\sqrt{n}}$$

$\bar{\bar{x}}$ is the mean of the sample means.
 \bar{s} is an average of the sample standard deviations.

Sample Size (n)	A ₂	D ₃	D ₄
2	1.880	0	3.267
3	1.023	0	2.575
4	0.729	0	2.282
5	0.577	0	2.115

- 3. Control Factors for calculating UCL and LCL have been developed by the **American Society for Testing and Materials (ASTM)** (See table).

$$UCL = \bar{\bar{x}} + A_2 \bar{R} \quad LCL = \bar{\bar{x}} - A_2 \bar{R}$$

A₂ is a factor used to relate the mean's confidence interval to the mean of the ranges. (see ASTM table)

\bar{R} is the mean of the sample ranges.

Sample	Weight of Each Part				\bar{x}	Range
1	51	50	50	49	50	2
2	54	49	51	50	51	5
3	49	49	50	48	49	2
Totals for N = 3 samples					150	9
$\bar{\bar{x}} = \frac{\sum \bar{x}}{N} = \frac{150}{3} = 50$					$\bar{R} = \frac{\sum R}{N} = \frac{9}{3} = 3$	

- B. Three random samples of four parts designed to be 50 mm long were collected at half hour intervals. Control limits for these parts, using a 99.74% confidence interval, are determined below.

$$UCL = \bar{\bar{x}} + A_2 \bar{R}$$

$$= 50 + .729(3)$$

$$= 50 + 2.187$$

$$= 52.187$$

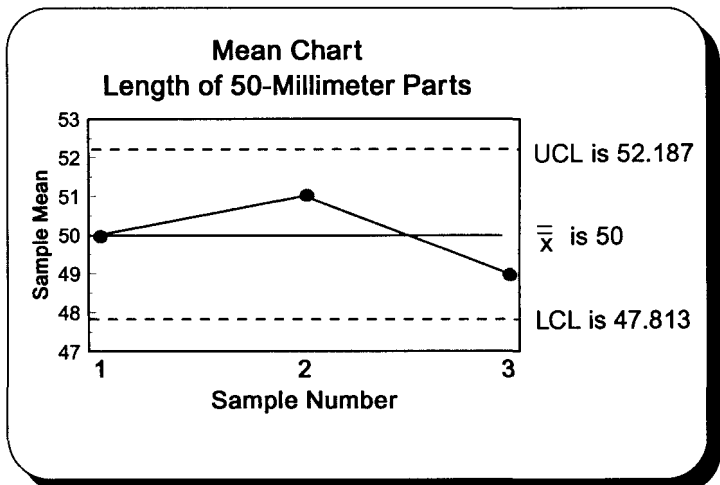
$$LCL = \bar{\bar{x}} - A_2 \bar{R}$$

$$= 50 - .729(3)$$

$$= 50 - 2.187$$

$$= 47.813$$

Note: Control limits should be set when a process is in control. Slight variations from a required tolerance (50 millimeters) are due to chance. With a mean control chart as a guide, future samples trending toward or beyond either control limit would indicate the process may be moving out of control. The answers to this chapter's Quick Questions introduce some of the methods used to judge whether a process is out of control.



III. The R chart

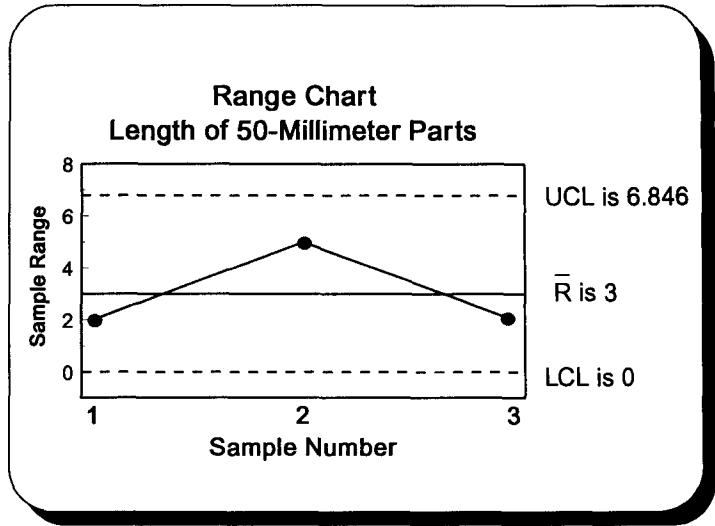
- A. A range chart, which measures variation, shows the confidence interval for the sample range.
- B. Simplified procedures for calculating UCL and LCL have been developed. Again the average range is multiplied by an ASTM factor. (See page 102)

$$UCL = D_4 \bar{R} \qquad LCL = D_3 \bar{R}$$

- C. Three sigma control limits for the page 102 data are determined below.

$$\begin{aligned} UCL &= D_4 \bar{R} \\ &= 2.282(3) \\ &= 6.846 \end{aligned}$$

$$\begin{aligned} LCL &= D_3 \bar{R} \\ &= 0(3) \\ &= 0 \end{aligned}$$



- D. These limits were determined when the process was in control. Three sigma (99.74%) control limits indicate 9,974 out of 10,000 sample ranges will be within these limits when the process is in control. Customer specifications (tolerances) may call for less variability. Statistics software may result in the more exact, but cumbersome, standard deviation replacing the range as the popular measure for determining variation.

IV. The p chart

- A. The p chart measures the proportion of some attribute (defective items) resulting from a process.
- B. It measures a qualitative attribute (being defective), rather than a quantitative characteristic (mean weight).
- C. Suppose we are interested in tracking the 50-millimeter part defects described on page 102. Daily random samples of 150 parts had the following defects. A 3 sigma p chart is constructed below.

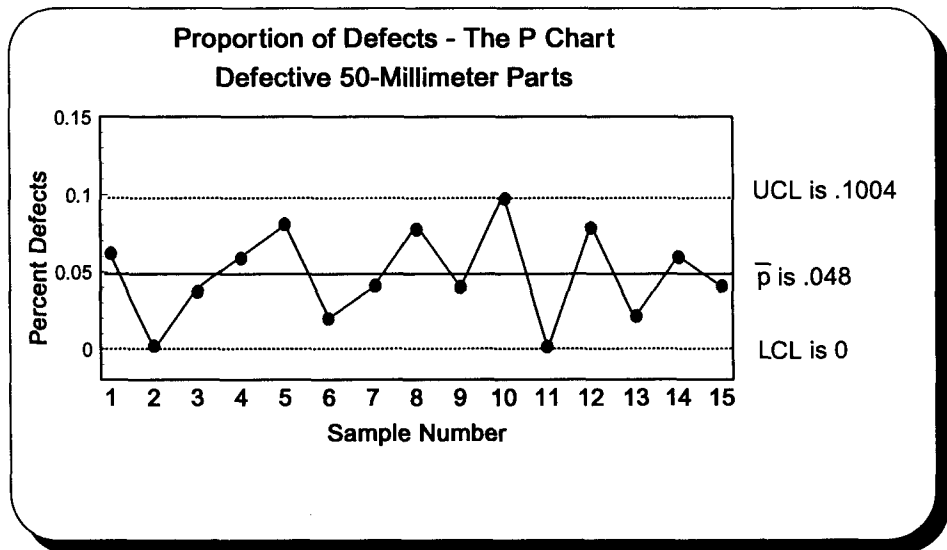
Quality Control Summary Sheet															
Date	1/3	1/4	1/5	1/6	1/7	1/10	1/11	1/12	1/13	1/14	1/17	1/18	1/19	1/20	1/21
Defects	9	0	6	9	12	3	6	12	6	15	0	12	3	9	6
Defect Proportions	.06	.00	.04	.06	.08	.02	.04	.08	.04	.10	.00	.08	.02	.06	.04

$$\bar{p} = \frac{\text{total defects}}{\text{total sampled}}$$

$$\begin{aligned} \bar{p} &= \frac{\text{total defects}}{\text{total sampled}} \\ &= \frac{108}{2,250} \\ &= .048 \end{aligned}$$

$$UCL \text{ and } LCL = \bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$\begin{aligned} UCL \text{ and } LCL &= \bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \\ &= .0480 \pm 3\sqrt{\frac{.048(1-.048)}{150}} \\ &= .0480 \pm .0564 \\ &= -.0044 \leftrightarrow +.1004 \\ &= -.0044 \text{ is rounded to zero.} \end{aligned}$$



Note: 1) This chart was prepared while the process was in control. An acceptable average proportion of defects is determined by the manufacturer and the customer. 2) The c chart is another attribute control chart. It measures counts such as the actual number of defects over time, defects per part, and complaints per period.

$$UCL \text{ and } LCL = \bar{c} \pm 3\sqrt{\bar{c}}$$

Practice Set 17 Statistical Quality Control

- I. Darin is doing a quality control study of the 30-milligram parts first analyzed in chapter 11. This data has been reproduced below. Assume the data consisted of 12 three-part samples. Also assume the process was in control when these samples were taken. Construct an \bar{X} chart and an R chart for this data using a 99.74% (3 sigma) confidence interval.

Sample #	1	2	3	4	5	6	7	8	9	10	11	12
	29.89	30.05	29.98	30.07	29.97	30.05	29.95	30.06	29.99	30.02	30.09	30.12
	29.96	29.97	30.06	30.05	29.95	29.95	29.99	29.89	29.99	30.08	30.06	30.16
	29.97	29.98	30.04	30.06	30.05	30.09	30.06	30.09	29.98	30.01	30.08	30.15
Sample Mean												
Sample Range												

- II. Darin wants to continue his study of the proportion of 30-milligram parts found to be defective in chapter 12. That study found 5 of 50 parts defective. This data and an additional 9 samples are summarized below. Construct a p chart for this data. Do not use the finite correction factor.

Defective 30-Milligram Parts										
Date	1/3	1/4	1/5	1/6	1/7	1/10	1/11	1/12	1/13	1/14
Sample #	1	2	3	4	5	6	7	8	9	10
Defects	5	4	6	3	5	4	7	4	3	7
Defects Proportion										

Quick Questions 17 Statistical Quality Control

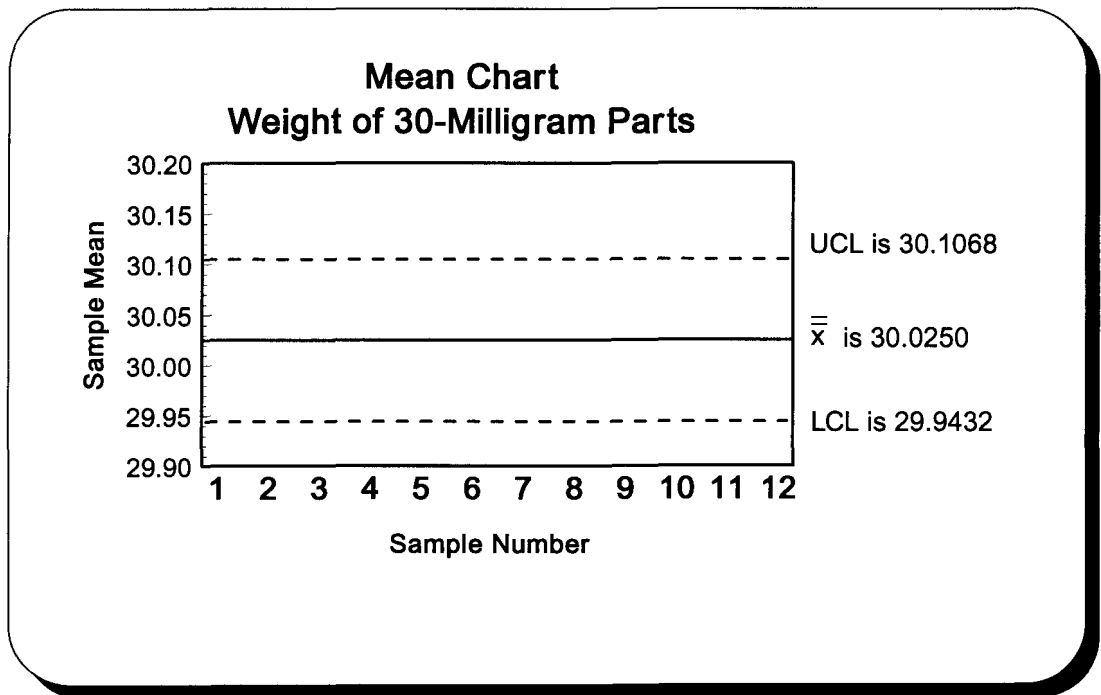
I. Place the number of the appropriate formula, expression, or term next to the concept it describes.

- A. A control chart _____
- B. Assignable variation _____
- C. Random variation _____
- D. An \bar{x} chart _____
- E. A range chart _____
- F. A p chart _____

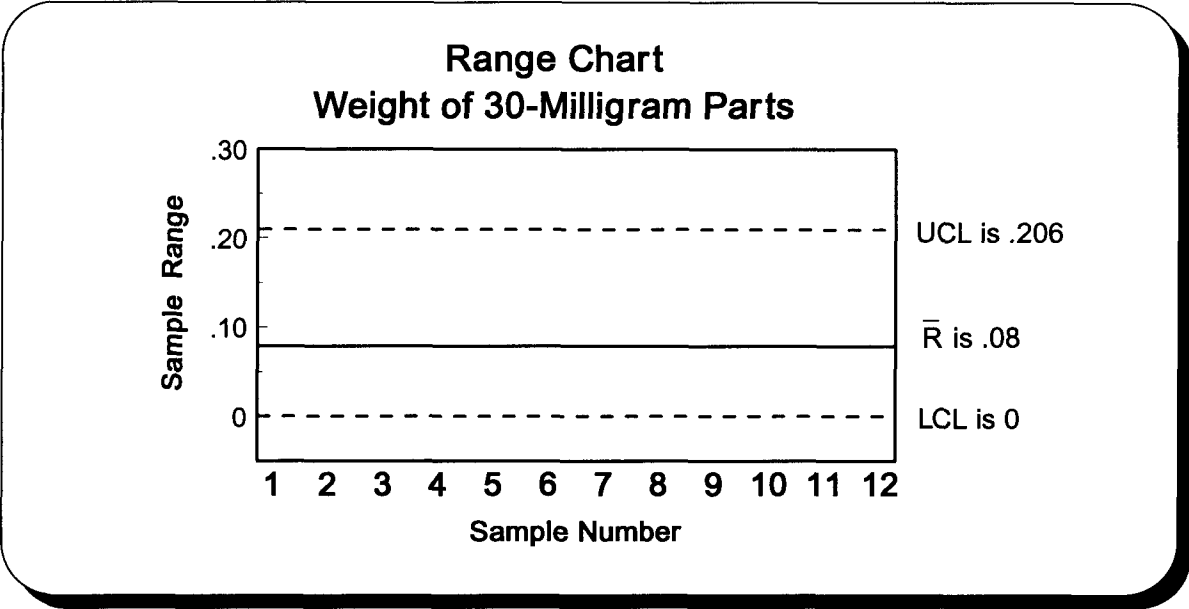
1. Measures whether the mean size, weight, or temperature, etc., is getting too high or too low.
2. Measures whether the proportion of some attribute (defects) is appropriate.
3. Results from an identifiable cause
4. Is due to chance
5. Measures a process value (statistic) sequentially over a period of time
6. Measures whether variation in size, weight, or temperature, etc., is too large.

II. Control charts developed in Practice Set 17 will now be used to determine whether the 30-milligram part manufacturing process is in control. Plot this data on the appropriate control chart and determine whether the process is in control.

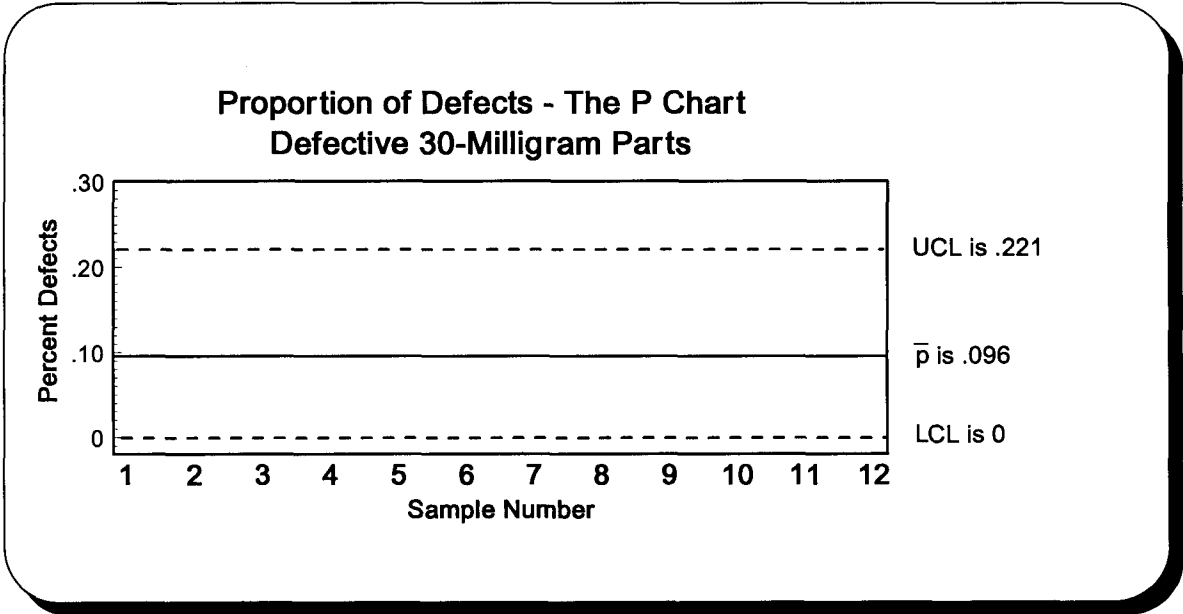
Sample	A	B	C	D	E	F	G	H	I	J	K	L
Sample Mean	29.98	30.04	30.08	30.02	29.97	30.04	30.09	30.15	30.10	30.12	30.14	30.16
Sample Range	0.07	0.09	0.11	0.13	0.10	0.09	0.08	0.14	0.18	0.21	0.22	0.21
Proportion of Defects	0.08	0.11	0.14	0.17	0.23	0.21	0.19	0.17	0.11	0.09	0.12	0.21



Analysis:



Analysis:

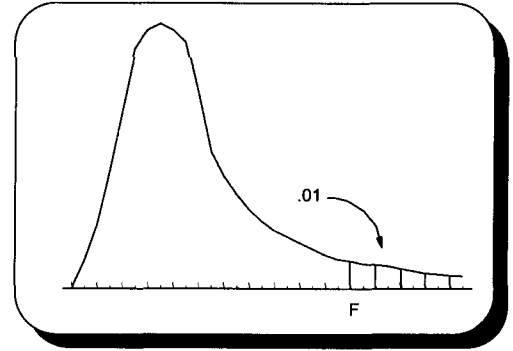


Analysis:

Chapter 18 Analysis of Variance

I. Introduction

- A. Understanding variability is important.
 - 1. Quality manufacturing, successful marketing, and effective training require an understanding of variability.
 - 2. Using a t-distribution (page 99, section V,C) requires determining the equality of population variances.
- B. Analysis of variance will require using the F distribution.
- C. Characteristics of the F distribution (named after originator R. Fisher)
 - 1. It is positively skewed, unimodal, continuous, and asymptotic.
 - 2. Basic assumptions of the F distribution
 - a. Experiments are of random design.
 - b. Variables are independent.
 - c. Populations are normally distributed with equal variances.
 - d. Both interval and ratio levels of data may be analyzed.
 - 3. Degrees of freedom for the numerator and degrees of freedom for the denominator determine the shape of this family of curves.



II. Testing two sample variances from normal populations

- A. Linda wants to compare sales variability of two stores. A sample of 5 from Store #1 measured mean daily sales at \$110. The standard deviation was \$16. A sample of 8 from Store #2 measured mean daily sales at \$125. The standard deviation was \$14. Test at the .02 level of significance whether these two stores have equal sales variances. This is a two-tail test. One-tail tests involve testing a difference in one direction.
- B. The 5-step approach to hypothesis testing
 - 1. These are the null hypothesis and alternate hypothesis.
 $H_0 : \sigma_1^2 = \sigma_2^2$ and $H_1 : \sigma_1^2 \neq \sigma_2^2$
 - 2. The level of significance will be .02 and $\alpha/2 = .01$.
 - 3. The relevant statistic F, is the ratio of the sample variances.
 - a. The larger variance is always put on top.
 - b. This means F is a positive number greater than one.
 - c. F's value increases as the difference in variability increases for this one-tail test.
 - 4. The decision rule will be, if F from the test statistic is large enough (beyond the critical value), the difference in variability is high and the null hypothesis is rejected.
 - a. Degrees of freedom is df and $df = n - 1$.
 - b. For the numerator, df is $n - 1 = 5 - 1 = 4$.
 - c. For the denominator, df is $n - 1 = 8 - 1 = 7$.
 - d. From the table, $f = 7.85$.
 - 5. Apply the decision rule.

$$F = \frac{s_1^2}{s_2^2}$$

$$F = \frac{s_1^2}{s_2^2} = \frac{16^2}{14^2} = 1.31 \quad \text{Accept } H_0 \text{ because } 1.31 < 7.85. \quad \text{Variances are equal.}$$

F Distribution, Critical Values for the Upper .01				
Denominator degrees of freedom	Numerator degrees of freedom			
	1	2	3	4
1	4052	4999	5403	5624
2	98.49	99.00	99.17	99.25
3	34.12	30.82	29.46	28.71
4	21.20	18.00	16.69	15.98
5	16.26	13.27	12.06	11.39
6	13.74	10.92	9.78	9.15
7	12.25	9.55	8.45	7.85
8	11.26	8.65	7.59	7.01
9	10.56	8.02	6.99	6.42

See pages ST 5A and 5B for more complete F tables.

III. Testing 3 or more sample means from normally-distributed populations

- A. These **analysis of variance** tests are called **ANOVA**.
- B. ANOVA measures whether a **treatment variable** has caused a change in a **response variable**.
- C. ANOVA is used to measure training effectiveness and product quality when 3 or more samples are involved.
- D. Basic procedures
 - 1. ANOVA uses two separate measures of population variance.
 - 2. Each is part of the f ratio.
 - 3. The numerator measures **between treatment variance**. This variance is due to differences among sample means.
 - 4. The denominator measures **within treatment variance**. This variance, which is only due to within group differences, is variation due to error.
 - 5. If the null hypothesis is true, the population means are equal, the expected value of the two measures of population variance will be equal, and F will be one. Otherwise, F will be larger than one.
 - 6. If, based upon some level of significance, the test statistic F is larger than the critical value of F, the means are not equal and the null hypothesis is rejected.

$$F = \frac{\text{Estimated variance between the treatments}}{\text{Estimated variance within the treatments}}$$

IV. Linda Smith is using ANOVA to measure whether there is a difference between the average weekly sales of her 3 salespeople. The test will be at the .05 level of significance.

A. These are the null hypothesis and alternate hypothesis. $H_0 : \mu_1 = \mu_2 = \mu_3$ $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

B. The level of significance for this single treatment, one-tail problem will be .05.

C. F is the test statistic.

$$F = \frac{\text{Estimated variance between the treatments}}{\text{Estimated variance within the treatments}}$$

Note: Salespeople is the treatment variable and sales is the response variable.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares (variance)	ANOVA
Between Treatments	t - 1	SS _T	$MS_T = \frac{SS_T}{t-1}$	$F = \frac{MS_T}{MS_E}$
Within Treatments (error)	N - t	SS _E	$MS_E = \frac{SS_E}{N-t}$	
Total Variance	N - 1	SS _{TOTAL}		

t is the number of treatments

n is the number of rows in a treatment

N is total observations

SS_T is the sum of the squares for treatments

SS_E is the sum of the squares for error

SS_{TOTAL} is the total sum of the squares

MS_T is the mean squares for treatments

MS_E is the mean squares for error

D. Reject the null hypothesis when F from the test statistic is beyond the critical value of F for the .05 level of significance.

E. Apply the decision rule.

df = t - 1 = 3 - 1 = 2 for the numerator
df = N - t = 12 - 3 = 9 for the denominator
F's critical value is 4.26.

Weekly Sales (x) in Thousands of Dollars for 3 Treatments (T)						Row Totals Required for Calculations
	Salesperson L is T ₁		Salesperson M is T ₂		Salesperson N is T ₃	
	Sales (X ₁)	X ₁ ²	Sales (X ₂)	X ₂ ²	Sales (X ₃)	X ₃ ²
	7	49	6	36	9	81
Column Totals	6	36	8	64	8	64
Required for Calculations	7	49	6	36	7	49
	<u>4</u>	<u>16</u>	<u>6</u>	<u>36</u>	<u>10</u>	<u>100</u>
$\sum X_T$	24		26		34	
$(\sum X_T)^2$	576		676		1156	
n	4		4		4	
$\frac{(\sum X_T)^2}{n}$	144		169		289	
$\sum X_T^2$		150		172		294
						$\sum x = 84$
						N = 12
						$\sum [\frac{(\sum X_T)^2}{n}] = 602$
						$\sum x^2 = 616$

$$SS_T = \sum \left[\frac{(\sum X_T)^2}{n} \right] - \frac{(\sum X)^2}{N}$$

$$= 602 - \frac{84^2}{12}$$

$$= 602 - 588 = 14$$

$$SS_E = \sum X^2 - \sum \left[\frac{(\sum X_T)^2}{n} \right]$$

$$= 616 - 602$$

$$= 14$$

Total variance equals SS_T + SS_E = 14 + 14 = 28. Total variance also equals

$$SS_{TOTAL} = \sum X^2 - \frac{(\sum x)^2}{N}$$

$$= 616 - 588 = 28$$

$$MS_T = \frac{SS_T}{t-1} = \frac{14}{3-1} = 7.0$$

$$MS_E = \frac{SS_E}{N-t} = \frac{14}{12-3} = \frac{14}{9} = 1.56$$

Note: Half the variability has been explained by the treatment variable.

Reject H₀ because F = $\frac{MS_T}{MS_E} = \frac{7.0}{1.56} = 4.49$ and $4.49 > 4.26$. Mean sales of these salespeople are not equal.

Practice Set 18 Analysis of Variance

- I. Darin wants to know whether the variance of 30-mg parts has increased. The standard deviation from a recent sample of 16 parts was .067 milligrams. The standard deviation from an earlier study of 14 parts was .062 milligrams. Test at the .01 level whether the population variance has increased.

Data Set For People Using Statistics Software	
Sample 1	Sample 2
29.91	29.89
29.93	30.09
29.96	29.96
29.95	29.96
29.94	29.98
29.95	29.99
29.97	30.05
30.09	29.99
30.04	30.07
29.96	30.06
30.09	29.97
30.06	30.09
29.95	30.04
29.91	30.09
30.09	
29.92	

- II. Time passed and the wonders of miniaturization have reduced the 30-mg parts to a weight of only 9 mg. Darin randomly selected samples of 9-mg parts from 3 departments with the following results. **People using statistics software should skip to part D.**

A. Complete this chart to begin an ANOVA study of the mean weight of parts produced by these 3 departments.

Weight Analysis of 9-mg Parts Produced by 3 Departments						Row Totals Required for Calculations
Parts Sample 1 is T_1		Parts Sample 2 is T_2		Parts Sample 3 is T_3		
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	
8.95	80.1025	9.05	81.9025	9.05	81.9025	
8.90	79.2100	9.05	81.9025	9.15	83.7225	
8.90	79.2100	9.10	82.8100	9.10	82.8100	
ΣX_T						$\Sigma x =$
$(\Sigma X_T)^2$						
n						$N = 9$
$\frac{(\Sigma X_T)^2}{n}$						$\Sigma \left[\frac{(\Sigma X_T)^2}{n} \right] =$
ΣX_T^2						$\Sigma x^2 =$

B. Using data from the previous page, calculate the following values.

$$SS_T = \Sigma \left[\frac{(\Sigma x_T)^2}{n} \right] - \frac{(\Sigma x)^2}{N}$$

$$SS_E = \Sigma x^2 - \Sigma \left[\frac{(\Sigma x_T)^2}{n} \right]$$

$$SS_{TOTAL} = \Sigma x^2 - \frac{(\Sigma x)^2}{N}$$

C. Complete the following chart using the data accumulated to this point.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	t - 1 =	SS _T =	MS _T =	F =
Within Treatments (error)	N - t =	SS _E =	MS _E =	
Total Variance	N - 1	SS _{TOTAL} =		

D. Using the 5-step approach to hypothesis testing and the above chart, test at the .05 level whether the sample means are from populations with equal means.

Quick Questions 18 Analysis of Variance

I. Copy the formulas and expressions on the right into this ANOVA summary chart.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments				
Within Treatments (error)				
Total Variance				

SS_T	$F = \frac{MS_T}{MS_E}$
$N - t$	SS_{TOTAL}
$MS_T = \frac{SS_T}{t-1}$	$t - 1$
$MS_E = \frac{SS_E}{N-t}$	SS_E
$N - 1$	

II. Answer the following fill in the blank questions.

- A. Analysis of variance requires the populations be _____ distributed.
- B. When using the F distribution, the numerator is always the _____ of the 2 variances.
- C. When doing ANOVA, the numerator of the F distribution measures variance _____ the treatments.
- D. When doing ANOVA, the denominator of the F distribution measures variance _____ the treatments.

III. Complete the following ANOVA study concerning grade point averages randomly selected by a local college. Those using statistics software should skip to part D.

A. Begin by completing this chart.

Analysis of College Grades Based Upon High School Grades						Row Totals Required for Calculations
	High H.S. Grades T_1		Medium H.S. Grades T_2		Low H.S. Grades T_3	
	College Grades(X_1)	X_1^2	College Grades(X_2)	X_2^2	College Grades(X_3)	X_3^2
	3.4		3.2		2.1	
	3.5		2.8		2.5	
	3.1		3.0		2.7	
$\sum X_T$						$\sum x =$
$(\sum X_T)^2$						
n						$N = 9$
$\frac{(\sum X_T)^2}{n}$						$\sum \left[\frac{(\sum X_T)^2}{n} \right] =$
$\sum X_T^2$						$\sum x^2 =$

B. Using the chart on the previous page, calculate the following values.

$$SS_T = \sum \left[\frac{(\sum x_T)^2}{n} \right] - \frac{(\sum X)^2}{N}$$

$$SS_E = \sum x^2 - \sum \left[\frac{(\sum x_T)^2}{n} \right]$$

$$SS_{TOTAL} = \sum x^2 - \frac{(\sum X)^2}{N}$$

$$MS_T = \frac{SS_T}{t-1} =$$

$$MS_E = \frac{SS_E}{N-t} =$$

C. Complete the following chart using data accumulated to this point.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	t - 1 =	SS _T =	MS _T =	
Within Treatments (error)	N - t =	SS _E =	MS _E =	F =
Total Variance	N - 1 =	SS _{TOTAL} =		

D. Using the 5-step approach to hypothesis testing, test at the .05 level whether these sample means come from populations with equal means.

E. Answer problem D at the .01 level of significance.

Chapter 19 Two-Factor Analysis of Variance

I. Sources of variability

- A. Variance between treatment variables exists because treatments are not alike.
- B. Variance within a treatment is unexplained and due to sampling error.
- C. Additional sources of variability (called factors or treatments) may be added to a study.
 1. Their variability may be used to reduce unaccounted for, within treatment variability (error).
 2. Additional treatments are called **blocking variables**.
 3. They represent a substantial source of inherent response variability.
 4. Treatments must not be independent. Treatment B may affect the factors of treatment A differently.
For example, weeks of experience may have a different affect on each of the recently hired sales people.
 5. Examples of blocking variables include age, gender, education, and time.

II. Two-factor variance analysis

- A. In chapter 18, Linda found that her 3 salespeople had different mean weekly sales and that half of the data's variability could be attributed to the salespeople treatment.
- B. Chapter 18 sales data was randomly assigned to each salesperson. Here, it has been arranged by weeks of experience. Using experience as a blocking variable may account for some of the unexplained variability. Treatments are not independent because weeks of experience may affect salespeople differently.
- C. $\sum X_B$ is the sales associated with each block (week). Number of treatments is now t, b is the number of blocks.

Weekly Sales (x) in Thousands of Dollars							Row Totals Required for Calculations		
Block(B_x)	Salesperson L is T_1		Salesperson M is T_2		Salesperson N is T_3				
Weeks	Sales(X_1)	X_1^2	Sales(X_2)	X_2^2	Sales(X_3)	X_3^2	$\sum X_B$	$(\sum X_B)^2$	$\frac{(\sum X_B)^2}{t}$
1	4	16	6	36	7	49	17	289	96.3
2	6	36	6	36	8	64	20	400	133.3
3	7	49	6	36	9	81	22	484	161.3
4	<u>7</u>	<u>49</u>	<u>8</u>	<u>64</u>	<u>10</u>	<u>100</u>	<u>25</u>	625	<u>208.3</u>
$\sum X_T$	24		26		34		84 = $\sum x$	$\sum [\frac{(\sum X_B)^2}{t}] = 599.3$	
$(\sum X_T)^2$	576		676		1156		84 = $\sum x$		
b	4		4		4		N = 12		
$\frac{(\sum X_T)^2}{b}$	144		169		289		$\sum [\frac{(\sum X_T)^2}{b}] = 602$		
$\sum X_T^2$		150		172		294	$\sum x^2 = 616$		

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	t - 1	SS_T	$MS_T = \frac{SS_T}{t-1}$	$F = \frac{MS_T}{MS_E}$
Block	b - 1	SS_B	$MS_B = \frac{SS_B}{b-1}$	
Within Treatments (error)	(t - 1)(b - 1)	SS_E	$MS_E = \frac{SS_E}{(t-1)(b-1)}$	$F = \frac{MS_B}{MS_E}$
Total Variance	N - 1	SS_{TOTAL}		

Note: This analysis is called "mean square" because it is based upon the variance.

- D. Linda wants to know the variability explained by the blocking variable experience at the .05 level of significance.
- E. The 5-step approach to hypothesis testing
- A check of each null hypothesis will be made.
 - $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$ for the treatment means.
 - $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ and $H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ for the block means.
 - The level of significance is .05.
 - The test statistic is F.
 - If F from the test statistic is beyond the critical value of F for the .05 level of significance, the null hypothesis will be rejected.
 - Apply the decision rule.

$$SS_T = \sum \left[\frac{(\sum x_T)^2}{b} \right] - \frac{(\sum x)^2}{N}$$

$$= 602 - \frac{84^2}{12}$$

$$= 602 - 588 = 14$$

$$SS_B = \sum \left[\frac{(\sum X_B)^2}{t} \right] - \frac{(\sum X)^2}{N}$$

$$= 599.3 - \frac{84^2}{12}$$

$$= 599.3 - 588 = 11.3$$

$$SS_{TOTAL} = \sum x^2 - \frac{(\sum x)^2}{N}$$

$$= 616 - 588 = 28$$

$$SS_E = SS_{TOTAL} - (SS_T + SS_B)$$

$$= 28.0 - (14.0 + 11.3) = 2.7$$

Unexplained variability is down from 14.0 to 2.7.

$$MS_T = \frac{SS_T}{t-1} = \frac{14}{3-1} = 7.0$$

$$MS_B = \frac{SS_B}{(b-1)} = \frac{11.3}{4-1} = 3.77$$

$$MS_E = \frac{SS_E}{(t-1)(b-1)} = \frac{2.7}{(3-1)(4-1)} = .45$$

Treatment hypothesis degrees of freedom
 $t - 1 = 3 - 1 = 2$ for numerator
 $(t - 1)(b - 1) = (3 - 1)(4 - 1) = 6$ for denominator
 $F = 5.14$

Reject H_0 because $F = \frac{MS_T}{MS_E} = \frac{7.0}{.45} = 15.56 > 5.14$.
 Average salesperson sales are not equal.

Block hypothesis degrees of freedom
 $b - 1 = 4 - 1 = 3$ for numerator
 $(t - 1)(b - 1) = (3 - 1)(4 - 1) = 6$ for denominator
 $F = 4.76$

Reject H_0 because $F = \frac{MS_B}{MS_E} = \frac{3.77}{.45} = 8.38 > 4.76$.
 Average weekly sales are not equal.

III. Comparing three or more treatment sample means for one-factor analysis

- Having proven that there is a difference in the average sales of the three treatments (salespeople) in chapter 18, determining whether treatment means differ from each of the other may be of interest.
- A range (confidence interval) will be found for the difference between 2 treatment means. A positive range for the difference of these means will indicate the difference could not be zero and the means are different.
- The t value for $\alpha/2$ will be used.

$$(\bar{X}_3 - \bar{X}_1) \pm t \sqrt{MS_E \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- We will determine whether average sales for the first and third salesperson are different at the .05 level of significance.

Salespersons #1 and #3 Average Sales (Data from page 109)

The number of observations within each treatment is n_1 and n_2 .

$$\bar{X}_1 = \frac{\sum x}{n_1} = \frac{24}{4} = 6.0$$

$$\bar{X}_3 = \frac{\sum x}{n_3} = \frac{34}{4} = 8.5$$

t for $\alpha/2$ and $N - t$ degrees of freedom is $12 - 3 = 9 \rightarrow t = 2.262$
 MS_E from page 109 is 1.56.

$$(\bar{X}_3 - \bar{X}_1) \pm t \sqrt{MS_E \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$(8.5 - 6.0) \pm 2.262 \sqrt{1.56 \left(\frac{1}{4} + \frac{1}{4} \right)}$$

$$2.5 \pm 2.262 \sqrt{.78}$$

$$2.5 \pm 2.0$$

A positive range of .5 \leftrightarrow 4.5 indicates the means are different.

Practice Set 19 Two-Factor Analysis of Variance

I. Practice Set 18 will be expanded by assuming the data was randomly collected at hourly intervals. Page 110 data has been arranged accordingly. Darin wants to determine whether samples taken later in a shift are less likely to pass inspection. **People using statistics software should skip to part D.**

A. Complete this chart to begin an ANOVA study of the production process producing these parts.

Weight Analysis of 9-mg Parts Produced by 3 Departments							Row Totals Required for Calculations		
Time	Parts Sample 1 is T ₁		Parts Sample 2 is T ₂		Parts Sample 3 is T ₃		ΣX_B	$(\Sigma X_B)^2$	$\frac{(\Sigma X_B)^2}{t}$
	X_1	X_1^2	X_2	X_2^2	X_3	X_3^2			
9:15 AM	8.90	79.2100	9.05	81.9025	9.05	81.9025			
10:20 AM	8.90	79.2100	9.05	81.9025	9.10	82.8100			
11:10 AM	<u>8.95</u>	<u>80.1025</u>	<u>9.10</u>	<u>82.8100</u>	<u>9.15</u>	<u>83.7225</u>			
ΣX_T	26.75		27.20		27.30		$\Sigma x =$	$\Sigma[\frac{(\Sigma X_B)^2}{t}] =$	
$(\Sigma X_T)^2$	715.5625		739.84		745.29		$\Sigma x = 81.25$		
b	3		3		3		$N = 9$		
$\frac{(\Sigma X_T)^2}{b}$	238.521		246.613		248.430		$\Sigma[\frac{(\Sigma X_T)^2}{b}] = 733.564$		
ΣX_T^2		238.5225		246.6150		248.4350	$\Sigma x^2 = 733.5725$		

B. Using the above data, calculate the following values.

$$\begin{aligned}
 SS_T &= \Sigma\left[\frac{(\Sigma X_T)^2}{b}\right] - \frac{(\Sigma X)^2}{N} \\
 &= 733.564 - \frac{81.25^2}{9} \\
 &= 733.564 - 733.507 = .057
 \end{aligned}$$

$$SS_B = \Sigma\left[\frac{(\Sigma X_B)^2}{t}\right] - \frac{(\Sigma X)^2}{N}$$

$$SS_{TOTAL} = \Sigma x^2 - \frac{(\Sigma X)^2}{N}$$

$$SS_E = SS_{TOTAL} - (SS_T + SS_B)$$

C. Complete the following chart using data accumulated to this point.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	$t - 1 =$	$SS_T =$	$MS_T =$	$F = \frac{MS_T}{MS_E} =$
Block	$b - 1 =$	$SS_B =$	$MS_B =$	
Within Treatments (error)	$(t - 1)(b - 1) =$	$SS_E =$	$MS_E =$	$F = \frac{MS_B}{MS_E} =$
Total Variance	$N - 1 =$	$SS_{TOTAL} =$		

D. Using the 5-step approach to hypothesis testing, determine at the .01 level of significance whether the sample treatment and block means come from populations with equal means.

II. Using information from page 111, determine at the .01 level of significance whether there is a difference between treatments 1 and 3.

Quick Questions 19 Two-Factor Analysis of Variance

I. Use the symbols to the right to complete the following ANOVA summary chart.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments				
Block				
Within Treatments (error)				
Total Variance				

SS_T	$F = \frac{MS_T}{MS_E}$
$(t-1)(b-1)$	SS_{TOTAL}
$MS_T = \frac{SS_T}{t-1}$	$t-1$
$MS_B = \frac{SS_B}{b-1}$	SS_E
$MS_E = \frac{SS_E}{(t-1)(b-1)}$	
$b-1$	SS_B
$F = \frac{MS_B}{MS_E}$	$N-1$

II. The analysis in the last set of Quick Questions will be expanded by rearranging the data in each row so it is based upon the amount of time students spend studying. Complete the following ANOVA study concerning college grades and study times collected by a local college. Begin by completing this chart. **People using statistics software should skip to part C.**

Analysis of College Grades Based Upon High School Grades and Time Spent Studying While in College							Row Totals Required for Calculations		
College Study Time	High H.S. Grades T_1		Medium H.S. Grades T_2		Low H.S. Grades T_3		$\sum X_B$	$(\sum X_B)^2$	$\frac{(\sum X_B)^2}{t}$
	College Grades(X_1)	X_1^2	College Grades(X_2)	X_2^2	College Grades(X_3)	X_3^2			
High	3.5	12.25	3.2	10.24	2.7	7.29			
Medium	3.4	11.56	3.0	9.00	2.5	6.25			
Low	3.1	9.61	2.8	7.84	2.1	4.41			
$\sum X_T$	10		9		7.3		$\sum X = 26.3$		
$(\sum X_T)^2$	100		81		53.29				
b	3		3		3		$N = 9$		
$\frac{(\sum X_T)^2}{b}$	33.33		27		17.76		$\sum [\frac{(\sum X_T)^2}{b}] = 78.09$		
$\sum X_T^2$		33.42		27.08		17.95	$\sum X^2 = 78.45$		

A. Using this chart, calculate the following values.

$$SS_T = \sum \left[\frac{(\sum X_T)^2}{b} \right] - \frac{(\sum X)^2}{N}$$

$$SS_B = \sum \left[\frac{(\sum X_B)^2}{t} \right] - \frac{(\sum X)^2}{N}$$

$$SS_{\text{TOTAL}} = \sum X^2 - \frac{(\sum x)^2}{N}$$

$$SS_E = SS_{\text{TOTAL}} - (SS_T + SS_B)$$

B. Complete the following chart using data accumulated to this point.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	t - 1 =	SS _T =	MS _T =	$F = \frac{MS_T}{MS_E} =$
Block	b - 1 =	SS _B =	MS _B =	
Within Treatments (error)	(t - 1)(b - 1) =	SS _E =	MS _E =	$F = \frac{MS_B}{MS_E} =$
Total Variance	N - 1 =	SS _{TOTAL} =		

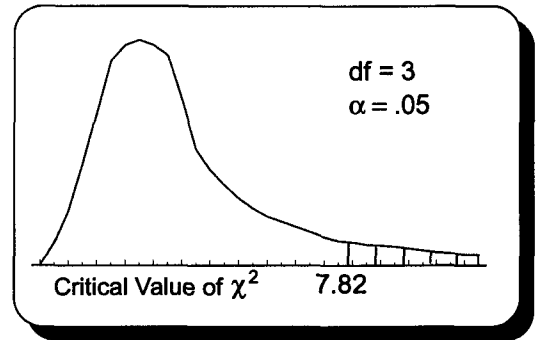
C. Using the 5-step approach to hypothesis testing, determine at the .05 level of significance whether these treatment and block means come from populations with equal means.

III. Using the chart data from pages 112 and 113, determine at the .01 level whether there is a difference between treatment means 1 and 2.

Chapter 20 Nonparametric Hypothesis Testing of Nominal Data

I. Introduction

- A. **Parametric statistics** is the name given to much of the material covered through chapter 19.
 1. Parametric tests involve a population parameter for which the test statistic has a known distribution (shape).
 2. Measurement (data) sophistication is of an interval or ratio level. (see page 2)
- B. **Nonparametric statistics** are used when the requirements of parametric statistics are not fulfilled.
 1. Data is considered **distribution-free** because the distribution of the sample statistic may be unknown.
 2. Nominal and ordinal data can be tested.
- C. **Count data (categorical data)**
 1. In this chapter, sample observations (counts) are grouped into categories and compared to some expected count (frequency). A small difference between the actual and expected frequencies indicates a match.
 2. Applications
 - a. Determining brand preference by age, gender, etc.
 - b. Measuring the success of an advertising campaign or training program.
- D. **The chi-square distribution** (pronounced "kigh" square)
 1. The chi-square distribution is like the t distribution because there is a family of curves, one for each degree of freedom.
 2. The distribution becomes more normal as the degrees of freedom increase. Chi-square is the ratio of $(n - 1)s^2$ to σ^2 .



II. Goodness of fit tests for a one categorical variable

A. Linda is interested in determining if consumers at her four stores are giving equal acceptance to the low sales price of a new hit music video.

B. The 5-step approach to hypothesis testing

1. H_0 : sales are equally distributed
 H_1 : sales are not equally distributed

Music Video Sales	Store A	Store B	Store C	Store D	Totals
Sample sales, f_o	8	22	19	11	60
Expected sales, f_e	15	15	15	15	60

2. The significance level is .05.
3. Chi-square is the test statistic.

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

f_o is an observed frequency of a category.
 f_e is an expected frequency of a category. It should be ≥ 5 when using the continuous chi-square distribution for a discrete problem.
 Equal acceptance means $f_e = 60/4 = 15$.
 k is the number of categories.
 There are $k - 1$ degrees of freedom for a goodness of fit problem.

4. The decision rule:
 If χ^2 from the test statistic is beyond the critical value, the difference is high and the null hypothesis is rejected.
5. Apply the decision rule for this one-tail test.

Store	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
A	8	15	-7	49	49/15 = 3.27
B	22	15	7	49	49/15 = 3.27
C	19	15	4	16	16/15 = 1.07
D	11	15	-4	16	16/15 = 1.07
			0		$\chi^2 = 8.68$

$df = k - 1 = 4 - 1 = 3 \rightarrow \chi = 7.82$
 Reject H_0 because $8.68 > 7.82$.
 Sales are not equally distributed.

Chi-Square					
Degrees of freedom	Right-tail area				
	.10	.05	.025	.01	.005
1	2.71	3.84	5.02	6.64	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.82	9.35	11.35	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75

Page ST 6 has a more complete chi-square table.

Note: This procedure can be used to test unequal expected frequencies. Suppose Store A usually has 40% of company sales and the 3 other stores each have 20%. Store A would be expected to have 24 sales (.40 x 60) and the other stores would be expected to have 12 sales (.20 x 60).

Note: Opinions vary on the exact lower limit for f_e .

III. Test for independence of two categorical variables with a contingency table

- A. The one category variable problem on page 120 tested one variable (sales) against some hypothesized frequency to determine if there was a good fit between the hypothesized frequency and the observed frequency.
- B. Here, a contingency table is used to determine if there is a relationship (statistical dependency) between two variables. That is, does knowledge of variable A's value provide knowledge of variable B's value. If so, variables are statistically dependent. Otherwise, they are independent.
- C. The idea of statistical dependency was first encountered in the probability chapter on page 46. At that time, advertising expenditures and sales revenue were said to be dependent. To be sure sales increased enough when advertising increased to indicate dependency, a statistical proof for dependency could be conducted. We must adjust the monthly data on page 46 to weekly data because cell values (f_o) must be ≥ 5 . Fifty weeks of data will be studied. The null hypothesis will again proclaim no difference (sales and advertising are independent). The test will measure whether the difference is large and did not happen by chance. That is, the variables are dependent.
- D. The 5-step approach to hypothesis testing
- H_0 : advertising and sales are independent
 H_1 : advertising and sales are dependent
 - The significance level is .01.
 - Chi-square is the test statistic.

Sales	Less than or equal to \$12,000	Greater than \$12,000	Totals
Advertising			
Less than or equal to \$1,000	20	5	25
Greater than \$1,000	5	20	25
Totals	25	25	50

Sales	Less than or equal to \$12,000		Greater than \$12,000		Totals	
Advertising	f_o	f_e	f_o	f_e	f_o	f_e
Less than or equal to \$1,000	20	12.5	5	12.5	25	25
Greater than \$1,000	5	12.5	20	12.5	25	25
Totals	25	25	25	25	50	50

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] \text{ where } f_e = \frac{f_r \times f_c}{n}$$

$df = (r - 1)(c - 1)$
 r is the number of rows,
 c is the number of columns

- If χ^2 from the test statistic is beyond the critical value, reject the null hypothesis.
- Applying the decision rule for this one-tail test.
 - Imagine the above contingency table has only f_o data and the f_e data cells and totals are blank.
 - Row and column totals for f_o are equal to those of f_e .
 - A table cell is completed by multiplying its row total by its column total and dividing by the grand total. For example, the first f_o cell has been calculated in the frame to the right.

Note: If 2 variables are independent, their cell values are in proportion. This formula is used to determine expected row and column cell values.

$$f_e = \frac{f_r \times f_c}{n} = \frac{25 \times 25}{50} = 12.5$$

$$df = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1 \rightarrow \chi^2 = 6.64 \text{ (see chart page 120)}$$

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = \sum \left[\frac{(20 - 12.5)^2}{12.5} + \frac{(5 - 12.5)^2}{12.5} + \frac{(5 - 12.5)^2}{12.5} + \frac{(20 - 12.5)^2}{12.5} \right]$$

$$= 4.5 + 4.5 + 4.5 + 4.5 = 18$$

The null hypothesis is rejected because $18 > 6.64$. Advertising expenditures affect sales revenue. These variables are dependent at the .01 level of significance.

Note: Chi-square analysis is used to test interesting relationships such as level of income (low, medium, and high) and frequency of purchase (often and not often).

Note: As demonstrated with this advertising/sales data, it is often necessary to regroup data to assure that f_o is ≥ 5 . Classes with a low frequency are combined until the requirement is observed.

Practice Set 20 Nonparametric Hypothesis Testing of Nominal Data

- I. Darin feels 20% of the 9-mg part defects are produced by the first shift, 30% by the second shift, and 50% by the third shift. Do an .01 level of significance test to determine whether this sample data follows Darin's proposed distribution. **People using statistics software do not need to fill out the second chart.**

Analysis of Defects				
	Shift 1	Shift 2	Shift 3	Totals
Shift defects, f_o	6	11	23	40
Expected defects, f_e				

Shift	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Totals					

- II. This is Darin's page 42 study of customer age and making a sale. Test at the .05 level whether customer age and making a sale are independent. **People using statistics software do not need to fill out the chart.**

Customer Age and Making A Sale			
Customer Age	Less than or equal to 20	Over 20	Totals
Making A Sale			
No	16	8	24
Yes	24	12	36
Totals	40	20	60

Contingency Table of Customer Age and Making A Sale						
Customer Age	Less than or equal to 20		Over 20		Totals	
Making A Sale	f_o	f_e	f_o	f_e	f_o	f_e
No	16		8			
Yes	24		12			
Totals	40		20			

Note: Tests similar to those conducted here can be used as follows:

- Test observed data to see if it follows a normal probability distribution.
- Test observed data to see if it follows a Poisson probability distribution.
- These tests are easy to perform with statistics software.

Quick Questions 20 Nonparametric Hypothesis Testing of Nominal Data

I. Place the number of the formula or expression next to the concept it defines.

A. $\chi^2 =$ _____

B. Expected frequency f_e must be _____

C. f_e for a contingency table equals _____

D. Chi-square is the ratio of _____

E. df for use with a contingency table _____

F. df for a goodness of fit problem _____

1. $(n - 1)s$ to σ^2	4. $k - 1$
2. $\frac{f_r \times f_k}{n}$	5. $\sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$
3. ≥ 5	6. $(r - 1)(c - 1)$

II. Last year, 40% of Linda's customers rented 1 tape, 30% rented 2 tapes, 20% rented 3 tapes, and 10% rented 4 or more tapes. Below is last week's tape rental distribution for Linda's stores. Using the 5-step approach to hypothesis testing, test at the .05 level of significance whether there has been a change in the distribution of tape rentals. Each expected frequency will be the total of 1,000 observations multiplied by last year's appropriate percentage.

Tape Rental Analysis					
	Observed Frequency (f_o)	Expected Frequency (f_e)			
1 tape	300				
2 tapes	250				
3 tapes	250				
4+ tapes	200				
Totals	1,000				

III. Is Linda happy with these test results? Why?

- IV. Using the 5-step approach to hypothesis testing and the .01 level of significance, test whether the number of math courses taken and success in statistics are independent. **People using statistics software do not need to fill out the table.**

Statistics Grades and Math Background at State University			
Grade	Less than B	Greater than or equal to B	Totals
Math courses taken			
Less than or equal to 2	15	5	20
Greater than 2	<u>5</u>	<u>25</u>	<u>30</u>
Totals	20	30	50

Contingency Table of Statistics Grades and Math Background						
Grade	Less than B		Greater than or equal to B		Totals	
	f_o	f_e	f_o	f_e	f_o	f_e
Math courses taken						
Less than or equal to 2	15		5		20	
Greater than 2	<u>5</u>		<u>25</u>		<u>30</u>	
Totals	20		30		50	

Chapter 21 Nonparametric Hypothesis Testing of Ordinal Data Part I

- I. A run test is used to determine randomness based upon order of occurrence.
- A. To be successful, an experiment often requires data be randomly collected.
 1. Inferential statistics often requires data be collected randomly.
 2. Quality control, studied in chapter 17, requires defect testing be done to randomly selected items.
 - B. Data studied pertains to a two category variable (male/female, pass/fail, etc.). The number of runs (similar observations) determines randomness. Too many or too few runs causes rejection of the null hypothesis.
 - C. Linda wants an .05 level test to determine whether the gender of people walking into her store is a random event.
 1. This gender data was collected from Linda's Saturday morning customers. Runs have been underlined.
 2. F F F, M M, F F F F, M, F F F F F, M M M M, F, M M M M, F F F F F, M M M M M, F F, M M, F F F

The sample size of either category is n_1 .
The sample size of the other category is n_2 .
The number of runs is r . The sampling distribution of r is approximately normal provided the sample size of either category (n_1 or n_2) is beyond 20. If both are ≤ 20 , tables containing the critical value of r should be used.
Here are the mean and standard error associated with the sampling distribution of r .
$\mu_r = \frac{2n_1n_2}{n_1+n_2} + 1 \quad \sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$
$Z = \frac{r - \mu_r}{\sigma_r}$ The test statistic is r . If z from the test statistic is beyond the critical value of z , the null hypothesis is rejected.

$n_1 = 23$ females	$n_2 = 18$ males	$r = 13$ runs
--------------------	------------------	---------------

$\mu_r = \frac{2n_1n_2}{n_1+n_2} + 1$ $= \frac{2(23)(18)}{23+18} + 1$ $= \frac{828}{41} + 1$ $= 21.195$	$\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$ $= \sqrt{\frac{2(23)(18)[(2(23)(18) - 23 - 18)]}{(23+18)^2(23+18-1)}}$ $= \sqrt{\frac{651,636}{67,240}}$ $= 3.113$
---	--

$Z = \frac{r - \mu_r}{\sigma_r}$ $= \frac{13 - 21.195}{3.113}$ $= -2.63$	For the .05 level of significance, z is ± 1.96 for this two-tail test. Reject H_0 because -2.63 is beyond -1.96 . Gender of customers walking into Linda's store is not random.
--	--

- D. Run tests may be done using the median. Runs consist of consecutive outcomes larger or smaller than the median. Outcomes equal to the median are ignored.
- II. One-tail testing of one sample median using the sign test
- A. This test is equivalent to a one-tail parametric test of 1 sample mean.
 - B. Data must be at least ordinal in nature and knowledge about the shape of the distribution is not required.
 - C. A (+) sign is assigned to values above the median of interest and a (-) sign to those below the median. Those equal to the median are dropped from the test and n is reduced accordingly.
 - D. Our study of inferential statistics began when Linda became concerned about a drop in the average customer purchase from \$7.75. If Linda does not know the shape of the distribution, she can do a sign test of this year's data against last year's median of \$7.70. Median hourly sale for 7 randomly selected periods will be tested at the .05 level of significance.
 1. If the median has decreased, the proportion of (-) signs should be greater than the proportion of (+) signs.
 2. $H_0: p \geq .50$ and $H_1: p < .50$ (H_1 must be less-than because this is the change being tested.)
 - a. For small samples, the binomial distribution is used to calculate the probability of the distribution tail (observations beyond the proposed median).
 - b. P (often called π) equals .5, n equals total observations, and x equals observations beyond the proposed median. If the probability of the tail is less than the level of significance (α), the null hypothesis is rejected. With a two-tail test, the probability calculation is doubled.
 3. Z is appropriate for large samples with p equal to .50 (see section IC of page 94).
 4. The p-value approach to hypothesis testing will be used with these sign tests.
 - a. Five median sales figures are below \$7.70 and n is 6 because of a tie.
 - b. The binomial table (ST 1) yields the following: $P(x \geq 5) = .094 + .016 = .11$.
 - c. Accept H_0 as .11 is greater than .05. Chance could have caused these decreases.
 - d. With samples of 6, all must decrease to reject H_0 . $P(x = 6) = .016$ and $.016 < .05$

Sample	Median	Sign
1	\$7.65	-
2	\$7.50	-
3	\$8.00	+
4	\$7.60	-
5	\$7.70	0
6	\$7.35	-
7	\$7.55	-

III. Two-tail testing of 2 sample medians from independent populations using the Mann-Whitney test

- A. This hypothesis test is used when populations are not symmetrical and do not have equal variances.
- B. Data must be at least ordinal in nature.
- C. Procedures
 1. Data from 2 samples will be combined into an ordered array. Sample size may differ.
 2. Beginning with the number 1, data will be ranked. Equal data, called ties, will be given their averaged rank.
 3. Ranks will be assigned to their respective sample and the mean rank of each sample calculated.
 4. If population medians are equal, there will be little difference between the mean rank of each sample.
 5. Either mean calculation, U_1 or U_2 , may be used.
 6. The sampling distribution of U will be approximately normal provided both samples n_1 and n_2 are ≥ 10 .
 7. Special procedures, not covered in **Quick Notes Statistics**, are used when either n is less than 10.

D. Twenty-three employees were randomly assigned to training method A or B. Distribution shapes are not known. Linda wants to determine the equality of training methods at the .05 level of significance.

n_1 is sample size #1.	n_2 is sample size #2.	$U = U_1$ or U_2
R_1 is sample 1's rank.	R_2 is sample 2's rank.	

$H_0 : \text{Median}_1 = \text{Median}_2 \quad H_1 : \text{Median}_1 \neq \text{Median}_2$

$z = \frac{U - \mu_U}{\sigma_U}$ U is the test statistic. If z from the test statistic is beyond the critical value of z, H_0 will be rejected. That is, the medians are not equal.

$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$
 or
 $U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$

Method		Rank		Ranked Scores	
A	B	Ordered Array and Method		Method	
Score				A	B
14	12	1.	12	B	
17	21	2.	13	A	2
27	28	3.	14	A	4
19	16	4.	14	B	4
13	30	5.	14	B	4
32	26	6.	16	B	6
22	14	7.	17	A	7
25	18	8.	18	A	8.5
18	28	9.	18	B	8.5
30	22	10.	19	A	10
24	14	11.	21	B	11
33		12.	22	A	12.5
		13.	22	B	12.5
		14.	24	A	14
		15.	25	A	15
		16.	26	B	16
		17.	27	A	17
		18.	28	B	18.5
		19.	28	B	18.5
		20.	30	A	20.5
		21.	30	B	20.5
		22.	32	A	22
		23.	33	A	23
		Totals		R = 155.5 or 120.5	

R_1 has been calculated using the chart.

$\mu_U = \frac{n_1 n_2}{2}$

$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$

$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$
 $= 12(11) + \frac{12(12+1)}{2} - 155.5$
 $= 132 + 78 - 155.5 = 54.5$

$\mu_U = \frac{n_1 n_2}{2}$
 $= \frac{12(11)}{2}$
 $= 66$

$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$
 $= \sqrt{\frac{12(11)(12+11+1)}{12}}$
 $= \sqrt{\frac{3,168}{12}} = 16.248$

$z = \frac{U - \mu_U}{\sigma_U}$
 $= \frac{54.5 - 66.0}{16.248} = -.71$
 This two-tail .05 test has a z of ± 1.96 . Accept H_0 because z of $-.71$ from the test statistic is not beyond -1.96 . There is not a difference between these median scores.

Practice Set 21 Nonparametric Hypothesis Testing of Ordinal Data Part I

- I. Darin wants to determine whether the page 68 computer components were drawn at random. The median of 30.045 mg is the standard for this test. Determine at the .05 level of significance whether this data was randomly collected. Data was recorded one column at a time starting at the top of each column. Columns were recorded from left to right.

29.89	30.05	29.98	30.07	29.97	30.05	29.95	30.06	29.99	30.02	30.09	30.12
29.96	29.97	30.06	30.05	29.95	29.95	29.99	29.89	29.99	30.08	30.06	30.16
29.97	29.98	30.04	30.06	30.05	30.09	30.06	30.09	29.98	30.01	30.08	30.15

- II. Darin first studied the number of defective 30-milligram parts on page 96. At that time he did a parametric study because he felt the data was normally distributed. The consistency of raw material inputs has changed and Darin isn't sure the distribution is still normal. Do a .05 level of significance sign test to determine whether defects have increased from last year's median of 5.

Sample	Median Defects
1	6
2	7
3	5
4	4
5	8
6	6
7	7

- iii. Darin wants to reexamine the number of sick days taken by employees based upon education. This data was first presented on page 100. At that time it was assumed the populations were approximately normal with the same variance. As a result, population means were compared. Assume these assumptions might not be true and use a Mann-Whitney .01 level of significance test to determine whether these samples come from populations with equal medians.

Graduates' sick days: 5, 4, 7, 2, 7, 7, 0, 3, 6, 8, 6 Non-graduates' sick days: 9, 13, 8, 6, 14, 6, 12, 16, 8, 10, 7, 11											
People Using Statistics Software should not use this chart.											
Complete this table by: (1) completing an ordered array, (2) assigning a G for graduates and an N for non-graduates to each element of the array, (3) assigning each rank to the appropriate category (non-graduate or graduate), (4) calculating each subtotal, and (5) calculating R_1 , which equals the sum of the 3 subtotals for non-graduates or R_2 which equals the sum of the 3 subtotals for graduates.											
Rank Ordered Array and Degree Status (1) (2)		Ranked Scores		Rank Ordered Array and Degree Status (1) (2)		Ranked Scores		Rank Ordered Array and Degree Status (1) (2)		Ranked Scores	
		Grads (3)	Non-grads (3)			Grads (3)	Non-grads (3)			Grads (3)	Non-grads (3)
1.				9.				17.			
2.				10.				18.			
3.				11.				19.			
4.				12.				20.			
5.				13.				21.			
6.				14.				22.			
7.				15.				23.			
8.				16.							
(4) Subtotal				(4) Subtotal				(4) Subtotal			
(5) $R_1 =$											

- iii. Darin wants to reexamine the delivery time of 2 suppliers first presented on page 90 and reproduced below. Parametric tests using z or t assume the populations are approximately normal and have equal variances. If these conditions are not met (or unknown) and the shape and dispersion of the distributions are similar, the nonparametric Mann-Whitney test of 2 medians is appropriate. Test at the .05 level of significance whether these samples come from a population with equal medians. For calculation convenience, only the first 11 pieces of data will be used from each data set. **People using statistics software do not need to complete this chart.**

Supplier A: 10, 22, 14, 39, 37, 40, 30, 29, 30, 16, 11 Supplier B: 14, 37, 20, 19, 12, 18, 22, 23, 26, 21, 19								
Complete this table by: (1) completing an ordered array, (2) assigning an A for supplier A and a B for supplier B to each element of the array, (3) assigning each rank to the appropriate category (supplier A or B), (4) calculating each subtotal, and (5) calculating R_1 , which equals the sum of the 3 subtotals for supplier A or R_2 , which equals the sum of the 3 subtotals for supplier B.								
Rank Ordered Array and Supplier (1) (2)	Supplier		Rank Ordered Array and Supplier (1) (2)	Supplier		Rank Ordered Array and Supplier (1) (2)	Supplier	
	A	B		A	B		A	B
1.			8.			15.		
2.			9.			16.		
3.			10.			17.		
4.			11.			18.		
5.			12.			19.		
6.			13.			20.		
7.			14.			21.		
						22.		
(4) Subtotal			(4) Subtotal			(4) Subtotal		
(5) $R_1 =$								

Chapter 22 Nonparametric Hypothesis Testing of Ordinal Data Part II

- I. **Two-tail testing of 2 sample medians from dependent populations using a paired difference sign test**
- A. This test is equivalent to a two-tail parametric test for statistical dependence. (see part VI page 99)
 - B. Data must be at least ordinal in nature. Knowledge of the sampling distribution's shape is not necessary.
 - C. The test uses a (+) sign to represent situations where the first variable is larger than the second variable. It uses a (-) sign to represent the opposite situation. Zero represents a situation where variables are equal. Zero values are excluded from the test. Each time this happens, the sample size is reduced by one.
 - D. If the medians are equal, the proportion of (+) signs should be approximately equal to the proportion of (-) signs.
 1. $H_0: p = .50$ and $H_1: p \neq .50$
 2. For small samples, we use the binomial distribution to determine the likelihood of one of the signs occurring a large number of times. P is equal to .5 and n is equal to the number of observations. If the probability, based upon the observed signs, is greater than the level of significance, the null hypothesis is accepted. Z may be used for large samples with $p = .50$. (see IC of page 94)
 - E. Weekly sales before and after a big promotion at three of Linda's stores were \$1,200, \$1,300 and \$1,400 and \$1,400, \$1,500 and \$1,500 respectively. This data was first studied on page 99. At that time, it was assumed the populations were normal. If this were not the case or unknown, a .10 level sign test of the median could have been conducted.
 - F. The p-value approach to hypothesis testing is used for these sign tests.
 1. This table indicates median sales increased at all 3 stores. Sample size is 3.
 2. The Binomial table (ST 1) yields the following: $P(x \geq 3) = .125$.
 3. For this two-tail test, $p = (.125)(2) = .250$. Accept the null hypothesis because $.25 > .10$. Medians are equal.
 4. This null hypothesis can't be rejected because the sample size is too small.
 - G. One- and two-tail brand preference tests can be done with a paired difference sign test.

Store	Sales Dollars		Sign
	Before	After	
1	1,200	1,400	+
2	1,300	1,500	+
3	1,400	1,500	+

- II. **Testing 3 or more sample medians from independent populations using the Kruskal-Wallis test**
- A. The ANOVA analysis of chapter 18 required populations be normally distributed with equal variances. If these requirements are not met or unknown, the parametric ANOVA test of several means is replaced with the nonparametric Kruskal-Wallis H test of several medians.
 - B. This test complements the Mann-Whitney test of 2 medians.
 - C. This test requires that data from independent random samples be at least ordinal in nature.
 - D. Data is ranked. Ties are assigned the average of their ranks. A true null hypothesis means average group ranks are approximately equal. Special tables, not provided here, should be used if $n < 5$.
 - E. The chapter 18 salesperson's sales data, with a week added so $n = 5$, will be tested for equality of medians at the .05 level of significance. We will not assume normal distributions and use the Kruskal-Wallis test.
 - F. Weekly sales data is ranked with this chart.

$$H = \frac{12}{N(N+1)} \left[\frac{(\sum R_1)^2}{n_1} + \frac{(\sum R_2)^2}{n_2} + \dots + \frac{(\sum R_k)^2}{n_k} \right] - 3(N+1)$$

Where: H is the designated statistic.
 k is the number of samples.
 N is the number of observations.
 n_k is a sample's size.
 R_k is a sample's rank total.
 $df = k - 1 = 3 - 1 = 2 \rightarrow \chi^2 = 5.99$

Weekly Sales (x) in Thousands of Dollars					
Salesperson L		Salesperson M		Salesperson N	
Sales	Rank R_1	Sales	Rank R_2	Sales	Rank R_3
7	9.5	6.0	5.5	9.0	14.0
6	5.5	8.0	12.5	8.0	12.5
7	9.5	6.0	5.5	7.0	9.5
4	2.0	6.0	5.5	10.0	15.0
3	<u>1.0</u>	5.0	<u>3.0</u>	7.0	<u>9.5</u>
$R_1 = 27.5$		$R_2 = 32.0$		$R_3 = 60.5$	

$$H = \frac{12}{15(15+1)} \left[\frac{(27.5)^2}{5} + \frac{(32)^2}{5} + \frac{(60.5)^2}{5} \right] - 3(15+1)$$

$$= .05[151.25 + 204.80 + 732.05] - 3(15+1)$$

$$= 54.405 - 48.000 = 6.405$$

Reject H_0 because $H_0 = 6.41 > 5.99$. Medians are not equal.

Notes: 1) An adjustment, not shown here, is required when there are many ties.
 2) Both the Mann-Whitney test and the Kruskal-Wallis test require populations be of similar shape and dispersion.

Practice Set 22 Nonparametric Hypothesis Testing of Ordinal Data Part II

- I. Darin conducted a training program for 5 recently-hired employees. This problem first appeared on page 100. At that time it was assumed that the population was approximately normal. If this assumption is not correct or unknown, a .01 level of significance paired difference sign test may be conducted to determine whether training increased worker efficiency.

Employee	Efficiency Rating	
	Before	After
1	8	9
2	6	8
3	7	8
4	7	9
5	8	10

- II. Darin wants to reexamine the ANOVA study conducted on page 110. That study assumed populations were normally distributed with equal variances. Those assumptions are not appropriate. Conduct a .01 level of significance Kruskal-Wallis test to determine whether the median weight of parts produced by these 3 departments are equal. Page 110 data has been increased to conform with the $n \geq 5$ test requirement.

Weight Analysis of 9-mg Parts Produced by 3 Departments		
Department 1	Department 2	Department 3
8.95	9.05	9.05
8.90	9.05	9.15
8.90	9.10	9.10
8.92	9.07	9.13
8.88	9.11	9.14

Quick Questions 22 Nonparametric Hypothesis Testing of Ordinal Data Part II

- I. Linda is tracking the number of work days missed by employees before and after taking part in the company-sponsored lunchtime physical fitness program. This problem first appeared on page 101. At that time it was assumed the populations were approximately normal. If this assumption is not correct, a paired difference sign test may be conducted at the .10 level of significance to determine whether median work days missed has changed.

Employee	A	B	C	D	E	F	G
Before	8	9	6	8	3	4	5
After	6	7	5	6	5	2	5

- II. The page 112 ANOVA high school and college grades study assumed the populations were normally distributed with equal variances. These assumptions are not true or unknown. Conduct a .05 level of significance Kruskal-Wallis test to determine the equality of treatment median grades. Page 112 data has been increased to conform with the $n \geq 5$ test requirement.

High H.S. Grades T_1		Medium H.S. Grades T_2		Low H.S. Grades T_3	
College Grades	Rank (R_1)	College Grades	Rank (R_2)	College Grades	Rank (R_3)
3.4		3.2		2.1	
3.5		2.8		2.5	
3.1		3.0		2.7	
3.3		3.1		2.3	
3.6		2.9		1.8	

Inferential statistics is very important so Fred and I made up this special review. Use it with the formula review beginning on the next page. Don't forget to look at cumulative review chapters 25 - 27.



Executive Summary of Inferential Statistics

Being Tested	Sampling Distribution is Known			Sampling Distribution is Unknown						
	Parametric Tests of the Mean and Proportion Using Interval and Ratio Data use with <table style="width: 100%; border: none;"> <tr> <td style="text-align: center; width: 33%;"><u>Normal Population</u></td> <td style="text-align: center; width: 33%;"><u>Skewed Population</u></td> <td style="width: 34%;"></td> </tr> <tr> <td style="text-align: center;">Large Sample σ is known or unknown</td> <td style="text-align: center;">Small Sample σ is unknown¹</td> <td style="text-align: center;">Large Sample σ is known or unknown</td> </tr> </table>			<u>Normal Population</u>	<u>Skewed Population</u>		Large Sample σ is known or unknown	Small Sample σ is unknown ¹	Large Sample σ is known or unknown	Nonparametric Tests of the Median Using Ordinal Data use with <u>Skewed Populations</u> Small Sample
<u>Normal Population</u>	<u>Skewed Population</u>									
Large Sample σ is known or unknown	Small Sample σ is unknown ¹	Large Sample σ is known or unknown								
One Sample	z	t	z	Sign Test						
Two Independent Samples	z	t	z	Mann-Whitney Test						
Two Dependent Samples (paired difference test)	z	t	z	Sign Test						
3 or More Independent Samples (ANOVA)	F	F	Not Applicable	Kruskal-Wallis Test						
1. If σ is known, z may be used in place of t.				Nonparametric Tests of Nominal Data Using χ^2						
One Categorical Variable				Goodness of Fit Test						
Two Categorical Variables (Statistical Dependency)				Contingency Tables						

Inferential Statistics Formula Review

I. Large sample hypothesis testing ($n \geq 30$)

A. One sample mean

1. One-tail testing determines if a mean is different than a given value in a particular direction.
2. Two-tail testing determines if a mean is different than a given value in either direction. Divide α by 2.
3. The test statistic is \bar{x} .

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$H_0: \mu \geq x \text{ and } H_1: \mu < x$$

$$H_0: \mu \leq x \text{ and } H_1: \mu > x$$

$$H_0: \mu = x \text{ and } H_1: \mu \neq x$$

x is the hypothesized population mean.

B. Two sample means

1. One-tail testing determines if one mean is larger or smaller than another.
2. Two-tail testing determines if 2 means are equal. Divide α by 2.
3. The test statistic is \bar{x} .

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$H_0: \mu_1 \geq \mu_2 \text{ and } H_1: \mu_1 < \mu_2$$

$$H_0: \mu_1 \leq \mu_2 \text{ and } H_1: \mu_1 > \mu_2$$

$$H_0: \mu_1 = \mu_2 \text{ and } H_1: \mu_1 \neq \mu_2$$

C. One sample proportion

1. One-tail testing determines if a proportion is different than a given value in a particular direction.
2. Two-tail testing determines if a proportion is different than a given value in either direction. Divide α by 2.
3. The test statistic is \bar{p} .

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$H_0: p \geq x \text{ and } H_1: p < x$$

$$H_0: p \leq x \text{ and } H_1: p > x$$

$$H_0: p = x \text{ and } H_1: p \neq x$$

p is the hypothesized population proportion.

D. Two sample proportions

1. One-tail testing determines if one proportion is larger or smaller than another.
2. Two-tail testing determines if 2 proportions are equal. Divide α by 2.
3. The test statistic is \bar{p} .

$$\bar{p} = \frac{x}{n}$$

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{\bar{p}_w(1-\bar{p}_w)}{n_1} + \frac{\bar{p}_w(1-\bar{p}_w)}{n_2}}}$$

$$\text{and } \bar{p}_w = \frac{\text{Total successes}}{\text{Total sampled}} = \frac{x_1 + x_2}{n_1 + n_2}$$

II. Small sample hypothesis testing ($n < 30$)

A. One sample mean

1. One-tail testing determines if a mean is different from a given value in a particular direction.
2. Two-tail testing determines if a mean is different from a given value in either direction. Dividing α by 2.
3. The test statistic is \bar{x} .

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \text{ and } df = n - 1$$

B. Two sample means from independent populations

1. One-tail testing determines if one mean is larger or smaller than another.
2. Two-tail testing determines if 2 means are equal. Divide α by 2.
3. The test statistic is \bar{x} .

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_w^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \text{ and } S_w^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \text{ and } df = n_1 + n_2 - 2$$

C. Two sample means from dependent populations (paired difference test)

1. One- and two-tail problems may be analyzed.
2. The test statistic is \bar{d} .

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \text{ and } s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}} \text{ and } \bar{d} = \frac{\sum d}{n} \text{ and } df = n - 1$$

3. $H_0 : \mu_d \geq 0$ and $H_1 : \mu_d < 0$ Note: μ_d is negative when H_1 involves testing for an increase.

III. Statistical quality control A. The \bar{x} chart B. The R chart C. The p chart

IV. Analysis of variance

A. Testing 2 sample variances from normal populations

1. One- and two-tail problems may be analyzed.
2. The test statistic is F. $F = \frac{s_1^2}{s_2^2}$ $df = n - 1$ for both the numerator and the denominator
Two-tail test requires dividing the level of significance by 2.

B. Analyzing 3 or more sample means from normally distributed populations (ANOVA)

1. Equality of the means will be tested. $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$
2. The test statistic is F. $F = \frac{MS_T}{MS_E}$
3. This is a one-tail test.

C. Two-factor variance analysis

1. Equality of 3 or more means will be tested for both a treatment variable and a blocking variable.
2. The test statistic is F. $F = \frac{MS_T}{MS_E}$ and $F = \frac{MS_B}{MS_E}$
3. This is a one-tail test.

D. Comparing three or more treatment means to each other

1. Having rejected the null hypothesis when comparing the means of three or more populations, treatment means can then be compared (2 at a time) to determine individual differences.
2. The test statistic is the range for the difference between the treatments.
If the range includes 0, conclude there is not a difference.
3. This is a two-tail test. $(\bar{X}_3 - \bar{X}_1) \pm t \sqrt{MS_E \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$

V. Nonparametric hypothesis testing

A. Goodness of fit tests for expected frequency of one categorical variable

1. Do expected frequencies (equal or proportional) match the observed frequency?
2. The test statistic is chi-square. $\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$ and $f_e \geq 5$ and $df = k - 1$

B. Measuring independence of two categorical variables with a contingency table test

1. Are two variables dependent?
2. The test statistic is chi-square. $\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$ and $f_e = \frac{f_r \times f_c}{n}$ $f_o \geq 5$, and $df = (r - 1)(c - 1)$

C. The run test for determining randomness based upon order of occurrence

$$Z = \frac{r - \mu_r}{\sigma_r} \text{ where } r \text{ is the number of runs, } \mu_r = \frac{2n_1n_2}{n_1+n_2} + 1 \text{ and } \sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

D. One- and two-tail testing of one sample median using a sign test.

E. One- and two-tail testing of 2 medians from independent populations using the Mann-Whitney test.

$$z = \frac{U - \mu_U}{\sigma_U} \text{ where } U_1 = n_1n_2 + \frac{n_1(n_1+1)}{2} - R_1 \text{ and } \mu_U = \frac{n_1n_2}{2} \text{ and } \sigma_U = \sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}$$

F. One- and two-tail testing of 2 medians from dependent populations using the paired difference sign test.

G. The Kruskal-Wallis test for the equality of 3 or more independent sample medians

$$H = \frac{12}{N(N+1)} \left[\frac{(\sum R_1)^2}{n_1} + \frac{(\sum R_2)^2}{n_2} + \dots + \frac{(\sum R_k)^2}{n_k} \right] - 3(N+1)$$

Inferential Statistics Test

- I. A sample of 36 out of 25,000 baseball fans attending a game revealed average refreshment spending of \$7.60. The population standard deviation was \$2.10. The makers of Dud beer will not distribute their product to a ballpark unless it is possible that the average fan spends at least \$8.00 on refreshments. Use the 5-step approach to hypothesis testing and a .01 level of significance to test whether this ballpark qualifies to receive Dud beer.

Data set for those using statistics software			
Refreshment Spending			
4.50	8.00	9.00	9.00
6.95	4.90	7.00	8.05
10.00	8.00	9.50	2.00
11.00	9.00	5.00	8.00
8.05	8.50	10.00	4.80
6.00	4.90	11.00	9.00
6.50	7.00	7.00	8.00
11.00	8.00	5.00	5.75
9.10	6.00	9.10	9.00

- II. A marketing test of chocolate flavored shaving cream revealed a favorable response from 35 of 50 test subjects. Test subjects were chosen at random from the company's 1,200 employees. This product will be manufactured if at least 80% of the potential market like the product.

- A. Using the 5-step approach to hypothesis testing and a .05 level of significance, determine whether the product will be manufactured.

Data set for those using statistics software				
Favorable and Unfavorable Attitudes Toward Chocolate Flavored Shaving Cream				
U	F	F	F	F
F	U	F	F	U
U	F	U	F	F
U	F	F	F	U
F	U	F	F	F
U	F	F	U	F
F	F	F	F	F
U	F	F	U	U
F	F	F	F	F
F	F	F	U	U

- B. What are the pros and cons of using company employees to test this product?

- III. ABC Company is questioning whether the quality of material coming from the company's three suppliers has something to do with the number of defective products. The number of defects from 20 production runs for each supplier were counted. Using a .05 level of significance, determine whether the number of defects and the company supplying materials are related (dependent).

Analysis of Material Suppliers and Defects								
	Company #1		Company #2		Company #3		Totals	
	f_o	f_e	f_o	f_e	f_o	f_e	f_o	f_e
High defects	6		9		15		30	
Low defects	14		11		5		30	
Totals	20		20		20		60	

- IV. Four people were given extensive sales training. Test whether their sales performance improved using a .05 level of significance. Assume normally distributed populations with unknown standard deviations.

Analysis of Sales Training Effectiveness				
Salesperson	Sales Performance			
	Before	After		
A	12	15		
B	13	17		
C	10	14		
D	11	12		
Totals				

- V. Owners of the Quick Chow Restaurant are concerned about the average time to serve customers at two of their stores. A sample of 32 customers at store A resulted in a mean service time of 80 seconds and a standard deviation of 8 seconds. A sample of 49 customers at store B resulted in a mean service time of 75 seconds and a standard deviation of 7 seconds. Test at the .02 level of significance whether the mean time to wait on customers at these two stores is the same.

Data set for those using statistics software			
Store A		Store B	
66	72	79	81
84	73	74	84
70	86	83	68
84	77	85	65
78	72	74	78
83	85	71	63
71	83	88	77
72	90	74	64
87	86	77	62
68	87	85	75
98	73	83	74
78	84	80	78
80	83	70	62
75	93	78	80
72	75	86	69
93	82	71	76
		66	74
		83	68
		80	81
		82	82
		70	75
		64	71
		68	78
		78	66
		75	

- VI. Before recent improvements, it took 36.4 minutes to assemble a part. After improvements, a sample of 16 had an average assembly time of 34 minutes. The sample standard deviation was 2.4 minutes. Test at the .01 level of significance whether improvements lowered assembly time.

Data set for those using statistics software			
Time After Improvements			
35.9	31.8	31.5	36.6
30.8	32.3	32.0	36.2
35.8	35.7	36.8	36.4
32.6	36.8	31.3	31.5

VII. Samples of 10 taken in 1985 and 1995 revealed the average time people spend grocery shopping decreased from 18 minutes to 14 minutes. Respective standard deviations were 5 minutes and 4 minutes. Test at the .10 level of significance whether there has been a change in shopping time variability.

Data set for those using statistics software	
Shopping Time	
1985	1995
19	10
17	17
18	20
16	9
18	8
14	16
23	13
9	18
28	15
18	14

VIII. Test at the .05 level of significance whether workplace accidents happen equally throughout the workweek.

Analysis of Workplace Accidents					
Day	Accidents				
Monday	9				
Tuesday	5				
Wednesday	6				
Thursday	5				
Friday	10				
Totals	35				

- IX. Three computer component assembly methods were compared by Insel Corporation. Employee efficiency was based upon production time and product quality.
- A. Use ANOVA analysis to test at the .05 level of significance whether mean employee efficiency of these assembly methods are equal.

ANOVA Analysis of Assembly Methods				
Employee Efficiency Ratings for 3 Treatments (T)				Row Totals Required for Calculations
	Method 1	Method 2	Method 3	
	Score	Score	Score	
	4	6	8	
	6	7	8	
	7	4	9	
	7	7	9	
$\sum X_T$				
$(\sum X_T)^2$				
n				
$\frac{(\sum X_T)^2}{n}$				
$\sum X_T^2$				

$$SS_T = \sum \left[\frac{(\sum x_T)^2}{n} \right] - \frac{(\sum x)^2}{N}$$

$$SS_E = \sum x^2 - \sum \left[\frac{(\sum x_T)^2}{n} \right]$$

$$SS_{TOTAL} = \sum x^2 - \frac{(\sum x)^2}{N}$$

- B. Determine at the .01 level of significance whether there is a difference in performance of those who received teaching methods (treatments) 1 and 3.

- X. Darin wants to compare assembly time of 30-milligram parts using method A and method B. It is not known whether these populations are approximately normal with the same variance. Use the Mann-Whitney test to determine at the .05 level of significance whether these samples come from populations with equal medians.

Time to Assemble 30-Milligram Parts in Seconds											
Method A	90	95	104	88	91	94	87	102	96	98	101
Method B	95	102	93	105	96	99	100	103	91	97	106

Rank Ordered Array and Assembly Method	Ranked Scores		Rank Ordered Array and Assembly Method	Ranked Scores		Rank Ordered Array and Assembly Method	Ranked Scores	
	Method A	Method B		Method A	Method B		Method A	Method B

XII. Oven temperature at Chewy Pizza restaurants was in control when these samples were taken. Construct an \bar{X} chart and an R chart for this data using a 99.74% confidence interval.

Sample #	1	2	3	4	5	6	Totals
Oven Readings	405	402	398	410	391	411	
	404	404	390	402	409	409	
	397	412	388	412	400	407	
Sample Mean							
Sample Range							

ASTM Control Factors for 99.74%			
Sample Size (n)	A ₂	D ₃	D ₄
2	1.880	0	3.267
3	1.023	0	2.575
4	0.729	0	2.282
5	0.577	0	2.115

XIII. Potential customers were asked to rate brand A and brand B. Little is known about population distributions. Test at the .10 level of significance whether these brands were viewed equally by these potential customers. A paired difference sign test may be conducted even though this is not a test for statistical dependency.

Brand Preference Test		
Customer	Brand A	Brand B
1	87	89
2	91	97
3	81	85
4	73	81
5	92	98
6	89	81

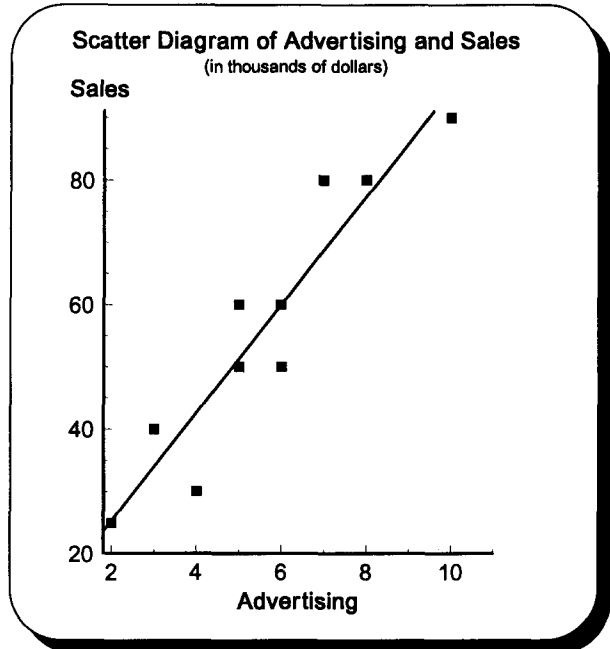
Chapter 23 Correlation Analysis

- I. Correlation analysis measures the strength of the arithmetic relationship between two variables.
- II. Correlation may be visually represented with a scatter diagram.
 - A. Linda Smith is interested in analyzing the relationship between monthly advertising expenditures and monthly sales revenue. Data on these variables was first presented in chapter 7.

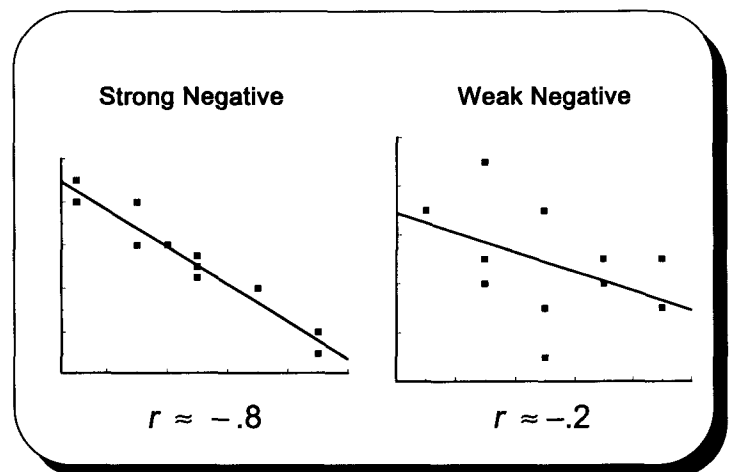
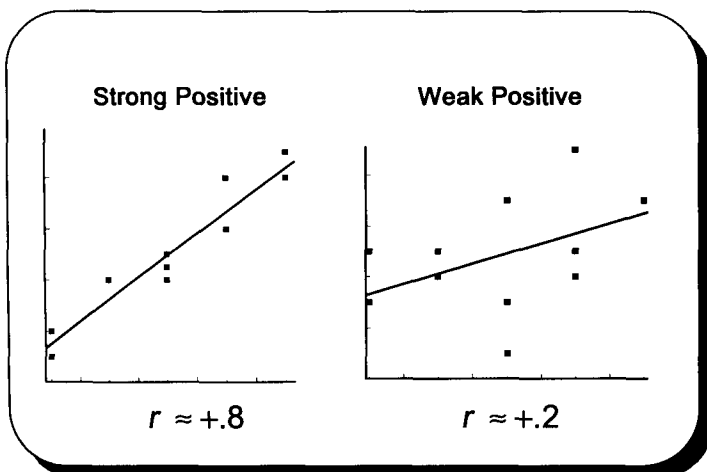
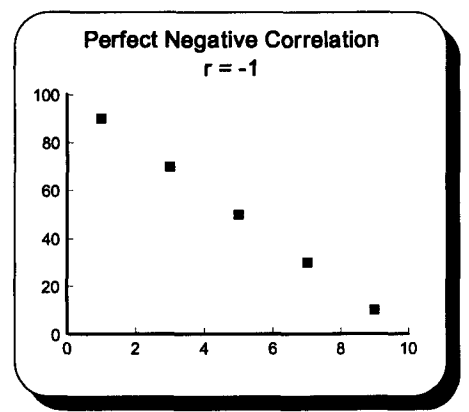
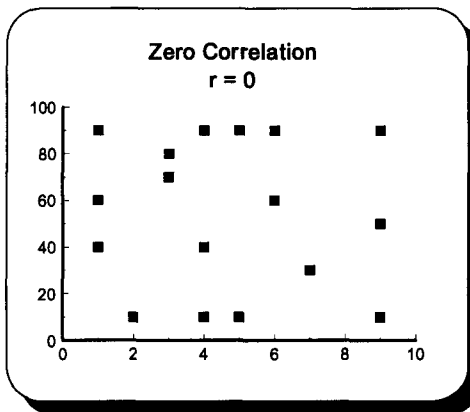
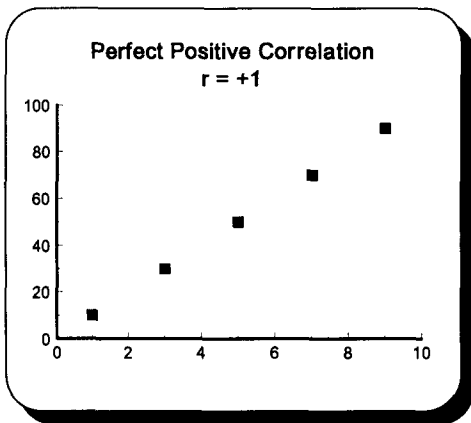
B.

Advertising expenditures (000)	5	2	7	6	10	4	6	5	3	8
Sales revenue (000)	50	25	80	50	90	30	60	60	40	80

- C. She began by making a **scatter diagram** of the data.
1. Sales is the dependent variable because sales revenue, to some degree, is dependent upon advertising expenditures. This dependency was verified on page 121. The dependent variable is graphed on the y-axis.
 2. The independent variable, advertising expenditures, is graphed on the x-axis (abscissa).
 3. In chapter 24, we will learn to draw a regression line through the middle of a scatter diagram.



- III. The sample coefficient of correlation (r)
- A. The coefficient of correlation (r) measures the strength of the relationship between 2 variables. It takes values between ± 1 inclusive. $-1 \leq r \leq +1$
 - B. The closer r is to either extreme, the higher (stronger) is the relationship (correlation).
 1. An r of about .8 or so is high positive correlation.
 2. An r of about .2 to -.2 is low correlation.
 3. An r of about -.8 or so is high negative correlation.



C.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{10(3,600) - (56)(565)}{\sqrt{[10(364) - (56)^2][10(36,225) - (565)^2]}}$$

$$= \frac{(36,000) - (31,640)}{\sqrt{[(3,640) - (3,136)][(362,250) - (319,225)]}}$$

$$= \frac{4,360}{\sqrt{[504][43,025]}}$$

r = .936

Advertising Expenditures (x) (000)	Sales Revenue (y) (000)	x ²	XY	y ²
5	50	25	250	2,500
2	25	4	50	625
7	80	49	560	6,400
6	50	36	300	2,500
10	90	100	900	8,100
4	30	16	120	900
6	60	36	360	3,600
5	60	25	300	3,600
3	40	9	120	1,600
<u>8</u>	<u>80</u>	<u>64</u>	<u>640</u>	<u>6,400</u>
56	565	364	3,600	36,225

IV. Coefficient of determination (r²)

- A. The coefficient of determination measures the total variation of the dependent variable (sales revenue) accounted for by variation of the independent variable (advertising expenditures).
- B. Approximately 88% of the variability in Linda's Video Showcase sales revenue is accounted for by advertising expenditure variability.

$$r^2 = (r)^2 = (.936)^2 = .876$$

V. Coefficient of nondetermination (r̄²)

- A. The coefficient of nondetermination measures the total variation of the dependent variable (sales revenue) not accounted for by variation of the independent variable (advertising expenditures).
- B. Approximately 12% of the variability in Linda's Video Showcase sales revenue is not accounted for by advertising expenditure variability.

$$\bar{r}^2 = 1 - r^2 = 1 - .876 = .124$$

Note: Advertising is not the only variable affecting sales. Multiple correlation and regression, not covered by Quick Notes, analyze the relationship between more than one independent variable and a dependent variable.

A note of caution. We have proven a high mathematical (linear) relationship between these 2 variables. We have not proven a cause-effect relationship.

VI. Measuring the significance of the coefficient of correlation

- A. To be significant, the population coefficient of correlation (ρ , the Greek letter for rho) cannot be zero.
- B. It must be determined whether r is large enough, given some level of significance, to indicate ρ is not zero.
- C. The 5-step approach to hypothesis testing
1. The null hypothesis and alternate hypothesis are $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$.
 2. The level of significance will be .05 for this two-tail problem with $n - 2$ degrees of freedom. Two is subtracted because two variables, x and y, are being estimated.
 3. The relevant statistic is r.

$$df = n - 2 = 10 - 2 = 8 \rightarrow t = 2.306$$

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Note: A large r leads to a large t and a large t leads to rejecting the null hypothesis. ρ is 0 because the H_0 is assumed to be true.

4. If t from the test statistic is beyond the critical value of t, the null hypothesis will be rejected.
5. Apply the decision rule.

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.936 - 0}{\sqrt{\frac{1 - (.936)^2}{10 - 2}}} = 7.52$$

Reject H_0 because $7.52 > 2.306$. This sample is not from a population with a coefficient of correlation equal to zero.

Practice Set 23 Correlation Analysis

I. Darin Jones wants to know whether age of sales personnel affects sales performance. Answer the following questions using this data.

A. Draw a scatter diagram.

Age	Sales Commissions (000)
23	30
25	25
34	20
29	24
21	35
32	22
23	34
24	33
27	27
<u>22</u>	<u>30</u>
260	280

B. Calculate the coefficient of correlation to 3 decimal places. Interpret your answer.

C. What is the coefficient of determination? Interpret your answer.

D. What is the coefficient of nondetermination? Interpret your answer.

E. Is the relationship between age of sales personnel and their sales commissions significant at the .01 level?

Quick Questions 23

Correlation Analysis

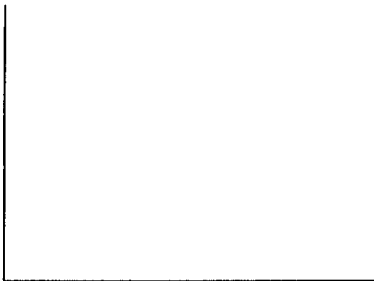
I. Place the number of the appropriate formula, expression, or term next to the appropriate concept.

- A. Coefficient of determination _____
- B. Coefficient of correlation _____
- C. A range for r _____
- D. Coefficient of nondetermination _____
- E. The test statistic (t) used to measure the significance of r _____

1.	$1 - r^2$, the variability in y that is not explained by x
2.	$\frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$
3.	$\frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$
4.	r^2 , the variability in y that is explained by x
5.	$-1 \leq r \leq +1$

II. Draw the following scatters and place the appropriate value for r in the space provided.

Perfect Positive Correlation



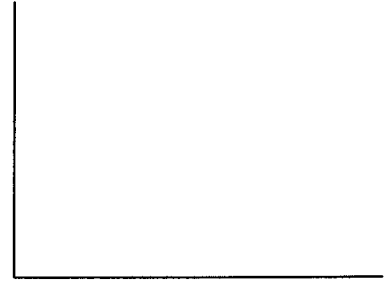
$r = \underline{\hspace{2cm}}$

Zero Correlation



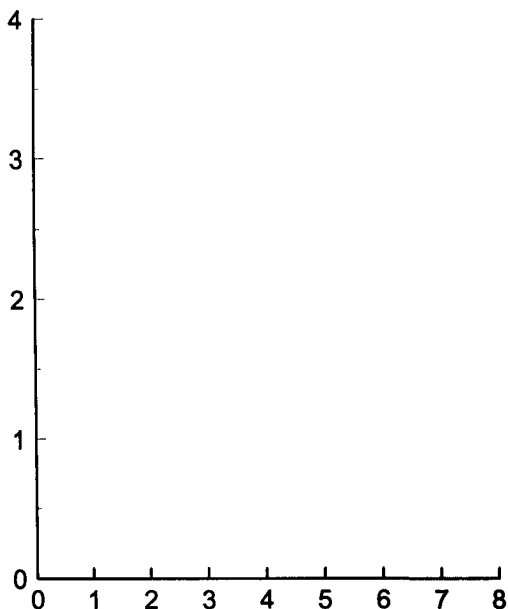
$r = \underline{\hspace{2cm}}$

Perfect Negative Correlation



$r = \underline{\hspace{2cm}}$

III. Draw a scatter diagram showing how hours studying per weekend affect grade point average.



Hours Studying per Weekend	Grade Point Average
3	3.0
2	2.0
6	3.8
3	2.6
4	3.2
8	3.7
2	2.1
3	2.8

IV. Using the data in question III, calculate the following:

A. Coefficient of correlation (to 3 decimal places)

B. Coefficient of determination

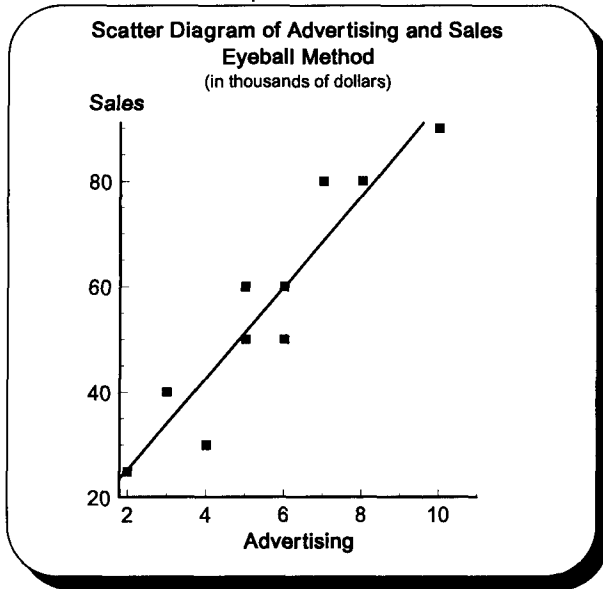
C. Coefficient of nondetermination

D. Interpret your answer to question IV B.

V. Could ρ (rho) be zero at the .01 level of significance?

Chapter 24 Simple Linear Regression Analysis

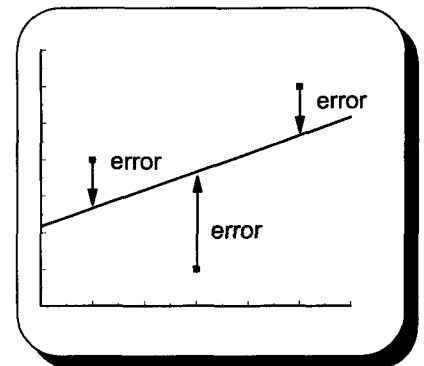
- I. Simple regression analysis defines the mathematical relationship between 2 variables.
- A. A scatter diagram depicts the relationship between the independent variable (advertising) on the x-axis and a dependent variable (sales) on the y-axis (see graph).
 - B. A line through the scatter plot can be used to mathematically define this relationship.
 1. The line can be estimated using the eyeball method by drawing a line with a ruler that divides the data in half.
 2. A regression equation may be used to more exactly define the relationship between two variables.



Linda's Video Showcase Advertising Expenditures and Sales Revenue				
Advertising Expenditures (X) (000)	Sales Dollars (Y) (000)	x^2	XY	y^2
5	50	25	250	2,500
2	25	4	50	625
7	80	49	560	6,400
6	50	36	300	2,500
10	90	100	900	8,100
4	30	16	120	900
6	60	36	360	3,600
5	60	25	300	3,600
3	40	9	120	1,600
<u>8</u>	<u>80</u>	<u>64</u>	<u>640</u>	<u>6,400</u>
56	565	364	3,600	36,225

- II. Determining a regression equation using the method of least squares
- A. Many different lines can be drawn through a scatter plot using a ruler.
 - B. The method of least squares gives more consistent results.
 - C. This technique results in a straight line that minimizes the sum of the squared vertical deviations between the resulting line and the individual data. These deviations may be thought of as error.
 - D. This is the general form of the regression equation.

$\hat{y}_{.x} = a + bx$	$\hat{y}_{.x}$ is the estimated value of y based upon a given value for x. The period next to \hat{y} is read "given" and this expression is read "y estimated given x."
where	a is the y-intercept (where the line crosses the y-axis). b is the slope of the line. It equals $\Delta y + \Delta x$.



- E. Determining the regression equation to 3 significant digits.

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$= \frac{10(3,600) - (56)(565)}{10(364) - (56)^2}$$

$$= \frac{4,360}{504} = 8.6507936$$

$$a = \bar{Y} - b\bar{X}$$

$$a = \frac{\sum Y}{n} - b\frac{\sum X}{n}$$

$$= \frac{565}{10} - 8.6507936\left(\frac{56}{10}\right)$$

$$= 8.055556$$

$$\hat{y}_{.x} = a + bx$$

$$\hat{y}_{.x} = 8.06 + 8.65x$$

- F. The example to the right uses the regression equation to calculate estimated monthly sales when advertising expenditures are \$9,000.

$$\hat{y}_{.x} = 8.06 + 8.65x$$

$$\hat{y}_{.9} = 8.06 + 8.65(9)$$

$$\hat{y}_{.9} = 8.06 + 77.85 = 85.91 \text{ or } \$85,910$$

III. Drawing a regression line

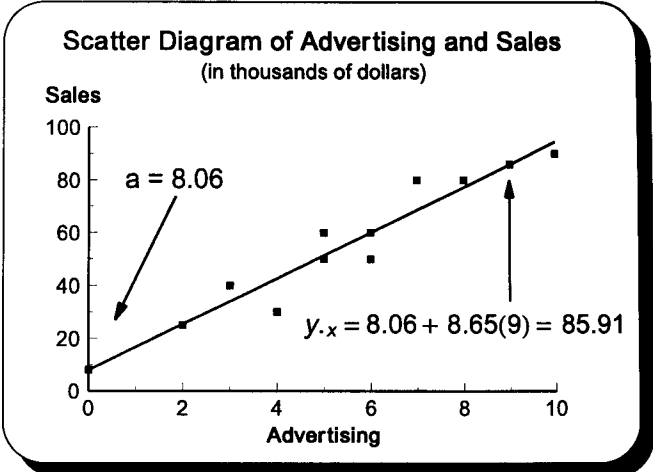
- A. Two points (x,y) may be used to draw a straight line.
- B. The y-intercept (0, \$8,060) will be one point.
- C. The estimated value of y for x of \$9,000 is \$85,910. It will be the second point (see page 152).

IV. The standard error of the estimate

- A. The standard error of the estimate measures the dispersion of the scatter (plots) around the regression line.
- B. It is the standard deviation of y given some value of x.
- C.

$$S_{y,x} = \sqrt{\frac{\sum(Y - \bar{Y})^2}{n-2}} = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n-2}}$$

$$S_{y,x} = \sqrt{\frac{36,225 - 8.05556(565) - 8.6507936(3,600)}{10 - 2}} = 8.145$$



V. An interval estimate for the conditional mean of y for some given value of x

- A. A confidence interval will be determined using the small sample t distribution.
- B.

$$\hat{y}_{.x} \pm ts_{y,x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

Note: The correction factor following the standard error of the estimate is needed because the sample is small and the scatter of sales data might not be normal.

- C. Linda Smith wants to determine the 95% confidence interval for expected sales for months when advertising expenditures are \$9,000.

Basic Assumptions Concerning Linear Regression Analysis

1. There are a number of y values for each value of x.
2. The conditional distributions of y given x are normal.
3. The variance of the conditional distributions are equal.
4. Predictions of y are limited to the existing range for x.
Note: Predicting an individual value (next month's sales) rather than the mean of Y (sales) requires inserting a +1 under the radical.

Problem Notes

$\hat{y}_{.x} = 8.06 + 8.65(x) = \$85,910$ when $x = 9$. See page 152.
Degrees of freedom for t will be $n - 2$ because both a and b were estimated in determining $\bar{y}_{.x}$. $df = n - 2 = 10 - 2 = 8$
$\alpha/2 = .05/2 = .025 \rightarrow 2.306$ for t
$\bar{x} = \frac{\sum x}{n} = \frac{56}{10} = 5.6$ $S_{y,x} = 7.89$ $n = 10$

$$\hat{y}_{.x} \pm ts_{y,x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

$$85.91 \pm 2.306(8.145) \sqrt{\frac{1}{10} + \frac{(9 - 5.6)^2}{364 - \frac{56^2}{10}}}$$

$$85.91 \pm 10.779$$

$$75.131 \leftrightarrow 96.689$$

- D. For regression analysis to be valid, the range for variables a and b must consist of realistic values. Here, the y-intercept cannot be negative because negative sales are not possible. But, determining the 95% confidence interval for the y-intercept (0,8.06) by recalculating acceptable error (E) results in a negative lower limit (8.06 - 15.96 = -7.90). This concern might be solved by lowering the standard error of the estimate with a larger sample. In addition, procedures exist for determining a confidence interval for the slope. The possibility of a negative slope would cause people to question the relationship between advertising and sales. A larger sample might also solve the problem of a negative slope.

Practice Set 24 Simple Linear Regression Analysis

I. Having determined that age affects sales performance, Darin Jones wants to estimate sales commissions using the data presented in the chapter 23 practice set.

A. Determine the regression equation to 3 significant digits.

Age	Sales Commissions (000)	xy	x ²	y ²
23	30	690	529	900
25	25	625	625	625
34	20	680	1,156	400
29	24	696	841	576
21	35	735	441	1,225
32	22	704	1,024	484
23	34	782	529	1,156
24	33	792	576	1,089
27	27	729	729	729
<u>22</u>	<u>30</u>	<u>660</u>	<u>484</u>	<u>900</u>
260	280	7,093	6,934	8,084

B. Estimate sales commissions for a group of 24-year-old salespeople.

C. Graph the regression line.

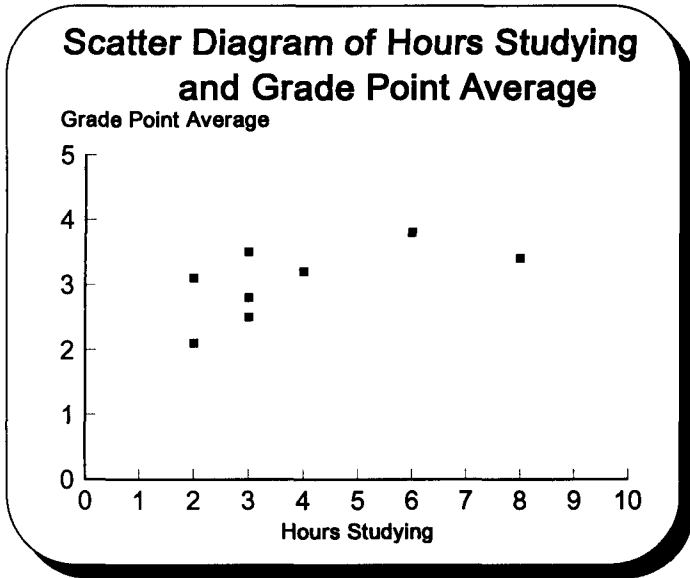
D. Determine the 99% confidence interval for the question B group.

E. What procedure should be followed if question D's range includes negative numbers?

Quick Questions 24 Simple Linear Regression Analysis

- I. Place the number of the appropriate formula, symbol, or expression next to the concept it describes.
- A. The standard error of the estimate _____
 - B. The y-intercept _____
 - C. The regression equation _____
 - D. The estimated value of y given x _____
 - E. The slope _____
 - F. An interval estimate for the conditional mean of Y _____
 - G. An interval estimate for an individual value of Y _____
- II. The following data was first presented in chapter 23. Estimate the regression line for this scatter using the eyeball method.

1.	$\hat{y}_{\cdot x} = a + bx$
2.	$\hat{y}_{\cdot x} \pm ts_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$
3.	$\bar{Y} - b\bar{x}$
4.	$\hat{y}_{\cdot x}$
5.	$\frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$
6.	$\hat{y}_{\cdot x} \pm ts_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$
7.	$\sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n - 2}}$



Hours Studying per Weekend	Grade Point Average	XY	X ²	Y ²
3	3.0	9.0	9	9.00
2	2.0	4.0	4	4.00
6	3.8	22.8	36	14.44
3	2.6	7.8	9	6.76
4	3.2	12.8	16	10.24
8	3.7	29.6	64	13.69
2	2.1	4.2	4	4.41
<u>3</u>	<u>2.8</u>	<u>8.4</u>	<u>9</u>	<u>7.84</u>
31	23.2	98.6	151	70.38

- III. Calculate the regression equation. Round the slope and y-intercept to three significant digits.

IV. Estimate the grade point average for people who studied 5 hours per weekend.

V. Draw the regression line on the page 156 scatter diagram.

VI. Calculate the 98% confidence interval for students who study 5 hours per weekend.

VII. What procedure should be followed if the range for your answer to question E includes negative numbers?

Correlation and Regression Formula Review

I. Correlation formulas

A. Coefficient of correlation $r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$

B. Coefficient of determination $r^2 = (r)^2$

C. Coefficient of nondetermination $\bar{r}^2 = 1 - r^2$

D. The value of t when determining the significance of the coefficient of correlation r $t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$ and df = n - 2

II. Regression formulas

A. The regression equation $\hat{y}_{\cdot x} = a + bx$

B. The slope of the regression equation $b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$

C. The y-intercept of the regression equation $a = \bar{Y} - b\bar{x} = \frac{\sum y}{n} - b \frac{\sum x}{n}$

D. The standard error of the estimate $S_{y \cdot x} = \sqrt{\frac{\sum (Y - \bar{Y})^2}{n - 2}} = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n - 2}}$

E. An interval estimate for the conditional mean of y for some given value for x

$$\hat{y}_{\cdot x} \pm ts_{y \cdot x} \quad \text{or} \quad \hat{y}_{\cdot x} \pm ts_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

Note: An interval estimate for an individual value of y, sales for a recently hired 24-year-old salesperson or grades for your roommate who studied 5 hours, would require adding a 1 under the radical. This makes the interval substantially larger.

$$\hat{y}_{\cdot x} \pm ts_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

Correlation and Regression Test

I. Place the number of the appropriate formula, expression, or term next to the appropriate concept.

- A. The independent variable _____
- B. The dependent variable _____
- C. Measures the strength in the relationship between two variables _____
- D. The variation of the dependent variable explained by the independent variable _____
- E. The variation of the dependent variable not explained by the independent variable _____
- F. Used when testing the significance of r _____
- G. The regression equation _____
- H. The slope of the regression line _____
- I. Where a regression line crosses the y-axis _____
- J. The standard error of the estimate _____

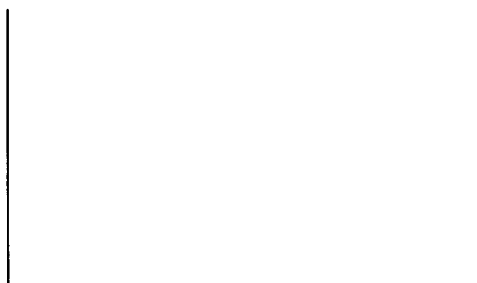
1.	r
2.	b
3.	$1 - r^2$
4.	x
5.	$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$
6.	$\sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n - 2}}$
7.	y
8.	a
9.	r^2
10.	$\hat{y}_{.x} = a + bx$

II. Draw the following scatters and place an appropriate value for r in the space provided.

High Positive Correlation $r \approx$ _____



Low Negative Correlation $r \approx$ _____



Zero Correlation $r =$ _____



Perfect Positive Correlation $r =$ _____



III. Answer the following questions using this data that was gathered to determine whether research and development expenditures affect profit.

R & D Expenditures Millions	Profits in Millions
5	30
3	40
7	60
6	60
10	80
4	40

A. The coefficient of correlation

B. The coefficient of determination and the coefficient of nondetermination

C. Could rho be zero at the .05 level of significance?

IV. Interpret your answers to question III.

V. Draw a scatter diagram of the above data and use the eyeball method to estimate the regression curve.

- VI. Answer the following questions using the data on the preceding page.
- A. Use the method of least squares to determine a regression equation.

B. Calculate the estimated profit for next year when R & D will be \$8,000,000.

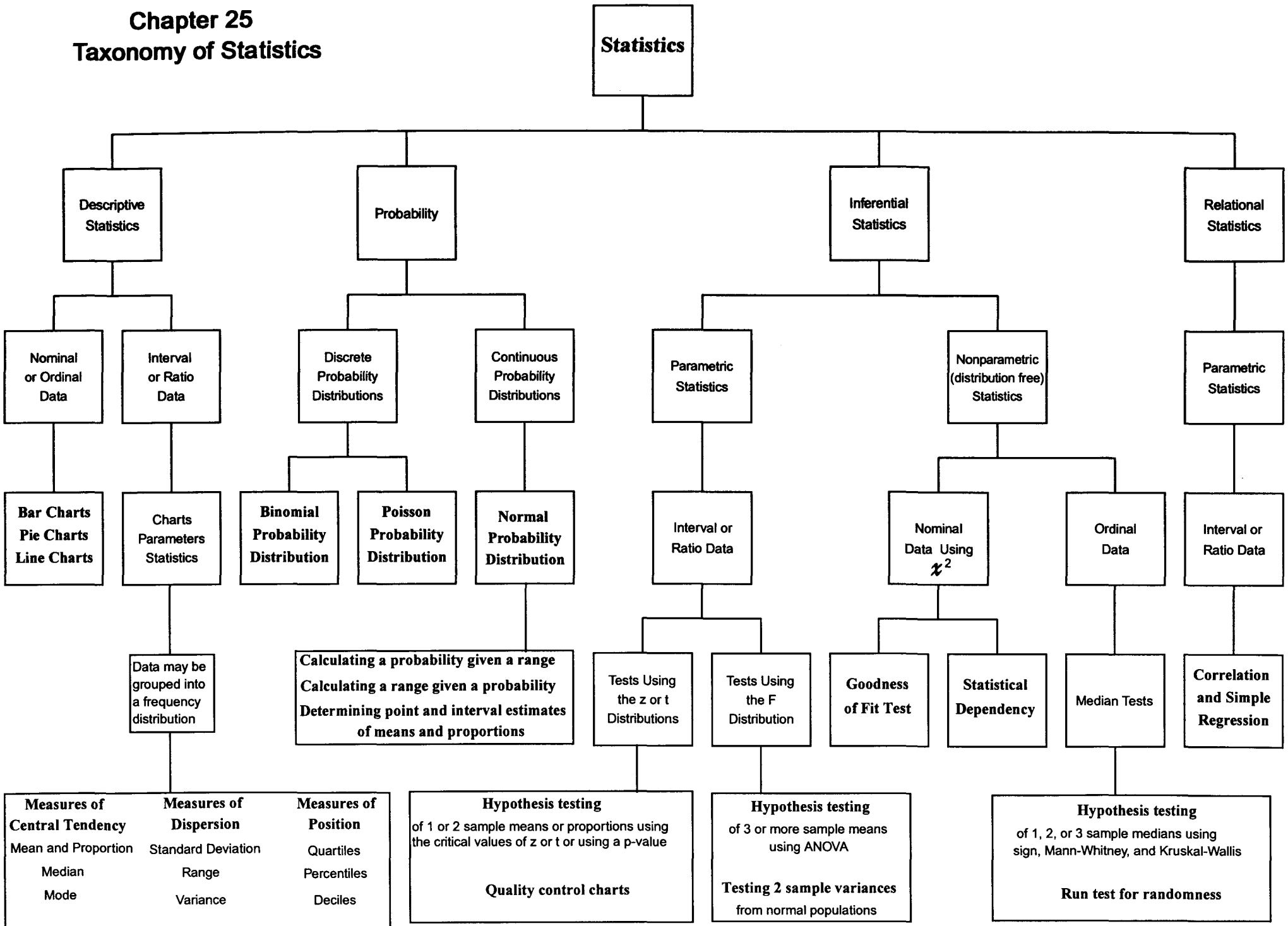
C. Draw the regression line on the page 160 scatter diagram.

D. Calculate the 99% confidence interval for question B.

E. What procedure should be followed if the range for the answer to question D includes zero or a negative number?

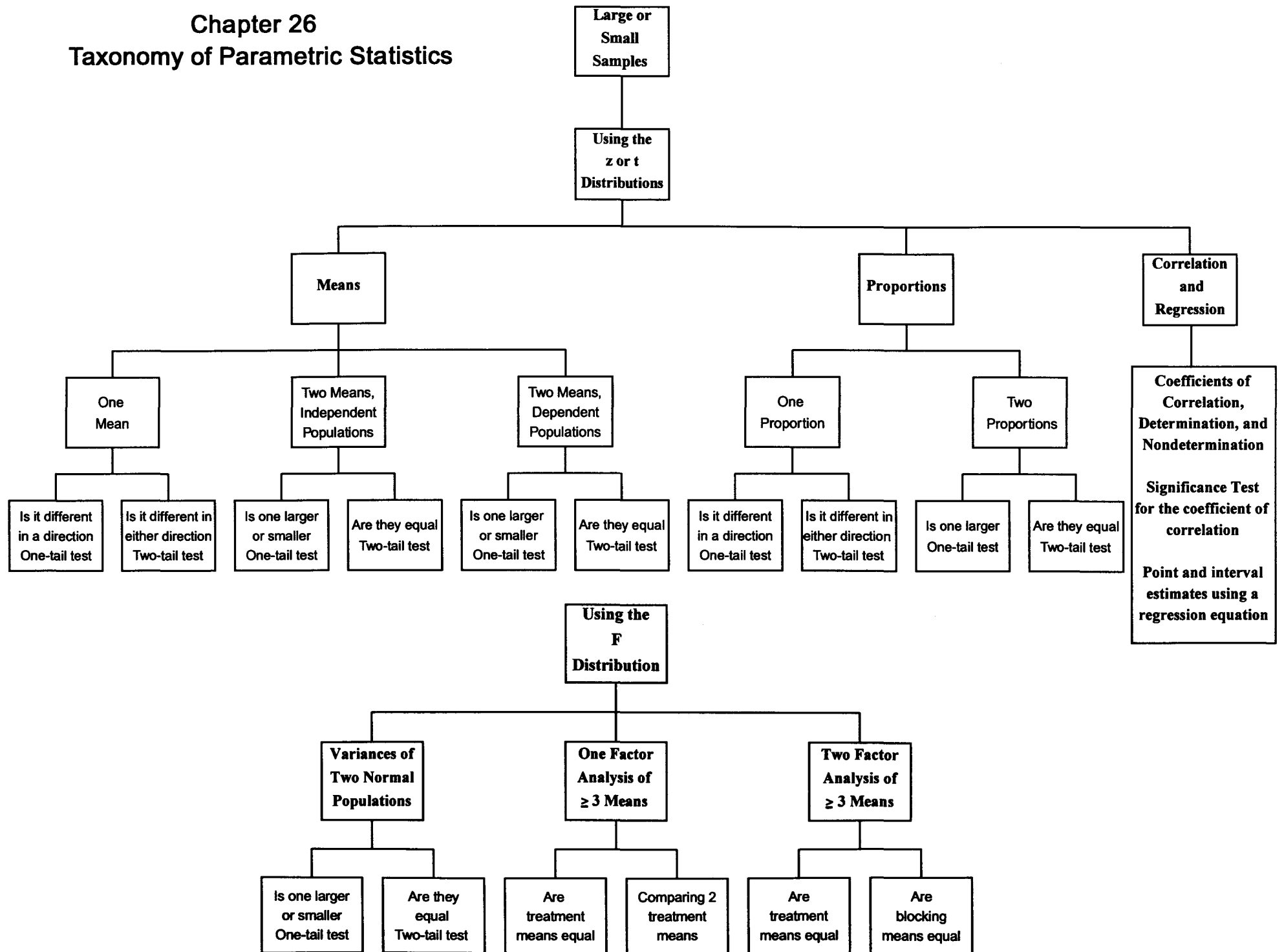
Chapter 25

Taxonomy of Statistics



Chapter 26

Taxonomy of Parametric Statistics



Chapter 27 Problem Review

Linda's Video Showcase

Descriptive Statistics (Chapters)	
2. Summarizing Data	The number of video rentals was summarized with an array, range, frequency distribution, relative frequency distribution, histogram, frequency polygon, relative frequency polygon, more-than ogive, and less-than ogive.
3. Measuring Central Tendency of Ungrouped Data	Last week's self-help tape rentals were used to estimate last year's self-help tape rentals. Population and sample measures included the mean, weighted mean, median, and mode. Measures of position calculated by Linda included quartiles, deciles, percentiles, and the inter-quartile range.
4. Measuring Dispersion of Ungrouped Data	The dispersion of self-help tape rentals was analyzed with a range, average deviation, variance, and standard deviation. The usefulness of the standard deviation was explained with a normal curve using the empirical rule, Chebyshev's rule, and the coefficient of variation.
5. Measuring Central Tendency of Grouped Data	Measures calculated above were recalculated for grouped data using the video tape rentals summarized in chapter 2. The skewness of nonsymmetrical data was defined, graphed, and measured with Pearson's coefficient of skewness.
6. Measuring Dispersion of Grouped Data	Measures calculated above were recalculated for grouped data. The first and third quartiles, the interquartile range, and percentiles were calculated. Kurtosis was used to describe the shape of a frequency polygon.
Probability, The Basis for Inferential Statistics	
7. Understanding Probability	The general and special rules for addition were used to study the relationship between advertising expenditures and sales revenue.
8. Probability Part II Multiplication Rules	The general and special rules for multiplication were used to study the relationship between advertising expenditures and sales revenue. The counting and factorial rules were used to determine Linda's options when visiting competitors. Permutations and combinations were used to determine Linda's options when displaying advertising posters.
9. Discrete Probability Distributions	The expected value of tape rentals was determined with a probability distribution. Flipping a coin was used to explain the binomial probability distribution. The average number of repair calls per 15-minute period were analyzed using a Poisson probability distribution. Average customer returns were analyzed using a Poisson approximation to the binomial probability distribution.
10. Continuous Normal Probability Distributions	Linda determined the probability of a store's population mean sales being within a given range. She also determined a range for a store's population mean sales given a probability.
11. The Sampling Distribution of the Means	The 99% confidence interval for population mean customer purchases was determined using a sample mean of \$7.50 from a sample size of 49 customers.
12. Sampling Distributions Part II	The 95% confidence interval for the population proportion of customers happy with service was determined using a sample proportion of .80 from a sample of 100 customers. An appropriate sample size was determined given an acceptable range for the population mean. An appropriate sample size was also determined given an acceptable range for the population proportion.

Inferential Statistics	
13. Large Sample Hypothesis Testing	The sample mean purchase of \$7.50 was causing some concern. Linda used hypothesis testing to determine that the population mean customer purchase had decreased from last year's \$7.75. A .01 level of significance was used. Having proved the mean purchase had decreased, she then used a two-tail test to prove the mean did not change in either direction.
14. Large Sample Hypothesis Testing Part II	Linda then used sample data and hypothesis testing to determine whether average sales of two of her stores were the same. A .01 level of significance was used. The two-tail problem, concerning a change in mean customer purchases described in chapter 13, was redone using a p-value test. A population mean of \$7.40 was found to have a type II error of 18.41%.
15. Hypothesis Testing of the Population Proportions	Linda used hypothesis testing to prove that a sample proportion measuring customer satisfaction with service of .80 was not low enough to conclude that the population proportion was below the .85 required to qualify for a Flopbuster Video franchise. She also determined that service at two stores was the same with a two-tail test. She used a one-tail test to determine that one store did not have better service than another store.
16. Small Sample Hypothesis Testing Using Student's t Test	Linda used a mean of 2.3 tape rentals from a sample of 9 customers to determine that the average number of tapes rented per customer had decreased from last year's population mean of 2.6 tapes. She also found the average customer waiting time at two of her stores (two independent populations) was the same. Linda used a paired difference test of normally distributed dependent populations to conclude a promotional campaign had increased weekly sales at three of her stores.
17. Statistical Quality Control	Linda didn't use statistical quality control to manage her retailing business. The chapter explained mean, range, and proportion of defects charts for parts that were designed to be 50 millimeters long.
18. Analysis of Variance Parts I and II	Linda found two of her stores had equal sales variance. She also used ANOVA to prove that average sales of three salespeople were not equal. Weeks of experience, the blocking variable, also indicated unequal sales. The treatment variable salespeople explained half (14 of 28) the total sales variability. Variability of 11.3 was explained by the blocking variable experience. Variability of 2.7 was unexplained. Salespersons one and three had different average weekly sales.
20. Nonparametric Hypothesis Testing of Nominal Data	Linda used a goodness of fit test to determine that sales of a new hit music video were not equally distributed among her four stores. Two categorical variables (advertising expenditures and sales revenue) were found to be statistically dependent using a contingency table.
21. Nonparametric Hypothesis Testing of Ordinal Data Part I	A run test proved the gender of customers entering a store was not a random event. The parametric test indicating an average customer purchase had decreased from \$7.75 was done again with a non-parametric sign test because Linda was not sure of the distribution's shape. A very small sample reversed the earlier result. Linda also found two instructional methods had equal test results at the .05 level of significance using a Mann-Whitney test of two independent population medians.
22. Nonparametric Hypothesis Testing of Ordinal Data Part II	In chapter 16, Linda found advertising expenditures and sales were dependent using a paired difference test that required populations be normally distributed. This assumption was dropped and the study was redone using a paired sign test of two dependent population medians. The earlier results were reversed. The sample size was only five and the study should be redone with more stores. The chapter 18 ANOVA test of three people's mean sales assumed the populations had equal variances. This assumption was dropped and a Kruskal-Wallis test of 3 medians had a similar result.
Correlation and Regression	
23. Correlation	A correlation coefficient of .936 was calculated for advertising expenditures and sales revenue. The coefficient of determination explained 87.6% of the variability between advertising expenditures and sales revenue. The coefficient of nondetermination showed 12.4% unexplained variability. A range for r proved rho, the population coefficient of correlation, could not be zero at the .05 level of significance.
24. Simple Linear Regression Analysis	The regression equation was $\hat{Y}_x = 8.06 + 8.65x$ (in thousands of dollars). It was used to estimate sales revenue of \$85,910 when \$9,000 is spent on advertising. A range of \$75,131 to \$96,689 was calculated for the \$85,910 sales estimate. A check of the y-intercept range resulted in a negative number. Negative sales are not possible and a larger sample must be taken.

Problem Review

Darin's Music Emporium and Future Horizons Corporation

Descriptive Statistics (Chapters)	
2. Summarizing Data	Walkman video recorder sales were summarized with an array, range, frequency distribution, relative frequency distribution, histogram, frequency polygon, and more-than ogive.
3. Measuring Central Tendency of Ungrouped Data	Measures of central tendency for Practice Set 2 data were calculated.
4. Measuring Dispersion of Ungrouped Data	Measures of dispersion for this data were calculated.
5. Measuring Central Tendency of Grouped Data	Grouped measures of central tendency for this data were calculated.
6. Measuring Dispersion of Grouped Data	Grouped measures of dispersion for this data were calculated.
Probability, The Basis for Inferential Statistics	
7. Understanding Probability	The general and special rules for addition were used to study the relationship between customer age and customer buying habits (making a sale).
8. Probability Part II Multiplication Rules	The general and special rules for multiplication were used to study the relationship between customer age and customer buying habits. The factorial rules for permutations and combinations were used to determine Darin's options when displaying advertising posters.
9. Discrete Probability Distributions	The expected unit value of walkman video recorder sales was determined using a probability distribution. The binomial probability distribution was used to analyze the probability of making a Walkman video recorder sale. The average number of customer complaints per 20-minute period were analyzed using a Poisson probability distribution. The average number of bounced checks was analyzed using the Poisson approximation to the binomial probability distribution.
10. Continuous Normal Probability Distributions	Darin determined the probability of population mean sales commissions being within a given range. He also determined a range for the population mean number of customer merchandise returns given a probability.
11. The Sampling Distribution of the Means	The 99% confidence interval for the population mean weight of computer parts was determined using a sample mean of 30.025 mg from a sample of 36 parts.
12. Sampling Distributions Part II	The 95% confidence interval for the population proportion of parts passing inspection was determined using a sample proportion of .90 from a sample of 50 parts. An appropriate sample size was determined given an acceptable range for the population mean weight of parts. An appropriate sample size was also determined given an acceptable range for the population proportion of parts passing inspection.

Inferential Statistics	
13. Large Sample Hypothesis Testing	Darin used hypothesis testing and a sample mean of 30.025 mg to determine whether the population mean weight of parts was above the required limit of 30 milligrams. Using a two-tail test, he determined that parts were not different from 30 mg at the .01 level of significance. However, parts were different from 30 mg at the .05 level of significance.
14. Large Sample Hypothesis Testing Part II	Darin used sample data and hypothesis testing to determine that delivery time for 2 of his suppliers was the same at the .05 level of significance. A chapter 13 test, which proved parts were not too heavy, was redone using a p-value test. Type II error for the weight of material containers hypothesis was determined and graphed with an operating characteristic curve. A power curve showing the probability of not making a type II error was estimated.
15. Hypothesis Testing of the Population Proportions	Darin used hypothesis testing to prove that the .90 sample proportion of parts passing inspection was not high enough to conclude an increase from the .86 population proportion recorded last year. Parts produced by the day and evening shifts had the same proportion of defects.
16. Small Sample Hypothesis Testing Using Student's t Test	Darin found average sick days taken by non-high school graduates were different than those taken by high school graduates (2 independent populations). Darin also found employee efficiency increased because of a training program (2 dependent populations).
17. Statistical Quality Control	Darin constructed mean, range, and proportion of defects charts for the 30 milligram parts first introduced in chapter 11.
18. Analysis of Variance Parts I and II	Darin found that the variance of 30-mg parts had not increased. He used ANOVA to prove the average weight of parts produced by 3 departments was not equal. A blocking variable, when produced, had equal means. Therefore, time of production did not affect part weight. The treatment variable department explained .057 variability. The blocking variable time explained .0071 variability. Variability of .0014 was unexplained. The average weight of parts produced by departments #1 and #3 was found to be different.
20. Nonparametric Hypothesis Testing of Nominal Data	Darin used a goodness of fit test to determine that defects from 3 shifts followed his .20, .30, and .50 expected distribution. Two categorical variables, customer age and making a sale, were found to be statistically independent using a contingency table.
21. Nonparametric Hypothesis Testing of Ordinal Data Part I	Darin used a run test to determine the 30-milligram parts first presented on page 68 were drawn at random. He then used a sign test to determine median defects had not increased from the median of 5 recorded last year. An earlier parametric test (page 100) indicating the inequality of mean sick days taken by graduates and non-graduates was confirmed using a Mann-Whitney median test.
22. Nonparametric Hypothesis Testing of Ordinal Data Part II	An earlier study (page 100) measuring the effects of employee training using a paired difference test of dependent population means was redone using a paired sign test of dependent populations. According to both studies, training improved efficiency. An earlier ANOVA test comparing the mean weight of 9-mg parts from three departments was redone with a Kruskal-Wallis test of 3 medians. The earlier test was confirmed.

Correlation and Regression

23. Correlation	A .908 coefficient correlation was calculated for a person's age and their sales commissions. The coefficient of determination explained 82.4% of the variability between age and sales commissions. The coefficient of nondetermination showed 17.6% unexplained variability. A range for r proved ρ , the population coefficient of correlation, could not be zero at the .01 level of significance.
24. Simple Linear Regression Analysis	The regression equation for age of sales people and their sales commissions was $\hat{Y}_x = 55.9 - 1.07x$ (in thousands of dollars). It was used to estimate commissions of \$30,200 for 24-year-old salespeople. A range of \$27,470 to \$32,930 was calculated for the \$30,200 commission's estimate. A check of the range for commissions (y) excluded the possibility of negative commissions. Therefore, a larger sample was not necessary.

Appendix I

Complete Solutions to Practice Sets

Practice Set complete solutions have been separated from Quick Question complete solutions to facilitate reading them in sequence, as if they were a business case. The result is an example of how statistical analysis may be used for decision making. This grouping also allows for an easy comparison of related items located in different chapters. Appendix I page numbers begin with the letters PS and match their corresponding Practice Set page number.

Reviewing Practice Set Solutions
is a great way to study for tests.



Note: Quick Question answers begin on page QQ 3.
Answers in **Quick Notes** may differ slightly from those
generated by statistics software.

Practice Set 2 Summarizing Data

I. Darin recently collected the following Walkman CD Recorder sales data.

Units sold per day: 17, 22, 17, 8, 12, 15, 14, 16, 21, 29, 16

A. Make an array and calculate the range of this data.

Array: 8, 12, 14, 15, 16, 16, 17, 17, 21, 22, 29
Range: H - L = 29 - 8 = 21

B. Calculate an appropriate class width for this data.

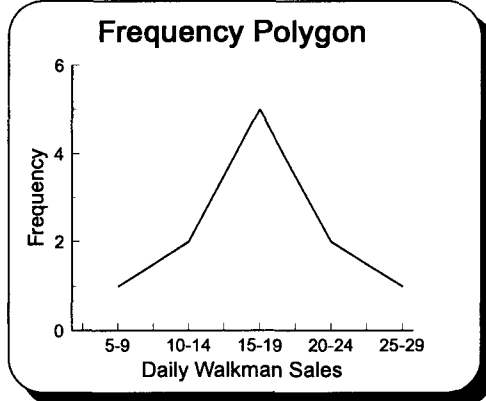
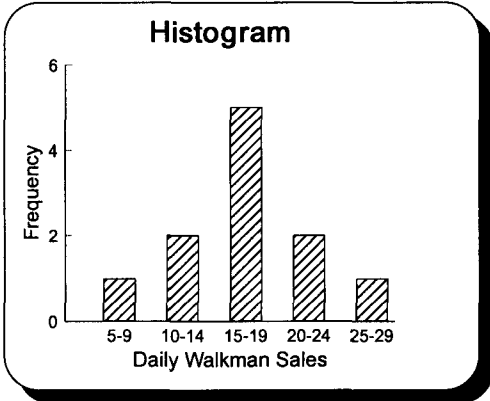
$$\frac{\text{range}}{\text{\# of classes}} = \frac{29 - 8}{5} = 4.2 \rightarrow 4 \text{ or } 5$$

II. Use the first three columns of this chart to make a 5-class frequency distribution. Use stated class limits for the first class of 5 - 9 sales units. Then answer the following questions.

Darin's Music Emporium Daily Walkman Sales Data					
Stated Class Limits	Tally	Frequency (f)	Relative Frequency $f \div n$	Cumulative Frequency	
				More-than	Less-than
5 - 9	I	1	0.09	11	0
10 - 14	II	2	0.18	10	1
15 - 19	III	5	0.46	8	3
20 - 24	II	2	0.18	3	8
25 - 29	I	1	0.09	1	10
Total frequency (n)		11	1.00	0	11

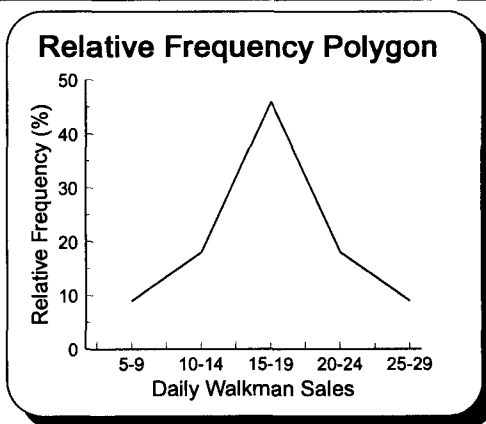
A. Draw or print a histogram.

B. Draw or print a frequency polygon.



C. Draw or print a less-than cumulative relative frequency polygon and a relative frequency polygon.

First answer requires dividing the less-than chart data by 11 or adding the relative frequencies.



Practice Set 3 Measuring Central Tendency of Ungrouped Data

- I. Darin Jones wants to know more about the sales of Walkman CD recorders/players described on page 6. Calculate the sample mean using this Walkman sales data from the last Practice Set. State the formula for the population mean.

Array of daily Walkman sales: 8, 12, 14, 15, 16, 16, 17, 17, 21, 22, 29

A. Sample mean

$$\bar{X} = \frac{\sum x}{n} = \frac{187}{11} = 17$$

B. Population mean formula

$$\mu = \frac{\sum X}{N}$$

- II. Darin sells three different Walkman CD recorders; one for \$149, one for \$159, and a third for \$169. Of the 187 machines sold during this eleven-day period; 43 were the least expensive, 90 were moderately priced, and 54 were the expensive model. Calculate the weighted mean sales price for these machines.

$$\bar{X}_w = \frac{\sum (W_x X_x)}{\sum W_x} = \frac{(43)(\$149) + (90)(\$159) + (54)(\$169)}{43 + 90 + 54} = \frac{\$6,407 + \$14,310 + \$9,126}{187} = \frac{\$29,843}{187} = \$159.59$$

- III. Using the data from question I, prove that the sum of the deviations from a mean is zero.

x	8	12	14	15	16	16	17	17	21	22	29
μ	17	17	17	17	17	17	17	17	17	17	17
$X - \mu$	-9	-5	-3	-2	-1	-1	0	0	4	5	12

$$\sum (x - \mu) = -9 + (-5) + (-3) + (-2) + (-1) + (-1) + 0 + 0 + 4 + 5 + 12 = 0$$

- IV. The median number of Walkman units sold is 16.

$$\frac{n}{2} + .5 = \frac{11}{2} + .5 = 5.5 + .5 = 6 \rightarrow 16$$

Note: Counting 6 positions from the left or right of the array yields 16 as the 6th number.

- V. The mode for this data is 16 and 17.

- VI. This data can be described as bimodal.

- VII. Calculate the following measures of position. Those using computer software should use a less-than cumulative relative frequency distribution to answer these questions.

A. $Q_1 \quad \frac{n}{4} + .5 = \frac{11}{4} + .5 = 2.75 + .5 = 3.25 \rightarrow 14.25$

B. $Q_3 \quad \frac{3n}{4} + .5 = \frac{33}{4} + .5 = 8.25 + .5 = 8.75 \rightarrow 20$

Note: $17 + .75(21-17) = 20$

C. Interquartile range $Q_3 - Q_1 = 20 - 14.25 = 5.75$

D. 6th decile $\frac{xn}{10} + .5 = \frac{6(11)}{10} + .5 = \frac{66}{10} + .5 = 6.6 + .5 \rightarrow 7.1 \rightarrow 17$

E. 95th percentile $\frac{xn}{100} + .5 = \frac{95(11)}{100} + .5 = \frac{1,045}{100} + .5 = 10.45 + .5 = 10.95 \rightarrow 28.65$

Practice Set 4

Measuring Dispersion of Ungrouped Data

- I. Darin is concerned about Walkman sales variability. First calculate the range for Walkman sales and then the average deviation, the standard deviation, and the variance.

Array of daily Walkman sales: 8, 12, 14, 15, 16, 16, 17, 17, 21, 22, 29

Sample mean: 17

A. Range $H - L = 29 - 8 = 21$

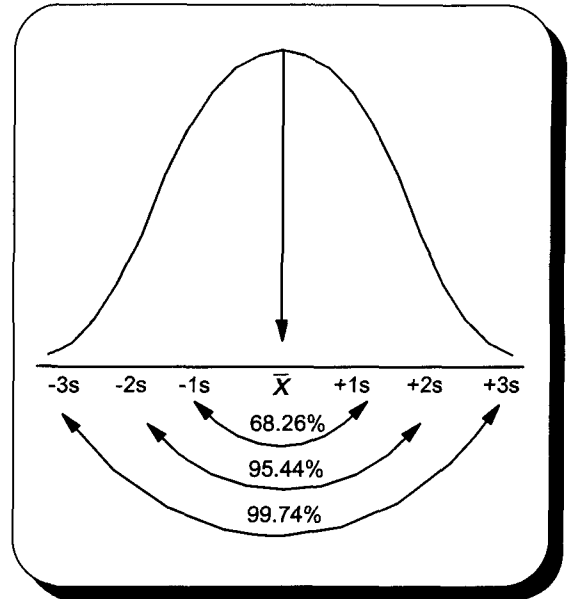
B. Sample average deviation $\frac{\sum |x - \bar{x}|}{n} = \frac{42}{11} = 3.8$

C. Sample variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{306}{10} = 30.6$$

D. Sample standard deviation

$$s = \sqrt{s^2} = \sqrt{30.6} = 5.53 \rightarrow 5.5$$



- II. Label this graph depicting the empirical rule.

- III. Last year's mean weekly Walkman sales were 16 and the standard deviation was 4. Use the empirical rule to determine a range for Walkman sales for one, two, and three sample standard deviations from the mean.

A. One Standard Deviation

$$16 \pm 1(4)$$

$$16 \pm 4$$

68.26% range: 12 ↔ 20

B. Two Standard Deviations

$$16 \pm 2(4)$$

$$16 \pm 8$$

95.44% range: 8 ↔ 24

C. Three Standard Deviations

$$16 \pm 3(4)$$

$$16 \pm 12$$

99.74% range: 4 ↔ 28

- IV. Use Chebyshev's rule to determine a range for Walkman sales being within two sample standard deviations of the mean (see question III).

$$1 - \frac{1}{k^2}$$

$$= 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} \rightarrow 75\%$$

- V. Darin read in a trade publication that the average Walkman sales and standard deviation for a store his size and type are 18 and 3 respectively. Using the sample data from page 18, are Darin's Walkman sales more or less variable than those of his industry? Use the standard deviation calculated in question 1.

Industry Sales Data

$$C.V. = \frac{s}{\mu}(100) = \frac{3}{18}(100) = 16.67\%$$

Darin's Music Emporium

$$C.V. = \frac{s}{\bar{x}}(100) = \frac{5.5}{17}(100) = 32.35\%$$

Sales from this small sample of only 11 days were twice as variable as industry population data.

Practice Set 5 Measuring Central Tendency of Grouped Data

- I. Label the top row of this Walkman sales data chart and calculate these measures of central tendency.

Array of Walkman sales data from page 6

8, 12, 14, 15, 16, 16, 17, 17, 21, 22, 29

- A. Grouped mean

$$\bar{X} = \frac{\sum fx}{n} = \frac{187}{11} = 17$$

- B. Grouped median

$$\frac{n}{2} = \frac{11}{2} = 5.5$$

$$\begin{aligned} L + \frac{\frac{n}{2} - CF_b}{f}(i) \\ = 14.5 + \frac{\frac{11}{2} - 3}{5}(5) \\ = 14.5 + \frac{2.5}{5}(5) \\ = 14.5 + 2.5 = 17 \end{aligned}$$

Darin's Music Emporium Walkman Sales Data			
Stated Class Limits	Frequency (f)	x	fx
5 - 9	1	7	7
10 - 14	2	12	24
15 - 19	5	17	85
20 - 24	2	22	44
25 - 29	1	27	27
	n = 11		Σ fx = 187

- C. Grouped mode

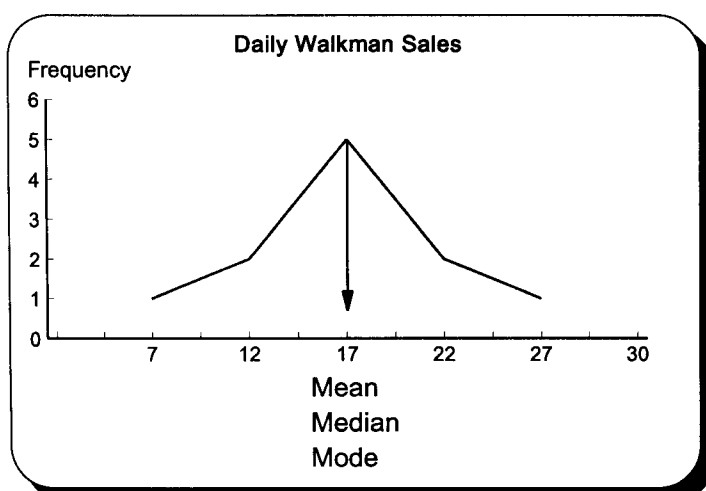
The midpoint of the class with the highest frequency is 17.

- II. Do your answers to question I differ from those calculated on pages 12 and 13? Is the difference large? Could the difference be large?

Measure	Ungrouped (page 12)	Grouped
Mean	17	17
Median	16	17
Mode	16 and 17	17

Difference will be minimal if the midpoint of a class adequately represents the data of that class. Here, the differences are minimal, even though the sample is small.

- III. Draw a frequency polygon of page 24 data and locate the mean, median, and mode.



- IV. Using the mean of 17 and median of 17 calculated on page 24, and a sample standard deviation of 5.5 to be calculated on page 30, calculate Pearson's coefficient of skewness.

$$\begin{aligned} \frac{3(\bar{x} - MD.)}{s} \\ = \frac{3(17 - 17)}{5.5} \\ = 0 \end{aligned}$$

Practice Set 6 Measuring Dispersion of Grouped Data

I. Label this chart of the page 24 frequency distribution and calculate the following measurements.

Darin's Music Emporium Walkman Sales Data					
Stated Class Limits	Frequency (f)	x	fx	x ²	fx ²
5 - 9	1	7	7	49	49
10 - 14	2	12	24	144	288
15 - 19	5	17	85	289	1,445
20 - 24	2	22	44	484	968
25 - 29	1	27	27	729	729
Totals		11	187		3,479

A. Range

$$H - L = 29.5 - 4.5 = 25$$

B. Sample variance

$$\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1} = \frac{3,479 - \frac{(187)^2}{11}}{11-1} = \frac{3,479 - 3,179}{10} = 30$$

C. Sample standard deviation

$$S = \sqrt{S^2} = \sqrt{30} = 5.5$$

D. Quartiles

1. First $\frac{n}{4}$

$$\begin{aligned} Q_1 &= L + \frac{\frac{n}{4} - CF_b}{f}(i) \\ &= 9.5 + \frac{\frac{11}{4} - 1}{2}(5) \\ &= 9.5 + \frac{1.75}{2}(5) \\ Q_1 &= 13.9 \end{aligned}$$

2. Second $\frac{n}{2}$

$$\begin{aligned} Q_2 &= L + \frac{\frac{n}{2} - CF_b}{f}(i) \\ &= 14.5 + \frac{\frac{11}{2} - 3}{5}(5) \\ &= 14.5 + \frac{2.5}{5}(5) \\ Q_2 &= 17.0 \end{aligned}$$

3. Third $\frac{3n}{4}$

$$\begin{aligned} Q_3 &= L + \frac{\frac{3n}{4} - CF_b}{f}(i) \\ &= 19.5 + \frac{\frac{33}{4} - 8}{2}(5) \\ &= 19.5 + \frac{2.25}{2}(5) \\ Q_3 &= 20.1 \end{aligned}$$

E. Interquartile range

$$Q_3 - Q_1 = 20.1 - 13.9 = 6.2$$

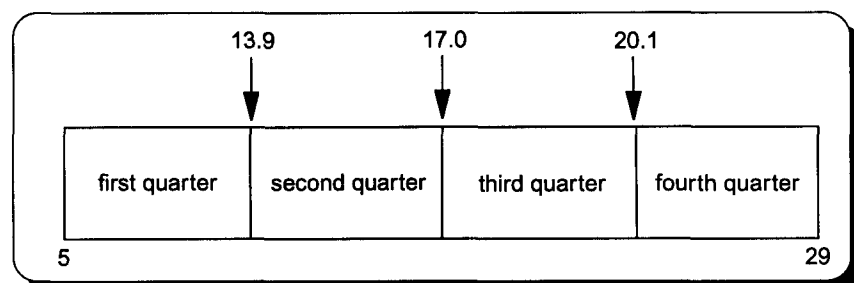
F. 80th percentile

$$P_x = L + \frac{\frac{xn}{100} - CF_b}{f}(i)$$

$$P_{80} = 19.5 + \frac{\frac{80(11)}{100} - 8}{2}(5) = 19.5 + \frac{8}{2}(5) = 19.5 + 2 = 21.5$$

$$\frac{xn}{100} = \frac{80(11)}{100} = 8.8$$

II. Locate the three quartile measures calculated above on this number line.



Practice Set 7 Understanding Probability

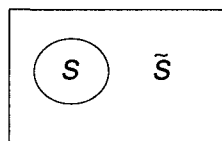
I. Darin collected the following information concerning customer age and making a sale. Please complete this chart.

Customer Age and Making A Sale			
Customer Age	Less than or equal to 20	Over 20	Totals
Making A Sale			
No	16	8	24
Yes	<u>24</u>	<u>12</u>	<u>36</u>
Totals	40	20	60

II. Solve the following problems using the data from question I. Be sure to use a formula and draw a Venn diagram.

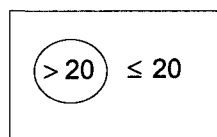
A. The probability of making a sale.

$$P(S) = \frac{S}{n} = \frac{36}{60} = .6 \rightarrow 60\%$$



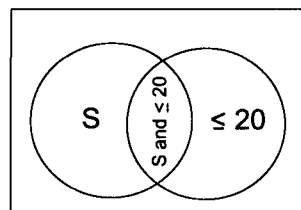
B. The probability of a customer being over 20.

$$P(> 20) = \frac{>20}{n} = \frac{20}{60} = .333 \rightarrow 33.3\%$$



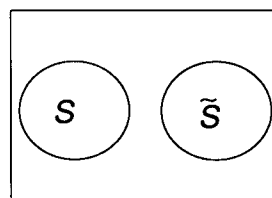
C. The probability of making a sale or a customer being less than or equal to 20.

$$\begin{aligned} P(S \text{ or } \leq 20) &= P(S) + P(\leq 20) - P(S \text{ and } \leq 20) \\ &= P\left(\frac{36}{60}\right) + P\left(\frac{40}{60}\right) - P\left(\frac{24}{60}\right) \\ &= \frac{52}{60} \\ &= .867 \\ &= 86.7\% \end{aligned}$$



D. The probability of making a sale or not making a sale.

$$\begin{aligned} P(S \text{ or } \tilde{S}) &= P(S) + P(\tilde{S}) \\ &= P\left(\frac{36}{60}\right) + P\left(\frac{24}{60}\right) \\ &= \frac{60}{60} \\ &= 1.00 \\ &= 100\% \end{aligned}$$



E. State the addition rule used to answer question C. What condition is necessary to apply this rule?

1. C was solved with the general rule of addition.
2. It is used when events are not mutually exclusive. The events intersect.

F. State the addition rule used to answer question D. What condition is necessary to apply this rule?

1. D was solved with the special rule for addition.
2. It is used when events are mutually exclusive. The events do not intersect.

Practice Set 8 Probability Part II Multiplication Rules

- I. Below is the data Darin Jones collected concerning sales to customers of different ages. (see page 42) Convert Table 1 to decimals and place the information into Table 2.

Analysis of Sales By Age of Customer (Table 1)				Decimal Analysis (Table 2)		
Customer Age Sale	Less than or equal to 20	Over 20	Totals	Less than or equal to 20	Over 20	Totals
No	16	8	24	0.267	0.133	0.40
Yes	24	12	36	0.400	0.200	0.60
Totals	40	20	60	0.667	0.333	1.00

- II. Use a formula to calculate the probability of these events and check your answers using Table 2.
- A. The probability of a customer being over 20 years old.
 $P(> 20) = \frac{>20}{n} = \frac{20}{60} = .333 \rightarrow 33.3\%$ Note how 33.3% can be read directly from Table 2.
- B. The probability of a customer being over 20 years old and not making a sale.
 $P(> 20 \text{ and } \bar{S}) = P(> 20) P(\bar{S}) = \frac{20}{60} \times \frac{24}{60} = \frac{480}{3,600} = .133 = 13.3\%$ See Table 2
- C. The probability of a customer being less than or equal to 20 years old and over 20 years old. These events are mutually exclusive, their intersection is empty. $P(\leq 20 \text{ and } > 20) = 0$
- D. Was the special rule of multiplication applicable to question B? Why or why not? Could the special rule of multiplication be used by Linda with the page 46 advertising data? Why or why not?
1. The special rule is appropriate because the events are independent. Age does not affect buying habits as demonstrated by the fact that 60% of both age groups make a purchase.
 2. The special rule for multiplication is not appropriate for the page 46 problem because sales and advertising are dependent.
- III. Use Bayes' theorem to calculate the probability of making a sale given a customer is less than or equal to 20 years of age.

$$P(S | \leq 20) = \frac{P(S \text{ and } \leq 20)}{P(\leq 20)} = \frac{P(S) \times P(\leq 20 | S)}{P(S) \times P(\leq 20 | S) + P(\bar{S}) \times P(\leq 20 | \bar{S})} = \frac{\frac{36}{60} \times \frac{24}{36}}{\frac{36}{60} \times \frac{24}{36} + \frac{24}{60} \times \frac{16}{24}} = \frac{.40}{.40 + .267} = .60 = 60\%$$

- IV. Recalculate your answer to question III using Table 2 on page 48. .400 ÷ .667 = .6 or 60%

- V. Use Linda's page 46 advertising data to calculate the possibility of having monthly advertising over \$5,000 and monthly sales over \$50,000.

$$P(A > \$5,000 \text{ and } S > \$50,000) = P(S > \$50,000) P(A > \$5,000 | S > \$50,000)$$

$$\frac{5}{10} \times \frac{4}{5} = \frac{20}{50} = 40\%$$

- VI. Answer these questions about 5 posters Darin has to advertise a new CD recorder/player.

- A. How many ways can he arrange these posters in a horizontal line across a wall?

$$N! = 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120 \text{ arrangements}$$

- B. How many ways can he arrange only 3 posters? Arrange implies that order counts. AB is not the same as BA and that both should be counted.

$${}_N P_R = \frac{N!}{(N-R)!}$$

$${}_5 P_3 = \frac{5!}{(5-3)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} = 5 \times 4 \times 3 = 60$$

- C. How many ways can he just hang them? (order doesn't count)

$${}_N C_R = \frac{N!}{(N-R)!(R!)}$$

$${}_5 C_3 = \frac{5!}{(5-3)!3!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = \frac{5 \times 4}{2 \times 1} = 10$$

Practice Set 9 Discrete Probability Distributions

- I. Darin sells three different Walkman CD recorders; one for \$149, one for \$159, and a third for \$169. Of the 187 machines sold during a recent period, 43 were the least expensive, 90 were moderately priced, and 54 were the expensive model.

A. Calculate the expected price of Walkman CD recorders.

Sales Price (x)	Number of Sales	Probability P(x)	x • P(x)
\$149	43	43/187 = .230	\$34.27
159	90	90/187 = .481	76.48
169	54	54/187 = .289	48.84
	187	1.00	\$159.59

B. Compare this answer to the page 12 weighted mean sales value of Walkman sales.

The answers are the same.

C. In theory, what is the difference between a weighted mean of variable x and the expected value of x?

A weighted mean concerns existing data and the expected value of x concerns data that could exist.

- II. When waiting on a customer, Darin's salespeople make a sale 60% of the time (see page 42). Use the binomial formula to calculate the probability of making exactly 3 sales to 5 customers.

Given

$$p = .6$$

$$q = 1 - p = .4$$

$$n = 5$$

$$x = 3$$

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$P(3) = \frac{5!}{3!(5-3)!} .6^3 .4^{5-3}$$

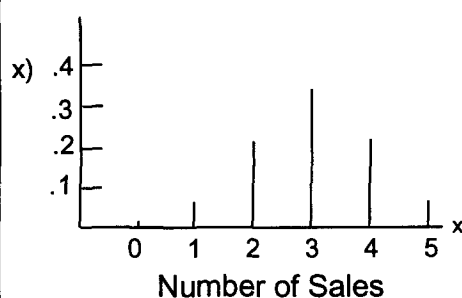
$$= \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 2 \times 1} \times .216 \times .16$$

$$= 10 \times .03456 = .3456 \text{ or } 34.6\%$$

- III. Using the appropriate table, complete the binomial distribution described by question II.

Binomial Probability Distribution n = 5, p = .6 and q = 1 - p = .4	
# of sales (x)	P(x)
0	.010
1	.077
2	.230
3	.346
4	.259
5	.078
Total	1.000

Note: Lulu thought a graph of this distribution might prove interesting.



IV. Using the answer to question III or statistics software, answer the following questions.

- A) .259 B) $.346 + .259 + .078 = .683$ C) $1 - .683 = .317$ D) $1 - .078 = .922$

V. Darin wants to know how busy his complaint department is during a 20-minute period. Data shows the expected number of calls is highly skewed with an average of only 1.0 calls per 20-minute period.

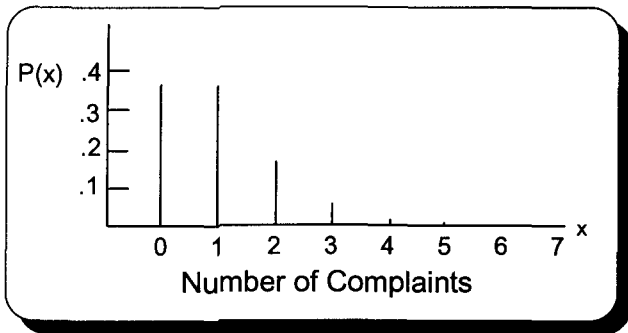
A. Assuming a Poisson probability distribution and using a formula or statistics software, is the probability of zero calls being received in a 20-minute period over or under 50%?

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

$$P(0) = \frac{(1.0^0)2.7183^{-1}}{0!} = \frac{(1)(0.3679)}{1} = 0.3679 = 36.8\%$$

Under

B. Using a table or statistics software, complete and draw this distribution.



Poisson Table	
x	$\mu = 1$
0	0.3679
1	0.3679
2	0.1839
3	0.0613
4	0.0153
5	0.0031
6	0.0005
7	0.0001

C. What is the probability of at least 3 calls being received in a 20-minute period?

$$P(\geq 3) = .0613 + .0153 + .0031 + .0005 + .0001 = .0803 = 8.03\%$$

or

$$P(< 3) = 1 - (.9197) = .0803 = 8.03\%$$

VI. Darin wants to know the number of customers who will bounce a check. Last year only .2% of the 1,000 checks deposited from customers did not clear. This year Darin expects 1,500 customers will pay by check and with the economy being about the same, the same percent of checks should bounce.

A. Can the Poisson approximation of the binomial be used to solve this problem? Why?

$$np = (1,500)(.002) = 3$$

Yes because $n \geq 30$ and $np < 5$

B. What is the expected number of bounced checks for this year?

$$E(x) = \mu = np = (1,500)(.002) = 3$$

C. What is the probability that no one will bounce a check this year?

From Table II, $P(x = 0) = .0498$ or 4.98%

D. What is the probability that at least 2 checks will bounce?

$$P(\geq 2) = [1 - (.0498 + .1494)] = 1 - .1992 = .8008 \text{ or } 80.08\%$$

E. What would you think if 5 checks had bounced by the end of May?

Five checks bouncing by the end of May is unlikely according to last year's data. Last year's data might not apply to this year.

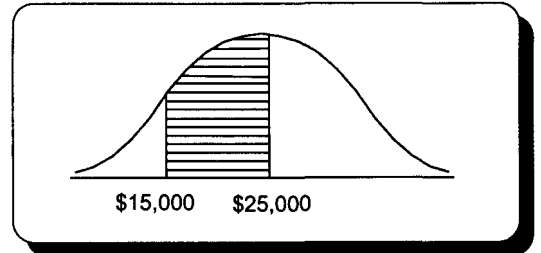
Poisson Table	
x	$\mu = 3$
0	0.0498
1	0.1494
2	0.2240
3	0.2240
4	0.1680
5	0.1008
6	0.0504
7	0.0216
8	0.0081
9	0.0027
10	0.0008
11	0.0002
12	0.0001

Practice Set 10 Continuous Normal Probability Distributions

- I. Sales commissions paid by Darin's Music Emporium are normally distributed with a mean of \$25,000 and a standard deviation of \$5,000. Solve the following being sure to draw a graph of each distribution.

A. $P(\$15,000 \leq x < \$25,000)$

$$Z = \frac{x - \mu}{\sigma} = \frac{\$15,000 - \$25,000}{\$5,000} = \frac{-\$10,000}{\$5,000} = -2 \rightarrow .4772$$

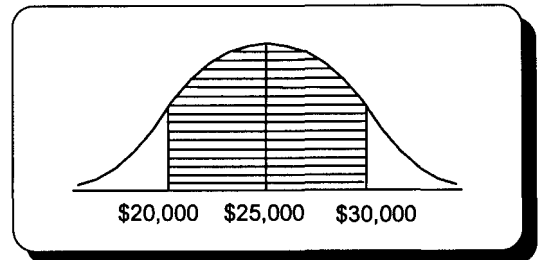


B. $P(\$20,000 \leq x < \$30,000)$

$$Z = \frac{x - \mu}{\sigma} = \frac{\$20,000 - \$25,000}{\$5,000} = \frac{-\$5,000}{\$5,000} = -1 \rightarrow .3413$$

$$Z = \frac{x - \mu}{\sigma} = \frac{\$30,000 - \$25,000}{\$5,000} = \frac{\$5,000}{\$5,000} = 1 \rightarrow .3413$$

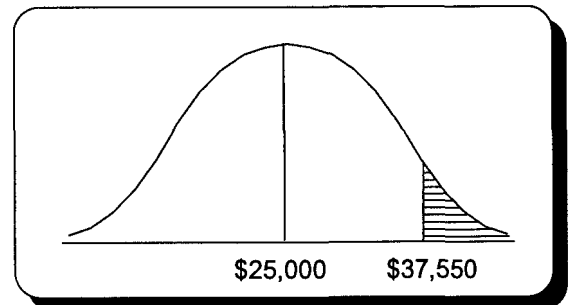
$$.3413 + .3413 = .6826 \rightarrow 68.26\%$$



C. $P(x \geq \$37,550)$

$$Z = \frac{x - \mu}{\sigma} = \frac{\$37,550 - \$25,000}{\$5,000} = \frac{\$12,550}{\$5,000} = 2.51 \rightarrow .4940$$

$$\begin{array}{r} .5000 \\ - .4940 \\ \hline .0060 \rightarrow .6\% \end{array}$$

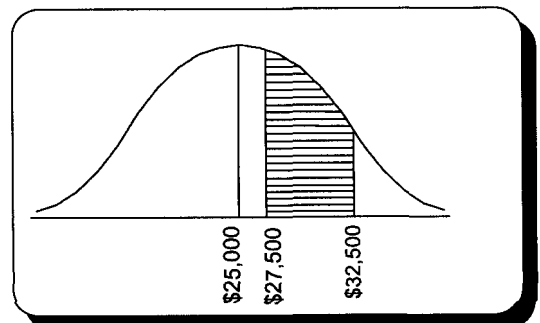


D. $P(\$27,500 \leq x < \$32,500)$

$$Z = \frac{x - \mu}{\sigma} = \frac{\$32,500 - \$25,000}{\$5,000} = \frac{\$7,500}{\$5,000} = 1.5 \rightarrow .4332$$

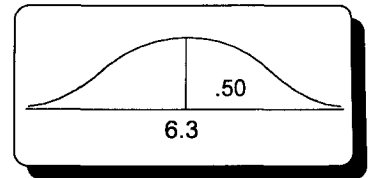
$$Z = \frac{x - \mu}{\sigma} = \frac{\$27,500 - \$25,000}{\$5,000} = \frac{\$2,500}{\$5,000} = .5 \rightarrow .1915$$

$$.4332 - .1915 = .2417 \rightarrow 24.17\%$$



II. The number of customers returning merchandise to Darin's Music Emporium is normally distributed with a mean of 6.3 per week and a standard deviation of 1.5. Given the following probabilities, calculate the appropriate value or values for x .

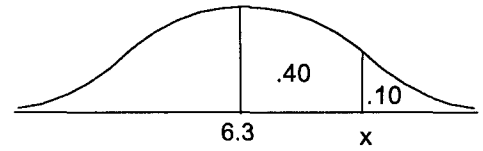
- A. Half of the time, returns will be above 6.3 per week.
 B. Ninety percent of the time returns will be below 8.2.



$$50\% - 10\% = 40\% \rightarrow Z = 1.28$$

Note: Use only the plus sign because the top 10% is to the right of 6.3.

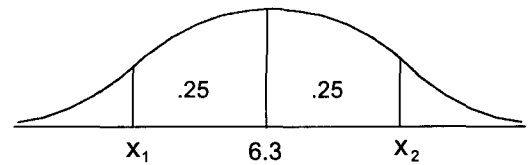
$$\begin{aligned} \mu \pm Z\sigma \\ 6.3 + 1.28(1.5) \\ 6.3 + 1.92 \\ = 8.2 \end{aligned}$$



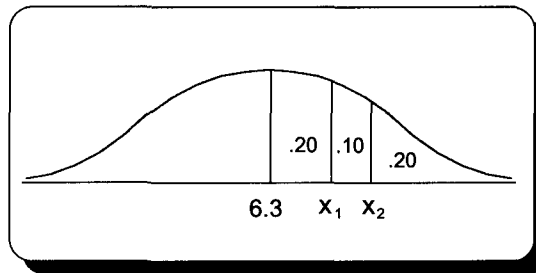
C. Find the interquartile range for returns to Darin's Music Emporium.

$$25\% \rightarrow Z = .67$$

$$\begin{aligned} \mu \pm Z\sigma \\ 6.3 \pm .67(1.5) \\ 6.3 \pm 1.005 \\ 5.3 \leftrightarrow 7.3 \end{aligned}$$



D. Draw a graph of the eighth decile for returns to Darin's Music Emporium.



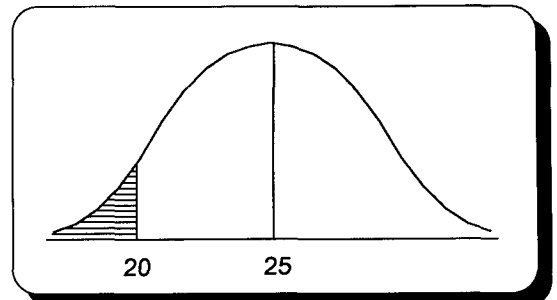
III. A recent study indicated 5% of Darin's customers return merchandise sold for credit. What is the probability of Darin having less than 20 returns for a 500 credit sales week?

The normal approximation of the binomial may be used when $n \geq 30$ and both np and nq are ≥ 5 .
 $n = 500$, $np = 500 \times .05 = 25$, and $nq = 500 \times .95 = 475$. The normal approximation may be used.

Given:

$$\begin{aligned} p &= .05 \\ n &= 500 \\ x &= 20 \end{aligned}$$

$$\begin{aligned} \mu &= np = (500)(.05) = 25 \\ \sigma &= \sqrt{npq} \\ &= \sqrt{500(.05)(.95)} = \sqrt{23.75} = 4.8734 \end{aligned}$$



$$\begin{aligned} Z &= \frac{x - \mu}{\sigma} = \frac{19.5 - 25.0}{4.8734} = -1.13 \rightarrow .3708 \\ 50.00\% - 37.08\% &= 12.92\% \end{aligned}$$

Note: Because the question's range does not include 20, 20's lower limit of 19.5 is used for x .

Practice Set 11 Sampling and the Sampling Distribution of the Means

- I. Darin's new company, Future Horizons Corporation, manufactures a component for computer chips. Darin wants to know the average weight of 1,000 recently produced components. A sample of 36 had a mean weight of 30.025 milligrams and a standard deviation of .065 milligrams. Calculate the 98% confidence interval for the population mean weight of these components.

Given: $n = 36$ | 98% CI | $\bar{X} = 30.025$ | $S = .065$

98% CI $\rightarrow z = 2.33$

$$\bar{x} \pm Z \frac{s}{\sqrt{n}}$$

$$30.025 \pm 2.33 \frac{.065}{\sqrt{36}}$$

$$30.025 \pm .0252$$

$$29.999 \leftrightarrow 30.050$$

I did not round the lower limit up because I wanted to show the population mean could be under 30 milligrams.

- II. Calculate the 95% confidence interval using problem I information.

95% CI $\rightarrow z = 1.96$

$$\bar{x} \pm Z \frac{s}{\sqrt{n}}$$

$$30.025 \pm 1.96 \frac{.065}{\sqrt{36}}$$

$$30.025 \pm .0212$$

$$30.004 \leftrightarrow 30.046$$

- III. What can Darin do to make this interval smaller?

Increase the sample size. This will lower $\frac{s}{\sqrt{n}}$, which is the point estimate of the standard error of the mean.

Practice Set 12 Sampling Distributions Part II

- I. Darin wants to know the proportion of page 68 parts passing inspection. Fifty parts were randomly selected from a recent production run of 1,000 parts and 45 passed inspection.

- A. Calculate the proportion of parts passing inspection.

$$\bar{p} = \frac{x}{n} = \frac{45}{50} = .90 \rightarrow 90\%$$

- B. Darin would like to use last week's data to predict a range for the proportion of future production runs passing inspection. Calculate the 95% confidence interval for the proportion of parts produced by this production process passing inspection.

$\frac{n}{N} = \frac{50}{1,000} = .05 \geq .05$
Finite correction factor applies.

Note: \bar{p} has been used as an estimate of p .

$n = 50 \geq 30$
 $np = 50 \times .9 = 45 \geq 5$
 $nq = 50 \times .1 = 5 \geq 5$
Normal approximation of the binomial applies.

$$\begin{aligned} \sigma_{\bar{p}} &= \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \sqrt{\frac{N-n}{N-1}} \\ &= \sqrt{\frac{.9(1-.9)}{50}} \sqrt{\frac{1,000-50}{1,000-1}} \\ &= .042(.975) = .041 \end{aligned}$$

$$\begin{aligned} \bar{p} \pm z\sigma_{\bar{p}} \\ .90 \pm 1.96(.041) \\ .90 \pm .080 \\ .82 \leftrightarrow .98 \end{aligned}$$

- C. What assumption is Darin making when using last week's data to predict future manufacturing quality?

Darin is assuming the factors affecting the weight of parts are stable. If tests soon to be explored in this part of Quick Notes indicate the proportion passing inspection is dropping, Darin will investigate these factors.

D. Calculate the 99% confidence interval for the proportion of parts passing inspection.

$$\begin{aligned} \bar{p} \pm z\sigma_{\bar{p}} \\ .90 \pm 2.58(.041) \\ .90 \pm .106 \\ .794 \leftrightarrow 1.006 \end{aligned}$$

Proportions over 100% are not possible. Darin needs to lower the point estimate of the sampling distribution's standard deviation with a larger sample.

E. What sample size is necessary to reduce acceptable error to $\pm 5\%$?

$$\begin{aligned} n = \bar{p}(1 - \bar{p})\left(\frac{z}{E}\right)^2 &= .90(1 - .90)\left(\frac{2.58}{.05}\right)^2 \\ &= .90(.10)(2662.56) \\ &= 239.630 \rightarrow 240 \end{aligned}$$

II. Darin is also concerned about the weight of page 68 parts. It must be possible for the mean weight of parts to be ≤ 30 mg with a 99% degree of confidence. As indicated on page 68 and reviewed below, a recent test was barely successful. Darin wants to reduce error from the current $\pm .0279$ mg to $\pm .025$ mg. What sample size is required?

Page 68 Problem Review

Given: $n = 36$, $z = 2.58$, $s = .065$ mg and $\bar{x} = 30.025$ mg

$$\begin{aligned} \bar{x} \pm zS_{\bar{x}} \\ 30.025 \pm .0279 \\ 29.997 \text{ mg} \leftrightarrow 30.053 \text{ mg} \end{aligned}$$

Note: This range indicates the population mean estimated from this sample could be under 30 mg.

The finite correction factor is not required because n/N is less than .05.

$$\begin{aligned} n &= \left(\frac{z\sigma}{E}\right)^2 \\ &= \left[\frac{(2.58)(.065)}{.025}\right]^2 \\ &= [6.708]^2 \\ &= 44.997 \rightarrow 45 \end{aligned}$$

III. Check your answer to problem II by calculating the 99% confidence interval using a sample size of 45 and a sample standard deviation of .065. Analyze the result.

$$\bar{x} \pm Z \frac{\sigma}{\sqrt{n}}$$

$$\begin{aligned} \bar{x} \pm 2.58 \frac{.065}{\sqrt{45}} \\ 30.025 \pm 2.58(.00969) \\ 30.025 \pm .025 \\ 30.000 \leftrightarrow 30.050 \end{aligned}$$

Note: Error equals $2.58(.00969) = .025$

IV. How would the solution to problem III change if the sample of 45 had been taken from a population of 500 items?

$$\frac{n}{N} = \frac{45}{500} = .09 > .05 \quad \text{The finite correction factor should be used.}$$

V. Recalculate the answer to problem III using the finite correction factor.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\begin{aligned} \bar{x} \pm 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \sqrt{\frac{N-n}{N-1}} \\ 30.025 \pm 2.58(.00969) \sqrt{\frac{500-45}{500-1}} \\ 30.025 \pm .0239 \\ 30.0011 \leftrightarrow 30.0489 \end{aligned}$$

Note: As expected, the answer became more exact.

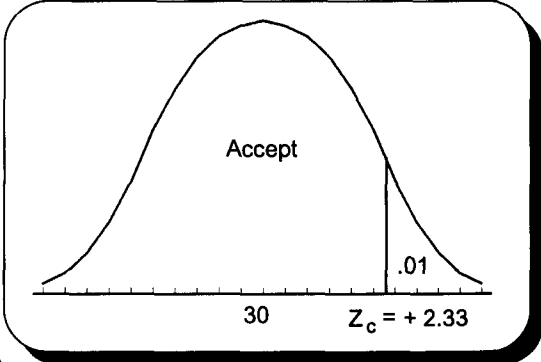
Practice Set 13 Large Sample Hypothesis Testing

- I. Darin Jones is very concerned that parts designed to weigh less than or equal to 30 mg may be too heavy and not pass inspection. From page 68, we know that a sample of 36 parts resulted in a sample mean of 30.025 mg and a sample standard deviation of .065 mg. Darin wants to control type I error (the probability of deciding the parts that are too heavy when they are not) to the .01 level of significance. Solve this problem using the 5-step approach to hypothesis testing.

1. $H_0 : \mu \leq 30$ mg and $H_1 : \mu > 30$ mg
2. $\alpha = .01$
3. \bar{x} is the test statistic.
4. The critical value for .01 is $z = 2.33$.
If the test Z is beyond 2.33, reject H_0 .
5. Apply the decision rule.

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{30.025 - 30.000}{\frac{.065}{\sqrt{36}}} = \frac{.025}{.0108} = 2.315$$

Accept H_0 because $2.315 < 2.33$. Parts are not too heavy.



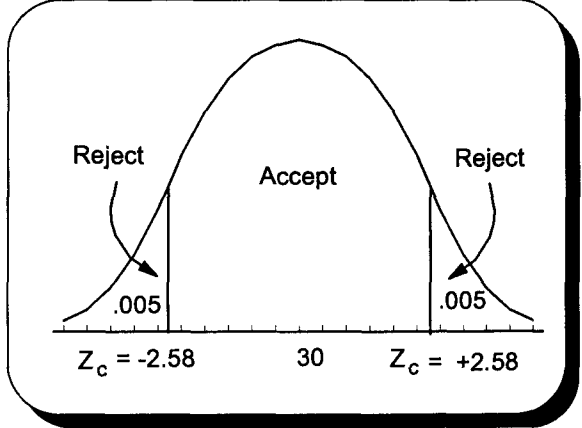
Note: For those using the p-value approach discussed in chapter 14, the p-values for problems 1, 2, and 3 using z are .0103, .0206, and .0206 respectively.

- II. Using the problem I data and a .01 level of significance, determine whether the population mean has changed from 30 milligrams.

Given:	$n = 36$	$\bar{x} = 30.025$	$s = .065$	$\alpha = .01$
---------------	----------	--------------------	------------	----------------

1. $H_0 : \mu = 30$ mg $H_1 : \mu \neq 30$ mg
2. $\alpha = .01$
3. \bar{x} is the test statistic.
4. The critical value of z for $\alpha/2 = .01/2 = .005$ is ± 2.58 .
If the test Z is beyond ± 2.58 , reject H_0 .
5. Apply the decision rule.

$2.315 < 2.58$, accept H_0



- III. Redo problem II using a .05 level of significance.

1. For steps 1 and 3, see problem two.
4. The critical value of z for $\alpha/2 = .05/2 = .025$ is ± 1.96 .
5. If z is beyond ± 1.96 , reject H_0 .

Reject H_0 because $2.315 > 1.96$. Parts are too heavy.

Practice Set 14 Large Sample Hypothesis Testing Part II

- I. Darin buys material for his 30-milligram parts from suppliers A and B. A sample of 30 orders placed with supplier A had a mean delivery time of 24 days and a standard deviation of 9 days. A sample of 40 orders placed with supplier B had a mean delivery time of 27 days and a standard deviation of 10 days. Using a .05 level of significance, determine whether these suppliers have different mean delivery times.

Supplier A	Supplier B
$n_1 = 30$	$n_2 = 40$
$\bar{x}_1 = 24$ days	$\bar{x}_2 = 27$ days
$s_1 = 9$ days	$s_2 = 10$ days

1. $H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$
2. $\alpha = .05$ and $.05/2 = .025$
3. \bar{x} is the test statistic.
4. The critical value of z for .025 is ± 1.96 .
If the test Z is beyond -1.96, reject H_0 .
5. Apply the decision rule.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{24 - 27}{\sqrt{\frac{(9)^2}{30} + \frac{(10)^2}{40}}} = \frac{-3}{2.280} = -1.32$$

Accept H_0 because -1.32 is not beyond -1.96. Delivery times are the same.

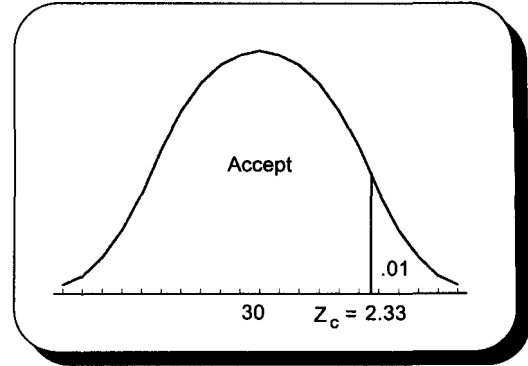
- II. Darin has decided to determine the p-value associated with the test of the 30-milligram parts conducted in problem 1 on page 86. This data was first analyzed on page 68.

Problem Review

Given: $\bar{x} = 30.025$ mg, $n = 36$, $s = .065$ mg, and $\alpha = .01$

$H_0 : \mu \leq 30.00$ mg $H_1 : \mu > 30.00$ mg

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{(30.025 - 30.000)}{\frac{.065}{\sqrt{36}}} = 2.315 < 2.33, \text{ accept } H_0$$



- A. Calculate the p-value associated with this study.

$$z = 2.315 \rightarrow .4897 \text{ and } .5000 - .4897 = .0103 = 1.03\%$$

Accept H_0 because $.0103 > .01$.

- B. Use this p-value to accept or reject the null hypothesis. Does your answer agree with the page 86 answer?

Yes

- C. What does this p-value indicate is the strength or validity of the decision made concerning the null hypothesis?

The low p-value indicates the hypothesis is barely accepted.

- III. Past experience indicates that the population mean weight of material containers used to make computer parts is 5,000 kilograms. The standard deviation is 28 kilograms. Type I error for a sample of 49 will be controlled to the .01 level of significance. The 99% confidence interval is 4,989.68 kilograms to 5,010.32 kilograms.

- A. Calculate the type II error for a two-tail problem using each of these possible population means.

$\mu_1 = 4,985$ kg

$\mu_2 = 4,995$ kg

$\mu_3 = 5,000$ kg

$\mu_4 = 5,005$ kg

$\mu_5 = 5,015$ kg

$$Z = \frac{x_c - \mu_1}{\frac{\sigma}{\sqrt{n}}} = \frac{4,989.68 - 4,985.00}{\frac{28}{\sqrt{49}}} = 1.17 \rightarrow .3790$$

$$.50 - .379 \rightarrow 12.1\%$$

$$Z = \frac{x_c - \mu_2}{\frac{\sigma}{\sqrt{n}}} = \frac{4,989.68 - 4,995.00}{\frac{28}{\sqrt{49}}} = 1.33 \rightarrow .4082$$

$$.50 - .4082 = .0918$$

$$.50 + .0918 = 59.18\%$$

There isn't any type II error as the null hypothesis is true. At a point just before 5,000 mg, type II error is 98+%.

$$Z = \frac{x_c - \mu_4}{\frac{\sigma}{\sqrt{n}}} = \frac{5,010.32 - 5,005.00}{\frac{28}{\sqrt{49}}} = 1.33 \rightarrow .4082$$

$$.50 - .4082 = .0918$$

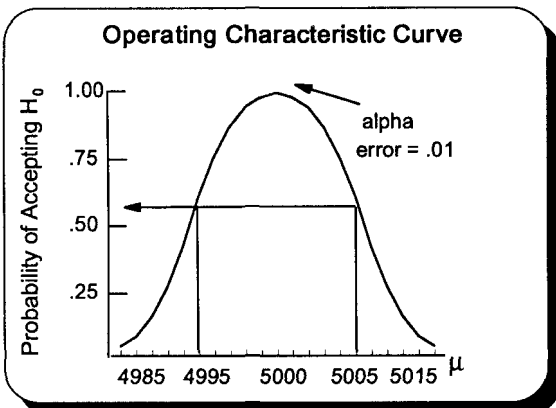
$$.50 + .0918 = 59.18\%$$

$$Z = \frac{x_c - \mu_5}{\frac{\sigma}{\sqrt{n}}} = \frac{5,010.32 - 5,015.00}{\frac{28}{\sqrt{49}}} = 1.17 \rightarrow .3790$$

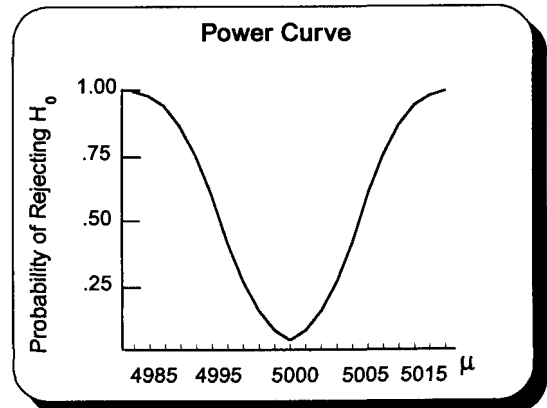
$$.50 - .379 \rightarrow 12.1\%$$

- B. Using the data calculated in problem A, sketch and label an operating characteristic curve.

- C. Using the data calculated in problem A, sketch and label a power curve.



Note: The x-axis on these two graphs is not drawn to scale.



Practice Set 15 Hypothesis Testing of Population Proportions

- I. Page 72 data showed that 90% (45 of 50) of the 30-milligram parts, taken from a lot of 1,000 parts, passed inspection. Darin wants a .01 level of significance test to determine whether the population proportion of parts passing inspection has increased from the 86% reported last year.

Given: n equals 50, $p = .86$, and $\bar{p} = .90$

The normal approximation to the binomial applies.

$$n = 50 > 30$$

$$np = 50(.86) = 43 \geq 5$$

$$nq = 50(1 - .86) = 7 \geq 5$$

1. The null hypothesis and alternate hypothesis are $H_0: p \leq .86$ and $H_1: p > .86$.
2. The level of significance will be .01.
3. The test statistic is \bar{p} .
4. If z from the test statistic is beyond the critical value of z , the null hypothesis will be rejected.
5. Apply the decision rule.

$$Z = \frac{\bar{p} - p}{\sigma_p} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.90 - .86}{\sqrt{\frac{.86(1-.86)}{50}}} = \frac{.04}{.0491} = .81$$

Accept H_0 because .81 is not beyond 2.33. The proportion of parts passing inspection is not higher than last year.

- II. Darin wants to determine at the .01 level of significance whether there is a difference in the proportion of defects produced during the day and night shifts. A sample of 100 parts was taken from each shift. The day shift had 5 defects and the night shift had 14 defects. Is there a difference in the proportion of defects produced by these two shifts?

Given: $n_1 = 100$ $n_2 = 100$ $x_1 = 5$ $x_2 = 14$ $\alpha = .01$ and $.01/2 = .005 \rightarrow z = \pm 2.58$

$$p_1 = \frac{x_1}{n_1} = \frac{5}{100} = .05$$

$$p_2 = \frac{x_2}{n_2} = \frac{14}{100} = .14$$

$$\bar{p}_w = \frac{x_1 + x_2}{n_1 + n_2} = \frac{5 + 14}{100 + 100} = .095$$

The 5-step approach to hypothesis testing

1. The null hypothesis and alternate hypothesis are: $H_0: p_1 = p_2$ and $H_1: p_1 \neq p_2$
2. The level of significance will be .01.
3. The test statistic will be \bar{p} .
4. If z from the test statistic is beyond the critical value of z , the null hypothesis will be rejected.
5. Apply the decision rule.

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{\bar{p}_w(1-\bar{p}_w)}{n_1} + \frac{\bar{p}_w(1-\bar{p}_w)}{n_2}}} = \frac{.05 - .14}{\sqrt{\frac{.095(1-.095)}{100} + \frac{.095(1-.095)}{100}}} = \frac{-.09}{.0415} = -2.17$$

Accept H_0 because -2.17 is not beyond -2.58.
The defects proportion is the same for these two shifts.

Practice Set 16 Small Sample Hypothesis Testing Using Student's t Test

- I. Darin wants to determine whether there is a difference in the number of sick days taken by employees based upon their education. A sample of 11 high school graduates had a mean of 5 sick days per year and a standard deviation of 2.5 days. Twelve non-graduates averaged 10 sick days per year. Their standard deviation was 3.25 days. Is there a difference in sick days taken based upon education? Use the .01 level of significance.

Given
$n_1 = 11$
$\bar{X}_1 = 5$
$S_1^2 = 2.5^2 = 6.25$
$n_2 = 12$
$\bar{X}_2 = 10$
$S_2^2 = 3.25^2 = 10.56$
$\alpha = .01$

$$S_w^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(11-1)6.25 + (12-1)10.56}{11 + 12 - 2}$$

$$= \frac{62.50 + 116.16}{21}$$

$$= 8.51$$

The 5-step approach to hypothesis testing

1. $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$
2. $\alpha = .01$
3. The test statistic is \bar{x} .
4. $df = n_1 + n_2 - 2 = 11 + 12 - 2 = 21$
 $\alpha = .01$ and $.01/2 = .005 \rightarrow t = \pm 2.831$
5. Apply the decision rule.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_w^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{5 - 10}{\sqrt{8.51 \left(\frac{1}{11} + \frac{1}{12} \right)}} = -4.11$$

Reject H_0 because -4.11 is beyond -2.831. Non-high school graduates took a different number of sick days than high school graduates.

- II. Darin conducted a training program for 5 recently-hired employees. Test at the .01 level whether the training program increased employee efficiency.

Employee	Efficiency Rating		d	d ²
	Before	After		
1	8	9	-1	1
2	6	8	-2	4
3	7	8	-1	1
4	7	9	-2	4
5	8	10	-2	4
			-8	14

$$\bar{d} = \frac{\sum d}{n} = \frac{-8}{5} = -1.6$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{14 - \frac{(-8)^2}{5}}{5-1}}$$

$$= \sqrt{\frac{14 - 12.8}{4}}$$

$$= .5477$$

The 5-step approach to hypothesis testing

1. $H_0: \mu_d \geq 0$ and $H_1: \mu_d < 0$
2. $\alpha = .01$
3. The test statistic is \bar{d} .
4. $df = n - 1 = 5 - 1 = 4$ and α of .01 $\rightarrow t = -3.747$
5. Apply the decision rule.

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{-1.6}{\frac{.5477}{\sqrt{5}}} = \frac{-1.6}{.245} = -6.53$$

Reject H_0 because -6.53 is beyond -3.747.

Training increased efficiency.

Note: H_0 points to the left and t is negative. Why? When scores increase, their difference is negative.

Practice Set 17 Statistical Quality Control

- I. Darin is doing a quality control study of the 30 milligram parts first analyzed in chapter 11. This data has been reproduced below. Assume the data consisted of 12 three part samples. Also assume the process was in control when these samples were taken. Construct an \bar{X} chart and an R chart for this data using a 99.74% (3 sigma) confidence interval.

Sample #	1	2	3	4	5	6	7	8	9	10	11	12	Totals
n = 3 N = 12	29.89	30.05	29.98	30.07	29.97	30.05	29.95	30.06	29.99	30.02	30.09	30.12	
	29.96	29.97	30.06	30.05	29.95	29.95	29.99	29.89	29.99	30.08	30.06	30.16	
	29.97	29.98	30.04	30.06	30.05	30.09	30.06	30.09	29.98	30.01	30.08	30.15	
Sample Mean	29.94	30.00	30.03	30.06	29.99	30.03	30.00	30.01	29.99	30.04	30.07	30.14	360.30
Sample Range	0.08	0.08	0.08	0.02	0.10	0.14	0.11	0.20	0.01	0.07	0.03	0.04	0.96

$$\bar{\bar{x}} = \frac{\sum \bar{x}}{N} = \frac{360.3}{12} = 30.0250$$

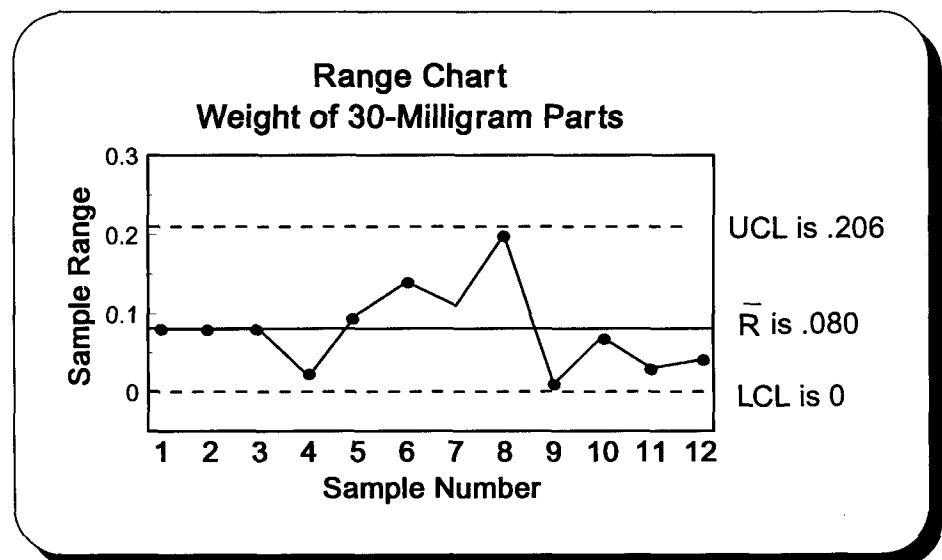
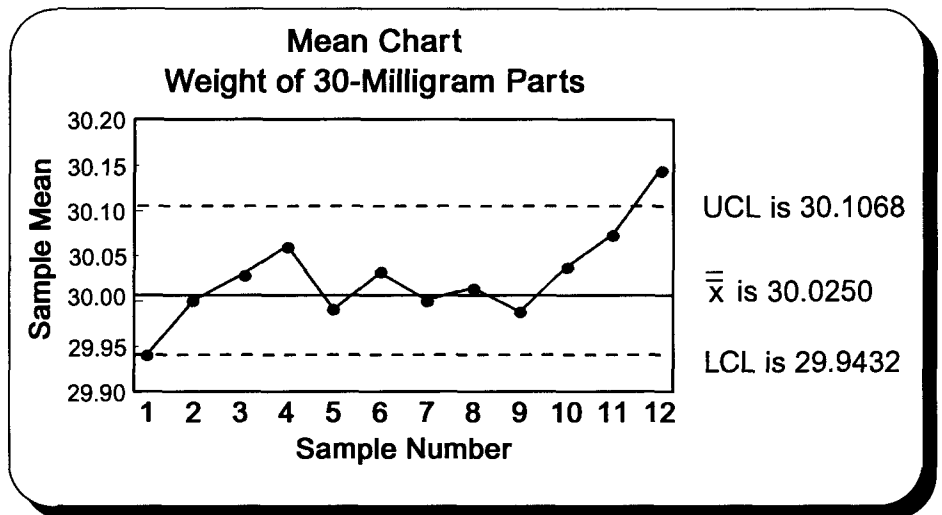
$$\bar{R} = \frac{\sum R}{N} = \frac{.96}{12} = .080$$

$$\begin{aligned} \text{UCL} &= \bar{\bar{x}} + A_2 \bar{R} \\ &= 30.025 + 1.023(.08) \\ &= 30.025 + .08184 \\ &= 30.10684 \end{aligned}$$

$$\begin{aligned} \text{LCL} &= \bar{\bar{x}} - A_2 \bar{R} \\ &= 30.025 - 1.023(.08) \\ &= 30.025 - .08184 \\ &= 29.94316 \end{aligned}$$

$$\begin{aligned} \text{UCL} &= D_4 \bar{R} \\ &= 2.575(.08) \\ &= 0.206 \end{aligned}$$

$$\begin{aligned} \text{LCL} &= D_3 \bar{R} \\ &= 0(.08) \\ &= 0 \end{aligned}$$



- II. Darin wants to continue his study of the proportion of 30-milligram parts found to be defective in chapter 12. This study found 5 of 50 parts were defective. This data and an additional 9 samples are summarized below. Construct a p chart for this data. Do not use the finite correction factor.

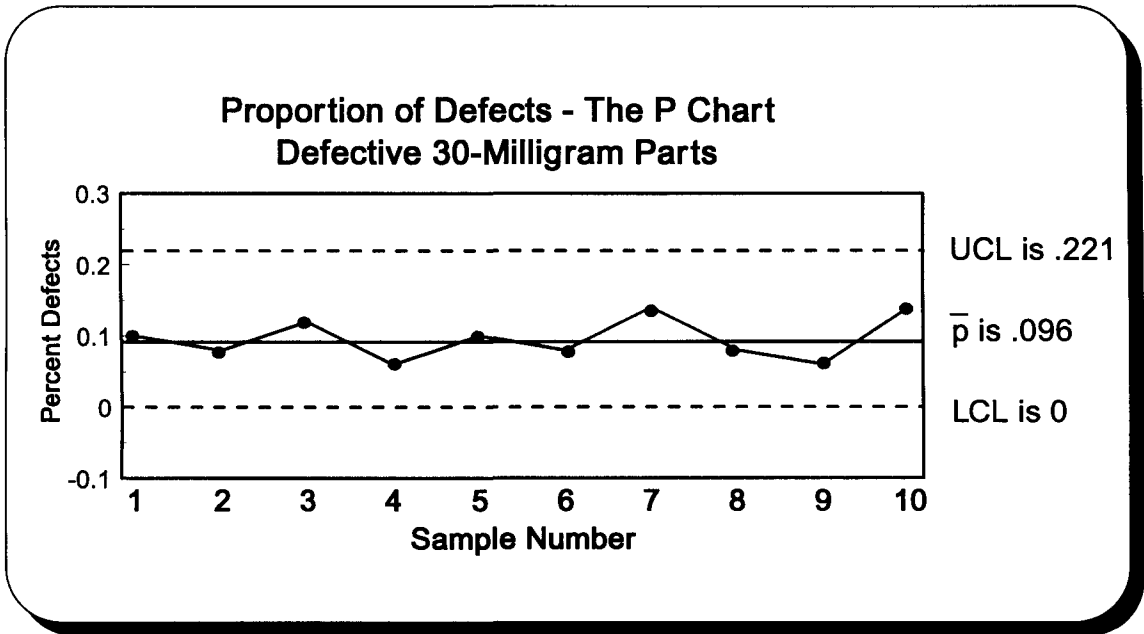
Defective 30-Milligram Parts										
Date	1/3	1/4	1/5	1/6	1/7	1/10	1/11	1/12	1/13	1/14
Sample #	1	2	3	4	5	6	7	8	9	10
Defects	5	4	6	3	5	4	7	4	3	7
Defects Proportion	.10	.08	.12	.06	.10	.08	.14	.08	.06	.14

$$\bar{p} = \frac{\text{total defects}}{\text{total sampled}}$$

$$\text{UCL and LCL} = \bar{p} \pm 3 \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$\begin{aligned} \bar{p} &= \frac{\text{total defects}}{\text{total sampled}} \\ &= \frac{48}{500} \\ &= .096 \end{aligned}$$

$$\begin{aligned} \text{UCL and LCL} &= .096 \pm 3 \sqrt{\frac{.096(1-.096)}{50}} \\ &.096 \pm 3(.04166) \\ &.096 \pm .125 \\ &-.029 \leftrightarrow +.221 \\ &-.029 \text{ is rounded to zero} \end{aligned}$$



Practice Set 18 Analysis of Variance

I. Darin wants to know whether the variance of 30-mg parts has increased. The standard deviation from a recent sample of 16 parts was .067 milligrams. The standard deviation from an earlier study of 14 parts was .062 milligrams. Test at the .01 level whether the population variance has increased.

1. These are the null hypothesis and alternate hypothesis.

$$H_0 : \sigma_1^2 \leq \sigma_2^2 \text{ and } H_1 : \sigma_1^2 > \sigma_2^2$$

2. The level of significance will be .01.

3. The test statistic is F.

4. H_0 will be rejected when F from the test statistic is greater than F's critical value.

a. df for the numerator is $16 - 1 = 15$

b. df for the denominator is $14 - 1 = 13$

c. From Table 5A, $F = 3.82$.

5. Apply the decision rule.

$$F = \frac{s_1^2}{s_2^2} = \frac{.067^2}{.062^2} = \frac{.004489}{.003844} = 1.17 \quad \text{Accept } H_0 \text{ because } 1.17 < 3.82. \\ \text{Variance has not increased.}$$

II. Time passed and the wonders of miniaturization have reduced the 30-mg parts to a weight of only 9 mg. Darin randomly selected samples of 9-mg parts from 3 departments with the following results. **People using statistics software should skip to part D.**

A. Complete this chart to begin an ANOVA study of the mean weight of parts produced by these 3 departments.

Weight Analysis of 9-mg Parts Produced by 3 Departments							Row Totals Required for Calculations
	Parts Sample 1 is T_1		Parts Sample 2 is T_2		Parts Sample 3 is T_3		
	X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	
	8.95	80.1025	9.05	81.9025	9.05	81.9025	
	8.90	79.2100	9.05	81.9025	9.15	83.7225	
	<u>8.90</u>	<u>79.2100</u>	<u>9.10</u>	<u>82.8100</u>	<u>9.10</u>	<u>82.8100</u>	
$\sum X_T$	26.75		27.20		27.30		$\sum x = 81.25$
$(\sum X_T)^2$	715.5625		739.84		745.29		
n	3		3		3		$N = 9$
$\frac{(\sum X_T)^2}{n}$	238.521		246.613		248.43		$\sum \left[\frac{(\sum X_T)^2}{n} \right] = 733.564$
$\sum X_T^2$		238.5225		246.6150		248.4350	$\sum x^2 = 733.5725$

B. Using data from the previous page, calculate the following values.

$$SS_T = \sum \left[\frac{(\sum x_T)^2}{n} \right] - \frac{(\sum X)^2}{N}$$

$$= 733.564 - \frac{81.25^2}{9}$$

$$= 733.564 - 733.507$$

$$= .0570$$

$$SS_E = \sum X^2 - \sum \left[\frac{(\sum x_T)^2}{n} \right]$$

$$= 733.5725 - 733.5640$$

$$= .0085$$

$$SS_{TOTAL} = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$= 733.5725 - 733.5070$$

$$= .0655$$

Note: Most of the variability (.0570 out of .0655) has been explained by the treatment variable.

C. Complete the following chart using the data accumulated to this point.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	$t - 1 = 3 - 1 = 2$	$SS_T = .057$	$MS_T = \frac{SS_T}{t-1} = \frac{.057}{3-1} = .0285$	$F = \frac{MS_T}{MS_E} = \frac{.0285}{.0014} = 20.36$
Within Treatments (error)	$N - t = 9 - 3 = 6$	$SS_E = .0085$	$MS_E = \frac{SS_E}{N-t} = \frac{.0085}{9-3} = .0014$	
Total Variance	$N - 1 = 9 - 1 = 8$	$SS_{TOTAL} = .0655$		

D. Using the 5-step approach to hypothesis testing and the above chart, test at the .05 level whether the sample means are from populations with equal means.

1. These are the null hypothesis and alternate hypothesis.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ and } H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

- The level of significance for this one-tail problem is .05.
- The test statistic is F.
- If F from the test statistic is beyond the critical value for the .05 level of significance, the null hypothesis will be rejected.

df for the numerator is 2.
df for the denominator is 6.
F's critical value is 5.14 (see Table 5B).

5. Apply the decision rule.

Reject H_0 because $20.36 > 5.14$. These populations have different means.

Practice Set 19 Two-Factor Analysis of Variance

I. Practice Set 18 will be expanded by assuming the data was randomly collected at hourly intervals. Page 110 data has been arranged accordingly. Darin wants to determine whether samples taken later in a shift are less likely to pass inspection. **People using statistics software should skip to part D.**

A. Complete this chart to begin an ANOVA study of the production process producing these parts.

Weight Analysis of 9-mg Parts Produced by 3 Departments							Row Totals Required for Calculations		
Time	Parts Sample 1 is T ₁		Parts Sample 2 is T ₂		Parts Sample 3 is T ₃		ΣX_B	$(\Sigma X_B)^2$	$\frac{(\Sigma X_B)^2}{t}$
	X_1	X_1^2	X_2	X_2^2	X_3	X_3^2			
9:15 AM	8.90	79.2100	9.05	81.9025	9.05	81.9025	27.00	729.0000	243.0000
10:20 AM	8.90	79.2100	9.05	81.9025	9.10	82.8100	27.05	731.7025	243.9008
11:10 AM	<u>8.95</u>	<u>80.1025</u>	<u>9.10</u>	<u>82.8100</u>	<u>9.15</u>	<u>83.7225</u>	<u>27.20</u>	739.8400	<u>246.6133</u>
							81.25 = Σx	$\Sigma[\frac{(\Sigma X_B)^2}{t}] = 733.5141$	
ΣX_T	26.75		27.20		27.30		81.25 = Σx		
$(\Sigma X_T)^2$	715.5625		739.84		745.29				
b	3		3		3		N = 9		
$\frac{(\Sigma X_T)^2}{b}$	238.521		246.613		248.430		$\Sigma[\frac{(\Sigma X_T)^2}{b}] = 733.564$		
ΣX_T^2		238.5225		246.6150		248.4350	$\Sigma X^2 = 733.5725$		

B. Using the above data, calculate the following values.

$$\begin{aligned}
 SS_T &= \Sigma\left[\frac{(\Sigma x_T)^2}{b}\right] - \frac{(\Sigma X)^2}{N} \\
 &= 733.564 - \frac{81.25^2}{9} \\
 &= 733.564 - 733.507 \\
 &= .057
 \end{aligned}$$

$$\begin{aligned}
 SS_B &= \Sigma\left[\frac{(\Sigma x_B)^2}{t}\right] - \frac{(\Sigma X)^2}{N} \\
 &= 733.5141 - \frac{81.25^2}{9} \\
 &= 733.5141 - 733.5070 \\
 &= .0071
 \end{aligned}$$

$$\begin{aligned}
 SS_{TOTAL} &= \Sigma X^2 - \frac{(\Sigma X)^2}{N} \\
 &= 733.5725 - 733.5070 \\
 &= .0655
 \end{aligned}$$

$$\begin{aligned}
 SS_E &= SS_{TOTAL} - (SS_T + SS_B) \\
 &= .0655 - (.057 + .0071) \\
 &= .0655 - .0641 \\
 &= .0014
 \end{aligned}$$

Unexplained variability is down from .0085 (see page PS 111) to .0014.

C. Complete the following chart using data accumulated to this point.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	$t - 1 = 3 - 1 = 2$	$SS_T = .057$	$MS_T = \frac{SS_T}{t-1} = \frac{.057}{2} = .0285$	$F = \frac{MS_T}{MS_E} = \frac{.0285}{.00035} = 81$
Block	$b - 1 = 3 - 1 = 2$	$SS_B = .0071$	$MS_B = \frac{SS_B}{b-1} = \frac{.0071}{2} = .0036$	
Within Treatments (error)	$(t - 1)(b - 1) = 2 \times 2 = 4$	$SS_E = .0014$	$MS_E = \frac{SS_E}{(t-1)(b-1)} = \frac{.0014}{4} = .00035$	$F = \frac{MS_B}{MS_E} = \frac{.0036}{.00035} = 10$
Total Variance	$N - 1 = 9 - 1 = 8$	$SS_{TOTAL} = .0655$		

D. Using the 5-step approach to hypothesis testing, determine at the .01 level of significance whether the sample treatment and block means come from populations with equal means.

- A check of each null hypothesis will be made.
 - $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$ for the treatment means
 - $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$ for the block means
- The level of significance is .01.
- The test statistic is F.
- The decision rule will be, if F from the test statistic is beyond the critical value of F for the .01 level of significance, the null hypothesis will be rejected.
- Apply the decision rule.

Degrees of freedom for the treatment hypothesis is 2 for the numerator and 4 for the denominator.

$$F = \frac{MS_T}{MS_E} = 81 \quad \text{Reject } H_0 \text{ because } 81 > 18.$$

F is 18.

Parts from these 3 departments do not have equal means.

Degrees of freedom for the block hypothesis is 2 for the numerator and 4 for the denominator.

$$F = \frac{MS_B}{MS_E} = 10 \quad \text{Accept } H_0 \text{ because } 10 < 18.$$

F is 18.

Parts produced at different times have equal means.

Note: With so much of the total variability explained by the treatment, there was little left to be explained by the block.

II. Using information from page 111, determine at the .01 level of significance whether there is a difference between treatments 1 and 3.

$$\bar{X}_1 = \frac{\sum x}{n_1} = \frac{26.75}{3} = 8.92$$

$$\bar{X}_3 = \frac{\sum x}{n_3} = \frac{27.30}{3} = 9.10$$

The t for .005 and df of 6 is 3.707.
 MS_E from page PS 111 is .0014.

$$(\bar{X}_3 - \bar{X}_1) \pm t \sqrt{MS_E \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$(9.10 - 8.92) \pm 3.707 \sqrt{.0014 \left(\frac{1}{3} + \frac{1}{3} \right)}$$

$$.18 \pm .113$$

The range of .067 \leftrightarrow .293 indicates the means are different.

Practice Set 20 Nonparametric Hypothesis Testing of Nominal Data

- I. Darin feels 20% of the 9-mg part defects are produced by the first shift, 30% by the second shift, and 50% by the third shift. Do an .01 level of significance test to determine whether this sample data follows Darin's proposed distribution. **People using statistics software do not have to fill out the second chart.**

Analysis of Defects				
	Shift 1	Shift 2	Shift 3	Totals
Shift defects, f_o	6	11	23	40
Expected defects, f_e	8	12	20	40

Shift	f _o	f _e	f _o - f _e	(f _o - f _e) ²	$\frac{(f_o - f_e)}{f_e}$
Shift 1	6	8	-2	4	4/8 = .50
Shift 2	11	12	-1	1	1/12 = .08
Shift 3	23	20	<u>3</u>	9	9/20 = <u>.45</u>
Totals			0		$\chi^2 = 1.03$

1. H_0 : defects follow Darin's distribution.
 H_1 : defects do not follow Darin's distribution.
2. The significance level is .01.
3. Chi-square is the test statistic.
4. The decision rule:
 If χ^2 from the test statistic is beyond the critical value,
 the difference is high and the null hypothesis is rejected.
5. Apply the decision rule.

$$df = k - 1 = 3 - 1 = 2 \rightarrow \chi^2 = 9.21$$

Accept H_0 because $1.03 < 9.21$.
 Defects follow Darin's distribution.

- II. This is Darin's page 42 study of customer age and making a sale. Test at the .05 level of significance whether customer age and making a sale are independent.

Customer Age and Making A Sale			
Customer Age	Less than or equal to 20	Over 20	Totals
Making A Sale			
No	16	8	24
Yes	24	12	36
Totals	40	20	60

Contingency Table of Customer Age and Making A Sale						
Customer Age	Less than or equal to 20		Over 20		Totals	
Making A Sale						
	f_o	f_e	f_o	f_e	f_o	f_e
No	16	16	8	8	24	24
Yes	24	24	12	12	36	36
Totals	40	40	20	20	60	60

Working left to right:

$$f_e = \frac{f_r \times f_c}{n} = \frac{24 \times 40}{60} = 16$$

$$f_e = \frac{f_r \times f_c}{n} = \frac{24 \times 20}{60} = 8$$

Alternate formula

$$f_e = \frac{f_r}{n} \times f_c = \frac{24}{60} \times 40 = 16$$

$$f_e = \frac{f_r}{n} \times f_c = \frac{24}{60} \times 20 = 8$$

- H_0 : customer age and making a sale are independent (not related).
 H_1 : customer age and making a sale are dependent (related).
- The significance level will be .05.
- Chi-square is the test statistic.
- The decision rule:
If χ^2 from the test statistic is beyond the critical value, reject the null hypothesis.
- Apply the decision rule.

$$df = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1 \rightarrow \chi^2 = 3.84$$

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = \sum \left[\frac{(16 - 16)^2}{16} + \frac{(8 - 8)^2}{8} + \frac{(24 - 24)^2}{24} + \frac{(12 - 12)^2}{12} \right] = 0 + 0 + 0 + 0 = 0$$

Accept H_0 because $0 < 3.84$. Customer age and making a sale are statistically independent at the .05 level of significance.

Note: These variables are independent because both types of buyers do not make a purchase 40% of the time and do make a purchase 60% of the time.

Practice Set 21 Nonparametric Hypothesis Testing of Ordinal Data Part I

- I. Darin wants to determine whether the page 68 computer components were drawn at random. The median of 30.045 mg is the standard for this test. Determine at the .05 level of significance whether this data was randomly collected. Data was recorded one column at a time starting at the top of each column. Columns were recorded from left to right.

29.89	30.05	29.98	30.07	29.97	30.05	29.95	30.06	29.99	30.02	30.09	30.12
29.96	29.97	30.06	30.05	29.95	29.95	29.99	29.89	29.99	30.08	30.06	30.16
29.97	29.98	30.04	30.06	30.05	30.09	30.06	30.09	29.98	30.01	30.08	30.15

$n_1 = 18$ below
 $n_2 = 18$ above
 $r = 18$ runs

-	+	-	+	-	+	-	+	-	-	+	+	
-	-	+	+	-	-	-	-	-	+	+	+	
-	-	-	+	+	+	+	+	+	-	-	+	+

$$\begin{aligned} \sigma_r &= \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}} \\ &= \sqrt{\frac{2(18)(18)[(2(18)(18) - 18 - 18)]}{(18 + 18)^2(18 + 18 - 1)}} \\ &= \sqrt{\frac{648[648 - 36]}{(36)^2(35)}} \\ &= \sqrt{\frac{396,576}{45,360}} = \sqrt{8.743} = 2.957 \end{aligned}$$

$$\begin{aligned} \mu_r &= \frac{2n_1n_2}{n_1 + n_2} + 1 \\ &= \frac{2(18)(18)}{18 + 18} + 1 \\ &= \frac{648}{36} + 1 \\ &= 19 \end{aligned}$$

$$\begin{aligned} Z &= \frac{r - \mu_r}{\sigma_r} \\ &= \frac{18 - 19}{2.957} \\ &= -.34 \end{aligned}$$

At .05 level of significance, z is ± 1.96 .
Accept H_0 because $-.34$ is not beyond -1.96 .
Parts were drawn at random.

- II. Darin first studied the number of defective 30-milligram parts on page 96. At that time he did a parametric study because he felt the data was normally distributed. The consistency of raw material inputs has changed and Darin isn't sure the distribution is still normal. Do a .05 level of significance sign test to determine whether defects have increased from last year's median of 5.

Sample	Median	Sign
1	6	+
2	7	+
3	5	0
4	4	-
5	8	+
6	6	+
7	7	+

- A. Five of the median defects were above 5. Let n equal 6 because of a tie.
 B. $\mu = .50$
 C. The Binomial table (ST 1) yields the following: $P(n \geq 5) = .094 + .016 = .11$
 D. Accepted H_0 because .11 is greater than .05.
 Median defects have not increased.

Note: Our small sample size continues to cause problems. All six samples must be above the median for the null hypothesis to be rejected at the .05 level of significance. A larger sample may be needed.

- iii. Darin wants to reexamine the number of sick days taken by employees based upon education. This data was first presented on page 100. At that time it was assumed the populations were approximately normal with the same variance. As a result, population means were compared. Assume these assumptions might not be true and use a Mann-Whitney .01 level of significance test to determine whether these samples come from populations with equal medians.

Graduates' sick days: 5, 4, 7, 2, 7, 7, 0, 3, 6, 8, 6 **Non-graduates' sick days:** 9, 13, 8, 6, 14, 6, 12, 16, 8, 10, 7, 11

People Using Statistics Software should not use this chart.

Complete this table by: (1) completing an ordered array, (2) assigning a G for graduates and an N for non-graduates to each element of the array, (3) assigning each rank to the appropriate category (non-graduate or graduate), (4) calculating each subtotal, and (5) calculating R_1 , which equals the sum of the 3 subtotals for non-graduates or R_2 which equals the sum of the 3 subtotals for graduates.

Rank			Ranked Scores		Rank			Ranked Scores		Rank			Ranked Scores	
Ordered Array and Degree Status			Grads	Non-grads	Ordered Array and Degree Status			Grads	Non-grads	Ordered Array and Degree Status			Grads	Non-grads
(1)	(2)		(3)	(3)	(1)	(2)		(3)	(3)	(1)	(2)		(3)	(3)
1.	0	G	1		9.	6	G	7.5		17.	9	N		17
2.	2	G	2		10.	7	N		11.5	18.	10	N		18
3.	3	G	3		11.	7	G	11.5		19.	11	N		19
4.	4	G	4		12.	7	G	11.5		20.	12	N		20
5.	5	G	5		13.	7	G	11.5		21.	13	N		21
6.	6	N		7.5	14.	8	N		15	22.	14	N		22
7.	6	N		7.5	15.	8	N		15	23.	16	N		23
8.	6	G	7.5		16.	8	G	15						
(4) Subtotal			22.5	15.0	(4) Subtotal			57.0	41.5	(4) Subtotal			0	140

$$(5) R_1 = 22.5 + 57 + 0 = 79.5$$

$$(5) R_2 = 15.0 + 41.5 + 140.0 = 196.5$$

$$\begin{aligned}
 U_1 &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \\
 &= 11(12) + \frac{11(11+1)}{2} - 79.5 \\
 &= 132 + 66 - 79.5 \\
 &= 118.5
 \end{aligned}$$

$$\begin{aligned}
 \mu_U &= \frac{n_1 n_2}{2} \\
 &= \frac{11(12)}{2} \\
 &= 66
 \end{aligned}$$

$$\begin{aligned}
 \sigma_U &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \\
 &= \sqrt{\frac{11(12)(11+12+1)}{12}} \\
 &= \sqrt{\frac{3,168}{12}} \\
 &= 16.248
 \end{aligned}$$

$$\begin{aligned}
 Z &= \frac{U - \mu_U}{\sigma_U} \\
 &= \frac{118.5 - 66.00}{16.248} \\
 &= 3.23
 \end{aligned}$$

The critical value for z is 2.575. H_0 is rejected because z of 3.23 is beyond 2.575.

Median sick days differ.

Practice Set 22 Nonparametric Hypothesis Testing of Ordinal Data Part II

- I. Darin conducted a training program for 5 recently-hired employees. This problem first appeared on page 100. At that time it was assumed that the population was approximately normal. If this assumption is not correct or unknown, a .01 level of significance paired difference sign test may be conducted to determine whether training increased worker efficiency.

Employee	Efficiency Rating		Sign
	Before	After	
1	8	9	+
2	6	8	+
3	7	8	+
4	7	9	+
5	8	10	+

- A. All 5 employee ratings increased. n is 5.
- B. The Binomial table (ST 1) yields the following: $p(x \geq 5) = .031$
- C. Accept H_0 because $.031 > .01$. Efficiency did not increase.
- D. **Note:** With a sample of only five and alpha of .01, the null hypothesis will not be rejected when $\mu = .50$.

- II. Darin wants to reexamine the ANOVA study conducted on page 110. That study assumed populations were normally distributed with equal variances. Those assumptions are not appropriate. Conduct a .01 level of significance Kruskal-Wallis test to determine whether the median weight of parts produced by these 3 departments are equal. Page 110 data has been increased to conform with the $n \geq 5$ test requirement.

Department 1		Department 2		Department 3	
Weight	Rank (R_1)	Weight	Rank (R_2)	Weight	Rank (R_3)
8.95	5	9.05	7	9.05	7
8.90	2.5	9.05	7	9.15	15
8.90	2.5	9.10	10.5	9.10	10.5
8.92	4	9.07	9	9.13	13
8.88	1	9.11	12	9.14	14
	$R_1 = 15.0$		$R_2 = 45.5$		$R_3 = 59.5$

H is the designated statistic.
N , the number of observations, is 15.
k , the number of samples, is 3.
n_k , a sample size, is 5.
R_k is a sample rank total.
$df = k - 1 = 3 - 1 = 2 \rightarrow \chi^2 = 9.21$

$$\begin{aligned}
 H &= \frac{12}{N(N+1)} \left[\frac{(\sum R_1)^2}{n_1} + \frac{(\sum R_2)^2}{n_2} + \dots + \frac{(\sum R_k)^2}{n_k} \right] - 3(N+1) \\
 &= \frac{12}{15(15+1)} \left[\frac{(15)^2}{5} + \frac{(45.5)^2}{5} + \frac{(59.5)^2}{5} \right] - 3(15+1) \\
 &= .05[45.00 + 414.05 + 708.05] - 48.00 = 10.355
 \end{aligned}$$

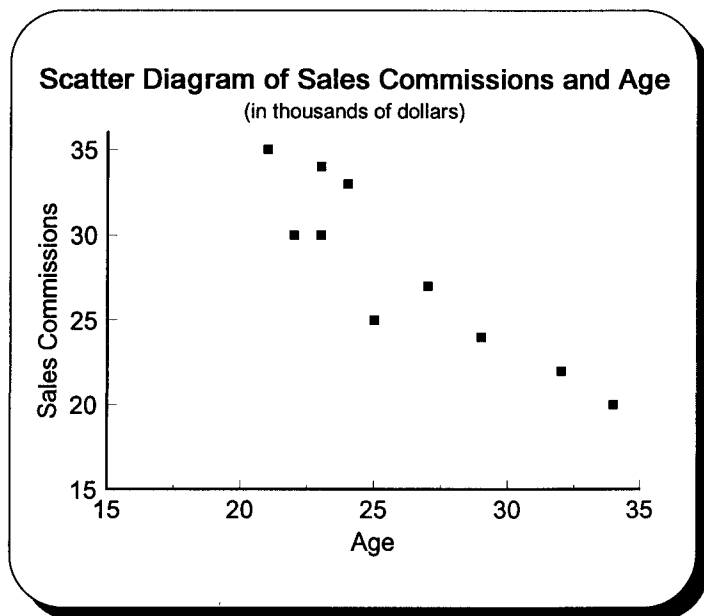
Reject H_0 because H of 10.355 is greater than 9.21. Medians are not equal.

Note: By leaving this page blank, the solutions to Practice Sets 23 and 24 can be on adjoining pages. This should make studying these two related Practice Set solutions easier.

Practice Set 23 Correlation Analysis

- I. Darin Jones wants to know whether age of sales personnel affects sales performance. Answer the following questions using the given data.
A. Draw a scatter diagram.

Age	Sales Commissions (000)	xy	x ²	y ²
23	30	690	529	900
25	25	625	625	625
34	20	680	1,156	400
29	24	696	841	576
21	35	735	441	1,225
32	22	704	1,024	484
23	34	782	529	1,156
24	33	792	576	1,089
27	27	729	729	729
<u>22</u>	<u>30</u>	<u>660</u>	<u>484</u>	<u>900</u>
260	280	7,093	6,934	8,084



- B. Calculate the coefficient of correlation to 3 decimal places. Interpret your answer.

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}} = \frac{10(7,093) - (260)(280)}{\sqrt{[10(6,934) - (260)^2][10(8,084) - (280)^2]}}$$

$$= \frac{(70,930) - (72,800)}{\sqrt{[(69,340) - (67,600)][(80,840) - (78,400)]}} = \frac{-1,870}{\sqrt{[1,740][2,440]}} = \frac{-1,870}{2,060} = -.908$$

There is a high negative correlation between age of sales people and sales performance.

- C. What is the coefficient of determination? Interpret your answer.

$$r^2 = (.908)^2 = .824 \text{ or } 82.4\%$$

Eighty-two and four tenths percent of the variability in sales performance is accounted for by age variability of salespeople.

- D. What is the coefficient of nondetermination? Interpret your answer.

$$\tilde{r}^2 = 1 - r^2 = 1 - .824 = .176 \text{ or } 17.6\%$$

Seventeen and six tenths percent of the variability in sales performance is not accounted for by age variability of salespeople.

- E. Is the relationship between age of sales personnel and their sales commissions significant at the .01 level?

The null hypothesis and alternate hypothesis are $H_0: \rho = 0$ and $H_1: \rho \neq 0$.

- The level of significance will be .01 for this two-tail problem with $n - 2$ degrees of freedom.
- The relevant statistic will be t .
- If t from the test statistic is beyond the critical value of t , the null hypothesis will be rejected.
- Apply the decision rule.

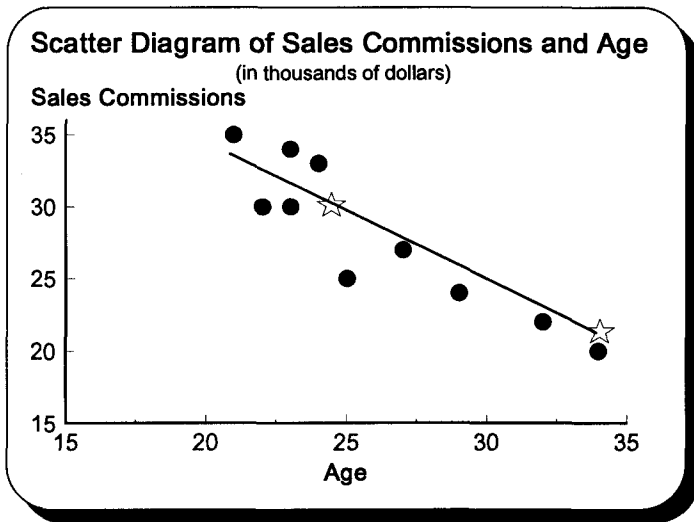
$$df = n - 2 = 10 - 2 = 8 \rightarrow t = 3.355$$

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.908 - 0}{\sqrt{\frac{1 - (.908)^2}{10 - 2}}} = 6.13$$

Reject H_0 because $6.13 > 3.355$. The population coefficient of correlation could not be zero at the .01 level of significance. **Note:** Because t may be positive or negative, the absolute value of r is used to calculate t .

Practice Set 24 Simple Linear Regression Analysis

- I. Having determined that age affects sales performance, Darin Jones wants to estimate sales commissions using the data presented in the chapter 23 practice set.



Age	Sales Commissions (000)	xy	x ²	y ²
23	30	690	529	900
25	25	625	625	625
34	20	680	1,156	400
29	24	696	841	576
21	35	735	441	1,225
32	22	704	1,024	484
23	34	782	529	1,156
24	33	792	576	1,089
27	27	729	729	729
<u>22</u>	<u>30</u>	<u>660</u>	<u>484</u>	<u>900</u>
260	280	7,093	6,934	8,084

- A. Determine the regression equation to 3 significant digits.

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} = \frac{-1,870}{1,740} = -1.0747126$$

Note: Data for b was taken from the previous page.

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y}{n} - b\frac{\sum X}{n} = \frac{280}{10} - (-1.0747126)\left(\frac{260}{10}\right) = 55.942527$$

$$\hat{y}_{\cdot x} = a + bx$$

$$\hat{y}_{\cdot x} = 55.9 - 1.07x$$

- B. Estimate sales commissions for a group of 24-year-old salespeople.

$$\hat{y}_{\cdot 24} = 55.9 - 1.07x = 55.9 - 1.07(24) = 30.2$$

- C. Graph the regression line.

$$\hat{y}_{\cdot 34} = 55.9 - 1.07x = 55.9 - 1.07(34) = 19.5$$

Two points (x,y) will be used to draw a straight line. We will use the coordinates from questions B and C.

- D. Determine the 99% confidence interval for the question B group.

$$S_{y,24} = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n-2}} = \sqrt{\frac{8,084 - 55.942527(280) - (-1.0747126)(7,093)}{10-2}} = 2.319$$

$$df = 10 - 2 = 8$$

$$\alpha/2 = .01/2 = .005 \rightarrow t = 3.355$$

$$\bar{x} = \frac{\sum x}{n} = \frac{260}{10} = 26$$

$$\hat{y}_{\cdot x} \pm ts_{y,x} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

$$\hat{y}_{\cdot 24} = 30.2 \pm 3.355(2.319) \sqrt{\frac{1}{10} + \frac{(24-26)^2}{6,934 - \frac{(260)^2}{10}}}$$

$$= 30.2 \pm 2.73$$

$$27.47 \leftrightarrow 32.93$$

- E. What procedure should be followed if question D's range includes negative numbers?

The standard error of the estimate may be lowered with a larger sample. Here, the standard error is low and the range for y does not include zero.

Appendix II

Complete Solutions to Quick Questions

Quick Questions 1

Statistics Is About Using Data in Decision Making

Place the number of the appropriate description next to the item it describes.

- | | |
|--------------------------------------|--|
| A. Statistic <u> 4 </u> | 1. Subset of a population |
| B. Parameter <u> 9 </u> | 2. Expressed numerically |
| C. Population <u> 10 </u> | 3. The use of sample statistics to estimate population parameters |
| D. Discrete <u> 5 </u> | 4. Characteristic of a sample |
| E. Quantitative variable <u> 2 </u> | 5. Only finite values can exist on the x-axis |
| F. Secondary source data <u> 8 </u> | 6. Published by the original collector |
| G. Sample <u> 1 </u> | 7. Measurement may assume any value associated with an uninterrupted scale |
| H. Inferential statistics <u> 3 </u> | 8. Published by a noncollector |
| I. Continuous <u> 7 </u> | 9. Characteristic of a population |
| J. Primary source data <u> 6 </u> | 10. Totality under study |

Note: Quick Question answers may differ slightly from computer generated answers.

Quick Questions 2 Summarizing Data

I. Place the number of the appropriate formula or phrase next to the item it describes.

- A. Mutually-exclusive events 2
- B. Relative frequency 4
- C. Class midpoint 3
- D. Approximate class width 1
- E. All-inclusive events (collectively exhaustive) 6
- F. Ogive 5

II. Complete the following using this data.

Data: 38, 48, 27, 14, 31, 23, 46, 38, 54, 26, 44, 33, 17, 34, 6, 37

A. Array: 6, 14, 17, 23, 26, 27, 31, 33, 34, 37, 38, 38, 44, 46, 48, 54

B. Range $R = H - L = 54 - 6 = 48$

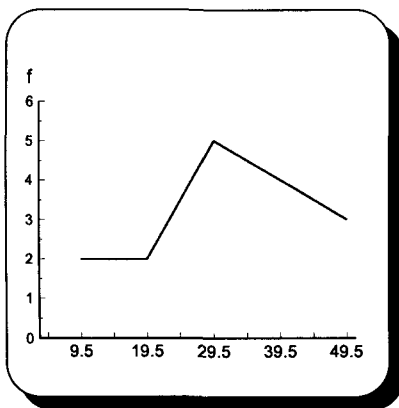
C. Approximate class width

$$\frac{\text{range}}{\# \text{ of classes}} = \frac{48}{5} = 9.6 \rightarrow 10$$

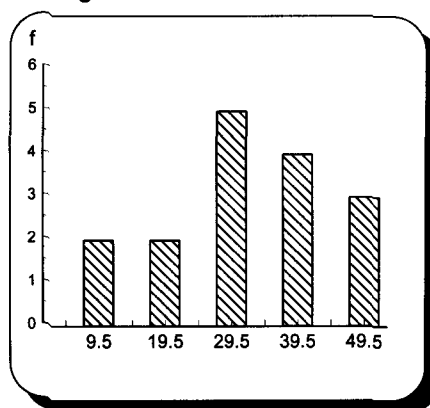
D.

Frequency Distribution							
Stated Class Limits	Real Class Limits	Tally	Frequency (f)	Relative Frequency $f \div n$	Cumulative Frequency		
					More-than	Less-than	
5 - 14	4.5 - 14.5		2	0.1250	16	0	
15 - 24	14.5 - 24.5		2	0.1250	14	2	
25 - 34	24.5 - 34.5		5	0.3125	12	4	
35 - 44	34.5 - 44.5		4	0.2500	7	9	
45 - 54	44.5 - 54.5		3	0.1875	3	13	
Total frequency (n)			16	1.0000	0	16	

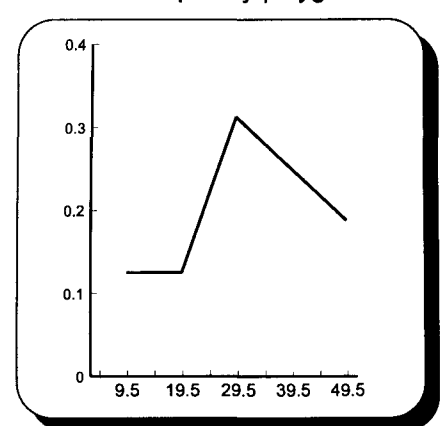
E. Frequency polygon



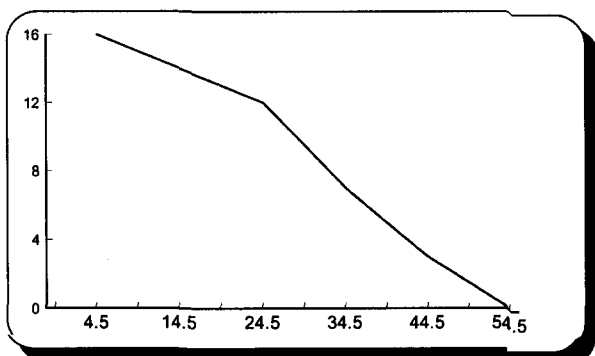
F. Histogram



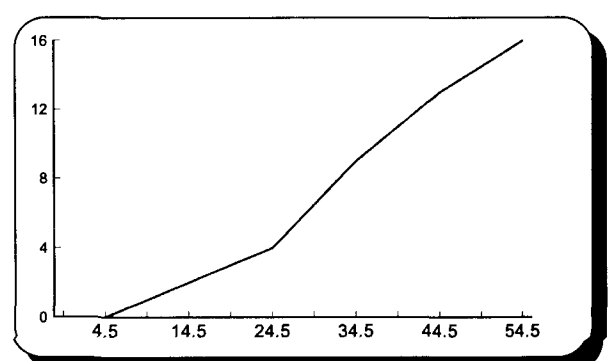
G. Relative frequency polygon



H. More-than cumulative frequency polygon



I. Less-than cumulative frequency polygon



Quick Questions 3 Measuring Central Tendency of Ungrouped Data

I. Write the number of the appropriate formula next to the item it describes.

- A. Sample mean 3
- B. Population mean 2
- C. Location of the median 4
- D. Location of Q_1 6
- E. Weighted mean 5
- F. Location of Q_3 1

II. List and calculate the 3 measures of central tendency.

Data: 5, 7, 3, 8, 6, 10, 9, 8

A. Mean

$$\bar{x} = \frac{\sum x}{n} = \frac{56}{8} = 7$$

B. Median

Array: 3, 5, 6, 7, 8, 8, 9, 10

$$\frac{n}{2} + .5 = \frac{8}{2} + .5 = 4.5 \rightarrow 7.5$$

C. Mode The number that appears most often is 8.

III. What is the primary disadvantage of the mean as a measure of central tendency?

The primary disadvantage of the mean as a measure of central tendency concerns it being severely affected by a few values at either extreme.

IV. Using this data, prove that the sum of the deviations around an arithmetic mean is zero.

Data: 3, 7, 5 The mean of this data is 5.

$$\begin{aligned} \sum(x - \mu) &= (3 - 5) + (7 - 5) + (5 - 5) \\ &= (-2) + (2) + (0) = 0 \end{aligned}$$

V. Calculate a weighted mean of parking tickets costing \$25, \$35, and \$45 with corresponding weights of 10, 20, and 10 respectively. Why must the answer be \$35?

$$\bar{X}_w = \frac{W_1X_1 + W_2X_2 + W_3X_3 + \dots + W_nX_n}{W_1 + W_2 + W_3 + \dots + W_n} = \frac{\sum(W_x X_x)}{\sum w_x}$$

$$\bar{X}_w = \frac{(10)(\$25) + (20)(\$35) + (10)(\$45)}{10 + 20 + 10} = \frac{\$250 + \$700 + \$450}{40} = \frac{\$1,400}{40} = \$35.00$$

Low- and high-priced tickets are each different from middle-priced tickets by \$10.00 and both have equal weights. In effect, they cancel each other.

VI. Calculate the following for the question II data.

A. Q_1

$$\frac{n}{4} + .5 = \frac{8}{4} + .5 = 2 + .5 = 2.5 \rightarrow 5.5$$

B. Q_3

$$\frac{3n}{4} + .5 = \frac{24}{4} + .5 = 6.5 \rightarrow 8.5$$

C. Interquartile range

$$Q_3 - Q_1 = 8.5 - 5.5 = 3.0$$

D. 2nd decile

$$\frac{xn}{10} + .5 = \frac{2(8)}{10} + .5 = 1.6 + .5 = 2.1 \rightarrow 5.1$$

E. 85th percentile

$$\frac{xn}{100} + .5 = \frac{85(8)}{100} + .5 = 6.8 + .5 = 7.3 \rightarrow 9.3$$

Quick Questions 4 Measuring Dispersion of Ungrouped Data

I. Place the number of the appropriate formula next to the parameter or statistic it describes.

- A. Population average deviation 1
- B. Population variance 2
- C. Population standard deviation 3
- D. Alternative population variance 4
- E. Alternative population standard deviation 5
- F. Chebyshev's rule 6
- G. Sample variance 7
- H. Sample standard deviation 8
- I. Alternative sample variance 9
- J. Alternative sample standard deviation 10

Note how the answers are in sequence. This was done to allow students to compare population formulas on the left with the corresponding sample formula on the right.

II. Calculate the following statistics using this sample data.

Data: 5, 7, 3, 8, 6, 10, 9, 8

$\bar{x} = 7$

A. Variance (use alternative formula)

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{428 - \frac{(56)^2}{8}}{8-1} = \frac{428 - 392}{8-1} = 5.1$$

B. Standard deviation

$$s = \sqrt{s^2} = \sqrt{5.1} = 2.3$$

C. Average deviation

$$A.D. = \frac{\sum |x - \bar{x}|}{n} = \frac{14}{8} = 1.8$$

III. Use Chebyshev's rule to calculate the percentage of question II outcomes that will be within 3 standard deviations of the mean. Was this prediction correct?

$$1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} \rightarrow 88.9\%$$

$$7 \pm 3(2.3)$$

$$7 \pm 6.9$$

$$.1 \leftrightarrow 13.9$$

A. Chebyshev predicts a minimum of 88.9% will be between .1 and 13.9.

B. Array of daily Walkman sales: 3, 5, 6, 7, 8, 8, 9, 10

C. All are between .1 and 13.9.

IV. A data set of grades is normally distributed and has a mean of 84 and a standard deviation of 4. Calculate a range of grades that will include the middle 95.44% of the data set.

The empirical rule states that 95.44% of normally distributed data will be within 2 standard deviations.

$$84 \pm 2(4)$$

$$84 \pm 8$$

$$76 \leftrightarrow 92$$

Quick Questions 5 Measuring Central Tendency of Grouped Data

I. Place the number of the appropriate formula next to the item it describes.

- A. Grouped sample mean 2
- B. Location of the grouped median 4
- C. Grouped median 3
- D. Class midpoint 1

II. The x values for this chart are 12, 17, and 22 respectively.

- A. The first class has real class limits of 9.5 and 14.5.
- B. The first class has stated class limits of 10 and 14.
- C. The class width is 5.
- D. The midpoint of the first class is 12.
- E. The range using real class limits is from 9.5 to 24.5.

III. Calculate the following statistics using this frequency distribution of exam grades.

A. Mean

$$\bar{X} = \frac{\sum fx}{n} = \frac{1,401}{18} = 77.8$$

B. Median

$$L + \frac{\frac{n}{2} - CF_b}{f}(i)$$

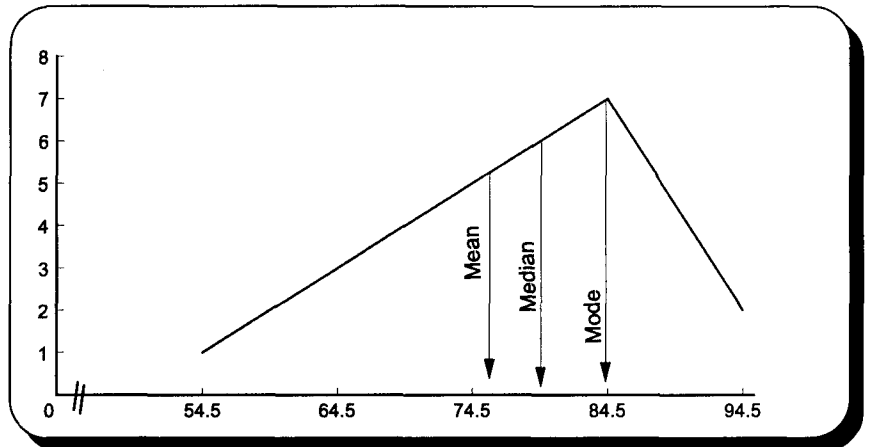
$$= 69.5 + \frac{\frac{18}{2} - 4}{5}(10)$$

$$= 69.5 + 10 = 79.5$$

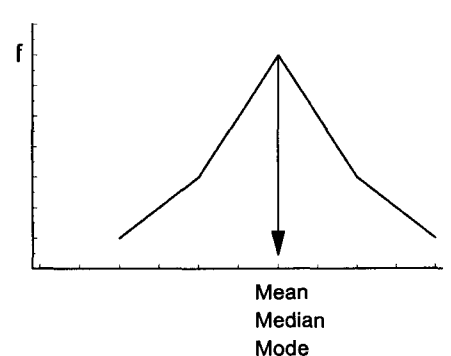
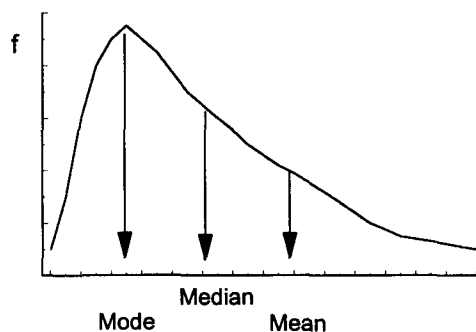
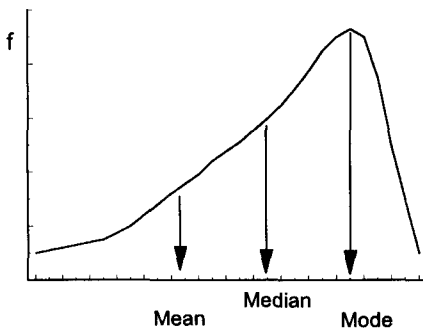
C. Mode

The midpoint of the class with the highest frequency is 84.5.

IV. Draw a frequency polygon for the question III data and locate the mean, median, and mode.



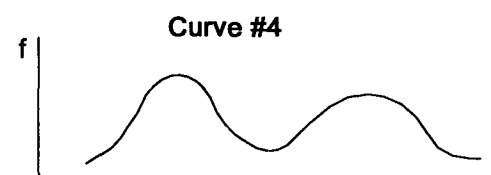
V. Show the approximate location of the mean, median, and mode on the x-axis of these frequency distributions.



VI. Answer these questions using Curves #1 to #4.

- A. Curve #1 is skewed to the left.
- B. Curve #3 is not skewed and is said to be symmetrical or normal.
- C. Curve #4 is bimodal.
- D. Mean
- E.

$$\frac{3(\bar{x} - Md.)}{s} = \frac{3(77.8 - 79.5)}{14.2} = \frac{3(-1.7)}{14.2} = \frac{-5.1}{14.2} = -.4$$



Quick Questions 6 Measuring Dispersion of Grouped Data

I. Place the number of the appropriate formula next to the item it describes.

- A. Grouped sample standard deviation 2
- B. First quartile 3
- C. Median (second quartile) 1
- D. Third quartile 4
- E. Interquartile range 6
- F. Percentile 5

II. Complete the first row of this table and calculate the following measurements.

Stated Class Limits	Frequency (f)	x	fx	x ²	fx ²
Totals	16.00	417.00	1,162.00	30,731.50	87,434.00

A. Range $H - L = 99.5 - 39.5 = 60$

B. Sample variance

$$S^2 = \frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1} = \frac{87,434 - \frac{(1,162)^2}{16}}{16-1} = \frac{87,434 - 84,390}{15} = 202.9$$

C. Sample standard deviation

$$s = \sqrt{s^2} = \sqrt{202.9} = 14.2$$

D. First Quartile $\frac{n}{4} = \frac{16}{4} = 4$

Median $\frac{n}{2} = \frac{16}{2} = 8$

Third Quartile $\frac{3n}{4} = \frac{(3)(16)}{4} = 12$

$$Q_1 = L + \frac{\frac{n}{4} - CF_b}{f}(i)$$

$$= 59.5 + \frac{\frac{16}{4} - 3}{3}(10)$$

$$= 59.5 + \frac{1}{3}(10)$$

$Q_1 = 62.8$

$$Q_2 = L + \frac{\frac{n}{2} - CF_b}{f}(i)$$

$$= 69.5 + \frac{\frac{16}{2} - 6}{5}(10)$$

$$= 69.5 + \frac{2}{5}(10)$$

$Q_2 = 73.5$

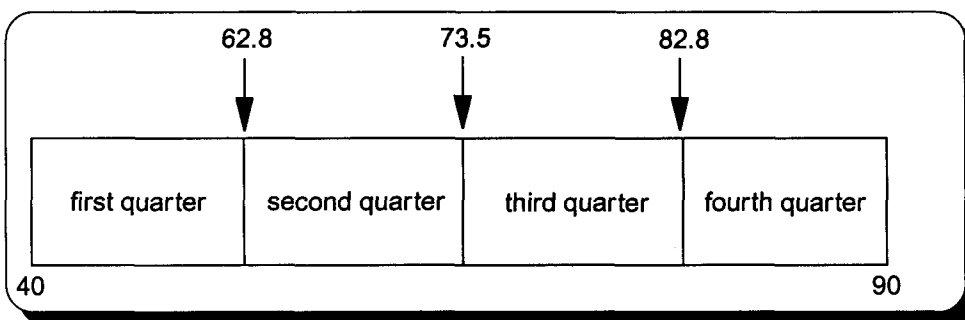
$$Q_3 = L + \frac{\frac{3n}{4} - CF_b}{f}(i)$$

$$= 79.5 + \frac{\frac{48}{4} - 11}{3}(10)$$

$$= 79.5 + \frac{1}{3}(10)$$

$Q_3 = 82.8$

E. Locate the three quartiles and the four quarters on this figure.



F. Interquartile range

$$Q_3 - Q_1 = 82.8 - 62.8 = 20.0$$

G. 95th percentile

$$\frac{xn}{100} = \frac{95(16)}{100} = 15.2$$

$$P_x = L + \frac{\frac{xn}{100} - CF_b}{f}(i) = P_{95} = 89.5 + \frac{\frac{95(16)}{100} - 14}{2}(10)$$

$$= 89.5 + \frac{1.2}{2}(10) = 95.5$$

Quick Questions 7 Understanding Probability

I. List the three types of probability.

- A. Classical
- B. Empirical
- C. Subjective

II. Place the letter of the appropriate definition, formula, or expression next to the concept it defines.

1. E 2. J 3. M 4. K 5. O 6. B 7. F 8. C 9. N 10. D 11. G 12. L 13. H 14. I 15. A

III. Identify these probability situations by placing in the space provided a C for Classical, E for Empirical, or S for Subjective.

1. C 2. C 3. E 4. S 5. S 6. E 7. C 8. S 9. E 10. S

IV. The following data concerns the buying habits of people entering a retail store in relation to their gender. Please complete the chart.

Customer Buying Habits and Gender			
Customer Gender Making a Sale	Male	Female	Totals
Yes	42	14	56
No	18	6	24
Totals	60	20	80

V. Using the above data, draw a Venn diagram and determine, using a formula, the probability of each of these events.

A. The probability of making a sale.

$$P(S) = \frac{S}{n} = \frac{56}{80} = .70 \rightarrow 70\%$$



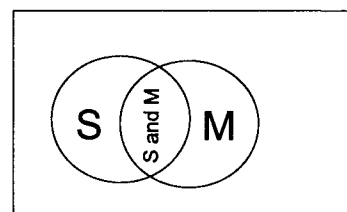
B. The probability of a customer being female.

$$P(F) = \frac{F}{n} = \frac{20}{80} = .25 \rightarrow 25\%$$



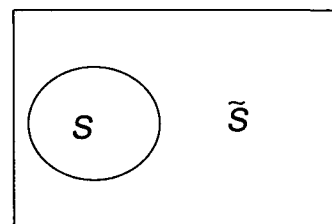
C. The probability of making a sale or a customer being male.

$$\begin{aligned} P(S \text{ or } M) &= P(S) + P(M) - P(S \text{ and } M) \\ &= P\left(\frac{56}{80}\right) + P\left(\frac{60}{80}\right) - P\left(\frac{42}{80}\right) = \frac{74}{80} = .925 = 92.5\% \end{aligned}$$



D. The probability of making a sale or not making a sale.

$$\begin{aligned} P(S \text{ or } \bar{S}) &= P(S) + P(\bar{S}) \\ &= P\left(\frac{56}{80}\right) + P\left(\frac{24}{80}\right) = \frac{80}{80} = 1.00 \rightarrow 100\% \end{aligned}$$



E. State the rule used to answer questions C and D. What condition is necessary to apply each rule?

1. C was done with the general rule of addition because the events are not mutually exclusive.
2. D was done with the special rule for addition because the events are mutually exclusive.

Quick Questions 8 Probability Part II Multiplication Rules

- I. Place the letter of the appropriate definition or formula next to the concept it defines.
 1. E 2. D or A 3. A or D 4. B 5. G 6. J 7. C 8. F 9. H 10. I
- II. Complete this chart concerning the number of hours students studied for a test and their exam grades.

Hours studying	<4	≥ 4	Total
Test score			
< 85	8	2	10
≥ 85	<u>2</u>	<u>8</u>	<u>10</u>
Totals	10	10	20

- III. Use a formula and the data in question II to answer the following questions.

A. The probability of earning a grade less than 85.

$$P(< 85) = \frac{<85}{n} = \frac{10}{20} = .50 \rightarrow 50\%$$

B. The probability of someone studying 4 or more hours and earning a grade of 85 or higher.

$$P(\geq 4 \text{ and } \geq 85) = P(\geq 4) P(\geq 85 | \geq 4) = \frac{10}{20} \times \frac{8}{10} = \frac{80}{200} = .40 = 40\%$$

C. Was the special rule of multiplication applicable to question B? Why or why not?

The special rule for multiplication was not used because the events are not independent. The high percentage of grades ≥ 85 in the group that studied at least 4 hours indicates that studying affects grades. Because studying affects grades, it is the condition or given variable.

D. Use Bayes' theorem to calculate the probability of someone who studied 4 or more hours scoring 85 or higher.

$$P(\geq 85 | \geq 4) = \frac{P(\geq 85 \text{ and } \geq 4)}{P(\geq 4)} = \frac{P(\geq 85) \times P(\geq 4 | \geq 85)}{P(\geq 85) \times P(\geq 4 | \geq 85) + P(< 85) \times P(\geq 4 | < 85)} = \frac{\frac{10}{20} \times \frac{8}{10}}{\frac{10}{20} \times \frac{8}{10} + \frac{10}{20} \times \frac{2}{10}} = \frac{\frac{80}{200}}{\frac{80}{200} + \frac{20}{200}} = \frac{.40}{.50} = 80\%$$

E. Prove your answer to question D using the chart on page 50.

$$P(\geq 85 | \geq 4) = \frac{P(\geq 85)}{P(\geq 4)} = \frac{8}{10} = 80\%$$

IV. How many stores will a salesperson visit if they must visit 3 locations in each of 4 cities?

$$\text{Number of stores} = MN = 3 \times 4 = 12$$

V. An advertising manager has 6 advertisements of equal size to place horizontally across a magazine page.

A. How many ways can the 6 ads be arranged?

$$N! = 6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720 \text{ possible arrangements}$$

B. How many ways can 4 of the 6 ads be arranged if order counts?

$${}_N P_R = \frac{N!}{(N-R)!}$$

$${}_6 P_4 = \frac{6!}{(6-4)!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} = 6 \times 5 \times 4 \times 3 = 360$$

C. How many ways can 4 of the 6 ads be arranged if order does not count and a,b,c,d and d,c,b,a are considered the same arrangement?

$${}_N C_R = \frac{N!}{(N-R)!(R!)}$$

$${}_6 C_4 = \frac{6!}{(6-4)!4!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 4 \times 3 \times 2 \times 1} = \frac{6 \times 5}{2 \times 1} = 15$$

Quick Questions 9 Discrete Probability Distributions

I. Place the letter of the appropriate definition or formula next to the concept or value it defines.

1. I 2. H or J 3. A 4. E 5. J or H 6. B 7. D 8. C 9. F 10. G

II. The sales manager of the XYZ Company made the following estimates of next year's sales.

Sales (x) (millions of \$)	P(x)	x • P(x)	x ²	x ² • P(x)
4	0.2	.80	16	3.20
5	0.4	2.00	25	10.00
5	<u>0.4</u>	<u>2.00</u>	25	<u>10.00</u>
Totals	1.0	4.80		23.20

A. What are expected sales for next year?

$$E(x) = \sum [x \cdot P(x)] = \$4.80$$

B. Calculate the variance for this probability distribution.

$$V(x) = [\sum x^2 \cdot P(x)] - [E(x)]^2$$

$$\begin{aligned} V(x) &= \$23.20 - (\$4.80)^2 \\ &= \$23.20 - \$23.04 \\ &= \$.16 \end{aligned}$$

III. Five percent of the parts coming off an assembly line are defective.

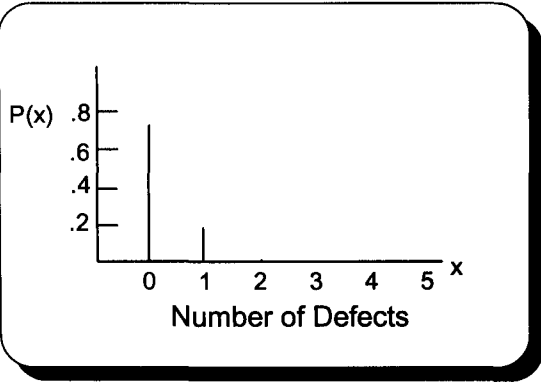
A. Using the binomial formula or your statistics software, calculate the probability of exactly 2 out of 5 parts being defective.

Given
 $p = .05$
 $q = 1 - p = .95$
 $n = 5$
 $x = 2$

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$\begin{aligned} P(2) &= \frac{5!}{2!(5-2)!} \cdot .05^2 \times .95^{5-2} \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} \times .0025 \times .857375 \\ &= 10 \times .0021 = .021 = 2.1\% \end{aligned}$$

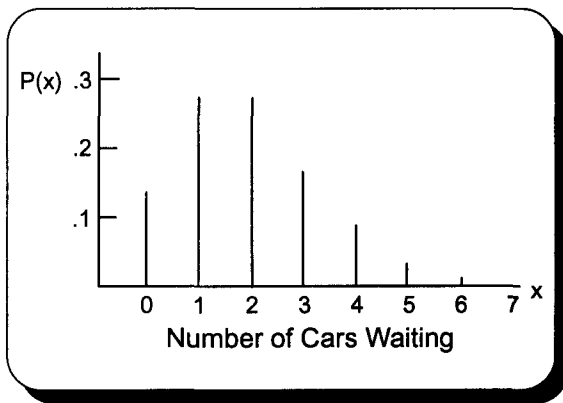
B. Determine the distribution of defective parts using a table in the back of this book. Graph the distribution.



Binomial Probability Distribution n = 5, p = .05, and q = 1 - p = .95	
# of sales (x)	P(x)
0	.774
1	.204
2	.021
3	.001
4	.000
5	<u>.000</u>
Total	1.000

IV. A bank found that the average number of cars waiting during the noon hour at a drive-up window follows a Poisson distribution with a mean of 2 cars. Make a chart of this distribution using a Poisson distribution table. Graph the distribution and answer these questions concerning the probability of cars waiting at the drive-up window.

A.



x	$\mu = 2$
0	0.1353
1	0.2707
2	0.2707
3	0.1804
4	0.0902
5	0.0361
6	0.0120
7	0.0034
8	0.0009
9	0.0002

B. No cars waiting

$$P(x = 0) = .1353 \rightarrow 13.53\%$$

C. Two cars waiting

$$P(x = 2) = .2707 \rightarrow 27.07\%$$

D. At least three cars waiting

$$P(x \geq 3) = [1 - (.1353 + .2707 + .2707)] = [1 - .6767] = .3233 \rightarrow 32.33\%$$

E. Not as many as 3 cars waiting

$$P(x \leq 2) = .1353 + .2707 + .2707 = .6767 = 67.67\%$$

Note: The events described by questions C and D are complements and their answers total to one.

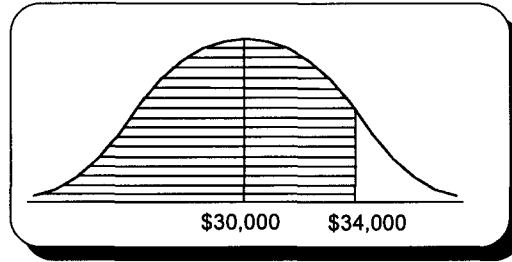
Quick Questions 10 Continuous Normal Probability Distributions

- I. The average income of 30-year-old college graduates from State University is normally distributed with a mean of \$30,000 and a standard deviation of \$4,000. Calculate the following being sure to graph each question.

A. $P(x < \$34,000)$

$$P(\$30,000 < x \leq \$34,000)$$

$$\begin{aligned} Z &= \frac{x - \mu}{\sigma} \\ &= \frac{\$34,000 - \$30,000}{\$4,000} \\ &= \frac{\$4,000}{\$4,000} \\ &= 1 \rightarrow .3413 \end{aligned}$$



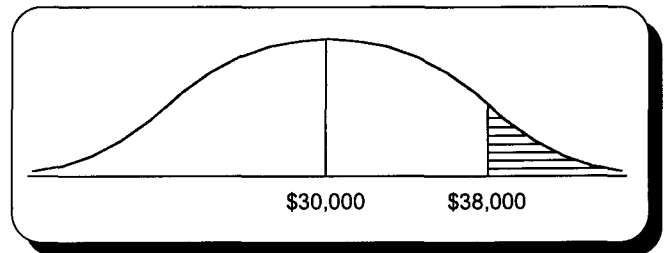
$$P(x < \$30,000) = .5000$$

$$P(x < \$34,000) = .3413 + .5000 = .8413 \rightarrow 84.13\%$$

B. $P(x > \$38,000)$

$$\begin{aligned} Z &= \frac{x - \mu}{\sigma} \\ &= \frac{\$38,000 - \$30,000}{\$4,000} \\ &= \frac{\$8,000}{\$4,000} \\ &= 2.0 \rightarrow .4772 \end{aligned}$$

$$\begin{aligned} &.5000 \\ &- .4772 \\ &----- \\ &.0228 \\ &\text{or } 2.28\% \end{aligned}$$

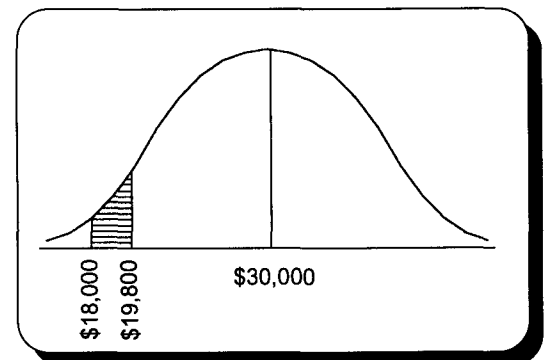


C. $P(\$18,000 \leq x < \$19,800)$

$$Z = \frac{x - \mu}{\sigma} = \frac{\$18,000 - \$30,000}{\$4,000} = \frac{-\$12,000}{\$4,000} = -3.00 \rightarrow .4987$$

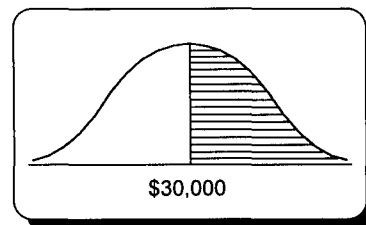
$$Z = \frac{x - \mu}{\sigma} = \frac{\$19,800 - \$30,000}{\$4,000} = \frac{-\$10,200}{\$4,000} = -2.55 \rightarrow .4946$$

$$.4987 - .4946 = .0041 \rightarrow .41\%$$



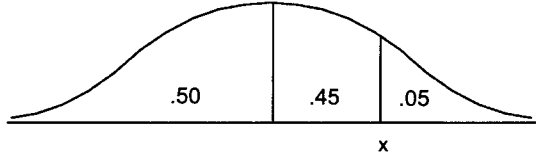
D. $P(x > \$30,000)$

$$P(x > \$30,000) = 50\%$$

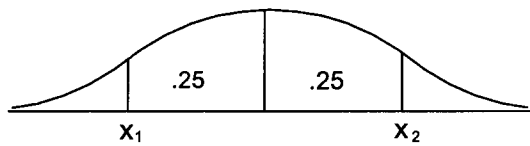


II. Grades of State University graduates are normally distributed with a mean of 3.0 and a standard deviation of .3. Calculate the following being sure to graph each question.

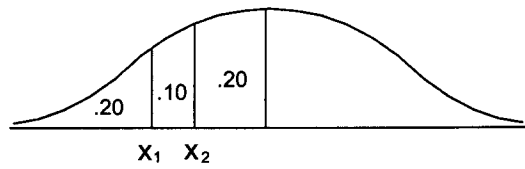
A. What grade point average is required to be in the top 5% of the graduating class?

$50\% - 5\% = 45\% \rightarrow Z = 1.65$	$\mu \pm Z\sigma$ $3.0 + 1.65(.3)$ $3.0 + .50$ 3.50	
--	--	--

B. Calculate the interquartile range.

$25\% \rightarrow Z = .67$	$\mu \pm Z\sigma$ $3.0 \pm .67(.3)$ $3.0 \pm .20$ $2.80 \leftrightarrow 3.20$	
----------------------------	--	--

C. An eccentric alumnus left scholarship money for students in the third decile from the bottom of their class. Determine the range for the third decile. Would a student with a 2.8 grade point average qualify for this scholarship?

$30\% \rightarrow Z = .84$ $3.0 - .84(.3)$ $3.0 - .25$ 2.75	$20\% \rightarrow Z = .52$ $3.0 - .52(.3)$ $3.0 - .16$ 2.84	
--	--	--

$2.75 \leftrightarrow 2.84$
 Yes!

D. What is the median grade point average of this class?

The median is 3.0 because with a normal distribution, the mean and median are equal.

Quick Questions 11 Sampling and the Sampling Distribution of the Means

I. Place the number of the appropriate formula next to the concept it defines.

- A. The 99% confidence interval 3
- B. Standard error of the mean 1
- C. Used when the population variance is unknown and the sample is large 5
- D. The 95% confidence interval 4
- E. The mean of the sampling distribution of the means 2

1.	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
2.	$\mu_{\bar{x}}$
3.	$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$
4.	$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
5.	$\bar{x} \pm 2.58 \frac{S}{\sqrt{n}}$

II. Answer the following true or false and fill in the blank questions.

- A. The primary cause of sampling error is poor collection techniques. T F
- B. The standard error of the mean is halved when the sample size is doubled. T F
- C. A one-number estimate of the population mean is called a point estimate of the mean.
- D. A range for a population parameter is called the confidence interval.
- E. A stratified random sample may be more accurate than a simple random sample because a small diverse section of the population might not be chosen with a simple random sample.

III. Calculate the 95% and 99% confidence intervals for the population mean given a sample of 36 resulted in a mean of 55 and a standard deviation of 18.

95% CI → z = 1.96

$$\bar{x} \pm Z \frac{S}{\sqrt{n}}$$

$$55 \pm 1.96 \frac{18}{\sqrt{36}}$$

$$55 \pm 1.96(3)$$

$$55 \pm 5.9$$

$$49.1 \leftrightarrow 60.9$$

99% CI → z = 2.58

$$\bar{x} \pm Z \frac{S}{\sqrt{n}}$$

$$55 \pm 2.58 \frac{18}{\sqrt{36}}$$

$$55 \pm 2.58(3)$$

$$55 \pm 7.7$$

$$47.3 \leftrightarrow 62.7$$

Quick Questions 12 Sampling Distributions Part II

I. Place the number of the appropriate formula next to the item it describes.

- A. Population proportion 5
- B. Standard error of the proportion 1
- C. Confidence interval for the population proportion 4
- D. Finite correction factor 2
- E. When to use the finite correction factor 3
- F. Sample size when predicting the population mean 7
- G. Sample size when predicting the population proportion 6

1.	$\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$
2.	$\sqrt{\frac{N-n}{N-1}}$
3.	$\frac{n}{N} \geq .05$
4.	$\bar{p} \pm z\sigma_{\bar{p}}$
5.	$\frac{x}{n}$
6.	$\bar{p}(1-\bar{p})\left(\frac{z}{E}\right)^2$
7.	$\left(\frac{z\sigma}{E}\right)^2$

II. A survey of 80 New York City voters revealed 60 planned to vote in the next election. Calculate both the 99% and 95% confidence interval for the population proportion.

$$n = 80 \geq 30$$

$$np = 80(.75) = 60 \geq 5$$

$$nq = 80(1.00 - .75) = 20 \geq 5$$

A. 99% confidence interval

$$\bar{p} = \frac{x}{n} = \frac{60}{80} = .75 \rightarrow 75\%$$

New York City has a very large population. n/N is less than .05 and the finite correction factor is not required.

$$\bar{p} \pm z\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$.75 \pm 2.58\sqrt{\frac{.75(1-.75)}{80}}$$

$$.75 \pm 2.58(.0484)$$

$$.625 \leftrightarrow .875$$

B. 95% confidence interval

$$\bar{p} \pm z\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$.75 \pm 1.96\sqrt{\frac{.75(1-.75)}{80}}$$

$$.75 \pm 1.96(.0484)$$

$$.655 \leftrightarrow .845$$

C. Using the same data, calculate the 99% confidence interval assuming the results came from a city of 1,500 voters.

$$\frac{n}{N} = \frac{80}{1,500} = .053 > .05$$

The finite correction factor is required.

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$= .0484\sqrt{\frac{1,500-80}{1,500-1}}$$

$$= .0471$$

$$\bar{p} \pm z\sigma_{\bar{p}}$$

$$.75 \pm 2.58(.0471)$$

$$.628 \leftrightarrow .872$$

III. Restaurant customers leave a tip approximately 70% of the time. A 95% confidence interval for the tip's proportion is desired. The answer should be correct within 5%. How many customers must be surveyed? Computer students set s to $\sqrt{pq} = \sqrt{.21} = .458$

$$n = \bar{p}(1-\bar{p})\left(\frac{z}{E}\right)^2 = .70(1-.70)\left(\frac{1.96}{.05}\right)^2 = .70(.30)(39.2)^2 = .21(1,537) = 322.77 \rightarrow 323$$

IV. Linda will consider opening a new video showcase in towns with average family income over \$35,000. She requires a 99% confidence interval. The estimate should be within \$1,000 of the population mean. Recently gathered data indicates the population standard deviation is \$4,000. What size sample is required?

$$n = \left(\frac{z\sigma}{E}\right)^2$$

$$= \left[\frac{(2.58)(4,000)}{1,000}\right]^2$$

$$= [10.32]^2 = 106.502 \rightarrow 107$$

Quick Questions 13 Large Sample Hypothesis Testing

- I. Complete the following chart and questions.
- Type I error is called alpha error.
 - Type II error is called beta error.
 - When z calculated from sample data is beyond the critical value (less than for left tail problems and greater than for right tail problems), the null hypothesis is rejected.
 - True

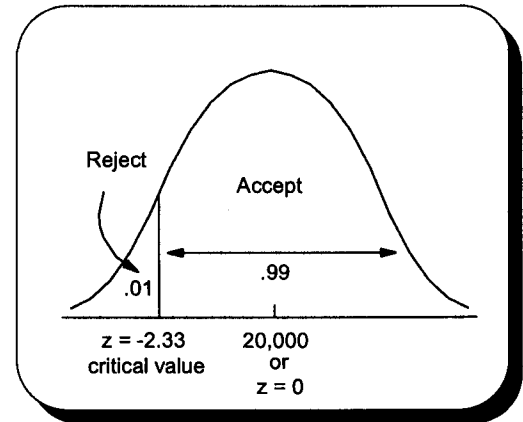
Error Summary		
Decision Concerning Null Hypothesis	Nature's True State	
	H ₀ is true	H ₀ is false
Accept H ₀	Correct	Type II error
Reject H ₀	Type I error	Correct

- II. Make these tests using the 5-step approach to hypothesis testing.
- A. A light bulb warranty states average bulb life is at least 20,000 hours. A sample of 49 bulbs had an average life of 19,000 hours. The population standard deviation is 1,400 hours. Test the warranty claim to the .01 level of significance.

- $H_0 : \mu \geq 20,000$ hours $H_1 : \mu < 20,000$ hours
- $\alpha = .01$ (Note: H_1 points to the area of rejection)
- \bar{x} is the test statistic.
- The critical value of z for .01 is -2.33. If the test Z is beyond -2.33, reject H₀.
- Apply the decision rule.

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{19,000 - 20,000}{\frac{1,400}{\sqrt{49}}} = \frac{-1,000}{200} = -5.0$$

Reject H₀ because -5.0 is beyond -2.33.
The claim is not substantiated.

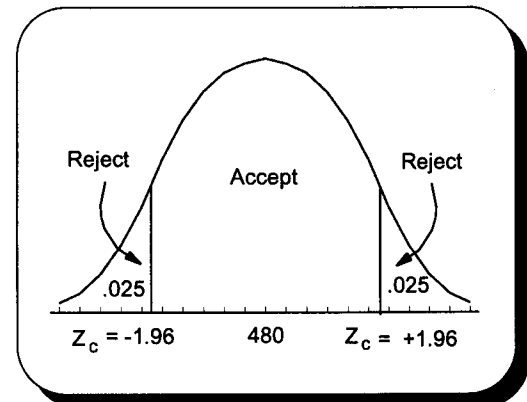


- B. Average weekly manufacturing earnings were \$480 and the standard deviation was \$72. A recent sample of 36 resulted in a mean of \$450. The standard deviation has not changed. Test to the .05 level whether average weekly earnings changed.

- $H_0 : \mu = \$480$ and $H_1 : \mu \neq \$480$
- $\alpha = .05$
- \bar{x} is the test statistic.
- The critical value of z for $\alpha/2 = .05/2 = .025$ is ± 1.96 . If the test Z is beyond ± 1.96 , reject H₀.
- Apply the decision rule.

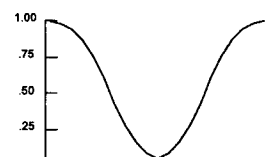
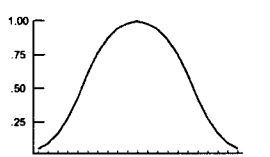
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{450 - 480}{\frac{72}{\sqrt{36}}} = \frac{-30}{12} = -2.50$$

Reject H₀ because -2.50 is beyond -1.96.
Weekly earnings changed.



Quick Questions 14 Large Sample Hypothesis Testing Part II

- I. Place the number of the description next to the item it describes.

1. Area beyond the test statistic	2. $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	3. 	4. 
-----------------------------------	---	---	--

- A. Power curve 3 B. P-value 1 C. Z for testing two means 2 D. Operating characteristics curve 4

- II. Ace Realty wants to determine whether the average time it takes to sell homes is different for its two offices. A sample of 40 from office #1 revealed a mean of 90 days and a standard deviation of 15 days. A sample of 50 from office #2 revealed a mean of 100 days and a standard deviation of 20 days. Use a .05 level of significance.

Office #1	$n_1 = 40$	$\bar{x}_1 = 90$ days	$s_1 = 15$ days
Office #2	$n_2 = 50$	$\bar{x}_2 = 100$ days	$s_2 = 20$ days

1.	$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$
2.	$\alpha = .05$ and $.05 / 2 = .025$
3.	\bar{X} is the test statistic.
4.	The critical value for .025 is ± 1.96 . If the test Z is beyond -1.96, reject H_0 .
5.	Apply the decision rule.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{90 - 100}{\sqrt{\frac{(15)^2}{40} + \frac{(20)^2}{50}}} = \frac{-10}{\sqrt{5.625 + 8}} = -2.71$$

Reject H_0 because -2.71 is beyond -1.96 .
Sales time is not the same at these two offices.

- III. Tough Tire Company is concerned that tread life of its new all weather tire may be below the 70,000 mile warranty. A sample of 36 revealed a mean of 69,800 miles and a standard deviation of 750 miles. Using a .05 level of significance and the p-value approach, test Tough Tire's warranty claim.

Given: $\bar{x} = 69,800$ miles, $n = 36$

$s = 750$ miles and $\alpha = .05$

$H_0 : \mu \geq 70,000$ miles $H_1 : \mu < 70,000$ miles

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{(69,800 - 70,000)}{\frac{750}{\sqrt{36}}} = \frac{-200}{125} = -1.60$$

$$z = -1.60 \rightarrow .4452 \text{ and } p = .5000 - .4452 = .0548$$

Accept H_0 because $.0548 > .05$. Warranty is substantiated.

- IV. The Easy Loan Company wants to determine whether the average length of car loans has increased from last year's population mean of 50 months. A sample of 49 had a mean of 53 months and a standard deviation of 14 months.

- A. Test $H_0 : \mu \leq 50$ and $H_1 : \mu > 50$ at the .05 level of significance.

Given: $\bar{x} = 53$ months, $n = 49$

$s = 14$ months and $\alpha = .05$

$\alpha = .05 \rightarrow z = 1.645$

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{(53 - 50)}{\frac{14}{\sqrt{49}}} = \frac{3}{2} = 1.50 \text{ Accept } H_0 \text{ because } 1.50 < 1.645$$

Loan length did not increase.

- B. Calculate the critical value of \bar{x} .

$$\bar{x} = \mu + z \frac{\sigma}{\sqrt{n}}$$

$$= 50 + 1.645 \frac{14}{\sqrt{49}}$$

$$= 50 + 3.29$$

$$= 53.29$$

- C. Calculate type II error for $\mu = 55$ months.

$$Z = \frac{\bar{x} - \mu_1}{\frac{\sigma}{\sqrt{n}}} = \frac{53.29 - 55.00}{\frac{14}{\sqrt{49}}} = \frac{-1.71}{2} = -.855 \rightarrow .3037$$

$$.50 - .3037 = 19.63\%$$

- D. What is the type II error for these population means?

54 months

$$Z = \frac{\bar{x} - \mu_2}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{53.29 - 54}{\frac{14}{\sqrt{49}}}$$

$$= \frac{-.71}{2} = -.355 \rightarrow .1387$$

$$.50 - .1387 = 36.13\%$$

53.31 months

$$Z = \frac{\bar{x} - \mu_3}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{53.29 - 53.31}{\frac{14}{\sqrt{49}}}$$

$$= \frac{-.02}{2} = -.01 \rightarrow .0040$$

$$.50 - .004 = 49.6\%$$

50.01 months

$$Z = \frac{\bar{x} - \mu_4}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{53.29 - 50.01}{\frac{14}{\sqrt{49}}}$$

$$= \frac{3.28}{2} = 1.64 \rightarrow .4495$$

$$.4495 + .5000 = .9495$$

Note: When the population mean is 50 months or less, the null hypothesis is true and type II error (accepting a false null hypothesis) does not exist. The maximum type II error is 95% for the 5% level of significance.

Quick Questions 15 Hypothesis Testing of Population Proportions

I. Place the number of the appropriate formula or expression next to the item it describes.

A. When using the normal approximation to the binomial distribution,

1. np and $n(1 - p)$ must be 4

2. n must be 2

B. A one population test 5

C. $\bar{p}_w =$ 3

D. A two population test 1

1.	$\frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{\bar{p}_w(1-\bar{p}_w)}{n_1} + \frac{\bar{p}_w(1-\bar{p}_w)}{n_2}}}$
2.	≥ 30
3.	$\frac{x_1 + x_2}{n_1 + n_2}$
4.	≥ 5
5.	$\frac{\bar{p} - p}{\sigma_{\bar{p}}}$

II. A national video publication stated long-term tape rentals average 20% of all tape rentals. A 150 customer study at Linda's Video Showcase revealed 24 long-term rentals. Test at the .05 level of significance whether Linda's long-term rentals are less than the national average.

$p = .20$	$x = 24$	$n = 150$	$\alpha = .05 \rightarrow z = \pm 1.645$
-----------	----------	-----------	--

$$n = 150 \geq 30$$

$$np = 150(.2) = 30 \geq 5$$

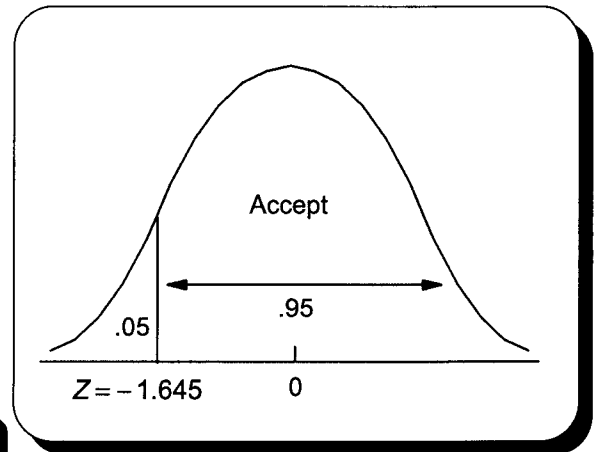
$$nq = 150(1 - .2) = 120 \geq 5$$

$$\bar{p} = \frac{x}{n} = \frac{24}{150} = .16$$

$$H_0 : P \geq .20 \text{ and } H_1 : P < .20$$

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.16 - .20}{\sqrt{\frac{.20(1-.20)}{150}}} = -1.22$$

Accept H_0 because - 1.22 is not beyond - 1.645. The proportion of customers renting tapes for longer than the minimum period is not less than the number stated in a national publication.



III. Linda Smith found that 70 out of 100 customers rented 2 or more tapes at one store and 44 out of 50 rented 2 or more tapes at a second store. Test at the .05 level of significance whether there is a difference between the proportion of customers at these two stores renting 2 or more tapes.

Given:	$X_1 = 70$	$n_1 = 100$	$X_2 = 44$	$n_2 = 50$	$\alpha = .05/2 = .025 \rightarrow z = \pm 1.96$
---------------	------------	-------------	------------	------------	--

$$p_1 = \frac{70}{100} = .70$$

$$p_2 = \frac{44}{50} = .88$$

$$\begin{aligned} \bar{p}_w &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{70 + 44}{100 + 50} \\ &= \frac{114}{150} \\ &= .76 \end{aligned}$$

$$\begin{aligned} Z &= \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{\bar{p}_w(1-\bar{p}_w)}{n_1} + \frac{\bar{p}_w(1-\bar{p}_w)}{n_2}}} \\ &= \frac{.70 - .88}{\sqrt{\frac{.76(1-.76)}{100} + \frac{.76(1-.76)}{50}}} \\ &= -2.43 \end{aligned}$$

Reject H_0 because - 2.43 is beyond - 1.96. The proportion of customers renting two or more tapes differs at these two stores.

Quick Questions 16 Small Sample Hypothesis Testing Using Student's t Test

I. Place the number of the appropriate definition or formula next to the concept it defines.

- A. Weighted or pooled estimate of the population variance 1
 B. Standard deviation of the differences 4
 C. t when comparing two dependent populations 5
 D. t when comparing two independent populations 2
 E. Used with one population 6
 F. Requires the use of the t distribution 3

1. $\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$	4. $\sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$
2. $\frac{x_1 - x_2}{\sqrt{s_w^2(\frac{1}{n_1} + \frac{1}{n_2})}}$	5. $\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$
3. the population is approximately normal, $n \leq 30$, and the population variance isn't known	6. $\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$

II. Linda is tracking the number of work days missed by employees before and after taking part in a company-sponsored lunchtime physical fitness program. Test at the .01 level of significance whether the average number of days missed went down for program participants.

Employee	A	B	C	D	E	F	G	
Before	8	9	6	8	3	4	5	
After	6	7	5	6	5	2	5	
d	2	2	1	2	-2	2	0	$\sum d = 7$
d ²	4	4	1	4	4	4	0	$\sum d^2 = 21$

$$\bar{d} = \frac{\sum d}{n} = \frac{7}{7} = 1.0$$

$$df = n - 1 = 7 - 1 = 6$$

$$\alpha \text{ of } .01 \rightarrow t = 3.143$$

$$H_0 : \mu_d \leq 0 \text{ and } H_1 : \mu_d > 0$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}} = \sqrt{\frac{21 - \frac{7^2}{7}}{7-1}} = 1.53$$

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{1.0}{\frac{1.53}{\sqrt{7}}} = 1.72 \quad \text{Accept } H_0 \text{ because } 1.72 < 3.143.$$

Days missed did not go down.

III. Eight men applying to State University had a sample mean and variance on college board tests of 1,050 and 2,500 respectively. The respective numbers for nine women were 1,075 and 3,600. Test at the .05 level of significance whether women did better than men on these tests.

$n_1 = 8$
$\bar{X}_1 = 1,050$
$S_1^2 = 2,500$
$n_2 = 9$
$\bar{X}_2 = 1,075$
$S_2^2 = 3,600$
$\alpha = .05$

1. $H_0 : \mu_2 \leq \mu_1$ and $H_1 : \mu_2 > \mu_1$
2. $\alpha = .05$
3. The test statistic is \bar{x}
4. $df = n_1 + n_2 - 2 = 8 + 9 - 2 = 15$ $\alpha \text{ of } .05 \rightarrow t = -1.753$
5. Apply the decision rule.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_w^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$= \frac{1,050 - 1,075}{\sqrt{3,086.7(\frac{1}{8} + \frac{1}{9})}}$$

$$= -.93$$

$$s_w^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(8-1)2,500 + (9-1)3,600}{8 + 9 - 2}$$

$$= \frac{17,500 + 28,800}{15} = 3,086.7$$

Accept H_0 because $-.93$ is not beyond -1.753 . Women's scores were not higher than men's scores.

Quick Questions 17 Statistical Quality Control

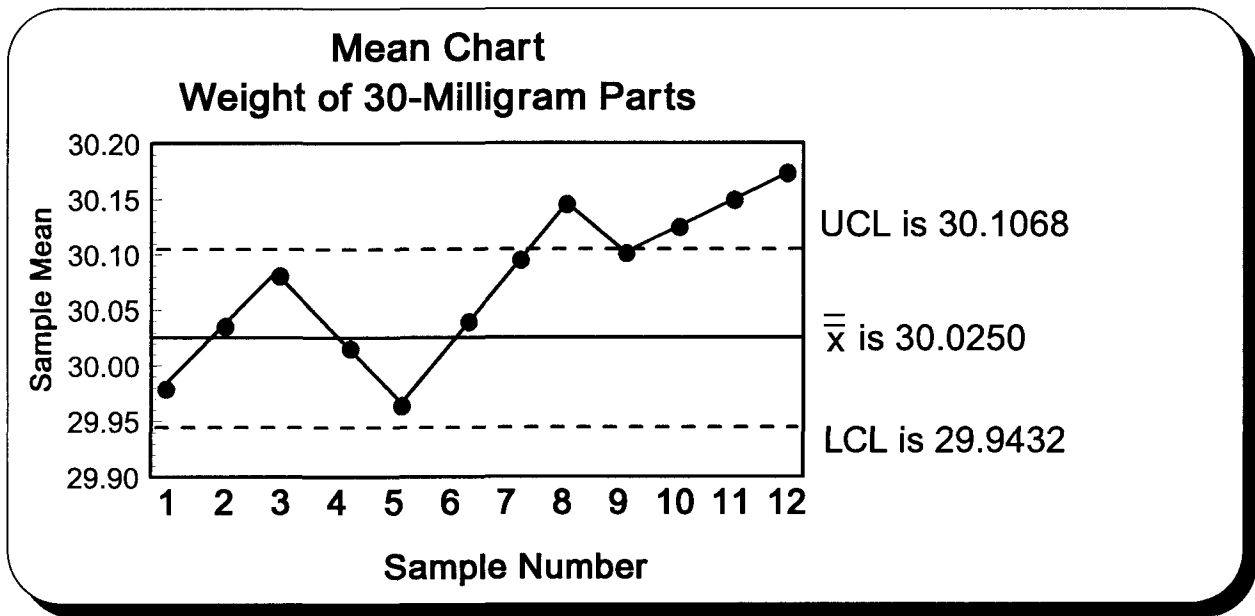
I. Place the number of the appropriate formula, expression, or term next to the concept it describes.

- A. A control chart 5
- B. Assignable variation 3
- C. Random variation 4
- D. An \bar{x} chart 1
- E. A range chart 6
- F. A p chart 2

1. Measures whether the mean size, weight, or temperature, etc., is getting too high or too low.
2. Measures whether the proportion of some attribute (defects) is appropriate.
3. Results from an identifiable cause
4. Is due to chance
5. Measures a process value (statistic) sequentially over a period of time
6. Measures whether variation in size, weight, or temperature, etc., is too large.

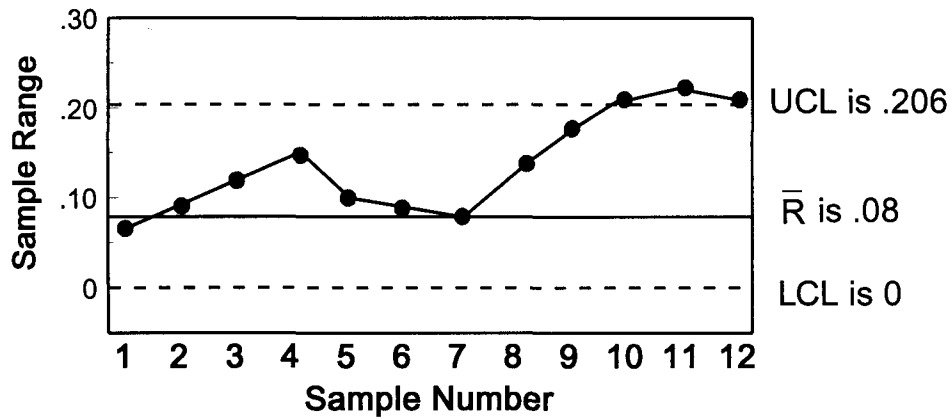
II. Control charts developed in Practice Set 17 will now be used to determine whether the 30-milligram part manufacturing process is in control. Plot this data on the appropriate control chart and determine whether the process is in control.

Sample	A	B	C	D	E	F	G	H	I	J	K	L
Sample Mean	29.98	30.04	30.08	30.02	29.97	30.04	30.09	30.15	30.10	30.12	30.14	30.16
Sample Range	0.07	0.09	0.11	0.13	0.10	0.09	0.08	0.14	0.18	0.21	0.22	0.21
Proportion of Defects	0.08	0.11	0.14	0.17	0.23	0.21	0.19	0.17	0.11	0.09	0.12	0.21



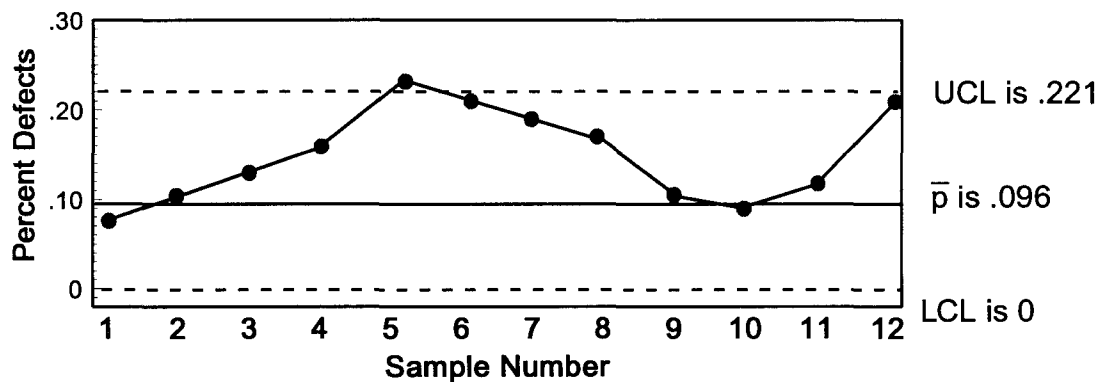
Analysis: The process appears out of control. With a 99.74% confidence level, one sample beyond a control limit should happen only 26 times out of 10,000 samples. Three samples in a row and four out of five beyond the 99.74% confidence interval is unlikely for a process under control.

Range Chart
Weight of 30-Milligram Parts



Analysis: The last 3 samples were beyond the control limit and the trend seems to be increasing. Most of the data is above the center line which is not a good sign. Deciding to shut down a line that is trending out of control requires knowledge and experience concerning the manufacturing process, an analysis of the cost of shutting down, and information concerning the nonmanufacturing costs associated with poor quality. Measuring the nonmanufacturing costs associated with poor quality is a very subjective process.

Proportion of Defects - The P Chart
Defective 30-Milligram Parts



Analysis: Data that is trending out of control often is cyclical in nature. This data appears to be cyclical.

Quick Questions 18 Analysis of Variance

I. Copy the formulas and expressions on the right into this ANOVA summary chart.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	t - 1	SS_T	$MS_T = \frac{SS_T}{t-1}$	$F = \frac{MS_T}{MS_E}$
Within Treatments (error)	N - t	SS_E	$MS_E = \frac{SS_E}{N-t}$	
Total Variance	N - 1	SS_{TOTAL}		

SS_T	$F = \frac{MS_T}{MS_E}$
N - t	SS_{TOTAL}
$MS_T = \frac{SS_T}{t-1}$	t - 1
$MS_E = \frac{SS_E}{N-t}$	SS_E
N - 1	

II. Answer the following fill in the blank questions.

- A. Analysis of variance requires the populations be normally distributed.
- B. When using the F distribution, the numerator is always the larger of the 2 variances.
- C. When doing ANOVA, the numerator of the F distribution measures variance between the treatments.
- D. When doing ANOVA, the denominator of the F distribution measures variance within the treatments.

III. Complete the following ANOVA study concerning grade point averages randomly selected by a local college.

A. Begin by completing this chart. Those using statistics software should skip to part D.

Analysis of College Grades Based Upon High School Grades							Row Totals Required for Calculations
	High H.S. Grades T_1		Medium H.S. Grades T_2		Low H.S. Grades T_3		
	College Grades(X_1)	X_1^2	College Grades(X_2)	X_2^2	College Grades(X_3)	X_3^2	
	3.4	11.56	3.2	10.24	2.1	4.41	
	3.5	12.25	2.8	7.84	2.5	6.25	
	<u>3.1</u>	<u>9.61</u>	<u>3.0</u>	<u>9.00</u>	<u>2.7</u>	<u>7.29</u>	
$\sum X_T$	10.0		9.0		7.3		$\sum x = 26.3$
$(\sum X_T)^2$	100.0		81.0		53.29		
n	3.0		3.0		3.0		$N = 9$
$\frac{(\sum X_T)^2}{n}$	33.33		27.0		17.76		$\sum [\frac{(\sum X_T)^2}{n}] = 78.09$
$\sum X_T^2$		33.42		27.08		17.95	$\sum x^2 = 78.45$

B. Using the chart on the previous page, calculate the following values.

$$SS_T = \Sigma \left[\frac{(\Sigma x_T)^2}{n} \right] - \frac{(\Sigma X)^2}{N}$$

$$= 78.09 - \frac{26.3^2}{9}$$

$$= 78.09 - 76.85$$

$$= 1.24$$

$$SS_E = \Sigma x^2 - \Sigma \left[\frac{(\Sigma x_T)^2}{n} \right]$$

$$SS_E = 78.45 - 78.09$$

$$= .36$$

$$SS_{TOTAL} = \Sigma x^2 - \frac{(\Sigma x)^2}{N}$$

$$SS_{TOTAL} = 78.45 - 76.85$$

$$= 1.60$$

Note: Unexplained variability is .36.

C. Complete the following chart using data accumulated to this point.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	$t - 1 = 3 - 1 = 2$	$SS_T = 1.24$	$MS_T = \frac{SS_T}{t-1} = \frac{1.24}{3-1} = .62$	$F = \frac{MS_T}{MS_E} = \frac{.62}{.06} = 10.33$
Within Treatments (error)	$N - t = 9 - 3 = 6$	$SS_E = .36$	$MS_E = \frac{SS_E}{N-t} = \frac{.36}{9-3} = .06$	
Total Variance	$N - 1 = 9 - 1 = 8$	$SS_{TOTAL} = 1.60$		

D. Using the 5-step approach to hypothesis testing, test at the .05 level whether these sample means come from populations with equal means.

1. These are the null hypothesis and alternate hypothesis.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ and } H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

2. The level of significance for this one-tail problem is .05.

3. The test statistic is F.

4. The decision rule will be, if F from the test statistic is beyond that of the critical value for the .05 level of significance, the null hypothesis will be rejected.

The numerator has a df of 2.
The denominator has a df of 6.
F is 5.14.

5. Apply the decision rule.

Reject H_0 because $10.33 > 5.14$. Average grades from these treatments are not equal at the .05 level.

E. Answer problem D at the .01 level of significance.

1. F's critical value is 10.92 (see Table 5A).

2.

Accept H_0 because $10.33 < 10.92$. Average grades from these treatments are equal at the .01 level.

Grades are the same at the .01 level because at this lower level of significance, the acceptance confidence interval is larger.

Quick Questions 19 Two-Factor Analysis of Variance

I. Use the symbols to the right to complete the following ANOVA summary chart.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	t - 1	SS _T	$MS_T = \frac{SS_T}{t-1}$	$F = \frac{MS_T}{MS_E}$
Block	b - 1	SS _B	$MS_B = \frac{SS_B}{b-1}$	
Within Treatments (error)	(t - 1)(b - 1)	SS _E	$MS_E = \frac{SS_E}{(t-1)(b-1)}$	$F = \frac{MS_B}{MS_E}$
Total Variance	N - 1 =	SS _{TOTAL}		

SS _T	$F = \frac{MS_T}{MS_E}$
(t - 1)(b - 1)	SS _{TOTAL}
$MS_T = \frac{SS_T}{t-1}$	t - 1
$MS_B = \frac{SS_B}{b-1}$	SS _E
$MS_E = \frac{SS_E}{(t-1)(b-1)}$	
b - 1	SS _B
$F = \frac{MS_B}{MS_E}$	N - 1

II. The analysis in the last set of Quick Questions will be expanded by rearranging the data in each row so it is based upon the amount of time students spend studying. Complete the following ANOVA study concerning college grades and study times collected by a local college. Begin by completing this chart. Those using statistics software should skip to part C.

Analysis of College Grades Based Upon High School Grades and Time Spent Studying While in College							Row Totals Required for Calculations		
College Study Time	High H.S. Grades T ₁		Medium H.S. Grades T ₂		Low H.S. Grades T ₃		Σ X _B	(Σ X _B) ²	$\frac{(\sum X_B)^2}{t}$
	College Grades(X ₁)	X ₁ ²	College Grades(X ₂)	X ₂ ²	College Grades(X ₃)	X ₃ ²			
High	3.5	12.25	3.2	10.24	2.7	7.29	9.4	88.36	29.45
Medium	3.4	11.56	3.0	9.00	2.5	6.25	8.9	79.21	26.40
Low	<u>3.1</u>	<u>9.61</u>	<u>2.8</u>	<u>7.84</u>	<u>2.1</u>	<u>4.41</u>	<u>8.0</u>	64.00	<u>21.33</u>
							26.3 = Σ x	$\sum [\frac{(\sum X_B)^2}{t}] = 77.18$	
Σ X _T	10		9		7.3		26.3 = Σ x		
(Σ X _T) ²	100		81		53.29				
b	3		3		3		N = 9		
$\frac{(\sum X_T)^2}{b}$	33.33		27		17.76		$\sum [\frac{(\sum X_T)^2}{b}] = 78.09$		
Σ X _T ²		33.42		27.08		17.95	Σ x ² = 78.45		

A. Using this chart, calculate the following values.

$$\begin{aligned}
 SS_T &= \sum \left[\frac{(\sum X_T)^2}{b} \right] - \frac{(\sum X)^2}{N} \\
 &= 78.09 - \frac{26.3^2}{9} \\
 &= 78.09 - 76.85 = 1.24
 \end{aligned}$$

$$\begin{aligned}
 SS_B &= \sum \left[\frac{(\sum X_B)^2}{t} \right] - \frac{(\sum X)^2}{N} \\
 &= 77.18 - 76.85 \\
 &= .33
 \end{aligned}$$

$$\begin{aligned}
 SS_{TOTAL} &= \sum x^2 - \frac{(\sum x)^2}{N} \\
 &= 78.45 - 76.85 \\
 &= 1.60
 \end{aligned}$$

$$\begin{aligned}
 SS_E &= SS_{TOTAL} - (SS_T + SS_B) \\
 &= 1.60 - (1.24 + .33) \\
 &= 1.60 - 1.57 \\
 &= .03
 \end{aligned}$$

Note: Unexplained variability is down from .36 (see page QQ 113) to .03.

B. Complete the following chart using data accumulated to this point.

Variance Analysis Summary Table				
Variance Sources	df	Sum of the Squares	Mean Squares	ANOVA
Between Treatments	$t - 1 = 3 - 1 = 2$	$SS_T = 1.24$	$MS_T = \frac{SS_T}{t-1} = \frac{1.24}{2} = .62$	$F = \frac{MS_T}{MS_E} = \frac{.62}{.0075} = 83$
Block	$b - 1 = 3 - 1 = 2$	$SS_B = .33$	$MS_B = \frac{SS_B}{b-1} = \frac{.33}{2} = .165$	
Within Treatments (error)	$(t - 1)(b - 1) = (2)(2) = 4$	$SS_E = .03$	$MS_E = \frac{SS_E}{(t-1)(b-1)} = \frac{.03}{4} = .0075$	$F = \frac{MS_B}{MS_E} = \frac{.165}{.0075} = 22$
Total Variance	$N - 1 = 9 - 1 = 8$	$SS_{TOTAL} = 1.6$		

C. Using the 5-step approach to hypothesis testing, determine with a .05 level of significance whether these treatment means and block means come from populations with equal means.

- A check of each null hypothesis will be made.
 - $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$ for the treatment means.
 - $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$ for the block means.
- The level of significance is .05.
- The test statistic is F.
- If F from the test statistic is beyond the critical value of F for the .05 level of significance, the null hypothesis will be rejected.
- Apply the decision rule.

Degrees of freedom for the treatment hypothesis is 2 for the numerator and 4 for the denominator. F will be 6.94.

Reject H_0 because $F = 83$ and $83 > 6.94$. Grades from these 3 samples do not come from populations with equal means.

Degrees of freedom for the block hypothesis is 2 for the numerator and 4 for the denominator. F will be 6.94.

Reject H_0 because $F = 22$ and $22 > 6.94$. These studying times do not come from populations with equal means.

III. Using the chart data on pages 112 and 113, determine at the .01 level whether there is a difference between treatment means 1 and 2.

$$\bar{X}_1 = \frac{\sum x}{n_1} = \frac{10}{3} = 3.33$$

$$\bar{X}_2 = \frac{\sum x}{n_2} = \frac{9}{3} = 3.0$$

t is 3.707 for $.01/2 = .005$ and df of $N - t = 9 - 3 = 6$
 $MS_E = .06$ (see page QQ 113)

$$(\bar{X}_1 - \bar{X}_2) \pm t \sqrt{MS_E \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$(3.33 - 3.0) \pm 3.707 \sqrt{.06 \left(\frac{1}{3} + \frac{1}{3} \right)}$$

$$.33 \pm .74$$

The range of $-.41 \leftrightarrow 1.07$ indicates zero is possible and the means are not significantly different.

Quick Questions 20 Nonparametric Hypothesis Testing of Nominal Data

I. Place the number of the formula or expression next to the concept it defines.

- A. $\chi^2 = \underline{5}$
 B. Expected frequency f_e must be 3
 C. f_e for a contingency table equals 2
 D. Chi-square is the ratio of 1
 E. df for use with a contingency table 6
 F. df for a goodness of fit problem 4

1. $(n - 1)s$ to σ^2	4. $k - 1$
2. $\frac{f_r \times f_k}{n}$	5. $\sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$
3. ≥ 5	6. $(r - 1)(c - 1)$

II. Last year, 40% of Linda's customers rented 1 tape, 30% rented 2 tapes, 20% rented 3 tapes, and 10% rented 4 or more tapes. Below is last week's tape rental distribution for Linda's stores. Using the 5-step approach to hypothesis testing, test at the .05 level of significance whether there has been a change in the distribution of tape rentals. Each expected frequency will be the total of 1,000 observations multiplied by last year's appropriate percentage.

Tape Rental Analysis					
	Observed Frequency (f_o)	Expected Frequency (f_e)	($f_o - f_e$)	($f_o - f_e$) ²	($f_o - f_e$) ² / f_e
1 tape	300	.4 x 1,000 = 400	-100	10,000	25.00
2 tapes	250	.3 x 1,000 = 300	-50	2,500	8.33
3 tapes	250	.2 x 1,000 = 200	50	2,500	12.50
4+ tapes	<u>200</u>	.1 x 1,000 = <u>100</u>	100	10,000	<u>100.00</u>
Totals	1,000	1,000			$\chi^2 = 145.83$

1. H_0 : defects follow Linda's distribution.
 H_1 : defects do not follow Linda's distribution.
2. The significance level is .05.
3. Chi-square is the test statistic.
4. The decision rule:

 If χ^2 from the test statistic is beyond the critical value, the difference is significant, reject the null hypothesis.
5. Apply the decision rule.

$$df = k - 1 = 4 - 1 = 3 \rightarrow \chi^2 = 7.81$$

Reject H_0 because $145.83 > 7.81$.
 Last week's distribution does not follow last year's distribution.

III. Is Linda happy with these test results? Why?

Yes. Customers are renting tapes for a longer period than last year. Other things being equal, this means more sales revenue and more profit.

- IV. Using the 5-step approach to hypothesis testing and the .01 level of significance, test whether the number of math courses taken and success in statistics are independent.

Statistics Grades and Math Background at State University			
Grade	Less than B	Greater than or equal to B	Totals
Math courses taken			
Less than or equal to 2	15	5	20
Greater than 2	<u>5</u>	<u>25</u>	<u>30</u>
Totals	20	30	50

Contingency Table of Statistics Grades and Math Background								
Math courses taken	Grade		Less than B		Greater than or equal to B		Totals	
	f_o	f_e	f_o	f_e	f_o	f_e	f_o	f_e
Less than or equal to 2	15	8	5	12	20	20		
Greater than 2	<u>5</u>	<u>12</u>	<u>25</u>	<u>18</u>	<u>30</u>	<u>30</u>		
Totals	20	20	30	30	50	50		

$$f_e = \frac{f_r \times f_c}{n} = \frac{20 \times 20}{50} = 8$$

$$f_e = \frac{f_r \times f_c}{n} = \frac{30 \times 20}{50} = 12$$

$$f_e = \frac{f_r \times f_c}{n} = \frac{20 \times 30}{50} = 12$$

$$f_e = \frac{f_r \times f_c}{n} = \frac{30 \times 30}{50} = 18$$

- H_0 : math courses taken and statistics grades are independent.
 H_1 : math courses taken and statistics grades are not independent.
- The significance level is .01.
- Chi-square is the test statistic.
- The decision rule:
If χ^2 from the test statistic is beyond the critical value, reject the null hypothesis.
- Apply the decision rule.

$$df = (r - 1)(c - 1) = (2 - 1)(2 - 1) = \rightarrow \chi^2 = 6.64$$

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = \sum \left[\frac{(15 - 8)^2}{8} + \frac{(5 - 12)^2}{12} + \frac{(5 - 12)^2}{12} + \frac{(25 - 18)^2}{18} \right]$$

$$= 6.13 + 4.08 + 4.08 + 2.72$$

$$= 17.01$$

Reject H_0 because $17.01 > 6.64$. Math courses taken and statistics grades are not statistically independent at the .01 level of significance.

- III. Darin wants to reexamine the delivery time of 2 suppliers first presented on page 90 and reproduced below. Parametric tests using z or t assume the populations are approximately normal and have equal variances. If these conditions are not met (or unknown) and the shape and dispersion of the distributions are similar, the nonparametric Mann-Whitney test of 2 medians is appropriate. Test at the .05 level of significance whether these samples come from a population with equal medians. For calculation convenience, only the first 11 pieces of data will be used from each data set. **People using statistics software do not need to complete this chart.**

Supplier A: 10, 22, 14, 39, 37, 40, 30, 29, 30, 16, 11 Supplier B: 14, 37, 20, 19, 12, 18, 22, 23, 26, 21, 19														
Complete this table by: (1) completing an ordered array, (2) assigning an A for supplier A and a B for supplier B to each element of the array, (3) assigning each rank to the appropriate category (supplier A or B), (4) calculating each subtotal, and (5) calculating R_1 , which equals the sum of the 3 subtotals for supplier A or R_2 , which equals the sum of the 3 subtotals for supplier B.														
Rank			Supplier		Rank			Supplier		Rank			Supplier	
Ordered Array and Supplier			A	B	Ordered Array and Supplier			A	B	Ordered Array and Supplier			A	B
(1)	(2)		(3)	(3)	(1)	(2)		(3)	(3)	(1)	(2)		(3)	(3)
1.	10	A	1		8.	19	B		8.5	15.	26	B		15
2.	11	A	2		9.	19	B		8.5	16.	29	A	16	
3.	12	B		3	10.	20	B		10	17.	30	A	17.5	
4.	14	A	4.5		11.	21	B		11	18.	30	A	17.5	
5.	14	B		4.5	12.	22	A	12.5		19.	37	A	19.5	
6.	16	A	6		13.	22	B		12.5	20.	37	B		19.5
7.	18	B		7	14.	23	B		14	21.	39	A	21	
										22.	40	A	22	
(4) Subtotal			13.5	14.5	(4) Subtotal			12.5	64.5	(4) Subtotal			113.5	34.5
(5) $R_1 = 13.5 + 12.5 + 113.5 = 139.5$														

$$\begin{aligned}
 U_1 &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \\
 &= 11(11) + \frac{11(11+1)}{2} - 139.5 \\
 &= 121 + 66 - 139.5 \\
 &= 47.5
 \end{aligned}$$

$$\begin{aligned}
 \sigma_U &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \\
 &= \sqrt{\frac{11(11)(11+11+1)}{12}} \\
 &= 15.229
 \end{aligned}$$

$$\begin{aligned}
 \mu_U &= \frac{n_1 n_2}{2} \\
 &= \frac{11(11)}{2} \\
 &= 60.5
 \end{aligned}$$

$$\begin{aligned}
 Z &= \frac{U - \mu_U}{\sigma_U} \\
 &= \frac{47.5 - 60.5}{15.229} \\
 &= -.85
 \end{aligned}$$

Z for this two-tail problem at the 05 level of significance is ± 1.96 . Accept H_0 because $-.85$ is not beyond -1.96 . There is not a significant difference between median delivery times.

Quick Questions 22 Nonparametric Hypothesis Testing of Ordinal Data Part II

- I. Linda is tracking the number of work days missed by employees before and after taking part in the company-sponsored lunchtime physical fitness program. This problem first appeared on page 101. At that time it was assumed the populations were approximately normal. If this assumption is not correct, a paired difference sign test may be conducted at the .10 level of significance to determine whether median work days missed has changed.

Employee	A	B	C	D	E	F	G
Before	8	9	6	8	3	4	5
After	6	7	5	6	5	2	5
Sign	+	+	+	+	-	+	0

- Employee G missed the same number of days and will be excluded from the study.
 - Five of six missed fewer days.
 - The Binomial table (ST 1) yields the following: $p(x \geq 5) = .094 + .016 = .11$. For this two-tail problem, $p = 2(.11) = .22$.
 - Accept H_0 because .22 is greater than .10. Employee absenteeism has not changed.
 - Note:** The null hypothesis would have been rejected if all 6 employees had changed their absenteeism ($p(x \geq 6) = .016$ and $2(.016) = .032 < .10$).
- II. The page 112 ANOVA high school and college grades study assumed the populations were normally distributed with equal variances. These assumptions are not true or unknown. Conduct a .05 level of significance Kruskal-Wallis test to determine the equality of treatment median grades. Page 112 data has been increased to conform with the $n \geq 5$ test requirement.

High H.S. Grades T_1		Medium H.S. Grades T_2		Low H.S. Grades T_3	
College Grades	Rank (R_1)	College Grades	Rank (R_2)	College Grades	Rank (R_3)
3.4	13	3.2	11	2.1	2
3.5	14	2.8	6	2.5	4
3.1	9.5	3.0	8	2.7	5
3.3	12	3.1	9.5	2.3	3
3.6	<u>15</u>	2.9	<u>7</u>	1.8	<u>1</u>
	63.5		41.5		15

H is the designated statistic.
N, the number of observations, is 15.
k, the number of samples, is 3.
n_k , a sample's size, is 5.
R_k is a sample's rank total.
$df = k - 1 = 3 - 1 = 2 \rightarrow \chi^2 = 9.21$

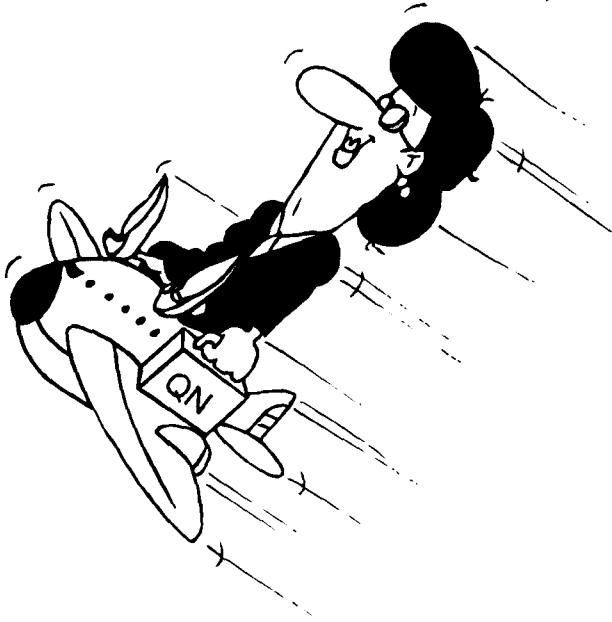
$$H = \frac{12}{N(N+1)} \left[\frac{(\sum R_1)^2}{n_1} + \frac{(\sum R_2)^2}{n_2} + \dots + \frac{(\sum R_k)^2}{n_k} \right] - 3(n+1)$$

$$= \frac{12}{15(15+1)} \left[\frac{(63.5)^2}{5} + \frac{(41.5)^2}{5} + \frac{(15)^2}{5} \right] - 3(15+1)$$

$$= .05(806.45 + 344.45 + 45.00) - 48.00 = 11.795$$

H_0 is rejected because $11.80 > 9.21$. Medians are not equal.

Statistics makes me dizzy.

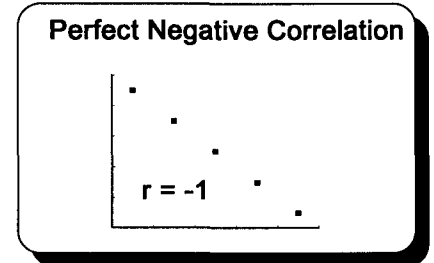
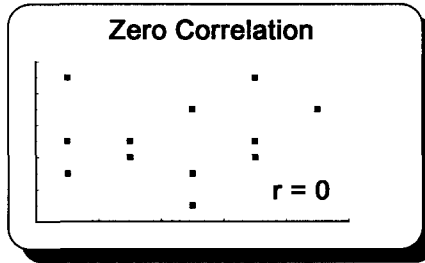
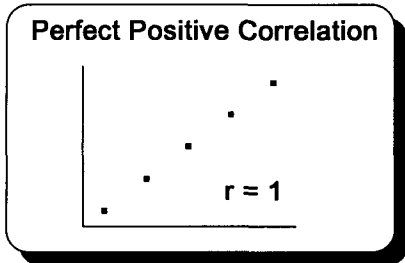


Quick Questions 23 Correlation Analysis

I. Place the number of the appropriate formula, expression, or term next to the appropriate concept.

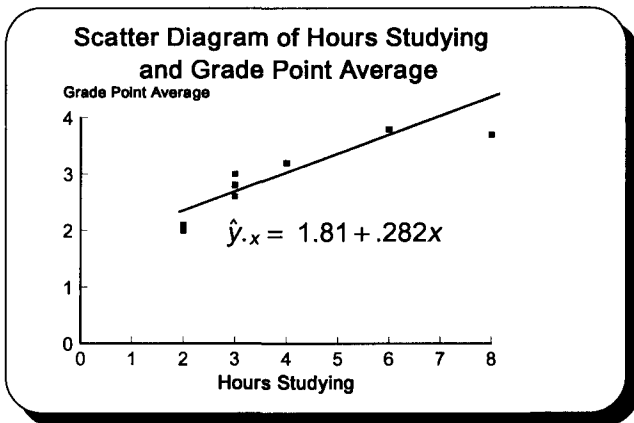
- A. Coefficient of determination 4
 B. Coefficient of correlation 2
 C. A range for r 5
 D. Coefficient of nondetermination 1
 E. The test statistic (t) used to measure the significance of r 3

II. Draw the following scatters and place the appropriate value for r in the space provided.



1. $1 - r^2$, the variability in y that is not explained by x	
2.	$\frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$
3.	$\frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$
4. r^2 , the variability in y that is explained by x	
5.	$-1 \leq r \leq +1$

III. Draw a scatter diagram showing how hours studying per weekend affect grade point average.



Hours Studying per Weekend(x)	Grade Point Average (y)	XY	X ²	Y ²
3	3.0	9.0	9	9.00
2	2.0	4.0	4	4.00
6	3.8	22.8	36	14.44
3	2.6	7.8	9	6.76
4	3.2	12.8	16	10.24
8	3.7	29.6	64	13.69
2	2.1	4.2	4	4.41
<u>3</u>	<u>2.8</u>	<u>8.4</u>	<u>9</u>	<u>7.84</u>
31	23.2	98.6	151	70.38

IV. Using the data in question III, calculate the following:

A. Coefficient of correlation (to 3 decimal places)

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}} = \frac{8(98.6) - (31)(23.2)}{\sqrt{[8(151) - (31)^2][8(70.38) - (23.2)^2]}} = \frac{69.6}{\sqrt{(247)(24.8)}} = .889$$

B. Coefficient of determination $r^2 = (.889)^2 = .790$ or 79.0%

C. Coefficient of nondetermination $\bar{r}^2 = 1 - r^2 = 1 - .790 = .210$ or 21.0%

D. Interpret the answer to question IV B. **Seventy-nine percent of grade variability is accounted for by study hour variability.**

V. Could ρ (rho) be zero at the .01 level of significance?

1. The null hypothesis and alternate hypothesis are $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$.
2. The level of significance will be .01 for this two-tail problem with $n - 2$ degrees of freedom.
3. The relevant statistic will be r.
4. If t from the test statistic is beyond the critical value of t, the null hypothesis will be rejected.
5. Apply the decision rule.

$$df = n - 2 = 8 - 2 = 6 \rightarrow t = 3.707$$

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.889 - 0}{\sqrt{\frac{1 - (.889)^2}{8 - 2}}} = 4.76 \text{ Reject } H_0 \text{ because } 4.76 > 3.707. \text{ Rho could not be zero.}$$

Quick Questions 24 Simple Linear Regression Analysis

- I. Place the number of the appropriate formula, symbol, or expression next to the concept it describes.
- A. The standard error of the estimate 7
 B. The y-intercept 3
 C. The regression equation 1
 D. The estimated value of y given x 4
 E. The slope 5
 F. An interval estimate for the conditional mean of Y 2
 G. An interval estimate for an individual value of Y 6

- II. The following data was first presented in chapter 23. Estimate the regression line for this scatter using the eyeball method.

See page QQ 150

- III. Calculate the regression equation to 3 significant digits.

Data from page QQ 150

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$= \frac{69.6}{247} = .2817813$$

$$a = \bar{Y} - b\bar{X}$$

$$= \frac{\sum Y}{n} - b \frac{\sum X}{n}$$

$$= \frac{23.2}{8} - (.2817813)\left(\frac{31}{8}\right)$$

$$= 1.8080975$$

$$\hat{y}_{.x} = a + bx$$

$$\hat{y}_{.x} = 1.81 + .282x$$

- IV. Estimate the grade point average for people who studied 5 hours per weekend.

$$y_{.5} = 1.81 + .282x = 1.81 + .282(5) = 1.81 + 1.41 = 3.22$$

- V. Draw the regression line on the page 156 scatter diagram.

Easy points to determine are the y-intercept (0, 1.81) and question IV coordinates (5, 3.22).

- VI. Calculate the 98% confidence interval for students who study 5 hours per weekend.

$$S_{y,x} = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n-2}} = \sqrt{\frac{70.38 - 1.8080975(23.2) - (.2817813)(98.6)}{8-2}} = .329$$

$$df = 8 - 2 = 6$$

$$\alpha/2 = .02/2 = .01 \rightarrow 3.143 \text{ for } t$$

$$\bar{x} = \frac{\sum x}{n} = \frac{31}{8} = 3.875$$

- VII. What procedure should be followed if the range by your answer to question E includes negative numbers?

A negative number is not possible. If the range expresses the possibility of a negative number, the confidence interval may be lowered with a larger sample. Even if someone studied only 2 hours, the lower limit of the data, y is positive. Why? The standard error is low and $\hat{y}_{.2}$ could not be zero.

$$\hat{y}_{.x} \pm ts_{y,x} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

$$\hat{y}_{.5} = 3.22 \pm 3.143(.329) \sqrt{\frac{1}{8} + \frac{(5-3.875)^2}{151 - \frac{(31)^2}{8}}}$$

$$= 3.22 \pm 3.143(.329)(.407421)$$

$$= 3.22 \pm .421$$

$$2.80 \leftrightarrow 3.64$$

1.	$\hat{y}_{.x} = a + bx$
2.	$\hat{y}_{.x} \pm ts_{y,x} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$
3.	$\bar{Y} - b\bar{X}$
4.	$\hat{y}_{.x}$
5.	$\frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$
6.	$\hat{y}_{.x} \pm ts_{y,x} \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$
7.	$\sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n-2}}$

Descriptive Statistics Test Solutions

I. Place the number of the appropriate definition next to the item it describes.

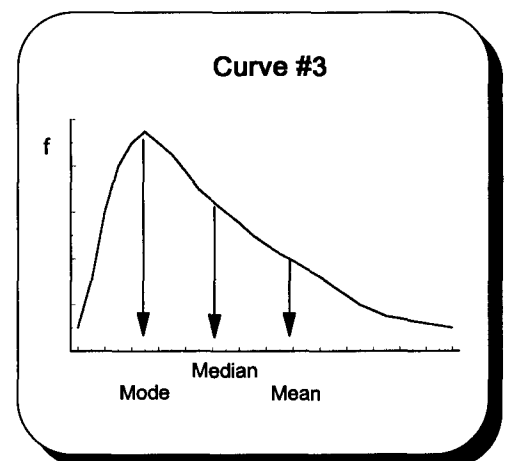
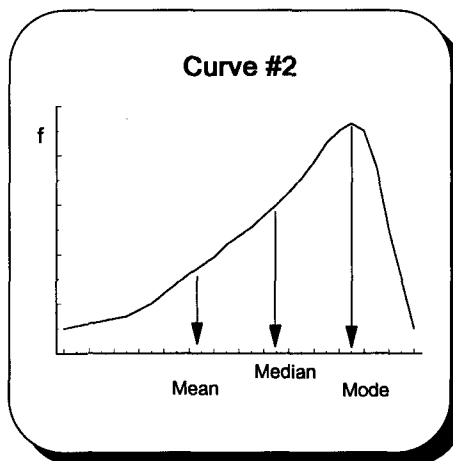
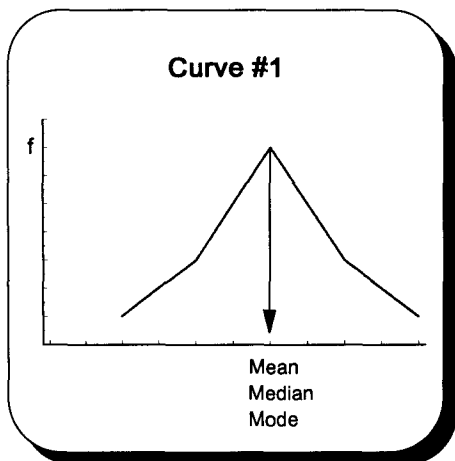
- | | |
|---|--|
| <p>A. Statistic <u> 4 </u></p> <p>B. Parameter <u> 9 </u></p> <p>C. All-inclusive <u> 1 </u></p> <p>D. Discrete <u> 5 </u></p> <p>E. Mutually exclusive <u> 2 </u></p> <p>F. Zero <u> 10 </u></p> <p>G. Continuous <u> 8 </u></p> <p>H. Inferential statistics <u> 3 </u></p> <p>I. Arithmetic mean <u> 7 </u></p> <p>J. Primary data <u> 6 </u></p> | <p>1. A place for every outcome</p> <p>2. Do not contain the same outcome</p> <p>3. The use of sample statistics to draw conclusions concerning the population</p> <p>4. A numerical characteristic of a sample</p> <p>5. Only finite values can exist on the x-axis</p> <p>6. Published by the original collector</p> <p>7. Severely affected by a few extreme values</p> <p>8. Measurement may assume any value associated with an uninterrupted scale</p> <p>9. A numerical characteristic of a population</p> <p>10. Sum of the deviations around a mean</p> |
|---|--|

II. Answer questions A - E using the information in this chart.

- A. The second class has real class limits of 24.5 and 39.5 .
- B. The first class has stated class limits of 10 and 24 .
- C. The class width is 15 .
- D. The midpoint of the third class is 47 .
- E. The range using real class limits is from 9.5 to 54.5 .

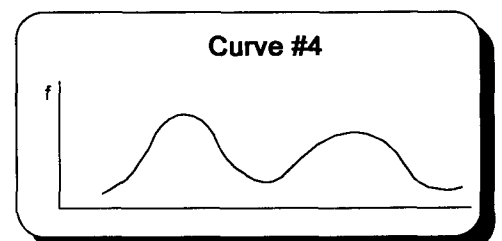
Stated Class Limits	x	Frequency (f)
10 - 24	17.0	2.0
25 - 39	32.0	3.0
40 - 54	47.0	5.0

III. Locate the approximate positions of the mean, median, and mode on these graphs.



IV. Answer questions A - E using Curves #1 to #4.

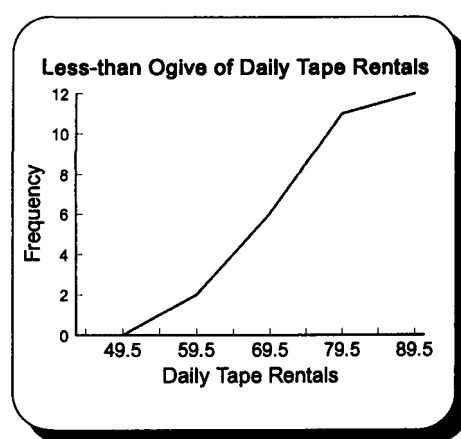
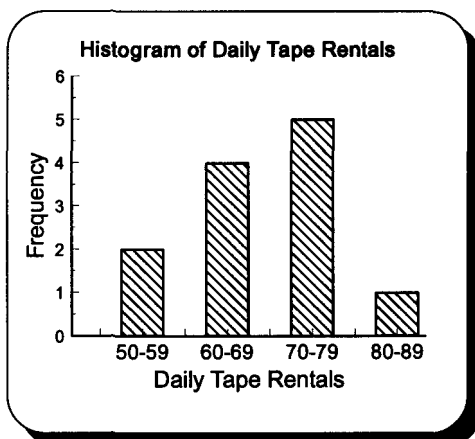
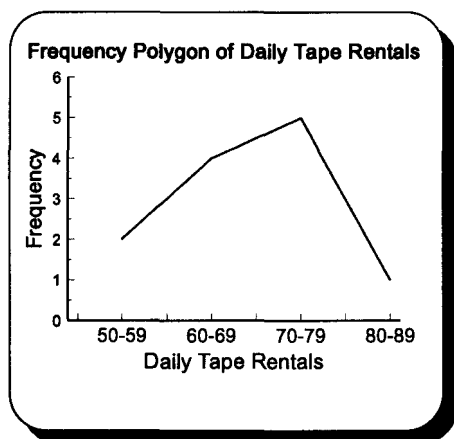
- A. Curve #1 is not skewed and is said to be symmetrical or normal .
- B. Curve #2 is skewed to the left .
- C. Curve #3 is skewed to the right .
- D. Curve #4 is bimodal .
- E. A curve with more than two peaks is multimodal .



- V. Using the following frequency distribution, construct and completely label a frequency polygon, histogram, and less-than ogive.

Linda's Video Showcase Daily Rental Figures	
Stated Class Limits	Frequency (f)
50 - 59	2.0
60 - 69	4.0
70 - 79	5.0
80 - 89	1.0

For People Using Statistics Software	
Data Set:	62, 66, 74, 58, 78, 71, 64, 84, 76, 53, 68, 75



- VI. Use this sample data when calculating the following statistics. Those not using statistics software may want to use the page 39 formulas.

Data: 4, 6, 3, 7, 6, 8, 17, 5 Array: 3, 4, 5, 6, 6, 7, 8, 17

A. Mean

$$\bar{x} = \frac{\sum x}{n} = \frac{56}{8} = 7$$

B. Median

$$\frac{n}{2} + .5 = \frac{8}{2} + .5 = 4.5 \rightarrow \frac{6+6}{2} = 6$$

C. Mode The number which happened most often is 6.

D. Variance

$$S^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{524 - \frac{(56)^2}{8}}{8-1} = \frac{132}{7} = 18.9$$

E. Standard deviation

$$s = \sqrt{s^2} = \sqrt{18.9} = 4.3$$

F. Use Chebyshev's rule to calculate the minimum proportion of items that will be within 3 standard deviations of the mean.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3)^2} = 1 - \frac{1}{9} = \frac{8}{9} \text{ or } 88.8\%$$

G. T F Chebyshev's rule only applies to normally distributed data. (true or false)

H. Calculate Pearson's coefficient of skewness.

$$\frac{3(\bar{x} - Md.)}{s} = \frac{3(7-6)}{4.3} = \frac{3}{4.3} = .698$$

For People Not Using Statistics Software	
x	x ²
3	9
4	16
5	25
6	36
6	36
7	49
8	64
<u>17</u>	<u>289</u>
56	524

VII. Label this chart. Calculate the following sample statistics being sure to state the symbol and formula for each measure. Formulas are given on page 39. This problem is only for people not using statistics software.

Stated Class Limits	Frequency (f)	x	fx	x ²	fx ²
40 - 49	1	44.5	44.5	1,980.25	1,980.25
50 - 59	2	54.5	109.0	2,970.25	5,940.50
60 - 69	3	64.5	193.5	4,160.25	12,480.75
70 - 79	5	74.5	372.5	5,550.25	27,751.25
80 - 89	5	84.5	422.5	7,140.25	35,701.25
90 - 99	2	94.5	189.0	8,930.25	17,860.50
Totals	18	417.0	1,331.0	30,731.50	101,714.50

A. Standard deviation

$$S = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}} = \sqrt{\frac{101,714.50 - \frac{(1,331)^2}{18}}{18-1}} = \sqrt{\frac{101,714.50 - 98,420.06}{17}} = \sqrt{193.8} = 13.9$$

B. Variance

$$S^2 = (S)^2 = \sqrt{193.8}^2 = 193.8$$

C. Median

$$\frac{n}{2} = \frac{18}{2} = 9$$

$$\begin{aligned} L + \frac{\frac{n}{2} - CF_b}{f}(i) \\ = 69.5 + \frac{\frac{18}{2} - 6}{5}(10) \\ = 69.5 + \frac{3}{5}(10) \\ = 75.5 \end{aligned}$$

D. 85th percentile

$$\frac{xn}{100} = \frac{85(18)}{100} = 15.3$$

$$\begin{aligned} P_x &= L + \frac{\frac{xn}{100} - CF_b}{f}(i) \\ P_{85} &= 79.5 + \frac{\frac{85(18)}{100} - 11}{5}(10) \\ &= 79.5 + \frac{4.3}{5}(10) \\ &= 88.1 \end{aligned}$$

VIII. Place the number of each formula next to the appropriate description of its function.

Ungrouped Measures

1. Population mean 7
2. Sample mean 1
3. Median 2
4. First quartile 13
5. Third quartile 8
6. x percentile 11
7. Interquartile range 9
8. Population average deviation 6
9. Population standard deviation 5
10. Sample standard deviation 14
11. Population variance 4
12. Sample variance 16
13. Chebyshev's rule 12
14. Coefficient of variation 10
15. Weighted mean 3
16. Pearson's coefficient of skewness 15

Ungrouped Formulas

1.	$\frac{\sum X}{n}$	2.	$\frac{n}{2} + .5$
3.	$\frac{\sum (W_x X_x)}{\sum w_x}$	4.	$\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2$
5.	$\sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2}$	6.	$\frac{\sum x-\mu }{N}$
7.	$\frac{\sum x}{N}$	8.	$\frac{3n}{4} + .5$
9.	$Q_3 - Q_1$	10.	$\frac{\sigma}{\mu}(100)$
11.	$\frac{xn}{100} + .5$	12.	$1 - \frac{1}{k^2}$
13.	$\frac{n}{4} + .5$	14.	$\sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$
15.	$\frac{3(\bar{x} - \text{md.})}{s}$	16.	$\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$

Grouped Measures

1. Approximate class width 4
2. Class midpoint 6
3. Population mean 3
4. Sample mean 9
5. Location of the median 1
6. Median 5
7. Range 10
8. Sample standard deviation 7
9. Sample variance 8
10. Relative frequency 2

Grouped Formulas

1.	$\frac{n}{2}$	2.	$\frac{\text{class frequency}}{\text{total frequencies}}$
3.	$\frac{\sum fx}{N}$	4.	$\frac{\text{range}}{\# \text{ of classes}}$
5.	$L + \frac{\frac{n}{2} - CF_b}{f}(i)$	6.	$\frac{X_1 + X_2}{2}$
7.	$\sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}}$	8.	$\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}$
9.	$\frac{\sum fx}{n}$	10.	$H - L$

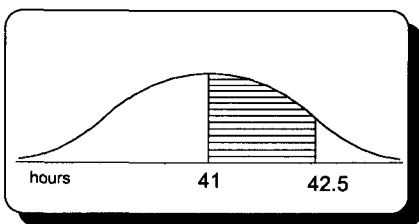
Probability Test Solutions

I. Average hours worked by manufacturing workers is normally distributed with a mean of 41 hours and a standard deviation of .5 hours. Graph and solve the following problems.

Given: $\mu = 41$ hours and $\sigma = .5$ hours

A. $P(41 \text{ hours} \leq x < 42.5 \text{ hours})$

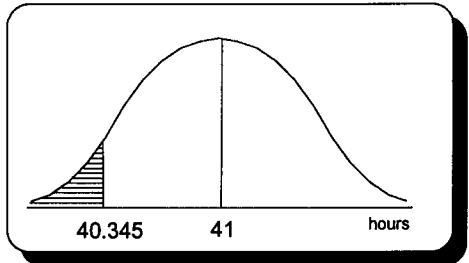
$$Z = \frac{x - \mu}{\sigma} = \frac{42.5 - 41.0}{.5} = \frac{1.5}{.5} = 3 \rightarrow .4987$$



B. $P(x < 40.345 \text{ hours})$

$$Z = \frac{x - \mu}{\sigma} = \frac{40.345 - 41.000}{.5} = \frac{-.655}{.5} = -1.31 \rightarrow .4049$$

$$\begin{array}{r} .5000 \\ - .4049 \\ \hline .0951 \rightarrow 9.51\% \end{array}$$

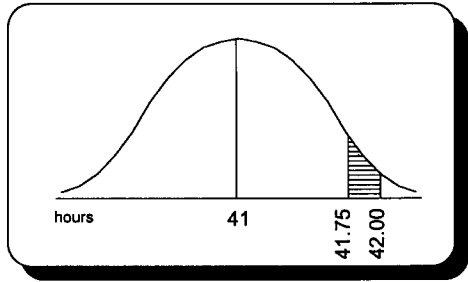


C. $P(41.75 \text{ hours} \leq x < 42 \text{ hours})$

$$Z = \frac{x - \mu}{\sigma} = \frac{42.00 - 41.00}{.5} = \frac{1}{.5} = 2.0 \rightarrow .4772$$

$$Z = \frac{x - \mu}{\sigma} = \frac{41.75 - 41.00}{.5} = \frac{.75}{.5} = 1.5 \rightarrow .4332$$

$$\begin{array}{r} .4772 \\ - .4332 \\ \hline .0440 \rightarrow 4.4\% \end{array}$$

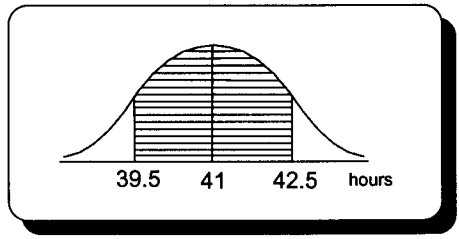


D. $P(39.5 \text{ hours} \leq x < 42.5 \text{ hours})$

$$Z = \frac{x - \mu}{\sigma} = \frac{39.5 - 41.00}{.5} = \frac{-1.5}{.5} = -3.0 \rightarrow .4987$$

$$Z = \frac{x - \mu}{\sigma} = \frac{42.5 - 41.00}{.5} = \frac{1.5}{.5} = 3.0 \rightarrow .4987$$

$$\begin{array}{r} .4987 \\ + .4987 \\ \hline .9974 \end{array}$$



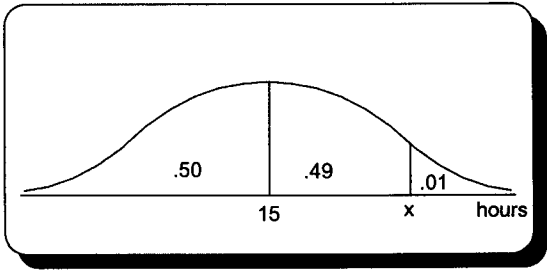
II. Study time at State University is normally distributed with a mean of 15 hours per week and a standard deviation of 3 hours. Graph and solve the following problems.

A. How many hours must a student study to be in the top 1% of the students attending State University?

Given:
 $\mu = 15$ hours
 $\sigma = 3$ hours

$$\begin{array}{l} \mu \pm z\sigma \\ 15 + 2.33(3) \\ 15 + 6.99 \\ 22 \text{ hours} \end{array}$$

$$\text{.50} - \text{.01} = \text{.49} \rightarrow z = 2.33$$



B. Calculate the fourth decile.

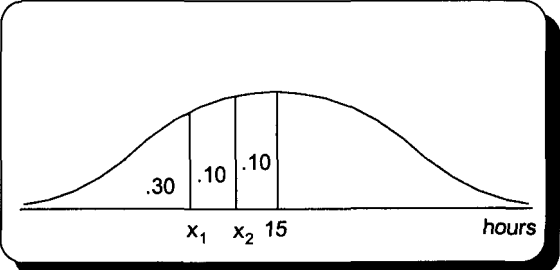
$.20 = \rightarrow z = .52$

$.10 = \rightarrow z = .25$

$\mu \pm z\sigma$
 $15 - .52(3)$
 $15 - 1.56$
 13.44

$\mu \pm z\sigma$
 $15 - .25(3)$
 $15 - .75$
 14.25

$13.44 \leftrightarrow 14.25$



III. Answer the following questions based upon this study of money spent on souvenirs at a virtual reality theme park.

	Money spent on souvenirs		Totals
Age	Under \$5	\$5 and over	
Under 22	5	15	20
22 and older	20	20	40
Totals	25	35	60

A. Use a formula to calculate the $P(\text{Age} < 22 \text{ or } \text{Age} \geq 22)$

$$P(< 22 \text{ or } \geq 22) = P(< 22) + P(\geq 22)$$

$$= P\left(\frac{20}{60}\right) + P\left(\frac{40}{60}\right) = \frac{60}{60} = 1.00 \rightarrow 100\%$$

B. The events in question A are mutually exclusive and therefore, the special rule for addition is applicable.

C. Use a formula to calculate the probability of someone being at least 22 years old and spending \$5 and over.

$$P(\geq 22 \text{ and } \geq \$5) = P(\geq 22) P(\geq \$5 | \geq 22) = \frac{40}{60} \times \frac{20}{40} = \frac{800}{2,400} = .333 = 33.3\%$$

D. Question C required the general rule for multiplication because the events are dependent.

E. Use Bayes' theorem to calculate the probability of someone at least 22 years old spending \$5 or more.

$$P(\geq \$5 | \geq 22) = \frac{P(\geq \$5 \text{ and } \geq 22)}{P(\geq 22)} = \frac{P(\geq \$5) \times P(\geq 22 | \geq \$5)}{P(\geq \$5) \times P(\geq 22 | \geq \$5) + P(< \$5) \times P(\geq 22 | < \$5)}$$

$$= \frac{\frac{35}{60} \times \frac{20}{35}}{\frac{35}{60} \times \frac{20}{35} + \frac{25}{60} \times \frac{20}{25}} = \frac{\frac{700}{2,100}}{\frac{700}{2,100} + \frac{500}{1,500}} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{3}} = \frac{\frac{1}{3}}{\frac{2}{3}} = .5 \rightarrow 50\%$$

F. Using the above chart, calculate the probability of someone at least 22 years old spending less than \$5.

$$P(< \$5 | \geq 22) = \frac{20}{40} = .5 \rightarrow 50\%$$

G. Why does your answer to question F make sense?

This answer makes sense because the answers to questions E and F are complements.

IV. Use a formula to calculate the probability of tossing a coin 3 times and getting exactly 3 heads. What is the probability of a head coming up on the fourth toss?

A. $P(\text{H and H and H}) = P(\text{H})P(\text{H})P(\text{H}) = .5 \times .5 \times .5 = .125$

B. $P(\text{H}) = 50\%$

V. Four customers have three branches and you will visit the manager and assistant manager at each branch. How many managers and assistant managers will you visit?

According to the counting rule: $MNO = 4 \times 3 \times 2 = 24$

- VI. A salesperson must visit 4 of 6 stores and order is important. That is, AB and BA represent different routes. How many routes are available to the salesperson?

$${}_N P_R = \frac{N!}{(N-R)!}$$

$${}_6 P_4 = \frac{6!}{(6-4)!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} = 6 \times 5 \times 4 \times 3 = 360$$

- VII. Redo problem VI assuming order does not count. AB and BA are the same and count as one route. Be sure to use a formula and show all work.

$${}_N C_R = \frac{N!}{(N-R)!(R!)}$$

$${}_6 C_4 = \frac{6!}{(6-4)!4!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 4 \times 3 \times 2 \times 1} = \frac{6 \times 5}{2 \times 1} = 15$$

- VIII. How many different 3-person subcommittees can be chosen from an 8-person committee?

$${}_N C_R = \frac{N!}{(N-R)!(R!)}$$

$${}_8 C_3 = \frac{8!}{(8-3)!3!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1 \times 3 \times 2 \times 1} = 8 \times 7 = 56$$

- IX. Three of 8 committee members must be chosen to give a speech. All 8 have very different personalities and order is important. How many different speaker arrangements are possible?

$${}_N P_R = \frac{N!}{(N-R)!}$$

$${}_8 P_3 = \frac{8!}{(8-3)!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1} = 8 \times 7 \times 6 = 336$$

- X. How many 4-place random numbers can be generated from 10 digits? Repeating digits is allowed.

This is a special adaptation of the counting rule because with each choice you have 10 digits to choose from.

$$\text{MNOP where all equal 10 is } 10 \times 10 \times 10 \times 10 = 10^4 = 10,000$$

- XI. Six parts are to be inspected from a production process designed to have approximately 5% defective parts. Using the binomial formula, determine the probability of zero defects. Use a table to determine the probability of at least 2 defective parts. State the entire probability distribution. What is the probability of 2 defective parts?

- A. The Poisson approximation of the binomial can not be used because n is not ≥ 30 .

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$P(0) = \frac{6!}{0!(6-0)!} .05^0 .95^{6-0} = \frac{1}{1}(1).95^6 = .7351 = 73.51\%$$

- B.

$$P(x \geq 2) = 1 - [P(0) + P(1)] = 1 - (.735 + .232) = 1 - .967 = .033 = 3.3\%$$

- C. See page ST 1 for the entire distribution.

$$D. P(x=2) = .031$$

- XII. Approximately 4% of the estimated 100 travelers driving on route 128 will stop for a snack between 11:10 PM and 11:20 PM. Is the Poisson approximation to the binomial distribution appropriate for the solution of this problem? Use a table to determine the probability that less than 2 travelers will stop for a snack. Draw a graph of this distribution.

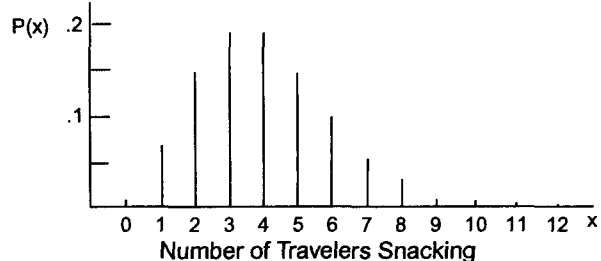
A Poisson approximation of the binomial is appropriate.

$$n \geq 30 \text{ as } n = 100$$

$$np < 5 \text{ as } np = .04 \times 100 = 4$$

$$\mu = np = (100)(.04) = 4$$

$$P(< 2) = P(0) + P(1) = .0183 + .0733 = .0916 = 9.16\%$$



XIII. Place each number next to the appropriate item.

- A. Standard error of the mean 2
- B. 99% confidence interval 1
- C. Standard error of the proportion 5
- D. Requires n be ≥ 30 3
- E. Acceptable error 4

1.	$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$
2.	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
3.	$\bar{x} \pm 2.58 \frac{s}{\sqrt{n}}$
4.	E
5.	$\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

XIV. Answer the following true or false and fill in the blank questions.

- A. The standard error of the mean will be halved if the sample size is doubled. F
- B. Sampling error exists because a nonrepresentative sample was taken in place of a census. T
- C. A one-number estimate of the population mean is called a point estimate of the mean.
- D. A range for a population parameter is called the confidence interval.
- E. A stratified random sample may be more accurate than a simple random sample because a small diverse section of the population might not be represented in a simple random sample.

XV. A sample of 36 out of 25,000 baseball fans attending a game revealed average refreshment spending of \$7.60. The standard deviation for the population is \$2.10. Calculate the 95% confidence interval for average refreshment spending by fans attending this game.

Data set for those using statistics software			
Refreshment Spending			
4.50	8.00	9.00	9.00
6.95	4.90	7.00	8.05
10.00	8.00	9.50	2.00
11.00	9.00	5.00	8.00
8.05	8.50	10.00	4.80
6.00	4.90	11.00	9.00
6.50	7.00	7.00	8.00
11.00	8.00	5.00	5.75
9.10	6.00	9.10	9.00

Given:
n = 36
N = 25,000
$\bar{x} = \$7.60$
$\sigma = \$2.10$

$\frac{n}{N} = \frac{36}{25,000} = .001 < .05$
The finite correction factor is not required.

$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$
 $\$7.60 \pm 1.96 \frac{\$2.10}{\sqrt{36}}$
 $\$7.60 \pm \$.686$
 $\$6.91 \leftrightarrow \8.29

XVI. A marketing test of chocolate flavored shaving cream revealed a favorable response from 35 of 50 test subjects. Test subjects were chosen at random from the company's 1,200 employees. Calculate the following:

- A. The 90% confidence interval for this market test.

Data set for those using statistics software				
Favorable (F) and Unfavorable (U) Attitudes Toward Chocolate Flavored Shaving Cream				
U	F	F	F	F
F	U	F	F	U
U	F	U	F	F
U	F	F	F	U
F	U	F	F	F
U	F	F	U	F
F	F	F	F	F
U	F	F	U	U
F	F	F	F	F
F	F	F	U	U

$\frac{n}{N} = \frac{50}{1,200} = .042 < .05$

$\bar{p} = \frac{35}{50} = .70$

$.7 \pm 1.645 \sqrt{\frac{.7(1-.7)}{50}}$
 $.7 \pm .107$
 $.593 \leftrightarrow .807$

$n = 50 \geq 30$
 $np = 50(.70) = 35 \geq 5$
 $nq = 50(.30) = 15 \geq 5$

$\bar{p} \pm z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

- B. The company is unhappy with the confidence interval calculated above and would like to lower acceptable error from 11% to 5%. How large a sample must be taken?

$n = \bar{p}(1-\bar{p}) \left(\frac{z}{E}\right)^2$

$n = .7(1-.7) \left(\frac{1.64}{.05}\right)^2$
 $= .21(1,075.84)$
 $= 225.9264 \rightarrow 226$

XVII. Match each item on the right with the concept it defines.

1. Bayes' theorem 20
2. Addition rule when events are mutually exclusive 1
3. Variance of a binomial probability distribution 19
4. Factorial rule for arranging all of the items of one event 16
5. Range for probability 5
6. Multiplication rule when the events are independent 15 or 7
7. Empirical probability 10
8. Subjective probability 23
9. General rule for addition 14
10. Permutation rule 2
11. To find a range given the probability 8
12. Classical probability 13
13. Mean of a probability distribution 25
14. Value of a complement 12
15. For independent events 7 or 15
16. Binomial distribution 18
17. To find the probability given a range 6
18. Combination rule 9
19. Poisson distribution 4
20. The complement of A 11
21. Variance of a probability distribution 17
22. The counting rule for multiple events 3
23. Is calculated for each value of x when determining a probability distribution 21
24. Mean of a binomial probability distribution 24
25. General rule for multiplication 22

1.	$P(A) + P(B)$
2.	$\frac{N!}{(N-R)!}$
3.	$M \times N$
4.	$\frac{\mu^x e^{-\mu}}{x!}$
5.	$0 \leq P(A) \leq 1$
6.	$\frac{x-\mu}{\sigma}$
7.	Joint probability is the product of the marginal probabilities
8.	$\mu \pm Z\sigma$
9.	$\frac{N!}{(N-R)!(R!)}$
10.	$\frac{A}{n}$
11.	\bar{A}
12.	$1 - P(A)$
13.	$\frac{A}{N}$
14.	$P(A) + P(B) - P(A \text{ and } B)$
15.	$P(A) \times P(B)$
16.	$N!$ ways
17.	$[\sum x^2 \cdot P(x)] - [E(x)]^2$
18.	$\frac{n!}{x!(n-x)!} p^x q^{n-x}$
19.	npq
20.	$\frac{P(A) \times P(B A)}{P(A) \times P(B A) + P(\bar{A}) \times P(B \bar{A})}$
21.	$x \cdot P(x)$
22.	$P(A) \times P(B A)$
23.	Use empirical formula assuming past data of similar events is appropriate
24.	np
25.	$\sum [x \cdot P(x)]$

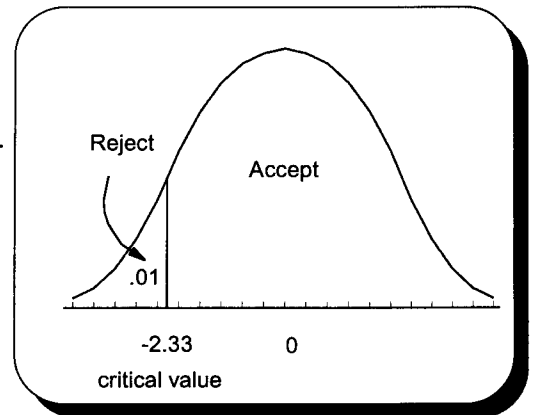
Inferential Statistics Test Solutions

- I. A sample of 36 out of 25,000 baseball fans attending a game revealed average refreshment spending of \$7.60. The population standard deviation was \$2.10. The makers of Dud beer will not distribute their product to a ballpark unless it is possible that the average fan spends at least \$8.00 on refreshments. Use the 5-step approach to hypothesis testing and a .01 level of significance to test whether this ballpark qualifies to receive Dud beer.

1. $H_0: \mu \geq \$8.00$ and $H_1: \mu < \$8.00$
2. Type I error of .01 $\rightarrow Z = \pm 2.33$
3. \bar{x} is the test statistic.
4. If z from the test statistic is beyond the critical value of z, reject H_0 .
5. Apply the decision rule.

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\$7.60 - \$8.00}{\frac{\$2.10}{\sqrt{36}}} = \frac{-.40}{.35} = -1.14$$

H_0 is accepted because -1.14 is not beyond -2.33.
The mean could be $\geq \$8.00$. Have a Dud beer.



- II. A marketing test of chocolate flavored shaving cream revealed a favorable response from 35 of 50 test subjects. Test subjects were chosen at random from the company's 1,200 employees. This product will be manufactured if at least 80% of the potential market like the product.

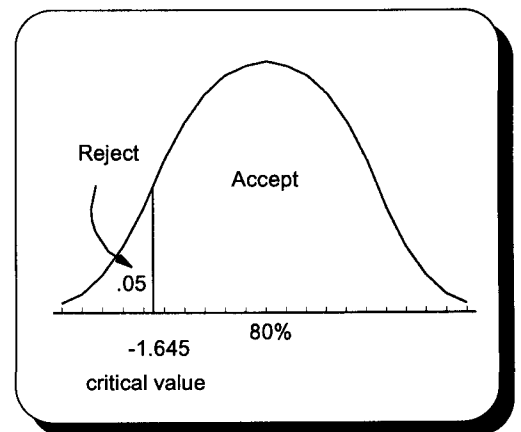
- A. Using the 5-step approach to hypothesis testing and a .05 level of significance, determine whether the product will be manufactured.

1. $H_0: p \geq 80\%$ and $H_1: p < 80\%$
2. Type I error is .05.
3. \bar{p} is the test statistic.
4. If z from the test statistic is beyond the critical value of z, reject H_0 .
5. Apply the decision rule.

$$\begin{aligned} n &= 50 \geq 30 \\ np &= 50(.80) = 40 \geq 5 \\ nq &= 50(.20) = 10 \geq 5 \end{aligned}$$

$$\begin{aligned} Z &= \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \\ &= \frac{.7 - .8}{\sqrt{\frac{.8(1-.8)}{50}}} \\ &= -1.77 \end{aligned}$$

Reject H_0 because -1.77 is beyond -1.645.
The mean could not be 80% at the .05 level of significance. Too bad, chocolate flavored shaving cream will not be produced.



- B. What are the pros and cons of using company employees to test this product?

Using company employees is convenient. Company employees have a vested interest in giving the survey adequate attention. On the other hand, some employees might be prejudice for or against the company.

- III. ABC Company is questioning whether the quality of material coming from the company's three suppliers has something to do with the number of defective products. The number of defects from 20 production runs for each supplier were counted. Using a .05 level of significance, determine whether the number of defects and the company supplying materials are related (dependent).

Analysis of Material Suppliers and Defects								
	Company #1		Company #2		Company #3		Totals	
	f_o	f_e	f_o	f_e	f_o	f_e	f_o	f_e
High defects	6	10	9	10	15	10	30	30
Low defects	14	10	11	10	5	10	30	30
Totals	20	20	20	20	20	20	60	60

H_0 : defects and supplier are independent

H_1 : defects and supplier are dependent

$$f_e = \frac{f_r \times f_c}{n}$$

$$f_e = \frac{30 \times 20}{60}$$

$$f_e = \frac{1}{2}(20) = 10$$

$$df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2 \rightarrow \chi^2 = 5.99$$

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

$$= \sum \left[\frac{(14 - 10)^2}{10} + \frac{(6 - 10)^2}{10} + \frac{(11 - 10)^2}{10} + \frac{(9 - 10)^2}{10} + \frac{(5 - 10)^2}{10} + \frac{(15 - 10)^2}{10} \right]$$

$$= \sum (1.6 + 1.6 + .1 + .1 + 2.5 + 2.5) = 8.4$$

Reject H_0 because $8.4 > 5.99$.

Material supplier and defects are dependent.

- IV. Four people were given extensive sales training. Test whether their sales performance improved using a .05 level of significance. Assume normally distributed populations with unknown standard deviations.

Analysis of Sales Training Effectiveness				
Salesperson	Sales Performance		d	d^2
	Before	After		
A	12	15	-3	9
B	13	17	-4	16
C	10	14	-4	16
D	11	12	-1	1
Total			-12	42

These are the null hypothesis and research hypothesis.

$$H_0 : \mu_d \geq 0 \text{ and } H_1 : \mu_d < 0$$

Note: An increase in performance results in a negative difference.

$$df = n - 1 = 4 - 1 = 3 \rightarrow t = -2.353$$

$$\bar{d} = \frac{\sum d}{n}$$

$$= \frac{-12}{4}$$

$$= -3$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{\sum 42 - \frac{(-12)^2}{4}}{4-1}}$$

$$= \sqrt{\frac{42 - 36}{3}}$$

$$= 1.414$$

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

$$= \frac{-3}{\frac{1.414}{\sqrt{4}}}$$

$$= -4.24$$

Reject H_0 because -4.24 is beyond -2.353 .

Training improved performance.

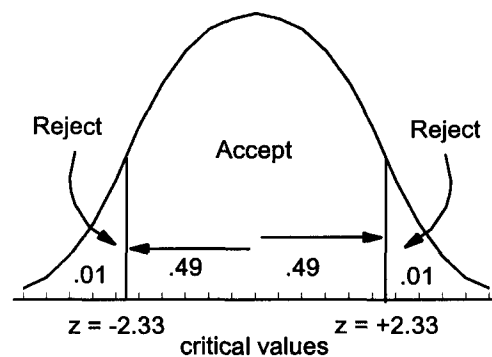
- V. Owners of the Quick Chow Restaurant are concerned about the average time to serve customers at two of their stores. A sample of 32 customers at store A resulted in a mean service time of 80 seconds and a standard deviation of 8 seconds. A sample of 49 customers at store B resulted in a mean service time of 75 seconds and a standard deviation of 7 seconds. Test at the .02 level of significance whether the mean time to wait on customers at these two stores is the same.

Given:	$n_1 = 32$	$\bar{X}_1 = 80$ seconds	$S_1 = 8$	$\alpha = .02$
	$n_2 = 49$	$\bar{X}_2 = 75$ seconds	$S_2 = 7$	

1. $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$
2. Type I error is .02 and $\alpha/.02 = .02/2 = .01 \rightarrow \pm 2.33$
3. \bar{X} is the test statistic.
4. If the z from the test statistic is beyond the critical value of z, reject H_0 .
5. Apply the decision rule.

$$\begin{aligned}
 Z &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \\
 &= \frac{80 - 75}{\sqrt{\frac{(8)^2}{32} + \frac{(7)^2}{49}}} \\
 &= \frac{5}{\sqrt{2 + 1}} \\
 &= 2.89
 \end{aligned}$$

Reject H_0 because 2.89 is beyond 2.33.
Mean time to wait on customers differs at these two stores.



- VI. Before recent improvements, it took 36.4 minutes to assemble a part. After improvements, a sample of 16 had an average assembly time of 34 minutes. The sample standard deviation was 2.4 minutes. Test at the .01 level of significance whether improvements lowered assembly time.

Given:	$\mu = 36.4$	$n = 16$	$\bar{X} = 34$	$S = 2.4$	$\alpha = .01$
---------------	--------------	----------	----------------	-----------	----------------

1. $H_0: \mu \geq 36.4$ and $H_1: \mu < 36.4$
2. Type I error is .01.
3. \bar{X} is the test statistic.
4. If t for the test statistic is beyond the critical value of t, reject H_0 .
 $df = n - 1 = 16 - 1 = 15 \rightarrow t = \pm 2.602$ for the .01 level of significance
5. Apply the decision rule.

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

$$t = \frac{34.0 - 36.4}{\frac{2.4}{\sqrt{16}}} = \frac{-2.4}{.6} = -4$$

Reject H_0 because -4 is beyond -2.602.
Assembly time went down.

VII. Samples of 10 taken in 1985 and 1995 revealed the average time people spend grocery shopping decreased from 18 minutes to 14 minutes. Respective standard deviations were 5 minutes and 4 minutes. Test at the .10 level of significance whether there has been a change in shopping time variability.

Given:
n_1 and $n_2 = 10$
$\bar{x}_1 = 18$ minutes
$\bar{x}_2 = 14$ minutes
$S_1 = 5$ minutes
$S_2 = 4$ minutes
.10 level of significance

- | | |
|---|--|
| <p>1. $H_0 : \sigma_1^2 = \sigma_2^2$ and $H_1 : \sigma_1^2 \neq \sigma_2^2$</p> <p>2. The level of significance is .10.</p> <p>3. F is the test statistic.
 $df = n_1 - 1 = 10 - 1 = 9$
 $df = n_2 - 1 = 10 - 1 = 9$
 $\alpha + 2 = .10 + 2 = .05 \rightarrow F = 3.18$</p> <p>4. If F for the test statistic is beyond the critical value of F, reject H_0.</p> | <p>5.</p> $F = \frac{s_1^2}{s_2^2}$ $= \frac{5^2}{4^2}$ $= 1.56$ <p>Accept H_0 because $1.56 < 3.18$.
Shopping time is not more variable.</p> |
|---|--|

VIII. Test at the .05 level of significance whether workplace accidents happen equally throughout the workweek.

Day	Accidents f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
Monday	9	7	2	4	4/7 = 0.571
Tuesday	5	7	-2	4	4/7 = 0.571
Wednesday	6	7	-1	1	1/7 = 0.143
Thursday	5	7	-2	4	4/7 = 0.571
Friday	<u>10</u>	<u>7</u>	<u>3</u>	9	9/7 = <u>1.286</u>
Totals	35	35	0		3.142

H_0 : accidents are equally distributed
 H_1 : accidents are not equally distributed

$df = k - 1 = 5 - 1 = 4$
 $\alpha = .05 \rightarrow \chi^2 = 9.49$

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = 3.142$$

Accept H_0 because $3.14 < 9.49$.
Accidents happen equally throughout the workweek.

IX. Three computer component assembly methods were compared by Insel Corporation. Employee efficiency was based upon production time and product quality.

A. Use ANOVA analysis to test at the .05 level of significance whether mean employee efficiency of these assembly methods are equal.

ANOVA Analysis of Assembly Methods							
Employee Efficiency Ratings for 3 Treatments (T)						Row Totals Required for Calculations	
	Method 1		Method 2		Method 3		
	Score X_1	X_1^2	Score X_2	X_2^2	Score X_3	X_3^2	
	4	16	6	36	8	64	
	6	36	7	49	8	64	
	7	49	4	16	9	81	
	<u>7</u>	<u>49</u>	<u>7</u>	<u>49</u>	<u>9</u>	<u>81</u>	
$\sum X_T$	24		24		34		$\sum x = 82$
$(\sum X_T)^2$	576		576		1156		
n	4		4		4		$N = 12$
$\frac{(\sum X_T)^2}{n}$	144		144		289		$\sum [\frac{(\sum X_T)^2}{n}] = 577$
$\sum X_T^2$		150		150		290	$\sum x^2 = 590$

- $H_0 : \mu_1 = \mu_2 = \mu_3$ $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$
- F is the test statistic and $\alpha = .05$.
- If F from the test statistic is beyond the critical value of F, the null hypothesis will be rejected.
- $df = t - 1 = 3 - 1 = 2$
 $df = N - t = 12 - 3 = 9$
f for .05 level of significance is 4.26.
- Apply the decision rule.

$$F = \frac{MS_T}{MS_E} = \frac{8.335}{1.44} = 5.79$$

Reject H_0 because $5.79 > 4.26$.
Training methods had different means.

$$SS_T = \sum \left[\frac{(\sum x_T)^2}{n} \right] - \frac{(\sum X)^2}{N}$$

$$= 577 - \frac{82^2}{12}$$

$$= 16.67$$

$$MS_T = \frac{SS_T}{t-1} = \frac{16.67}{3-1} = 8.335$$

$$SS_E = \sum x^2 - \sum \left[\frac{(\sum x_T)^2}{n} \right]$$

$$= 590 - 577$$

$$= 13.00$$

$$MS_E = \frac{SS_E}{N-t} = \frac{13}{12-3} = 1.44$$

$$SS_{TOTAL} = \sum x^2 - \frac{(\sum x)^2}{N} = 590 - 560.33 = 29.67$$

B. Determine at the .01 level of significance whether there is a difference in performance of those who received teaching methods (treatments) 1 and 3.

$$\bar{X}_1 = \frac{\sum x}{n_1} = \frac{24}{4} = 6.0$$

$$\bar{X}_3 = \frac{\sum x}{n_3} = \frac{34}{4} = 8.5$$

The t for $\alpha/2$ and $N - t$ degrees of freedom ($12 - 3 = 9$) is 3.25.

$$(\bar{X}_1 - \bar{X}_3) \pm t \sqrt{MS_E \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$(8.5 - 6.0) \pm 3.25 \sqrt{1.44 \left(\frac{1}{4} + \frac{1}{4} \right)}$$

$$2.5 \pm 2.758$$

$$-.258 \leftrightarrow 5.258$$

This range indicates the difference between these means could be zero.

- X. Darin wants to compare assembly time of 30-milligram parts using method A and method B. It is not known whether these populations are approximately normal with the same variance. Use the Mann-Whitney test to determine at the .05 level of significance whether these samples come from populations with equal medians.

Time to Assemble 30-Milligram Parts in Seconds											
Method A	90	95	104	88	91	94	87	102	96	98	101
Method B	95	102	93	105	96	99	100	103	91	97	106

Rank Ordered Array and Assembly Method	Ranked Scores		Rank Ordered Array and Assembly Method	Ranked Scores		Rank Ordered Array and Assembly Method	Ranked Scores	
	Method A	Method B		Method A	Method B		Method A	Method B
1. 87 A	1		8. 95 A	8.5		15. 100 B		15
2. 88 A	2		9. 95 B		8.5	16. 101 A	16	
3. 90 A	3		10. 96 A	10.5		17. 102 A	17.5	
4. 91 A	4.5		11. 96 B		10.5	18. 102 B		17.5
5. 91 B		4.5	12. 97 B		12	19. 103 B		19
6. 93 B		6	13. 98 A	13		20. 104 A	20	
7. 94 A	7		14. 99 B		14	21. 105 B		21
						22. 106 B		22
Subtotals	17.5	10.5		32.0	45.0		53.5	94.5

$$R_1 = 17.5 + 32.0 + 53.5 = 103$$

$$\begin{aligned}
 U_1 &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \\
 &= 11(11) + \frac{11(11+1)}{2} - 103 \\
 &= 121 + 66 - 103 \\
 &= 84
 \end{aligned}$$

$$\begin{aligned}
 \mu_U &= \frac{n_1 n_2}{2} \\
 &= \frac{11(11)}{2} \\
 &= 60.5
 \end{aligned}$$

$$\begin{aligned}
 \sigma_U &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \\
 &= \sqrt{\frac{11(11)(11+11+1)}{12}} \\
 &= \sqrt{\frac{2,783}{12}} \\
 &= 15.23
 \end{aligned}$$

$$\begin{aligned}
 Z &= \frac{U - \mu_U}{\sigma_U} \\
 &= \frac{84.0 - 60.5}{15.23} \\
 &= 1.543
 \end{aligned}$$

This two-tail problem has a z of ± 1.96 for the .05 level of significance. H_0 is accepted because $1.54 < 1.96$. Median assembly times are equal.

- XI. A third assembly method has recently been proposed for the 30-milligram parts examined in problem 10. Use a .01 level Kruskal-Wallis test to determine whether these samples come from populations with equal medians.

Time to Assemble 30-Milligram Parts in Seconds											
Method C	86	99	84	85	92	93	82	81	96	83	94

Assembly Time for 30-Milligram Parts					
Method A		Method B		Method C	
Time	Rank	Time	Rank	Time	Rank
90	9	95	17.5	86	6
95	17.5	102	28.5	99	24.5
104	31	93	13.5	84	4
88	8	105	32	85	5
91	10.5	96	20	92	12
94	15.5	99	24.5	93	13.5
87	7	100	26	82	2
102	28.5	103	30	81	1
96	20	91	10.5	96	20
98	23	97	22	83	3
101	<u>27</u>	106	<u>33</u>	94	<u>15.5</u>
	197.0		257.5		106.5

H is the designated statistic.
N, the number of observations, is 33.
k, the number of samples, is 3.
n_k , a sample's size, is 11.
R_k is a sample's rank total.
$df = k - 1 = 3 - 1 = 2 \rightarrow \chi^2 = 9.21$

$$\begin{aligned}
 H &= \frac{12}{N(N+1)} \left[\frac{(\sum R_1)^2}{n_1} + \frac{(\sum R_2)^2}{n_2} + \dots + \frac{(\sum R_k)^2}{n_k} \right] - 3(N+1) \\
 &= \frac{12}{33(33+1)} \left[\frac{(197)^2}{11} + \frac{(257.5)^2}{11} + \frac{(106.5)^2}{11} \right] - 3(33+1) \\
 &= .0106951[3,528.091 + 6,027.841 + 1,031.114] - 102 \\
 &= 11.23
 \end{aligned}$$

H of 11.23 is greater than 9.21, the value of χ^2 for the .01 level of significance.

Reject H_0 , assembly time medians are not equal.

XII. Oven temperature at Chewy Pizza restaurants was in control when these samples were taken. Construct an \bar{x} chart and an R chart for this data using a 99.74% confidence interval.

Sample #	1	2	3	4	5	6	Totals
Oven Readings n = 3 N = 6 samples	405	402	398	410	391	411	
	404	404	390	402	409	409	
	397	412	388	412	400	407	
Sample Mean	402	406	392	408	400	409	2,417
Sample Range	8	10	10	10	18	4	60

Sample Size (n)	A ₂	D ₃	D ₄
2	1.880	0	3.267
3	1.023	0	2.575
4	0.729	0	2.282
5	0.577	0	2.115

$$\bar{\bar{x}} = \frac{\sum \bar{x}}{N} = \frac{2,417}{6} = 402.83 \approx 402.8$$

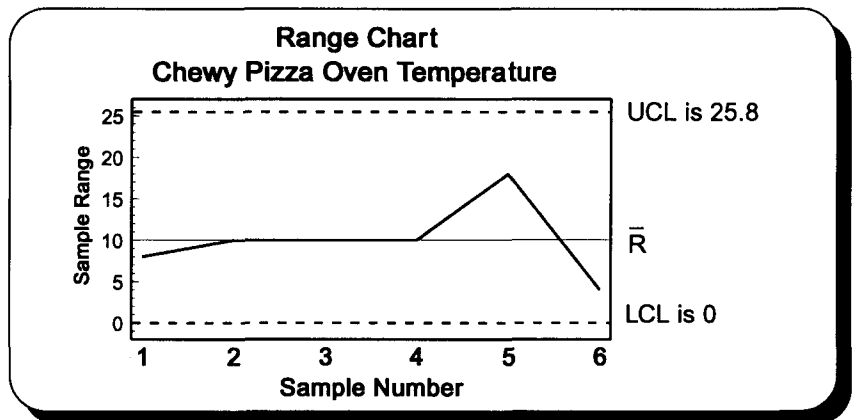
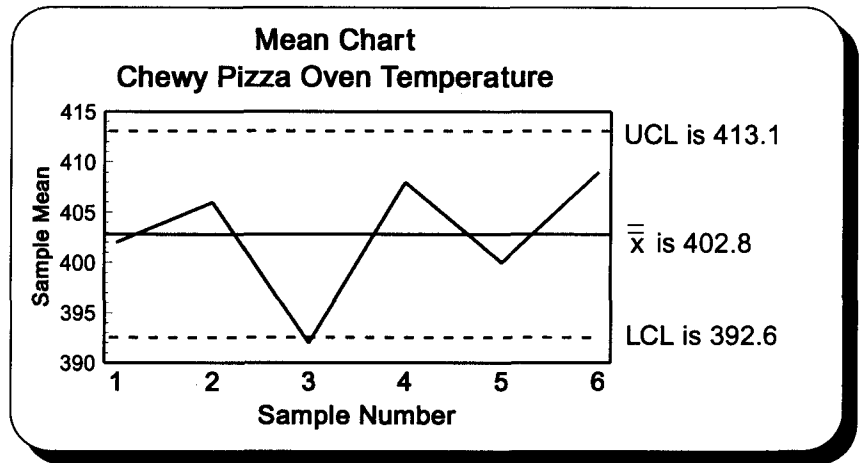
$$\bar{R} = \frac{\sum R}{N} = \frac{60}{6} = 10.0$$

$$\begin{aligned} \text{UCL} &= \bar{\bar{x}} + A_2 \bar{R} \\ &= 402.83 + 1.023(10) \\ &= 402.83 + 10.23 \\ &= 413.06 \\ &\approx 413.1 \end{aligned}$$

$$\begin{aligned} \text{LCL} &= \bar{\bar{x}} - A_2 \bar{R} \\ &= 402.83 - 1.023(10) \\ &= 402.83 - 10.23 \\ &= 392.6 \end{aligned}$$

$$\begin{aligned} \text{UCL} &= D_4 \bar{R} \\ &= 2.575(10) \\ &= 25.75 \\ &\approx 25.8 \end{aligned}$$

$$\begin{aligned} \text{LCL} &= D_3 \bar{R} \\ &= 0(10) \\ &= 0 \end{aligned}$$



XIII. Potential customers were asked to rate brand A and brand B. Little is known about the population distributions. Test at the .10 level of significance whether these brands were viewed equally by these potential customers. A paired difference sign test may be conducted even though this is not a test for statistical dependency.

Brand B was liked better by 5 of 6 customers. Sample size was n = 6. The Binomial table (ST 1) yields the following: $p(x = \geq 5) = .094 + .016 = .11$ and $(.11)(2) = .22 > .10$. The null hypothesis is accepted at the .10 level of significance. Customers rate the brands equally.

Potential Customer	Brand A	Brand B	Sign
1	87	89	+
2	91	97	+
3	81	85	+
4	73	81	+
5	92	98	+
6	89	81	-

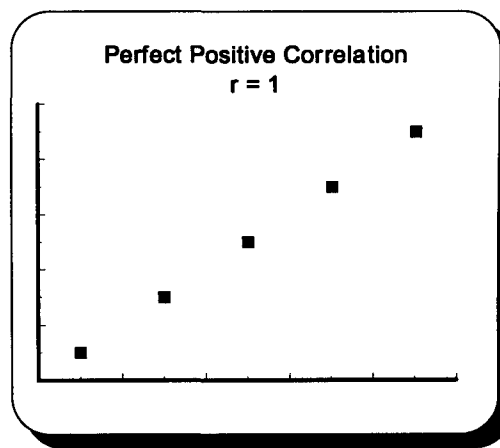
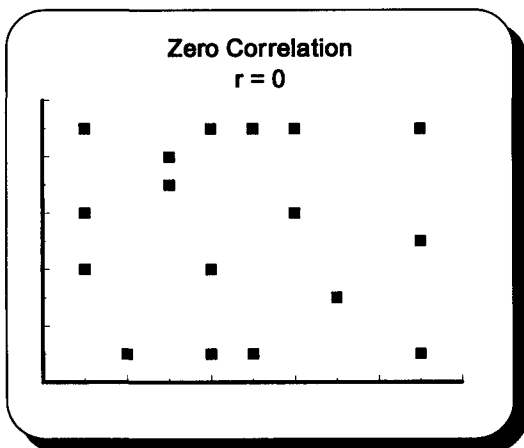
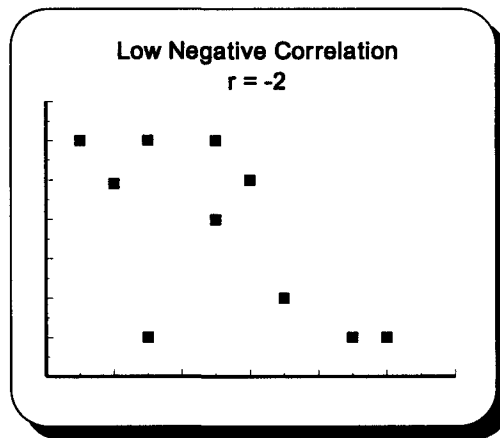
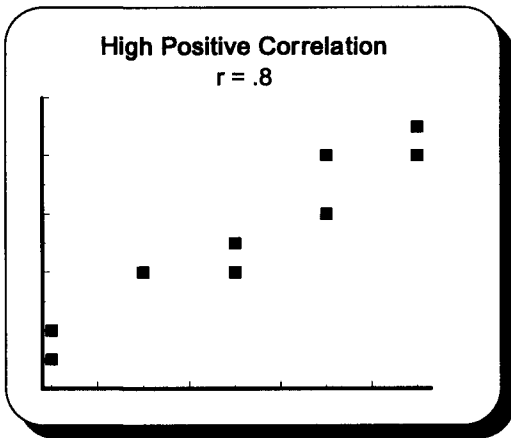
Correlation and Regression Test Solutions

I. Place the number of the appropriate formula, expression, or term next to the appropriate concept.

- A. The independent variable 4
- B. The dependent variable 7
- C. Measures the strength in the relationship between two variables 1
- D. The variation of the dependent variable explained by the independent variable 9
- E. The variation of the dependent variable not explained by the independent variable 3
- F. Used when testing the significance of r 5
- G. The regression equation 10
- H. The slope of the regression line 2
- I. Where a regression line crosses the y-axis 8
- J. The standard error of the estimate 6

1.	r
2.	b
3.	$1 - r^2$
4.	x
5.	$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$
6.	$\sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n - 2}}$
7.	y
8.	a
9.	r^2
10.	$\hat{y}_{.x} = a + bx$

II. Draw the following scatters and place an appropriate value for r in the space provided.



III. Answer the following questions using this data that was collected to determine whether research and development expenditures affect profit.

A. The coefficient of correlation

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$$

$$= \frac{6(2,010) - (35)(310)}{\sqrt{[6(235) - (35)^2][6(17,700) - (310)^2]}}$$

$$= \frac{(12,060) - (10,850)}{\sqrt{[(1,410) - (1,225)][(106,200) - (96,100)]}}$$

$$= \frac{1,210}{\sqrt{[185][10,100]}} = \frac{1,210}{1,367} = .885$$

R & D Expenditures Millions (x)	Profits in Millions (y)	xy	x ²	y ²
5	30	150	25	900
3	40	120	9	1,600
7	60	420	49	3,600
6	60	360	36	3,600
10	80	800	100	6,400
<u>4</u>	<u>40</u>	<u>160</u>	<u>16</u>	<u>1,600</u>
35	310	2,010	235	17,700

B. The coefficient of determination and the coefficient of nondetermination

$$r^2 = (.885)^2 = .783 \text{ or } 78.3\%$$

$$\bar{r}^2 = 1 - r^2 = 1 - .783 = .217 \text{ or } 21.7\%$$

C. Could rho be zero at the .05 level of significance?

- The null hypothesis and alternate hypothesis are $H_0: \rho = 0$ and $H_1: \rho \neq 0$.
- The level of significance will be .05 for this two-tail problem with $n - 2$ degrees of freedom.
- The test statistic is r .
 $df = n - 2 = 6 - 2 = 4 \rightarrow t$ of 2.776
- If t from the test statistic is beyond the critical value of t , the null hypothesis will be rejected.
- Apply the decision rule.

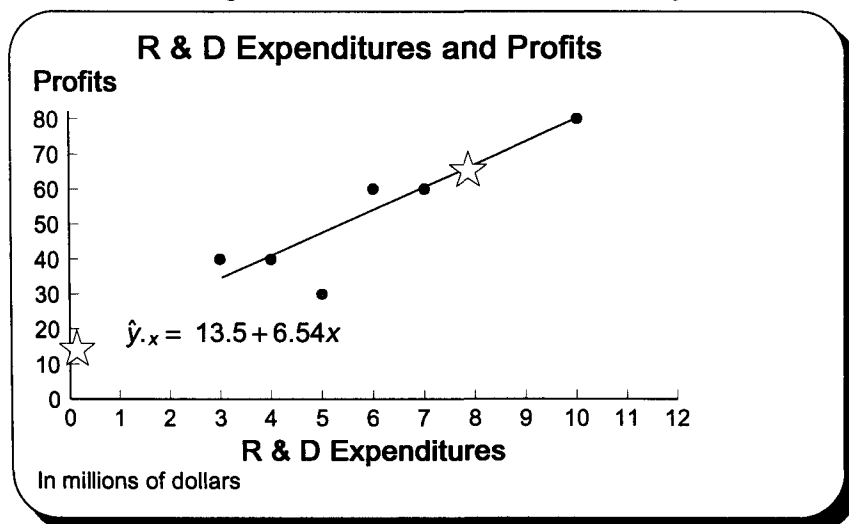
$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.885 - 0}{\sqrt{\frac{1 - (.885)^2}{6 - 2}}} = 3.80$$

Reject H_0 because $3.80 > 2.776$. The population coefficient of correlation could not be zero at the .05 significance level.

IV. Interpret your answers to question III.

- A. An r of .885 represents a high positive correlation. B. Profit variability not explained by R & D is 21.7%.
 B. Profit variability explained by R & D is 78.3%. C. The population coefficient of correlation is not 0.

V. Draw a scatter diagram of the above data and use the eyeball method to estimate the regression curve.



Note: Stars indicate coordinates determined using the regression equation from question VIC.

The line is not extended to the y-intercept because 3 is the lowest recorded R & D expenditure.

VI. Answer the following questions using the data on the preceding page.

A. Use the method of least squares to determine a regression equation.

Data from page T160

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$= \frac{1,210}{185}$$

$$= 6.5405405$$

$$a = \bar{Y} - b\bar{x}$$

$$= \frac{\sum Y}{n} - b\frac{\sum X}{n}$$

$$= \frac{310}{6} - 6.5405405\left(\frac{35}{6}\right)$$

$$= 13.513515$$

$$\hat{y}_{.x} = a + bx$$

$$\hat{y}_{.x} = 13.5 + 6.54x$$

When using a regression equation, values for x should be limited to the actual data range. Here, 3 to 10.

B. Calculate the estimated profit for next year when R & D will be \$8,000,000.

$$\hat{y}_{.x} = 13.5 + 6.54x$$

$$\hat{y}_{.8} = 13.5 + 6.54(8)$$

$$= 13.5 + 52.32$$

$$= \$65.82 \text{ million}$$

C. Draw the regression line on the page 160 scatter diagram.

Two points (x,y) may be used to draw a straight line. Here, 8 and 65.82 from question B, and the y-intercept (0,13.5) are used.

D. Calculate the 99% confidence interval for question B.

$$S_{y.x} = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n-2}} = \sqrt{\frac{17,700 - 13.513515(310) - (6.5405405)(2,010)}{6-2}} = 9.54$$

$$df = 6 - 2 = 4$$

$$\alpha/2 = .01/2 = .005 \rightarrow t = 4.604$$

$$\bar{x} = \frac{\sum x}{n} = \frac{35}{6} = 5.83$$

Note the use of +1 under the radical. This is necessary because the question concerns a particular value (the next value) of y, and not the mean value of y. Predicting a particular value of y increases the confidence interval.

$$\hat{y}_{.x} \pm tS_{y.x} \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

$$\hat{y}_{.8} = 65.82 \pm 4.604(9.54) \sqrt{1 + \frac{1}{6} + \frac{(8-5.83333)^2}{235 - \frac{(35)^2}{6}}}$$

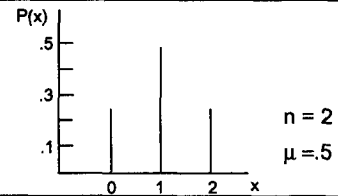
$$= 65.82 \pm 50.44$$

$$15.38 \leftrightarrow 116.26$$

E. What procedure should be followed if the range for the answer to question D includes zero or a negative number?

If the range expresses the possibility of a negative number, the confidence level may be lowered with a larger sample. This happens because a larger sample lowers t. Here, if only 3 million dollars is invested in R & D, the average value y is only $13.5 + 6.54(3) = 33.12$. For the 99% confidence level, acceptable error is approximately 52.47 (calculations for this number are not shown). This means profit could be negative ($33.12 - 52.47$). However, because profits can be negative, a larger sample is not required. But, this range is very large and may not be useful.

Table 1
Binomial Probability Distribution



Probability of x successful outcomes given the following population means (μ) and trials (n)

n = 1		μ									
x	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
0	0.950	0.900	0.800	0.700	0.600	0.500	0.400	0.300	0.200	0.100	0.050
1	0.050	0.100	0.200	0.300	0.400	0.500	0.600	0.700	0.800	0.900	0.950

n = 2		μ									
x	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
0	0.903	0.810	0.640	0.490	0.360	0.250	0.160	0.090	0.040	0.010	0.003
1	0.095	0.180	0.320	0.420	0.480	0.500	0.480	0.420	0.320	0.180	0.095
2	0.003	0.010	0.040	0.090	0.160	0.250	0.360	0.490	0.640	0.810	0.903

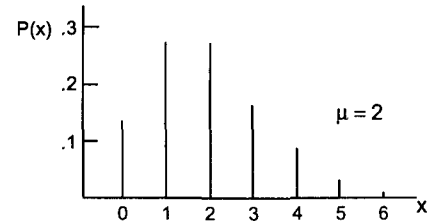
n = 3		μ									
x	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
0	0.857	0.729	0.512	0.343	0.216	0.125	0.064	0.027	0.008	0.001	0.000
1	0.135	0.243	0.384	0.441	0.432	0.375	0.288	0.189	0.096	0.027	0.007
2	0.007	0.027	0.096	0.189	0.288	0.375	0.432	0.441	0.384	0.243	0.135
3	0.000	0.001	0.008	0.027	0.064	0.125	0.216	0.343	0.512	0.729	0.857

n = 4		μ									
x	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
0	0.815	0.656	0.410	0.240	0.130	0.063	0.026	0.008	0.002	0.000	0.000
1	0.171	0.292	0.410	0.412	0.346	0.250	0.154	0.076	0.026	0.004	0.000
2	0.014	0.049	0.154	0.265	0.346	0.375	0.346	0.265	0.154	0.049	0.014
3	0.000	0.004	0.026	0.076	0.154	0.250	0.346	0.412	0.410	0.292	0.171
4	0.000	0.000	0.002	0.008	0.026	0.063	0.130	0.240	0.410	0.656	0.815

n = 5		μ									
x	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
0	0.774	0.590	0.328	0.168	0.078	0.031	0.010	0.002	0.000	0.000	0.000
1	0.204	0.328	0.410	0.360	0.259	0.156	0.077	0.028	0.006	0.000	0.000
2	0.021	0.073	0.205	0.309	0.346	0.313	0.230	0.132	0.051	0.008	0.001
3	0.001	0.008	0.051	0.132	0.230	0.313	0.346	0.309	0.205	0.073	0.021
4	0.000	0.000	0.006	0.028	0.077	0.156	0.259	0.360	0.410	0.328	0.204
5	0.000	0.000	0.000	0.002	0.010	0.031	0.078	0.168	0.328	0.590	0.774

n = 6		μ									
x	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
0	0.735	0.531	0.262	0.118	0.047	0.016	0.004	0.001	0.000	0.000	0.000
1	0.232	0.354	0.393	0.303	0.187	0.094	0.037	0.010	0.002	0.000	0.000
2	0.031	0.098	0.246	0.324	0.311	0.234	0.138	0.060	0.015	0.001	0.000
3	0.002	0.015	0.082	0.185	0.276	0.313	0.276	0.185	0.082	0.015	0.002
4	0.000	0.001	0.015	0.060	0.138	0.234	0.311	0.324	0.246	0.098	0.031
5	0.000	0.000	0.002	0.010	0.037	0.094	0.187	0.303	0.393	0.354	0.232
6	0.000	0.000	0.000	0.001	0.004	0.016	0.047	0.118	0.262	0.531	0.735

Table 2
Poisson Probability Distribution



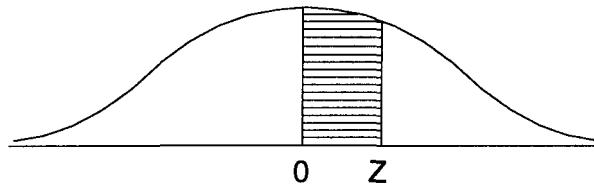
Probability of x outcomes given the following population means

x	.10	.20	.30	.40	.50	.60	.70	.80	.90
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494
4	0.0000	0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Probability of x outcomes given the following population means

x	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
0	0.3679	0.1353	0.0498	0.0183	0.0067	0.0025	0.0009	0.0003	0.0001
1	0.3679	0.2707	0.1494	0.0733	0.0337	0.0149	0.0064	0.0027	0.0011
2	0.1839	0.2707	0.2240	0.1465	0.0842	0.0446	0.0223	0.0107	0.0050
3	0.0613	0.1804	0.2240	0.1954	0.1404	0.0892	0.0521	0.0286	0.0150
4	0.0153	0.0902	0.1680	0.1954	0.1755	0.1339	0.0912	0.0573	0.0337
5	0.0031	0.0361	0.1008	0.1563	0.1755	0.1606	0.1277	0.0916	0.0607
6	0.0005	0.0120	0.0504	0.1042	0.1462	0.1606	0.1490	0.1221	0.0911
7	0.0001	0.0034	0.0216	0.0595	0.1044	0.1377	0.1490	0.1396	0.1171
8	0.0000	0.0009	0.0081	0.0298	0.0653	0.1033	0.1304	0.1396	0.1318
9	0.0000	0.0002	0.0027	0.0132	0.0363	0.0688	0.1014	0.1241	0.1318
10	0.0000	0.0000	0.0008	0.0053	0.0181	0.0413	0.0710	0.0993	0.1186
11	0.0000	0.0000	0.0002	0.0019	0.0082	0.0225	0.0452	0.0722	0.0970
12	0.0000	0.0000	0.0001	0.0006	0.0034	0.0113	0.0263	0.0481	0.0728
13	0.0000	0.0000	0.0000	0.0002	0.0013	0.0052	0.0142	0.0296	0.0504
14	0.0000	0.0000	0.0000	0.0001	0.0005	0.0022	0.0071	0.0169	0.0324
15	0.0000	0.0000	0.0000	0.0000	0.0002	0.0009	0.0033	0.0090	0.0194
16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0014	0.0045	0.0109
17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0021	0.0058
18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0009	0.0029
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0014
20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0006
21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003
22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

Table 3
The Standard Normal Distribution



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2832	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Table 4 The Student t Distribution	Degrees of freedom df	Area from the mean to a critical value				
		0.400	0.450	0.475	0.490	0.495
		α for a one-tail problem				
		0.100	0.050	0.025	0.010	0.005
		α for a two-tail problem				
		0.200	0.100	0.050	0.020	0.010
1	3.078	6.314	12.706	31.821	63.657	
2	1.886	2.920	4.303	6.965	9.925	
3	1.638	2.353	3.182	4.541	5.841	
4	1.533	2.132	2.776	3.747	4.604	
5	1.476	2.015	2.571	3.365	4.032	
6	1.440	1.943	2.447	3.143	3.707	
7	1.415	1.895	2.365	2.998	3.499	
8	1.397	1.860	2.306	2.896	3.355	
9	1.383	1.833	2.262	2.821	3.250	
10	1.372	1.812	2.228	2.764	3.169	
11	1.363	1.796	2.201	2.718	3.106	
12	1.356	1.782	2.179	2.681	3.055	
13	1.350	1.771	2.160	2.650	3.012	
14	1.345	1.761	2.145	2.624	2.977	
15	1.341	1.753	2.131	2.602	2.947	
16	1.337	1.746	2.120	2.583	2.921	
17	1.333	1.740	2.110	2.567	2.898	
18	1.330	1.734	2.101	2.552	2.878	
19	1.328	1.729	2.093	2.539	2.861	
20	1.325	1.725	2.086	2.528	2.845	
21	1.323	1.721	2.080	2.518	2.831	
22	1.321	1.717	2.074	2.508	2.819	
23	1.319	1.714	2.069	2.500	2.807	
24	1.318	1.711	2.064	2.492	2.797	
25	1.316	1.708	2.060	2.485	2.787	
26	1.315	1.706	2.056	2.479	2.779	
27	1.314	1.703	2.052	2.473	2.771	
28	1.313	1.701	2.048	2.467	2.763	
29	1.311	1.699	2.045	2.462	2.756	
30	1.310	1.697	2.042	2.457	2.750	
35	1.306	1.690	2.030	2.438	2.724	
40	1.303	1.684	2.021	2.423	2.705	
45	1.301	1.679	2.014	2.412	2.690	
50	1.299	1.676	2.009	2.403	2.678	
60	1.296	1.671	2.000	2.390	2.660	
70	1.294	1.667	1.995	2.381	2.648	
80	1.292	1.664	1.990	2.374	2.639	
90	1.291	1.662	1.987	2.368	2.632	
100	1.290	1.660	1.984	2.364	2.626	
160	1.287	1.655	1.975	2.350	2.607	
200	1.286	1.653	1.972	2.345	2.601	
Z	1.282	1.645	1.960	2.326	2.576	

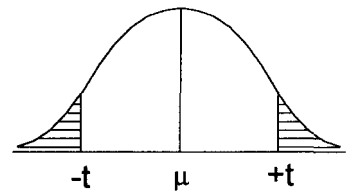
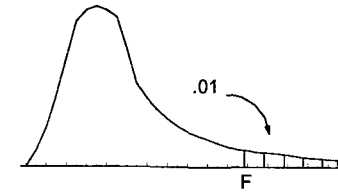


Table 5A

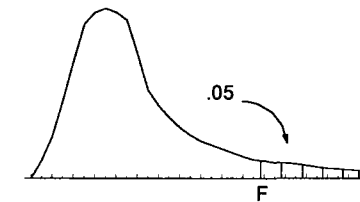
F Distribution, Critical Values for Upper .01



		Numerator degrees of freedom																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
Denominator degrees of freedom	1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339
	2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49
	3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22
	4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56
	5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11
	6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97
	7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74
	8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95
	9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40
	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84
	17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	

Table 5B

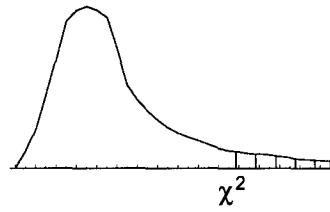
F Distribution, Critical Values for Upper .05



		Numerator degrees of freedom																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
Denominator degrees of freedom	1	161	199	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	

Table 6

Chi-Square Distribution



Degrees of freedom	Right tail area				
	.10	.05	.025	.01	.005
1	2.71	3.84	5.02	6.64	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.82	9.35	11.35	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.95
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.72	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.68	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
16	23.54	26.30	28.85	32.00	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.20	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40.00
21	29.62	32.67	35.48	38.93	41.40
22	30.81	33.92	36.78	40.29	42.80
23	32.01	35.17	38.08	41.64	44.18
24	33.20	36.42	39.36	42.98	45.56
25	34.38	37.65	40.65	44.31	46.93
26	35.56	38.89	41.92	45.64	48.29
27	36.74	40.11	43.19	46.96	49.64
28	37.92	41.34	44.46	48.28	50.99
29	39.09	42.56	45.72	49.59	52.34
30	40.26	43.77	46.98	50.89	53.67

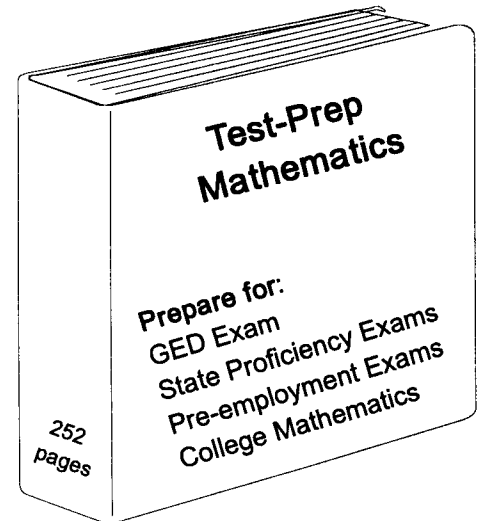
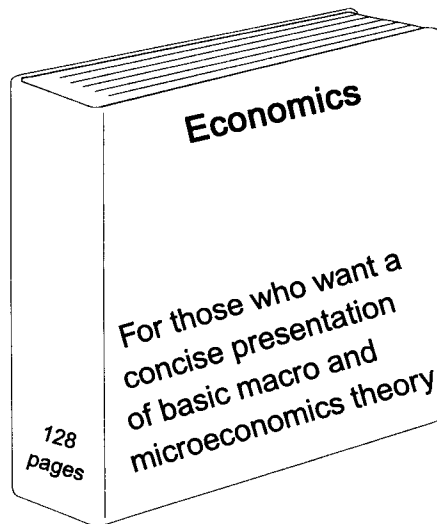
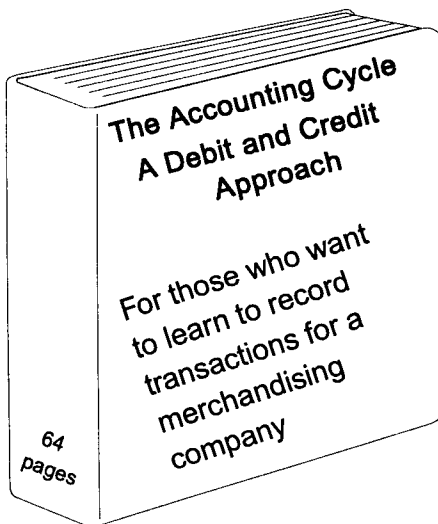
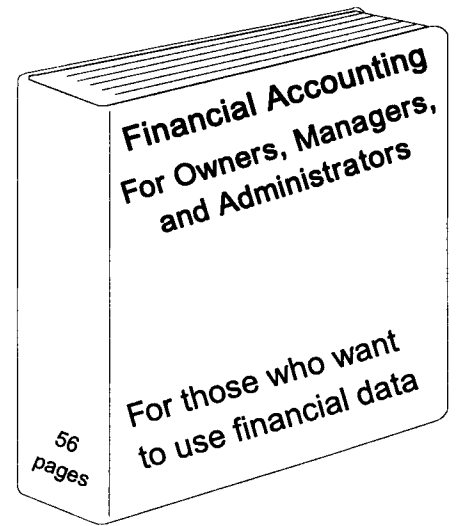
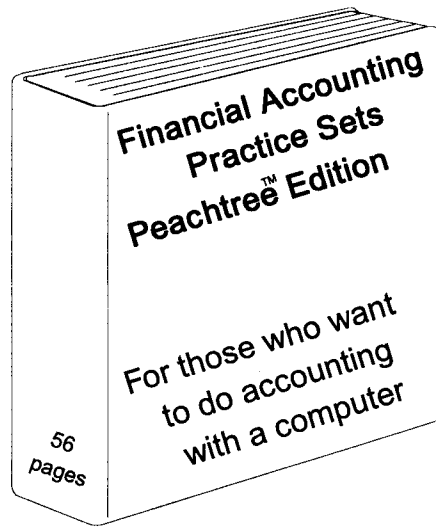
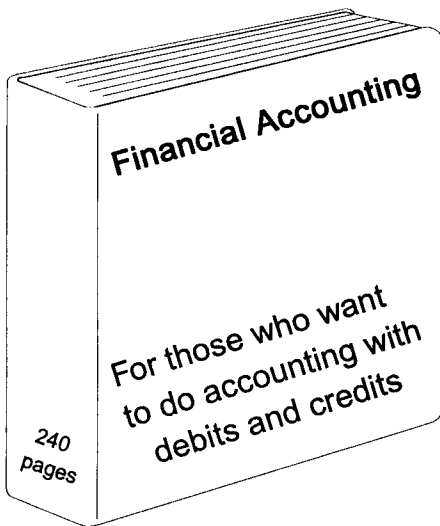
INDEX

Acceptable error	71	Data	3
Addition rules for probability	41	Deciles	11
All-inclusive	4	Degree of confidence	71
(α) Alpha error	84	Degrees of freedom	98
Alternate hypothesis	84	Descriptive statistics	2
Analysis of variance	108	Discrete probability distribution	52
ANOVA	108	Discrete variable	2
Arithmetic mean	10	Dispersion	16
Array	4	Distribution free data	120
Asymptotic probability distribution	58	Empirical probability	40
Average deviation	16	Empirical rule	17
Bayes' theorem	47	Error: Type I and II	84
Bell-shaped probability distribution	58	Estimates	67
(β) Beta error	84	Event	40
Bias	17	Experiment	3
Bimodal distribution	23	F distribution	108
Binomial distribution	52	Factorial rule	47
Binomial table	53	Finite correction factor	70
Blocking variable	114	Frequency	4
C chart	103	Frequency distribution	5
Census	3	Frequency polygon	5
Central limit theorem	67	General rule for addition	41
Central tendency	10	General rule for multiplication	46
Chebyshev's rule	17	Goodness of fit	120
Chi-square distribution	120	Grouped median	22
Class	4	Grouped mode	23
Class interval	22	Grouped sample mean	22
Class limits	4	Histogram	5
Class width	4	Hypothesis testing	84
Classical probability	40	Inferential statistics	2
Coefficient of correlation	146	Interquartile range	11
Coefficient of determination	147	Intersection	41
Coefficient of skewness, Pearson's	23	Interval estimate	67
Coefficient of variation	17	Joint probability	41
Collectively exhaustive	4	Kruskal-Wallis test	132
Combination rule	47	Kurtosis	29
Complement	41	Leptokurtic curve	29
Compound event	40	Less-than cumulative frequency distribution	5
Conditional probability	46	Level of significance	84
Confidence interval	67	Linear regression	153
Contingency table	46	Mann-Whitney test	132
Continuity correction factor	61	Marginal probability	46
Continuous variable	2	Mean	10
Control chart	102	Measurement scales	3
Correlation analysis	146	Measures of dispersion	11
Counting rule	47	Median	11
Critical values	84	Mesokurtic curve	29
Cumulative frequency distribution	5	Method of least squares	152

Midpoint formula, grouped	22	Relative probability	40
Midpoint formula, ungrouped	11	Response variable	108
Mode	11	(ρ) Rho	147
More-than ogive cumulative frequency distribution	5	Sample	2
(μ) Mu	10	Sample mean	10
Multiplication rule for probability	46	Sample proportion	94
Mutually exclusive	4	Sample size	71
Nonparametric statistics	120	Sample space	40
Nonsampling error	3	Sample standard deviation	17
Nonsymmetrical distribution	23	Sample statistics	10
Normal approximation to the binomial	61	Sample variance	17
Normal probability distributions	58	Sampling distribution of the means	66
Null hypothesis	84	Sampling error	3
Ogive	4	Scatter diagram	146
One-tail test	85	Scatter plot	152
Operating characteristic curves	89	Secondary source data	3
Outcome	40	(Σ) Sigma	10
P chart	103	Simple event	40
P-values	88	Simple random sample	66
Paired difference test, nonparametric	132	Skewness	23
Paired difference test, parametric	99	Special rule for addition	41
Parameter	2	Special rule for multiplication	46
Parametric statistics	120	Standard deviation	16
Pearson's coefficient of skewness	23	Standard error of the estimate	153
Percentiles	11	Standard error of the mean	70
Permutation rule	47	Standard error of the proportion	70
Platykurtic curve	29	Standard normal distribution	58
Point estimates	67	Stated class limits	4
Poisson approximation	53	Statistic	2
Poisson distribution	53	Stratified random sample	66
Population	2	Student t distribution	98
Population mean	10	Subjective probability	41
Population parameters	2	Sum of the deviations	10
Population proportion	70	Sum of the squares	109
Population standard deviation	16	Survey	3
Population variance	16	Systematic random sample	66
Power curves	89	t distribution	98
Primary source data	3	Tally	4
Probability	40	Test statistic	84
Probability distribution	52	Treatment means	108
Probability rules	41	Treatment variable	108
Proportions	94	Tree diagram	46
Qualitative variable	2	Two-factor ANOVA	114
Quantitative variable	2	Two-tail problem	85
Quartiles	11	Type I and Type II error	84
R chart	103	Variable	2
Random variable	52	Variances	16
Range	16	Venn diagram	41
Ratio scaled data	3	Weighted mean	10
Real class limits	4	Z table	59
Regression analysis	152		
Regression line	153		
Relative frequency polygon	5		

The Quick Notes Series

Designed To Make Learning Faster and Easier



Quick Notes Are User Friendly

For price information using our **Quick Delivery System**
call 603-424-4665 or write antonw@ix.netcom.com