Contents

Course Description **3**

Daily Class Schedule . . . 7

How To Learn Statistics . . . 9

Week 1 **11**

Introduction to marketing research . . . 11

Research designs . . . 17

Your Turn: Research Design Practice Questions . . . 23

Introduction to statistics and descriptive research . . . 26

Comparative Advertising, Measurement Scales and Data Analysis . . . 31

Week 2 **33**

Descriptive statistics for nominal-level variables . . . 33

Frequency distributions and histograms . . . 34

Measures of central tendency and position . . . 47

Measures of dispersion . . . 52

Boxplots and outliers . . . 58

Rescaling data . . . 62

Univariate Descriptive Statistics Class Exercises . . . 66

Suspicious customer case . . . 71

Week 3 **73**

Graphical excellence . . . 73

Crosstabs . . . 78

Presenting findings and graphical perception . . . 81

Dependence of a numerical variable on categorical variables . . . 89

Multivariate Descriptive Statistics Class Exercises . . . 93

Whodunit? . . . 100

Week 4 **101**

Discrete random variables . . . 101

Binomial random variables . . . 105

Normal random variables . . . 112

The normal approximation to the binomial . . . 118

Sampling distributions for means and totals . . . 125

Standard errors 134

Week 5 **137**

Sampling designs and errors 137

Can This Survey Be Saved? 148

Questionnaire Design 152

Practice Midterm **163**

Week 6 **167**

Confidence intervals for large samples 167

The t distribution 176

Determining sample sizes 180

Confidence Intervals and Sample Sizes (reading) 185

Week 7 **192**

Single-sample t test for means and proportions 192

Week 8 **201**

Evaluating Normality 201

Paired-sample t test 204

Independent-sample t test 208

Independent-sample Z test for proportions 215

Chi-Squared Test of Independence 220

How large is sufficiently large? 228

Week 9 **234**

Correlations 234

Simple linear regression 243

Click ball point pens case R 248

Week 10 **260**

Causality, experimentation and validity 261

Pre-experimental designs and threats to validity 265

True- and Quasi-experimental designs 274

Experimental Research Practice Questions 283

Measuring the Effects of Marketing and Showing ROI (reading) 286

Practice Final Exam Questions **289**

IMC 451: Statistics for Marketing Decisions
Fall Quarter, 2017
Tuesdays 9:00–10:50, MTC Forum
Thursdays 9-10:50, 11:00–12:50, or 2-3:50, MTC 3-119

Instructor: Edward C. Malthouse
Office: Fisk 304D, 1845 Sheridan Road
Phone: 847-467-3376
Fax: 847-491-5925
email: ecm@northwestern.edu

Objectives

The purpose of this course is to develop the data-analysis skills that you will need to be a successful marketer in the future. Marketing managers have unprecedented amounts of data available to assist them in making decisions. Individuals who understand how to use data to make better decisions and support their arguments with evidence have a great advantage. This class will focus on developing the following skills.

1. This course is critical to the IMC process because it teaches step 4 (measurement) and step 1 (customer understanding). The measurement part is an important differentiator of the IMC program, and you will learn the fundamentals of descriptive and causal research designs in this class.

2. You should be able to communicate with members of a market research department. You will learn a vocabulary that can be used to describe data, e.g., average, median, quartiles, deciles, percentiles, standard deviations, margins of error, statistical significance. You will also learn how market researchers think about marketing problems. Learning these skills will enable you to express your research needs in a language that a market researcher will understand, which will allow you to work more productively and efficiently with them.

3. You will learn to read, interpret, and use marketing research reports correctly. For example, you will learn how to interpret a cross tabulation. You will learn how to create visual displays of marketing data so that you can make more persuasive marketing presentations and reports. You will also learn some of the background theory that supports these methods so that you will have a deeper understanding of them.

4. You will learn how to use statistical software to run elementary market research reports from a database. Today marketing managers have computers, access to data sources, and statistical software. They are no longer dependent upon market research departments for reports.

5. A higher-level goal of this course is to make you more *numerate*. Marketing and advertising professionals are expected to quantify the effects of their actions. This course will make you more comfortable with integrating numbers into your reasoning. This class will help you use quantitative evidence to inform your marketing decisions.

6. Another higher-level goal is to improve your confidence and ability at being an *independent problem solver*. You will improve your ability to learn new things on your own ("learn how to learn").

There are no specific prerequisites for this course, but previous exposure to statistics in college-level courses will be helpful. This course covers many of the topics that are typically included in undergraduate statistics courses, but in greater depth and with an emphasis on marketing communication applications. Statistics usually does not "sink in" the first time. If you have had it before, then this course will improve your mastery of the basic concepts and your ability to apply them to real problems. Students who have already completed *three or more* previous statistics courses with good grades and are comfortable using statistical software may discuss waiving the course with me in person.

Course Materials

1. Course packet.

2. Siegel, Andrew, *Practical Business Statistics*, 3rd, 4th, 5th or 6th edition. You can receive 40% of ebooks and 25% off the print version using the discount code "PRF13" at http://tinyurl.com/elsevier-PRF13.

Here is a list of other texts that you may wish to consult:

1. (Suggested) Iacobucci and Churchill (2010), *Marketing Research: Methodological Foundations*, 10th edition, Thomson. Or 5th–9th editions, which were "Churchill and Iacobucci."

2. Cleveland, *The Elements of Graphing Data* and *Visualizing Data*.

3. Dillman, *Mail and Internet Surveys: The Tailored Design Method*, Wiley.

4. Edward Tufte, *The Visual Display of Quantitative Information*, Graphics Press. Also his other books.

Course Policies

Grades. I expect "B" students to be able to work problems similar to those in the assigned homework from the text. "A" students should be able to solve problems that are not exactly like the assigned problems from the text and combine concepts from multiple chapters to solve a problem. Roughly half of the questions on the examinations will be taken from the assigned problems (I may change the numbers or add/delete parts). The other half of the questions will be original, with varying levels of difficulty.

Your course grade will be determined as follows:

1. *Computer quizzes* (25%)

2. *Midterm exam* (30%)

3. *Final exam* (45%)

4. *Class participation.* If you say nothing in class your "class participation" will not affect your course grade. If you contribute to class by consistently asking constructive questions or making insightful comments, I may increase your course grade by one third of a letter (e.g., if the numbers indicate your are a B+ student, I may assign you a course grade of A−). If you are consistently disruptive during class, I may decrease your course grade by one third of a letter. Disruptive activity includes, but is not limited to, talking to fellow students *during* the lecture. **You should act like a business professional at all times.**

Exam policies.

- Always report at least two accurate digits when giving an answer. Carry out intermediate calculations with full precision (e.g., doing computations in Excel or R without rounding).
- I do not give makeup exams. Please complete the exam at the scheduled time.
- You may use one 8.5 by 11 inch sheet of notes on the quizzes and two sheets of this size on the final. I will provide a normal table. Bring a calculator and an English dictionary (if English is not your native language).
- When answering essay questions, be concise and get to the point. **Answer the question and nothing else.** Lead with the headline. Good answers prioritize what is most important. Think before you start to write and don't do a "brain dump." You should get into these habits in your professional life too.
- You are not responsible for doing the independent-sample t-test by hand on the final exam, but are responsible for using computer output in a sensible way. This includes knowing whether the assumptions of the test are met, discussing threats to internal and external validity of the test, stating the null and alternative hypotheses of each P-value in the output, interpreting what the result means, and suggesting marketing actions based on the results.

Office Hours.

- The best way to contact me is via email. I usually respond within hours.
- I have approximately 140 students this quarter. I have arranged for three student assistants who will have 6 hours of office hours each week on Fridays. The student assistants should be your first option for more help. I will also be in my office every Tuesdays from 1:30–4:30 for normal office hours. *Send me an email and I will assign you a 10-minute appointment.* If you come as a group of two or more students I will give longer slots. There will also be a disussion section led by a student assistant Monday afternoons from 4:15–5:45 in MFC 3-119, starting October 2.

Honesty, Plagiarism, and Cheating. All students are required to adhere to the Medill Integrity Code as well as Northwestern University's academic integrity policies. Academic dishonesty can result in penalties ranging from letters of warning to dismissal from the university. Instructors may give a failing grade in a course for academic dishonesty. It is also university policy that instructors can require students to submit their work electronically to be analyzed for possible plagiarism.

Northwestern University works to provide a learning environment for students with disabilities that affords equal access and reasonable accommodation. Any student who has a documented disability and needs accommodations for classes and/or course work is requested to speak directly to the Office of Services for Students with Disabilities (847)-467-5530) and the instructor as early as possible in the quarter (preferably within the first two weeks of class). All discussions will remain confidential. Accommodations can be made by instructors once OSSD has met with the student and verified the disability.

How to be successful in this class. **Practice daily**! Start with the problems assigned in the packet and textbook, and then do the problems in the practice exams in the back of the course packet. **The only way to learn this material is to practice.**

Statistical software. You will learn how to use SPSS, Tableau and Excel for analyzing statistical data in the class. You may use any software you like (e.g., R), but I will teach SPSS in class. Unless you are using another package you will need to buy SPSS here. You need the "**Statistics base grad pack**" version. The current version is 24, but older ones work too. A six-month license will be sufficient, since you will use SAS, R and Python in the more advanced statistics classes. All students will need SPSS for the winter IMC Process class (IMC 460). Here is a list of other vendors: thinkedu, studentdiscounts.com, creationengine.com and studica.com. If you have problems installing SPSS click here for support. You can download a student copy of Tableau for free.

If you do not plan to take any more statistics courses you can buy the "Statistics Base Grad-Pack," but if you plan to take the second course in statistics you should buy the "Statistics Standard GradPack," which includes the "Regression" module. By the end of the quarter you should be able to do the following in SPSS:

1. You will understand how to set up a data set using labels, value labels, and setting formatting options (e.g., numeric formats, dates, strings).

2. Run basic descriptive statistics using "frequencies," "descriptives" and "explore." This includes using many of the options to request percentiles and various graphs.

3. You should be comfortable with the graphics functions and be able to generate bar charts, box plots, error bar plots, scatterplots and scatterplot matrices. You should understand how to use all of the "conditioning" options including multi-panel displays and encoding other variables with color or symbols. You should also know how to label cases and edit graphs.

4. You will know how to modify a data set using the "compute" and "recode" statements. You will also know how to use "rank cases" to create quartiles, deciles, etc.

5. You will understand how to handle dates. You will be able to create dates and compute the differences between them. You will also know how to change the format used to display them.

6. You will know how to use "select cases" to restrict the universe of your analysis and sample cases.

7. You are responsible for knowing all of the options under "Analyze/Compare means," including basic comparisons (independent variables, dependent variables and "layers"), single sample t-tests, paired-sample tests, independent-sample tests and one-way ANOVAs.

8. You are responsible for knowing most of the options under "crosstabs," including the various types of counts, percents and residuals under "cells," the chi-square test and the "layers" box.

Those who are interested in marketing analytics should also learn R. This is strictly optional. You may work computer assignments using SPSS, R, Excel or any other program. You can download a copy of R from CRAN. The course packet gives commands in both Excel and R with the following typographical conventions:

- Excel commands are set in green, courier, all-caps font, e.g., `AVERAGE`.
- R commands are set in red, courier, lower-case font, e.g., `mean`. Red font will not be used when an entire page is devoted to R (usually indicated in heading).

Data sets are available on Canvas each week. Some are in Excel format and must be converted to SPSS or R. This is good practice for you, since you will often be sent data in Excel format in the workplace. Here is a video demonstrating the conversion.

Daily Class Schedule

- Week 1 (Sep 21, 25): Introduction to Marketing Research and Research Designs
  - Read: "How to learn statistics" (course packet page 9)
  - Read: Malthouse and Li, "Opportunities for and pitfalls of using big data in advertising research," *Journal of Advertising* article **before coming to class** (Canvas)
  - Read: Andreasen, "Backwards Marketing Research"
  - Read: Siegel chapter 2
- Week 2 (Sep 26, 28): Descriptive Statistics For One Variable
  - Topics: Frequency distributions, histograms, means, medians, modes, quantiles, standard deviations, variances, IQRs, boxplots, SPSS
  - Read: Siegel chapters 3–5
- Week 3 (Oct 3, 5): Presenting Results and Studying Relationships
  - Topics: Cross tabs, comparing means, graphs showing relationships
  - Read: Discussion of crosstabs in Siegel (chapter 17, ignore χ^2 test for now)
  - Read: Tufte, *The Visual Display of Quantitative Information* (optional) and his other books on visual displays
- Week 4 (Oct 10, 12): Probability Review
  - Topics: Simple discrete random variables, binomial and normal distributions, sampling distriubtions of the mean and total
  - Read: Siegel sections 7.1–7.4, 8.3
- Week 5 (Oct 17, 19): Sampling Designs and Errors
  - Topics: Standard errors, probability versus nonprobability samples, non-sampling errors/biases, questionnaire design

  - Read Siegel chapter 8

- Week 6 (Oct 24, 26): Confidence Intervals and Sample Sizes
  - **Midterm (October 24)** covers weeks 1–5 (Siegel Chapters 2–5, 7, 8).
  - Read: Siegel chapters 9
  - Read: "Confidence Intervals and Sample Sizes" (course packet page 185)
  - Thursday topics: Confidence intervals for means and percents, t distribution, formula for sample size, prediction intervals (as time permits)

- Week 7 (Oct 31, Nov 2): Hypothesis Testing I
  - Topics: Logic of hypothesis testing, null/alternative hypotheses, type I/II errors, P-values, confidence interval approach, tests for a single mean or proportion
  - Read Siegel §10.1–10.5

- Week 8 (Nov 7, 9): Hypothesis Testing II
  - Topics: Paired-sample test, independent-sample test for means and proportions, χ^2 test of independence
  - Read Siegel §10.6, 17.3 (week 9)

- Week 9 (Nov 14, 16, 21): Introduction to Correlation and Regression
  - Topics: Scatterplots, correlations, simple linear regression, tests and confidence intervals for slope and intercept, standard error of the estimate and R^2, prediction intervals
  - Read Siegel Chapter 11

- Week 10 (Nov 28, 30): Causal Designs
  - Topics: Causality, observational studies, true- and quasi-experimental designs, internal and external validity, THIS MESS.
  - Read: "Measuring the Effects of Marketing and Showing ROI" (course packet page 286)

- **Final exam (Wednesday, Dec 6 from 9–12)** is comprehensive, but will emphasize weeks 6–10 (Siegel chapters 9–11, 17 and causal designs). See page 289 for practice problems. **Do not plan to leave campus for vacation before noon on Dec 6**.

How to learn statistics

By Edward Malthouse

The ultimate goal of this course is to teach you to incorporate statistical reasoning and techniques into your process of making marketing and business decisions. Meeting this goal requires two main steps: (1) you must be exposed to statistical ideas and techniques and (2) you must practice applying them in different situations to different problems. Your grade will be determined by how well you can apply the ideas and techniques.

Learning statistics is like training for a marathon in that you must "train" a little bit every day (or at least several times a week). If you put off training until the day before the race (or exam), you will not do very well. Each week will introduce a few new ideas that build on what was covered in previous week. You must understand the ideas and practice applying them. The course builds on itself and if you fall behind you will be in trouble.

Step 1: exposure to ideas techniques. There are many ways to accomplish the first step:

- The textbook (Siegel) is well-written and students in the past have liked the way he covers the material. My lectures stick close to Siegel.

- There are many other books that cover the same material. It is often good to read another author's perspective, although you should beware that some authors use different notation. Some of my favorite introductory statistics books are listed in the syllabus.

- I have summarized the main ideas from each chapter in the course packet and have made instructional videos that explain the main ideas in my words.

- You may find other videos that explain the concepts, such as Kahn Academy.

You should find a combination of learning resources that works for you.

Step 2: practice. The only way for the concepts to "sink in" is to practice them—often. Reading a technical book is not a passive activity and readers must engage with the material by working through the examples and additional problems, and applying the techniques to their own work. It is only through these activities that readers will fully understand and remember the material. The more problems you do, the more comfortable and fluent you will feel with the material. I believe that there are several levels of practice, and you should go through these levels with each section, chapter and unit (note that each chapter has multiple sections, and several chapters aggregate to a unit).

1. Siegel (and other authors) discuss an idea and then work an illustrative problem. I will work problems in my videos. You should be able to work the example problems that Siegel and I do on your own. Make sure you can work these problems on your own before attempting other problems. You may have to read/view the solution first, but after reading the solution, take out a clean sheet of paper and try it on your own.

2. You should be able to work problems that are similar to the examples. After each section, I recommend specific problems in Siegel and give you the answers. In some cases you can watch a video that shows my solution the problems.

3. You should be able to work problems out of context where you must first determine which technique is appropriate, and then figure out how to apply it in the new situation or context. The discussion boards are designed to help you do this. I have

provided extra problems after chapters or units, along with solutions.

Move on to the next level when you feel comfortable. Some of you may be able to skip level 1 entirely; you will watch my video and/or read the examples in the book and jump to level 2 problems. Others will need to work the level 1 problems several times before moving to level 2 problems. Some of you will "get" the material after doing a small number of level 2 problems, while others may have to work dozens of level 2 problems from each section.

Definition

According to the American Marketing Association, *marketing research* is "the function that links the consumer, customer, and public to the marketer through information—information used to identify and define marketing opportunities and problems; generate, refine, and evaluate marketing actions; monitor marketing performance; and improve understanding of marketing as a process. Marketing research specifies the information required to address these issues, designs the method for collecting information, manages and implements the data collection process, analyzes the results, and communicates the findings and their implications."

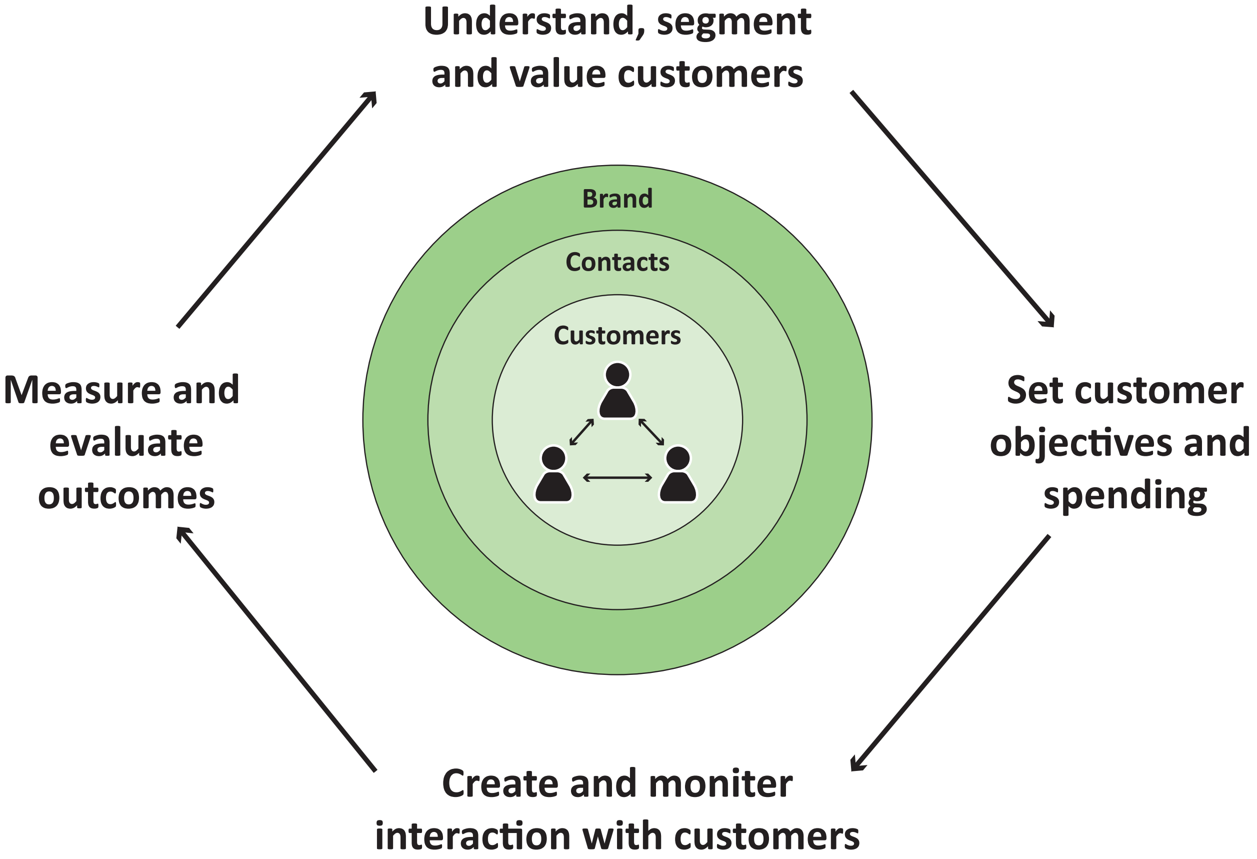

Understand, segment and value customers
Set customer objectives and spending
Create and moniter interaction with customers
Measure and evaluate outcomes
Brand
Contacts
Customers

Figure made by Karen Wang

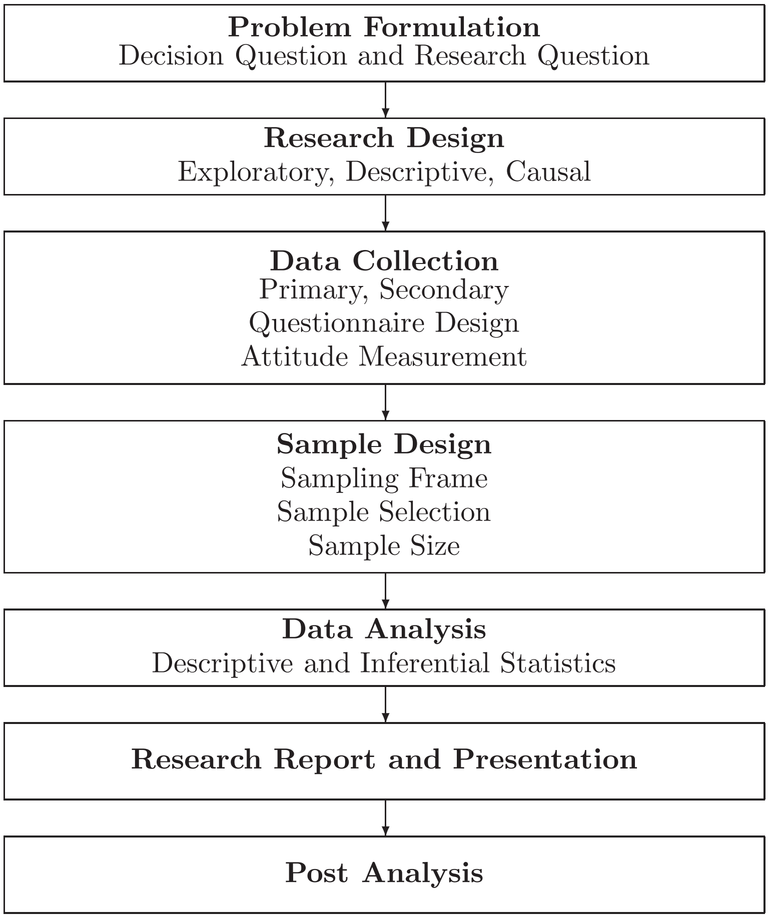
Role of Research in the Marketing Management Process

- Opportunity Identification
  - Segmentation, targeting, and positioning
  - Buyer analysis
  - Competitive and environmental analysis
- The Marketing Plan
  - Product
  - Price/price promotion
  - Promotion (non-price)
  - Channels
  - Service
- Evaluation and Control
  - Performance monitoring and evaluation
  - Refining the marketing program

Big Data Discussion

1. Where do big data come from?
2. What roles can big data play in IMC research?
3. How are big data different from the surveys, experiments and focus groups traditionally used in advertising research?
4. What is the jigsaw puzzle and how is it different from the traditional approach?
5. What are opportunities and pitfalls in using big data?

The Marketing Research Process



Problem Formulation
Decision Question and Research Question

Research Design
Exploratory, Descriptive, Causal

Data Collection
Primary, Secondary
Questionnaire Design
Attitude Measurement

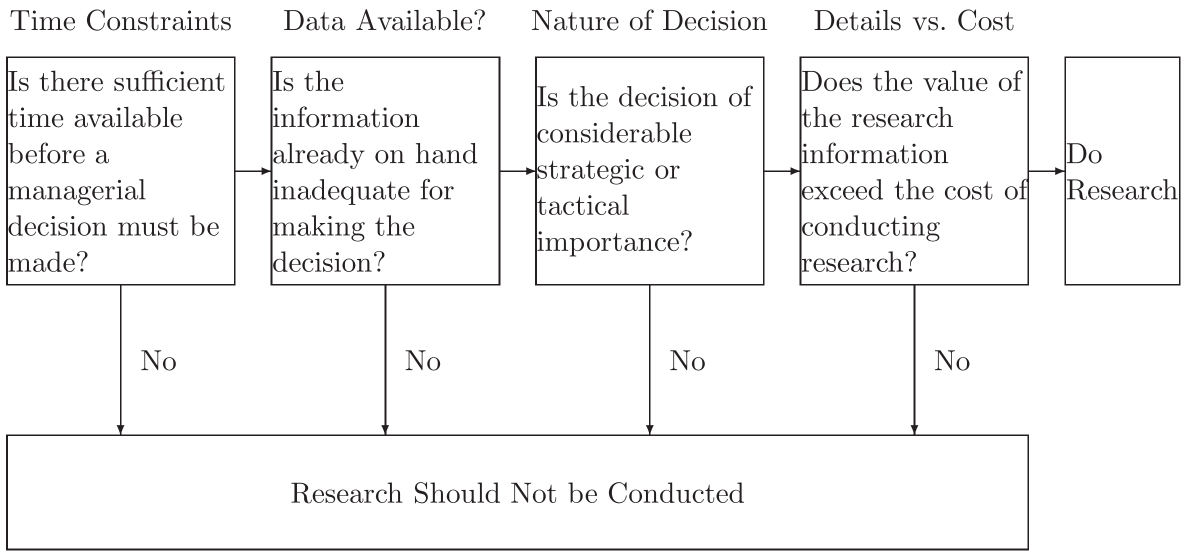
Sample Design
Sampling Frame
Sample Selection
Sample Size

Data Analysis
Descriptive and Inferential Statistics

Research Report and Presentation

Post Analysis

Determining when Marketing Research Should be Conducted



Time Constraints

Is there sufficient time available before a managerial decision must be made?

Data Available?

Is the information already on hand inadequate for making the decision?

Nature of Decision

Is the decision of considerable strategic or tactical importance?

Details vs. Cost

Does the value of the research information exceed the cost of conducting research?

Do Research

No

No

No

No

Research Should Not be Conducted

Problem Formulation: Backward Marketing Research

Close collaboration between researcher and corporate decision maker is the single most important factor predicting good outcome.

Start process where it usually ends and work backwards.

- Determine how research results will be implemented. *Decision question:* What decision must you make?
- To ensure the implementation of the results, determine what the final report should contain and how it should look. *Research question:* What information do you need to make the decision?
- Begin with and end in mind.

“Data is like garbage; you better know what you are going to do with it before you collect it.”

Research Design

Research design: the detailed blueprint used to guide the implementation of a research study toward the realization of its objectives.

Three types of research designs:

- *Exploratory research* is used to gain ideas and insight into the research problem, e.g., beer manufacturer faced with decreased sales might conduct an exploratory study to generate possible explanations.
- *Descriptive research* is used when the research objectives and questions are clearly defined and summary measures are needed to address the research questions, e.g., an investigation of the trends in the consumption of a brand of beer with respect to such characteristics as age, sex, occupation, etc. You are describing some population using measures of certain attributes.
- *Causal research* is used to establish cause and effect. To understand the connections between management actions and observed outcomes, e.g., want to know how people react to a particular promotional plan for our brand of beer.

Your Turn: Example Problems

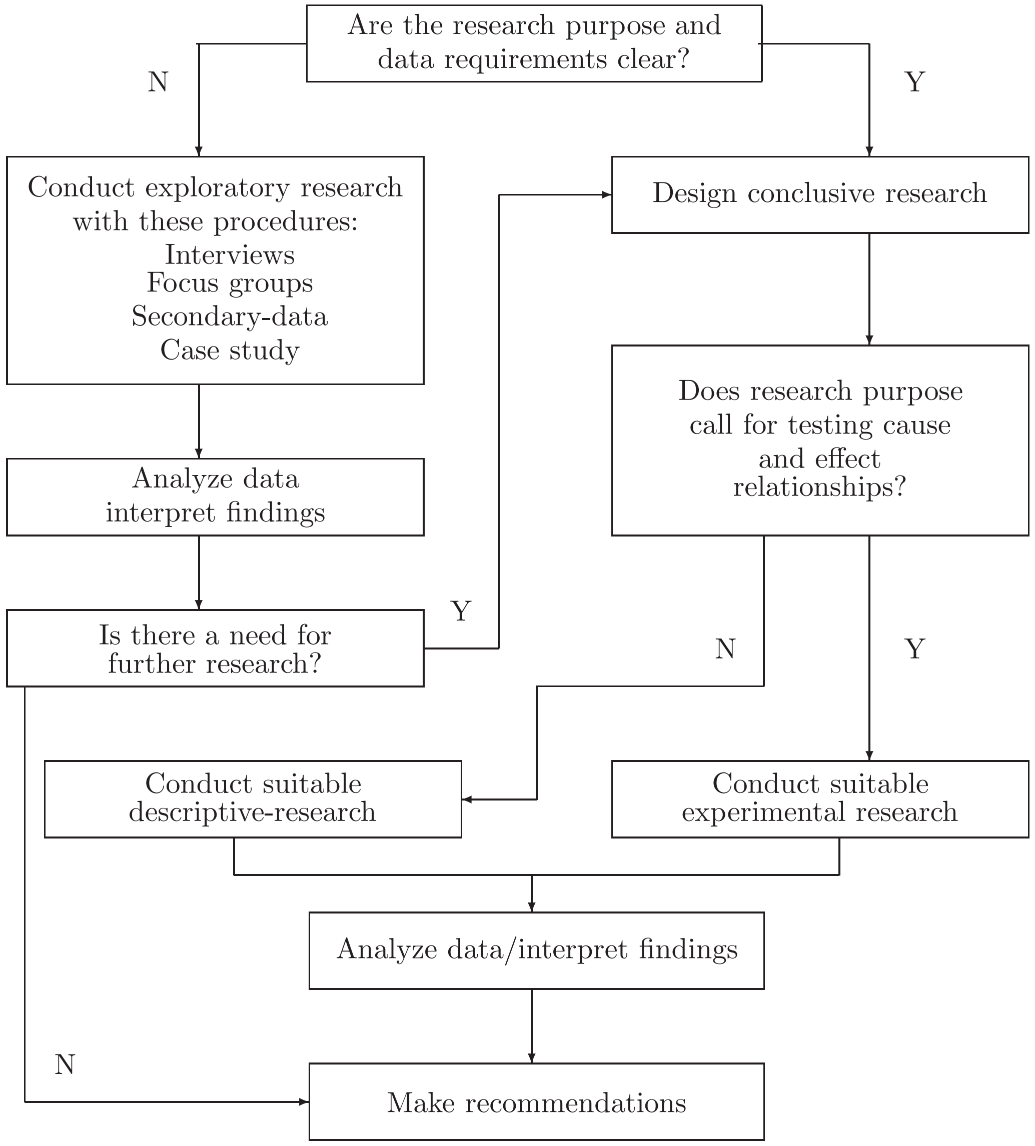
Why type of research design would you recommend?

1. Modern Office Products, Inc. (MOP) manufactures a broad line of office equipment and supplies. It sells its products to a variety of organizations through its sales force. Despite a healthy growth in industry sales, MOP's sales and profits have declined during the past two years, much to the concern of MOP executives.

2. Saver's National Bank (SNB) has grown rapidly since its inception a few years ago. Its growth is apparently due to a unique set of financial services it offers. While SNB management is pleased with the bank's performance so far, it is worried about growing competition from a variety of financial institutions. To consolidate SNB's current market position, the bank's executives want to ascertain the demographic composition of customers and their perceptions about the bank's strengths and weaknesses.

3. An auto-repair chain recently suffered major losses as well as embarrassment in its auto service business due to a temptation to oversell, which was built in the commission structure of "service advisers." The company now wants to expand the use of incentives and is considering two alternative compensation plans. Both plans include linking part of worker's incomes to customer satisfaction. Plan 1 would shoot for 60% of sellers' incomes to come from salary and 21% from commissions. The remaining 19% would be based on customer satisfaction. Plan 2 would make it 60% salary and 40% customer satisfaction, eliminating commissions. The selected plan would be introduced into the company's 799 auto repair centers.

Flow Diagram For Selecting the Appropriate Research Type



Are the research purpose and data requirements clear?
N
Y
Conduct exploratory research with these procedures: Interviews Focus groups Secondary-data Case study
Design conclusive research
Analyze data interpret findings
Does research purpose call for testing cause and effect relationships?
Is there a need for further research?
Y
N
Y
Conduct suitable descriptive-research
Conduct suitable experimental research
Analyze data/interpret findings
N
Make recommendations

Exploratory Research

Used to gain ideas and insights about the research problem.
Characteristics:

1. Less structured / more flexible
2. Number of respondents is often small — only partially representative of the population
3. Fast
4. Usually less expensive than descriptive research

Types:

1. Literature search
2. Individual (in-depth) interviews (IDI) and focus groups
3. Ethnography, netnography
4. Projective tests

Descriptive Research

- Major objective is to describe something . . . usually market characteristics or functions.
- Presupposes much prior knowledge about the phenomenon being studied. Less flexible / more rigid.
- Requires a clear specification of
  - *Who:* the population to be studied
  - *What:* characteristics that need to be measured to draw conclusions
  - *When, where, how:* details of research design including type of survey, questionnaire, sampling procedure, and analysis.
  - *Why:* what specific objectives will the research achieve?
- *Cross-sectional Design.* Sample of elements is measured only once. Provides snap-shot of variables of interest at a point in time.
- *Longitudinal/Panel Design.* A fixed sample of respondents measured repeatedly over time. Types of Panels:
  1. True — same variables measured each time
  2. Omnibus — information collected varies from study to study, e.g., Nielsen and Ipsos

Your Turn: Descriptive Research Examples

1. Return to the Bank example 3 on Page 18. Identify the who, what, where, when and how of a descriptive study to measure the demographic composition of customers and their perceptions about the bank's strengths and weaknesses.
2. Return to the automotive service center problem 3 on Page 18. Suppose that five stores have Plan 1 and a different five stores have Plan 2. Identify the who, what, where, when and how of a descriptive study to measure customer satisfaction in all 10 stores during a particular week after the plans have been implemented.

Your Turn: Research Design Practice Questions

1. The president of Jamaican Specialties, a specialty food marketing firm, was convinced that the target audience for her line of Caribbean conserves in mango, lime, passion fruit and papaya consisted of women, ages 25-44, with household incomes of $30,000 and up. Jamaican Specialties' major competitor's market segment appeared to be more widely dispersed with respect to both age and income. The president of Jamaican attributed this difference to the type of magazines in which their competitor advertised. She decided to conduct a study of women in other age groups and with other income levels to determine their interest in her products. She accepted a marketing research plan in which a panel of 800 women ages 18 and up would be mailed questionnaires. One month after receiving all questionnaires, the company would again send similar questionnaires to all the panel members. Critique this design.[1]

2. A medium-sized manufacturer of high-speed copiers and duplicators was introducing a new desktop model. The vice-president of communications had to decide between two advertising programs for this product. He preferred advertising program gamma and was sure it would generate more sales than its counterpart, advertising program beta. The next day he was to meet with the senior vice-president of marketing and planning to decide on an appropriate research design for a study that would aid in the final decision about which advertising program to implement. What research design would you recommend? Justify your choice.[2]

3. Greg Martin is the owner of a pizza restaurant that caters to college students. Through informal conversations with his customers, Greg has begun to suspect that a video-rental store specifically targeting college students would do quite well in the local market. Although his informal conversations with students have revealed an overall sense of dissatisfaction with existing rental outlets, he hasn't been able to isolate specific areas of concern. Thinking back to a marketing research course he took in school, Greg has decided that focus group research would be an appropriate method to gather information that

[1] The research design is not appropriate. Jamaican Specialties management team was resorting to longitudinal analysis when cross-sectional analysis would have been more appropriate. Longitudinal analysis is useful in the identification of changes that occur over a period of time. Further, a representative sample is required, and panel data are frequently not representative.

[2] In this situation causal research would be appropriate as the vice-president of communications is interested in determining which of the two contemplated advertising programs is likely to be more effective.

[3] (a) The decision problem concerns whether to proceed with further development of the idea of a student-oriented video rental store. The research problem involves identifying specific areas of dissatisfaction with existing video rental businesses. Some students may identify the decision problem as whether to open the store or not. This problem "jumps the gun" in that further research may be required to determine the feasibility of eliminating areas of dissatisfaction discovered in the focus group sessions. (b) Participants should be drawn from the population of college students who rent videos. A likely source is the group of customers who initially gave Martin the idea of opening the store. (c) The session should be conducted at a place convenient to the respondents. Possibilities include the library, student union, or dorm meeting rooms. (d) Most students will suggest that Martin should be the moderator. A better answer would be to use a properly qualified "third party" moderator. Martin may not be totally impartial, and could bias the results based on his personal feelings about the proposed business venture. When the moderator is not impartial, it is easy for the person to become argumentative rather than a good listener. (e) Discussion outlines

might be useful in deciding whether to pursue further development of his idea (e.g., a formal business plan, store policies, etc.).[3]

(a) What is the decision problem and resulting research problem apparent in this situation?

(b) Who should Greg select as participants in the focus group?

(c) Where should the focus group session be conducted?

(d) Who should be the moderator of the focus group?

(e) Develop a discussion outline for the focus group. Write out a list of questions that should be discussed during the focus group.

4. Gettings and Gettings, a father-and-son insurance agency in Lafayette, Indiana, was concerned with improving its service. In particular, the firm wanted to assess if customers were dissatisfied with current service and that nature of this dissatisfaction. What research design would you recommend? Justify your choice.[4]

5. The Federal Reserve (the Fed) controls currency in the United States. Recently, the Fed has been considering some changes in the currency that circulates. One change involves paper money, which is currently all the same size size and shape (no matter what the denomination of the bill). Some members of the Fed believe that money would be easier for consumers to handle if it came in different colors. For example, the one-dollar bill could remain green, the five-dollar bill could be printed on blue paper, the ten-dollar bill could be red, and so on. In addition, the Fed is considering changing the size of the bills, so that the five-dollar bill would be larger than the one, the ten would be larger than the five, and so on. Before making these changes, the Fed believes that it might be useful to conduct some marketing research. What research design would you recom-

are likely to vary widely among students, depending on which attributes are important to them. Classroom discussion of the outline should bring out the importance of covering as many attributes as possible in the outline, while still being open to those that surface as a result of group discussions. Following is a listing of possible outline topics: hours of operation, physical location, type of videos stocked, length of rental period, cost of rental, availability of laser disc formats, distribution methods (delivery, drive-up, etc.), and ancillary products (equipment, snacks, etc.).

[4] Exploratory research would be appropriate in this situation. It is justified since there is no well-defined hypothesis and the partnership wants to gain insight as to the existence and nature of customer dissatisfaction. They could conduct experience surveys and depth interviews with a cross section of their customers. After isolating some sources of dissatisfaction, Gettings & Gettings might then want to conduct a sample survey to determine how extensive each source of dissatisfaction might be. It is dangerous to proceed without the exploratory study though, as some main sources of dissatisfaction could be overlooked.

[5] There are a variety of actors who influence the potential success of the proposed currency changes. These include: (a) the American public (consumers); (b) bank employees (who influence which currency remains in circulation; (c) store clerks (who dispense currency from cash drawers; (d) store managers (who decide which currency to send to the bank and which to retain in the cash drawers to make change for the coming day's sales). Among the research designs which the Fed could implement are:

(a) *Secondary research.* As a starting point, the Fed could examine the findings of pre-existing research studies. For example, psychologists have completed dozens of studies on human perceptions and beliefs about color. In addition, foreign countries (e.g., Canada) use paper money which comes in a variety of colors (depending upon the denomination), so the Fed could obtain information about consumer reactions in these countries.

(b) *Case studies.* The U.S. currency has been altered several times in the recent past. For example, the two-dollar bill was introduced with only moderate success. The Susan B. Anthony dollar coin was a failure in the late 70s. Both of these situations could be examined as case studies; and the Fed could attempt to learn from past successes and/or failures.

(c) *Focus groups interviews.* There is some secrecy that surrounds the Fed's decision making process. Thus, a mend? Justify your choice.[5]

6. The leadership of the Boy Scouts of America (BSA) is concerned about several issues related to their membership. These issues include low retention rates among members (e.g., many scouts quit after only one or two years); high turnover rates among leaders; and declining membership in some regions (e.g., in large urban areas). Design a research program to assist BSA in assessing and reversing the trends.[6]

large scale survey may not be appropriate. A focus group interview provides a controlled setting, where the major actors involved in currency decisions could be allowed to express their opinions. It would probably be beneficial to interview each group of decision makers separately: (1) consumers; (2) bank employees; (3) store personnel. It would also make sense to conduct focus groups separately for men and women, as these groups handle currency quite differently (e.g., purses versus pockets).

[6] To study low retention rates, BSA could implement a policy of debriefing scouts who quit. This debriefing could take a variety of forms. For example, BSA could ask scout masters to talk with departing members and then file a short form with headquarters. Alternatively, if it is felt that this form would add too much work to already busy scout masters, then BSA could selectively interview departing scouts. These interviews could be in the form of surveys or could be conducted as focus group interviews. Here, BSA may also want to interview scouts who do not drop out, for comparison purposes. A major thrust of interviews with departing scouts would be to find out "why" the scouts have decided to end their membership.

Again, BSA may consider following up on all scout masters who resign. That is, a formal debriefing process could be implemented, and scout masters' reasons for resigning could be analyzed in great detail. Of course, such a debriefing program would be rather expensive. Alternatively, BSA may want to interview a selective number of scout master who resign. A face-to-face interview is one possibility. Focus group interviews or a mail-in survey would also be possible designs. A key question to ask departing scout masters is "Why did you resign?" This could be asked in an open-ended format, or (as BSA gains more knowledge about this phenomenon) closed-end options could be used in this question. Again, BSA may want to interview both scout masters who quit and those who remain on the job for a number of years, so as to have a basis for comparison.

Membership may be declining in urban areas because of high turn-over rates, so the research described in (1) and (2) above may provide some insight relative to this question. In addition, declining membership may indicate a recruiting problem. Thus, BSA may want to interview boys who join the counts and contrast this profile with boys who decide not to join. For example, BSA could ask non members: (a) if they are aware of scouting; (b) if they know personally any scouts or scout masters; (c) if they ever were scouts; (d) if they would consider joining the scouts; (e) if they have no interest in joining, why not?

Inferential Statistics: Important Terms

- *Unit* of analysis — the entity we wish to study, e.g., individuals, households, items, stores, corporations, divisions)
  - Household versus phone number example
- *Population* or *Universe* — collection of all units we wish to study, e.g., individuals who are 18 years or older living in Japan
- *Parameter* — a numerical fact about a population, e.g., average age of "everyone," average purchase intention
- *Survey* — collecting information from units in a population
  - *Census* — a survey of *all* units in the population
  - *Sample* — a survey of *some* units in the population
- *Statistic* — a numerical fact *about a sample* that approximates a parameter, e.g., average age computed from a sample
- **Your turn**: suppose we will conduct a poll to determine who will win an election next week. In this context, what specifically is the unit, population, parameter and statistic?

Parameters and Statistics

Name	Parameter	Statistic	Defined
Mean	μ	$\bar{x}$	47, 56
Total	T	$\hat{T}$	128
Proportion (percent)	π	p	108, 118
Standard deviation	σ	s	56, 52
Variance	σ^2	s^2	56, 52
Correlation	ρ	r	Ch. 11
Intercept	α	a	Ch. 11
Slope	β	b	Ch. 11

- Prepend the word "population" or "sample" before each name
  - Parameter μ is the "population mean" and statistic $\bar{x}$ is the "sample mean"
  - Parameter π is the "population percent" and statistic p is the "sample percent"
- Unfortunately, there are two conventions for proportions. Some books, including Siegel and Churchill, use π for the parameter and p for the statistic, while others use p for the parameter and $\hat{p}$ for the statistic. Beware if you are reading from other sources!

Variable Classifications

- *Categorical variable*[7] — takes "nonnumeric" values, e.g., race, brand choice, marital status, postal code
  - *Nominal variable* — values can be divided into categories
  - *Ordinal variable* — values can be divided into categories and ordered
- *Numerical variable*[8]— can be measured with numbers, e.g., age, income, sales, number of children
  - *Interval variable* — values be ranked and for which the difference between two values can be calculated and interpreted (but zero arbitrary)
  - *Ratio variable* — interval data with a meaningful zero

Other useful terms:

- *Dichotomous variable* (also *Binary variable*) — takes only two possible values, e.g., yes or no
- *Count variable* — values can be 0, 1, 2, . . . , e.g., number of purchases, number of kids
- *Amount variable* — a quantity of something. Amounts can never be negative, e.g., dollars last year, weight

[7] Statisticians also use the term *qualitative* variable synonymously. I avoid this usage because it could be confused with *qualitative research* methods.

[8] SPSS uses the term *scale*, which I avoid because scale has a very specific meaning in psychometrics and many numerical variables are not "scales."

Measurement Level Determines Methods

- Nominal: mode
- Ordinal: mode, median
- Interval: mode, median, mean, standard deviation
- Ratio: all, including ratios (e.g., percent change, coefficient of variation)

Example: temperature (degree F or C) is interval, but not ratio. We are allowed to compute means and standard deviations, but not percent changes, etc.

Yesterday	Today	Percent Change?
50°F	68°F	$\cancel{\frac{68-50}{50} = 36\%}$
10°C	20°C	$\cancel{\frac{20-10}{10} = 100\%}$

Preview (rest of course):

Independent Variable	**Dependent Variable**	
	Categorical	**Numerical**
Categorical	Crosstabs	*t*-tests / ANOVA
Numerical	Logistic regression Discrete choice	Correlations Linear regression
Mixed	Generalized linear model	General linear model

Your Turn: Restaurant Survey

Identify the types of variable that each question measures.

1. Please rate your overall dining experience (5-point scale, 1=poor, . . . , 5=excellent)
2. Was your visit today for lunch or dinner?
3. What main-menu item did you order today (fill in blank)?
4. Have you ordered this item before (Yes or no)?
5. What *one* area would you like to see improved?

   (a) Decor
   (b) Taste
   (c) Service
   (d) New items
   (e) Value
   (f) Other (specify)

6. How many times have you eaten at this chain?

   (a) First time
   (b) 2–4 times
   (c) 5–10 times
   (d) 11–20 times
   (e) 21+ times

7. How many people are in your party?
8. Are you (check either male or female)?
9. What is your zip code?

Comparative Advertising, Measurement Scales and Data Analysis

Suppose you see an advertisement that claims that *Vital* capsules are 50% more effective in easing tension than the leading tranquilizer, *Restease*. As research director of the firm that produces Restease, you immediately begin comparison tests. Using large sample sizes and a well-designed experiment, you have one group of individuals use Vital capsules and a second group use Restease. You then have each individual in each group rate the effectiveness of the brand they tried on a five-point scale as follows: "For easing tension, I found Vital (Restease) to be ("Very Effective," "Effective," "Neither Effective nor Ineffective," "Ineffective," or "Very Ineffective.")

For analysis, you decide to code the "very effective" response as +2; the "effective" response as +1; the "neither-nor" response as 0; the "ineffective" response as −1; and the "very ineffective" response as −2. This is a common way of coding data of this nature. You calculate the average response for Vital and Restease and obtain scores of 1.2 and 0.8, respectively. Because the 0.4 difference is 50% more than the 0.8 level obtained by your brand, you conclude that the claims for Vital are valid.

Shortly after reaching this conclusion, one of your assistants, who was also analyzing the data, enters your office with the good news that Vital was viewed as only 10.5% more effective than Restease. Immediately you examine his figures. He used the same data and made no computational mistakes. The only difference was that he assigned a "very ineffective" response at +1 and continued up to a +5 for the "very effective response," so that the average scores for Vital and Restease were 4.2 and 3.8, respectively. This is also a widely used procedure.

Then, as you are puzzling over these results, another member of your department enters. She used the same approach as your assistant but assigned a +5 to "very ineffective" and a +1 to "very effective," giving averages for Vital and Restease of 1.8 and 2.2, respectively. Again, with no computational errors, she found Vital to be 18.2% more effective. What do you conclude? Whose results are correct?

Your Turn: Siegel Chapter 2
Homework Solutions

8. (a) secondary; (b) primary; (c) secondary

11. (a) employee; (b) multivariate; (c) salary and years experience are quantitative, gender and education are qualitative; (d) education; (e) cross-sectional. For part (d) experience and salary are ratio, which means that they have an ordering, in addition to meaningful differences and zero, but we would not call them "qualitative."

12. (a) production facilities; (b) multivariate; (c) part and quality are qualitative (group ID is the key variable that uniquely identifies a facility and key variables, such as student id or account number, are usually not considered measured variables); (d) yes, quality; (e) cross-sectional.

13. (a) months; (b) bivariate; (c) both are quantitative; (d) time-series

Summaries for Nominal Data

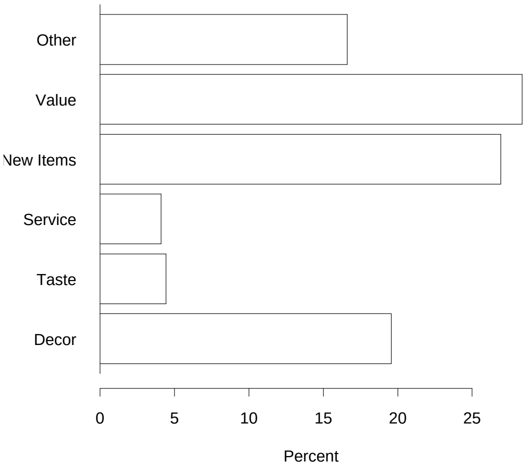
"What *one* area would you like to see improved?"

- *Frequency distributions* tell how often and how likely each possible value occurs
- Area and angles of *pie charts* proportional to counts and percents
- End position of bars, lengths, and areas of *bar charts* proportional to counts and percents
- Word Clouds. R: see wordcloud package.

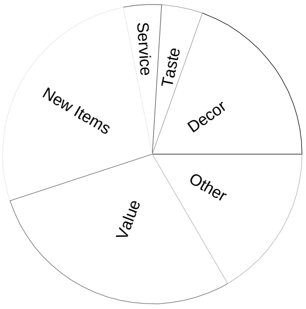
Frequency Distribution

Area to improve	Count	Percent
Decor	2297	19.5689
Taste of food	521	4.4386
Service	482	4.1063
New Items	3160	26.9211
Value	3329	28.3609
Other	1949	16.6042
Total	11738	100.00

Bar Chart



Pie Chart

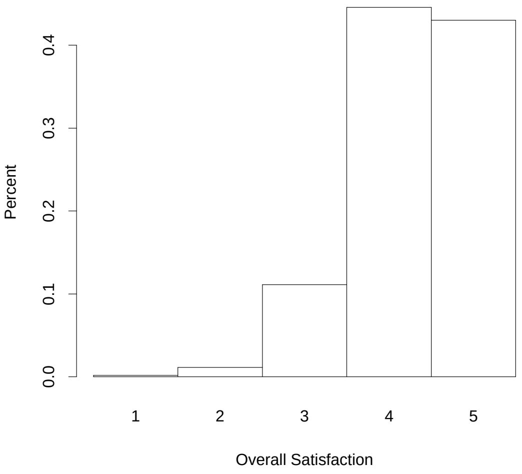


Restaurant Example

“Please rate your overall experience.”

```
   Frequency Distribution
                                Cum    Cum
# Q1          Freq   Pct   Freq    Pct
------------------------------------
1 Poor          27   0.2     27    0.2
2 Fair         178   1.1    205    1.3
3 Good        1779  11.1   1984   12.4
4 Very Good   7137  44.6   9121   57.0
5 Excellent   6889  43.0  16010  100.0

Single-number summaries
N              16010
Mean         4.29188
Std Dev     0.718262
Range              4
Q3-Q1              1
Mode               4
```

Histogram: Area of the bars proportional to the percents



```
                   Cum    Cum                                 Cum    Cum
AGE Freq    Pct   Freq    Pct             AGE Freq    Pct   Freq    Pct
-----------------------------             -----------------------------
14     3    0.0      3    0.0             56   266    1.4  13897   74.0
15     4    0.0      7    0.0             57   287    1.5  14184   75.5
18    40    0.2     47    0.3             58   222    1.2  14406   76.7
19    78    0.4    125    0.7             59   226    1.2  14632   77.9
20    71    0.4    196    1.0             60   219    1.2  14851   79.1
21   185    1.0    381    2.0             61   241    1.3  15092   80.4
22   177    0.9    558    3.0             62   239    1.3  15331   81.7
23   223    1.2    781    4.2             63   216    1.2  15547   82.8
24   242    1.3   1023    5.4             64   247    1.3  15794   84.1
25   246    1.3   1269    6.8             65   221    1.2  16015   85.3
26   259    1.4   1528    8.1             66   184    1.0  16199   86.3
27   328    1.7   1856    9.9             67   197    1.0  16396   87.3
28   350    1.9   2206   11.7             68   213    1.1  16609   88.5
29   366    1.9   2572   13.7             69   175    0.9  16784   89.4
30   371    2.0   2943   15.7             70   199    1.1  16983   90.5
31   394    2.1   3337   17.8             71   170    0.9  17153   91.4
32   457    2.4   3794   20.2             72   182    1.0  17335   92.3
33   494    2.6   4288   22.8             73   184    1.0  17519   93.3
34   492    2.6   4780   25.5             74   126    0.7  17645   94.0
35   469    2.5   5249   28.0             75   150    0.8  17795   94.8
36   519    2.8   5768   30.7             76   118    0.6  17913   95.4
37   532    2.8   6300   33.6             77   114    0.6  18027   96.0
38   525    2.8   6825   36.4             78   105    0.6  18132   96.6
39   532    2.8   7357   39.2             79   101    0.5  18233   97.1
40   541    2.9   7898   42.1             80    82    0.4  18315   97.5
41   560    3.0   8458   45.0             81    80    0.4  18395   98.0
42   439    2.3   8897   47.4             82    61    0.3  18456   98.3
43   486    2.6   9383   50.0             83    69    0.4  18525   98.7
44   461    2.5   9844   52.4             84    46    0.2  18571   98.9
45   531    2.8  10375   55.3             85    35    0.2  18606   99.1
46   403    2.1  10778   57.4             86    42    0.2  18648   99.3
47   378    2.0  11156   59.4             87    29    0.2  18677   99.5
48   356    1.9  11512   61.3             88    15    0.1  18692   99.6
49   374    2.0  11886   63.3             89    19    0.1  18711   99.7
50   373    2.0  12259   65.3             90    17    0.1  18728   99.7
51   299    1.6  12558   66.9             91    16    0.1  18744   99.8
52   257    1.4  12815   68.3             92    16    0.1  18760   99.9
53   308    1.6  13123   69.9             93     3    0.0  18763   99.9
54   243    1.3  13366   71.2             94     5    0.0  18768  100.0
55   265    1.4  13631   72.6             95     2    0.0  18770  100.0
                                          96     5    0.0  18775  100.0
```

Key Facts About Frequency Distributions

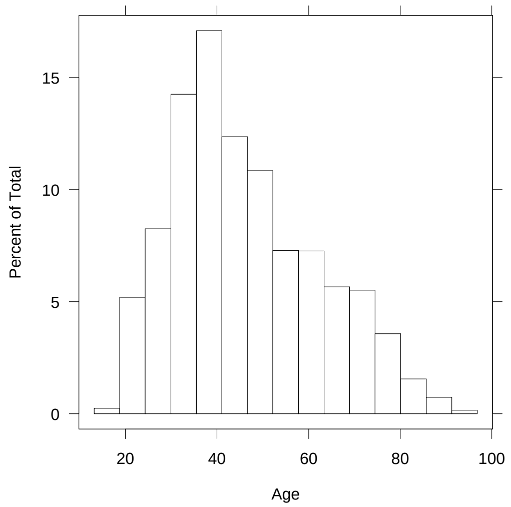
- *Freq* gives the *number* of people in the sample with a particular value, e.g., 3 people have age 14, 4 have age 15, etc.
- *Percent* gives the *percentage* of people in the sample with a particular value, e.g., 223/18775=1.2% of the sample has age 23
- *Cum Freq* gives the *number* of people *less than or equal* to a particular value, e.g., 47 people are 18 years old or younger (you can add up the frequencies 3+4+40=47)
- *Cum Pct* gives the *percentage* of people *less than or equal* to a particular value, e.g., 2.0% of the sample is 21 or younger
- The last entry in the "Cum Freq" column gives the sample size, e.g., 18,775 for the age example
- Note: the cumulative percents are used in the same way as the values from a normal table
- Your Turn:
  1. What percentage of people are 40 or younger?
  2. How many people are 65 or older?
  3. What percentage of people are 65 or older?
  4. What percentage of people are between 40 and 60 (inclusive[9])?
  5. What percentage of people are between 25 and 40 (inclusive)?

[9] Inclusive means including the endpoints, e.g., ages 40, 41, 42, . . . , 59, 60

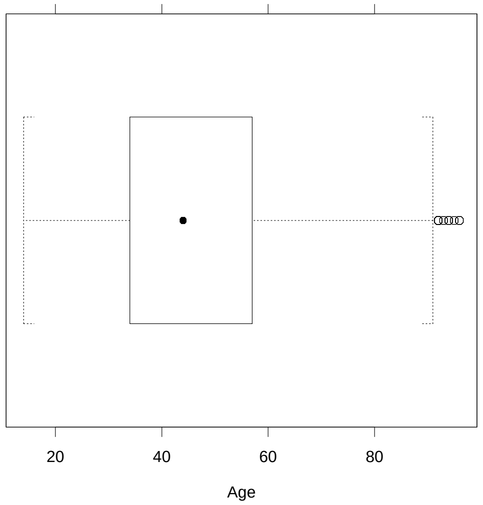
Age Example

```
Single-number summaries
    N               18775
    Mean            46.37
    Std Dev         15.84
    Skewness         0.55
    Range              82
    Q3-Q1              23

Percentiles     Value
    100% Max      96
     99%          85
     95%          76
     90%          70
     75% Q3       57
     50% Med      44
     25% Q1       34
     10%          28
      5%          24
      1%          20
      0% Min      14
```

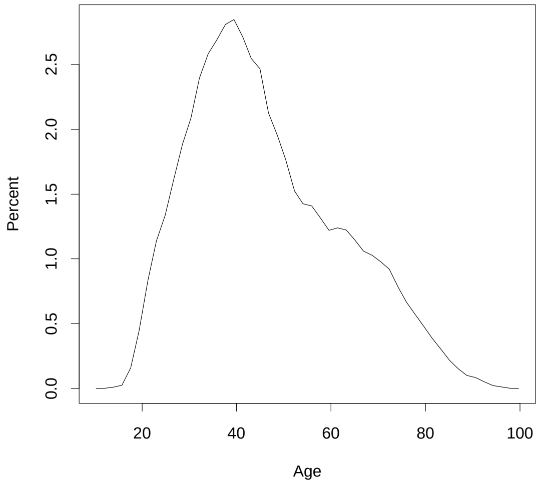
Histogram



Box Plot



Density Estimate

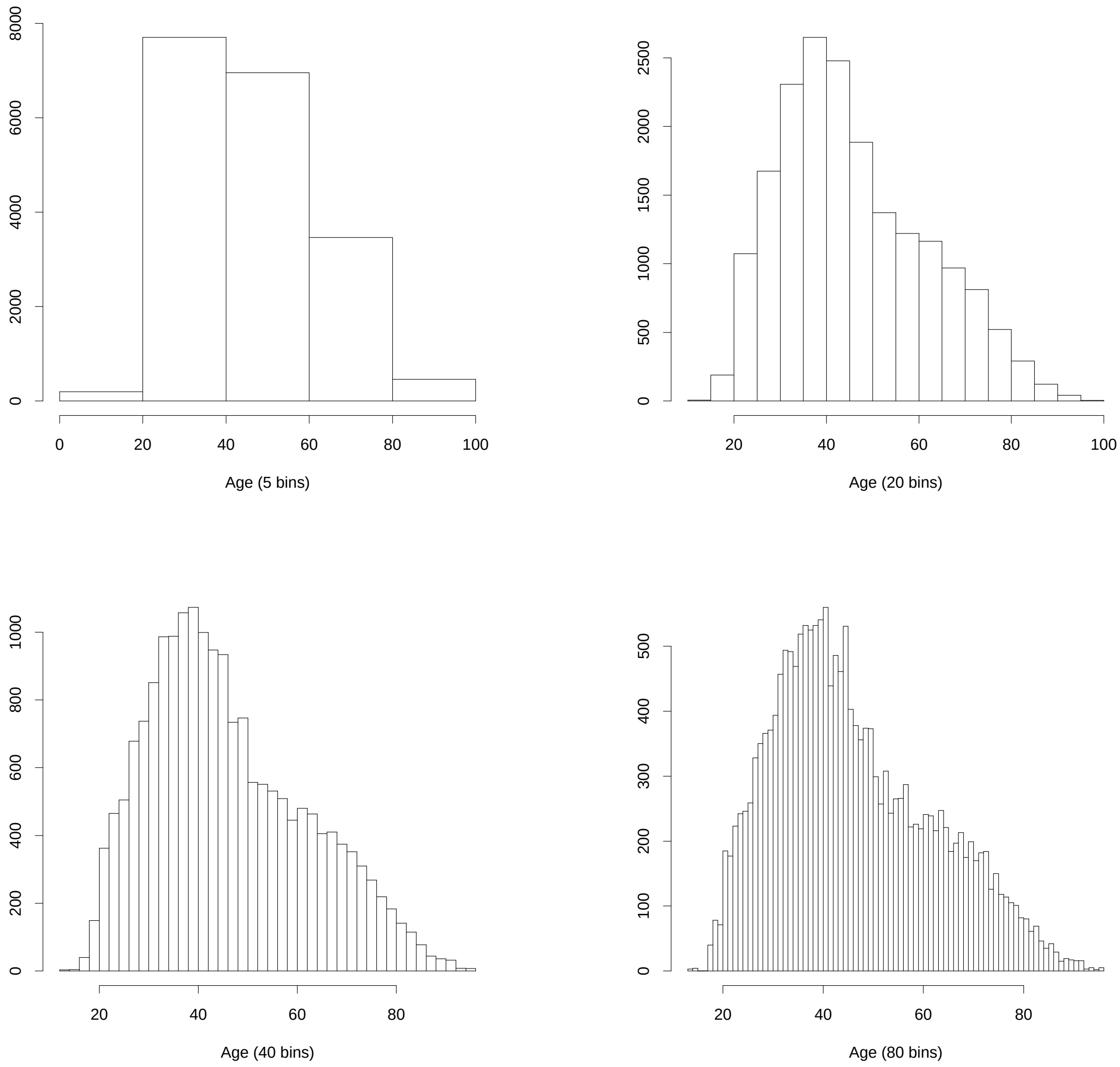


Key Facts About Histograms

- To make a histogram "by hand"
  1. Form bins and compute frequency distribution for bins
  2. Plot a bar for each bin so that the *area* of each bar is proportional to the count/percent of observations
- Note: if all bins have the same width, the height and end of the bars also represent the count/percents, although in general the area is the only interpretable quantity
- In chapter 7 we will find the area under portions of a normal distribution
- Limitations of histograms: the shape depends on the number of bins, the starting position of the bins, and the specific widths of the bins — if you change any of these the histogram will change
- You (or software) select the bin width — narrow bins give a jagged histogram and wide bins give a smooth one
- The bin width requires some judgement. See Wikipedia, "Number of bins and width"

Age Example

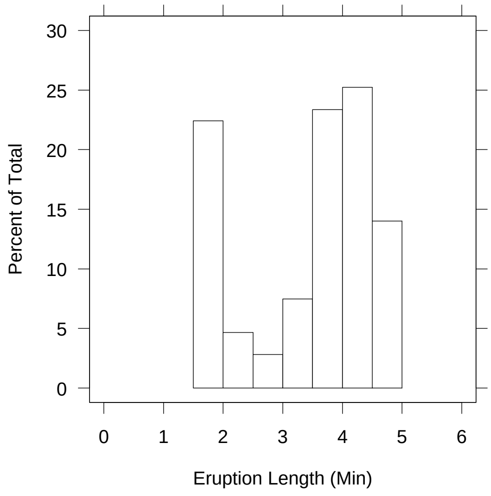
How many bins?

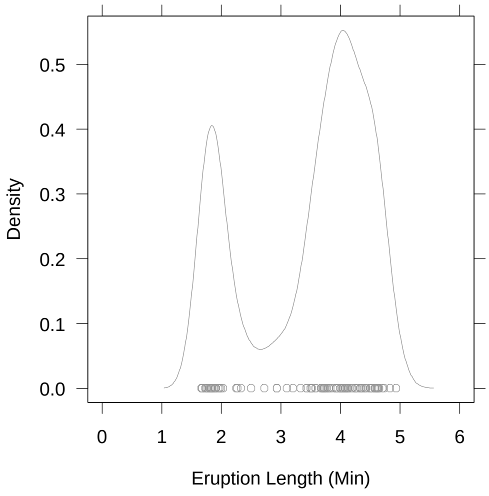


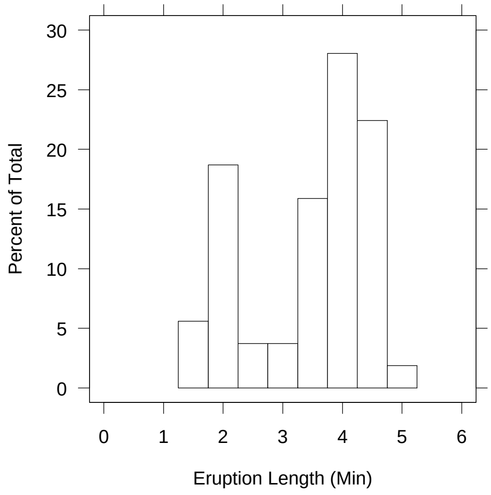
Key point: the more bins the more jagged the histogram

Old Faithful Geyser Example

4.37	4.70	1.68	1.75	4.35	1.77	4.25	4.10	4.05	1.90	4.00	4.42	1.83	1.83
3.95	4.83	3.87	1.73	3.92	3.20	2.33	4.57	3.58	3.70	4.25	3.58	3.67	1.90
4.13	4.53	4.10	4.12	4.00	4.93	3.68	1.85	3.83	1.85	3.80	3.80	3.33	3.73
1.67	4.63	1.83	2.03	2.72	4.03	1.73	3.10	4.62	1.88	3.52	3.77	3.43	2.00
3.73	4.60	2.93	4.65	4.18	4.58	3.50	4.62	4.03	1.97	4.60	4.00	3.75	4.00
4.33	1.82	1.67	3.50	4.20	4.43	1.90	4.08	3.43	1.77	4.50	1.80	3.70	2.50
2.27	2.93	4.63	4.00	1.97	3.93	4.07	4.50	2.25	4.25	4.08	3.92	4.73	3.72
4.50	4.40	4.58	3.50	4.33	4.13	1.95							







- This is an example of a *bimodal distribution*

Number-of-Times Example

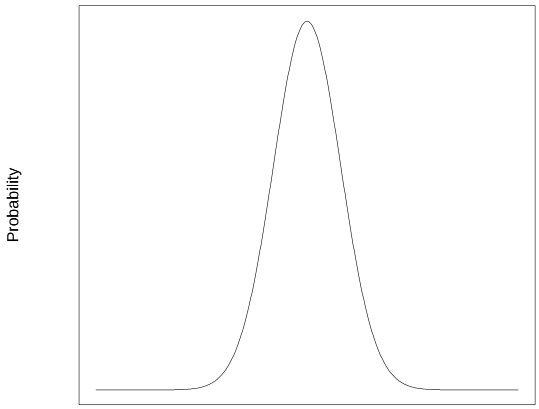
“How many times have you visited any other casual dining restaurant in the past 30 days?” What features of the distribution do not make sense?

```
       Frequency Distribution
                        Cum     Cum
Q14A Freq     Pct    Freq     Pct
------------------------------
 0      49    0.4      49     0.4
 1     473    3.7     522     4.1
 2     935    7.4    1457    11.5
 3     898    7.1    2355    18.6
 4    1051    8.3    3406    26.9
 5    1708   13.5    5114    40.3
 6     681    5.4    5795    45.7
 7     295    2.3    6090    48.0
 8     621    4.9    6711    52.9
 9     110    0.9    6821    53.8
10    2533   20.0    9354    73.8
11      30    0.2    9384    74.0
12     342    2.7    9726    76.7
13      21    0.2    9747    76.9
14      31    0.2    9778    77.1
15    1045    8.2   10823    85.4
16      39    0.3   10862    85.7
17      19    0.1   10881    85.8
18      54    0.4   10935    86.3
19      12    0.1   10947    86.3
20     981    7.7   11928    94.1
21      10    0.1   11938    94.2
22      12    0.1   11950    94.3
23       4    0.0   11954    94.3
24      12    0.1   11966    94.4
25     170    1.3   12136    95.7
```

```
                        Cum     Cum
Q14A Freq     Pct    Freq     Pct
------------------------------
26      10    0.1   12146    95.8
27       8    0.1   12154    95.9
28      22    0.2   12176    96.0
29      16    0.1   12192    96.2
30     340    2.7   12532    98.8
31       3    0.0   12535    98.9
32       4    0.0   12539    98.9
34       2    0.0   12541    98.9
35      12    0.1   12553    99.0
36       2    0.0   12555    99.0
37       2    0.0   12557    99.0
40      36    0.3   12593    99.3
41       1    0.0   12594    99.3
43       1    0.0   12595    99.3
45      15    0.1   12610    99.5
49       1    0.0   12611    99.5
50      32    0.3   12643    99.7
54       1    0.0   12644    99.7
55       1    0.0   12645    99.7
56       1    0.0   12646    99.7
59       1    0.0   12647    99.8
60      21    0.2   12668    99.9
62       1    0.0   12669    99.9
68       1    0.0   12670    99.9
70       3    0.0   12673   100.0
75       1    0.0   12674   100.0
99       4    0.0   12678   100.0
```

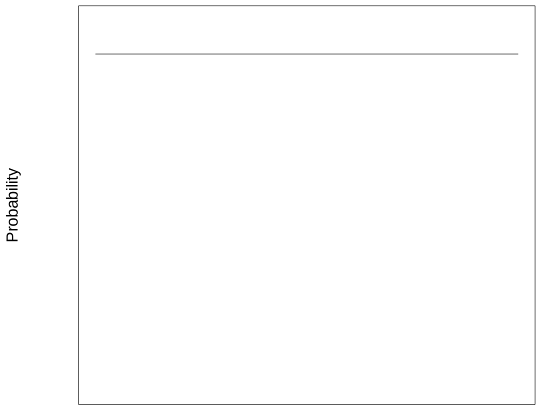
Shapes of Distributions

Normal distribution



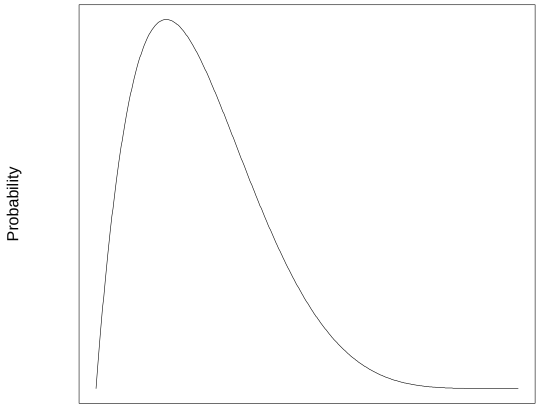
Skewness=0

Uniform distribution



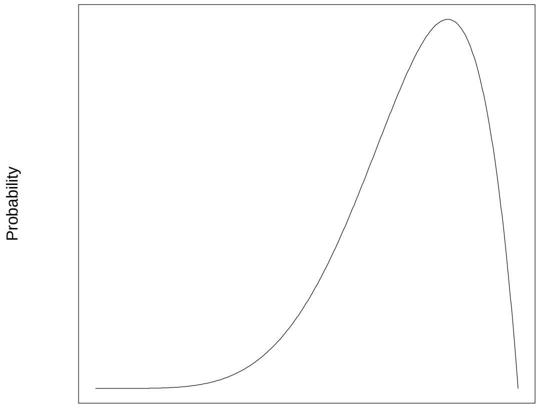
Skewness=0

Right (positive) skew



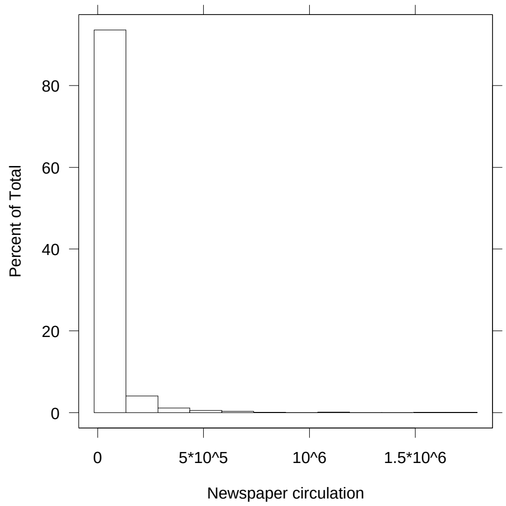
Skewness=0.7

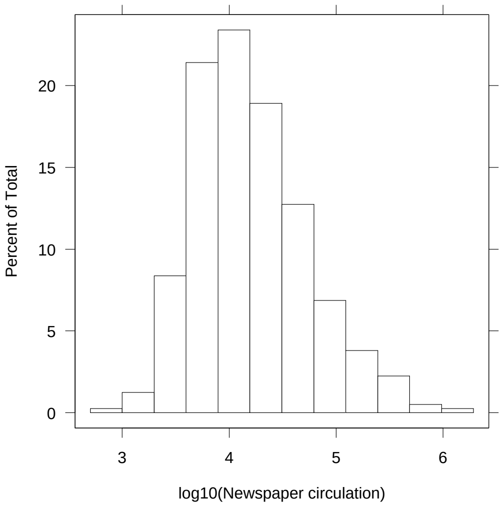
Left (negative) skew



Skewness=−0.7

Total Circulation of 1602 Newspapers in US





- Problem: almost all the data occupy a tiny fraction of the graph — can't tell much about the distribution

- Solution: take *logarithms* to bring in tail

- Definition: $x = b^y \Longleftrightarrow \log_b x = y$

x	$\log_{10} x$
$10^{-2} = 0.01$	-2
$10^{-1} = 0.1$	-1
$10^0 = 1$	0
$10^1 = 10$	1
$10^2 = 100$	2
$10^3 = 1000$	3
$10^4 = 10,000$	4
$10^5 = 100,000$	5

x	$\log_2 x$
$2^{-2} = 0.25$	-2
$2^{-1} = 0.5$	-1
$2^0 = 1$	0
$2^1 = 2$	1
$2^2 = 4$	2
$2^3 = 8$	3
$2^4 = 16$	4
$2^5 = 32$	5

- Properties of logs: $\log(xy) = \log x + \log y$, $\log x^a = a \log x$, $\log_b x = \log_a x / \log_a b$

- See Kahn Academy's introduction to logs.

Your Turn: Practice Logarithm Problems

Compute the following logarithms without using a calculator or computer.

1. $\log_{10} 100$
2. $\log_{10} 0.01$
3. $\log_2 8$
4. $\log_2 256$
5. $\log_2 1/2$
6. $\log_2 1/8$
7. $\log_{1/2} 8$
8. $\log_{1/2} 1/8$
9. $\log_5 25$
10. $\log_{1/5} 25$
11. $\log_{10} 10^7$
12. $\log_2 8^8$

Answers: (1) 2; (2) -2; (3) 3; (4) 8; (5) -1; (6) -3; (7) -3; (8) 3; (9) 2; (10) -2; (11) 7; (12) 24.

A student emailed me the following question about logs. My response is below.

Question: Why do we take logs of variables and how can we use the logged version to draw conclusions about unlogged data?

Answer:

> "One of the problems with skewness in data is that, as mentioned earlier, many of the most common statistical methods ... require at least an approximately normal distribution. When these methods are used on skewed data, [(1)] the answers may well be misleading or just plain wrong. [(2)] Even when the answers are basically correct, there is often some efficiency lost; essentially, the analysis has not made the best use of all the information in the data set (Siegel, p. 68, Fifth Edition)."

Siegel gives two reasons, which are both correct, but he does not elaborate. This note elaborates further.

Different pieces of information often have different levels of reliability. Some data are more trustworthy than others. In statistical terms this means that the "variance" of one observation (which is a realization of some random variable) is larger than another — the variance of the random variable generating the former observation is larger than the latter). When this is the case, giving the observations equal weight is not as "efficient" (i.e., the final estimate will have larger variance) as giving the more reliable observation more weight.

With amount and count variables, larger numbers tend to be less stable than smaller numbers. Statisticians would say that the variance of a count/amount random variable increases with its mean.[10] For example, consider two customers, one who orders \$100,000 on average each year and another that orders \$1000. Which customer would have larger variance? Common sense tells us that the larger (100,000) customer is likely to have larger variance. It would not be surprising if orders had a standard deviation 1000 (note the coefficient of variation is only 1% of the mean) from year to year or even more. A standard deviation of \$1000 for the smaller customer would give a CV=100%. The point is that larger customers tend to have larger fluctuations as measured by variance or standard deviation than smaller customers. Observations with larger variation are less reliable, which implies weighting the observation less in analyses.

Statisticians have two approaches for dealing with this problem: transformations and weighting.[11] Both approaches have the same effect of reducing the influence of less reliable observations. The logarithm and square root are sometimes called *variance stabilizing transformations* because the intent is to make the variance of observations equal. We'll call this *homoscedasticity* when we get to t-tests, ANOVA, and regression. Statisticians[12] recommend using the square root whenever the variance of a random variable is proportional to its mean, and the logarithm whenever the standard deviation is proportional to the mean.

[10] The Poisson distribution, covered in §7.5, is a classic example. We will not discuss this distribution in class, but it turns out that the mean of a Poisson random variable equals its variance.

[11] For example, see Carroll and Ruppert (1988), *Transformation and Weighting in Regression*, Chapman and Hall.

[12] e.g., see Tamhane and Dunlop (2000), *Statistics and Data Analysis*, Prentice Hall, pp. 377–379.

Your Turn: Siegel Chapter 3
Homework Solutions

Note: SPSS cannot read percent signs. If you copy/paste from Excel, change the format so that there no percent sign before copying. If you type the numbers directly, type, e.g., .0523 instead of 5.23%.

1. approximately normal
2. approximately normal with an outlier
3. right skewed
4. bimodal
5. (a) 2; (b) 3; (c) No, you can only tell that it lost more than 50% of its value but not more than 100%; (d) 24; (e) right skewed with 3 outliers
6. (R video) (a) Use SPSS; (b) between 5.1% and 5.3%; (c) approximately normal with two outliers
7. (a) Use SPSS; (b) omit; (c) Shares have appreciated typically by about 10% to 20%
9. (a) Use SPSS; (b) right skewed with most under 50,000 but a few outliers; (c) right skewed; (d,e) use SPSS; (f) symmetric
19. Use SPSS; right skewed
20. (a,b) use SPSS; (c) approximately normal; (d) log makes it symmetrical
21. (a) Use SPSS; (b) approximately normal with two outliers; (c) outliers are 25 and 41; (d) Use SPSS; (e) approximately normally distributed, extending from 0 to 7 with a typical value around 3.
22. (a) Use SPSS; (b) approximately normal
23. (a) Use SPSS; (b) approximately normal
24. (a) Use SPSS; (b) right skewed
25. (a) Use SPSS; (b) right skewed
26. (a) Use SPSS; (b) right skewed
27. (a) Use SPSS; (b) approximately normal
28. (a) Use SPSS; (b) right skewed
29. (video) (a) Use SPSS; (b) approx normal with an outlier
30. (a) Use SPSS; (b) approx normal; it seems to matter very much where you have a prescription filled, in terms of how much you will pay for the same medication.

Measures of Central Tendency

Let $x_1, \ldots, x_N$ be measures of a quantitative variable

- *Measures of central tendency* describe the location of the *center* of a distribution and represent a "typical" value.
- The *population mean* is (Excel: `AVERAGE`)

$$\mu = \frac{\text{Sum of values}}{\text{Number of values}} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- *Median*: a number such that half the values are greater than or equal to the number, and half are less than or equal to the number (Excel: `MEDIAN`)
  - If N is odd, the median is the $(N+1)/2$ largest value, e.g., 2 is the median of 1, 2, 3
  - If N is even, the median is (by convention) the average of the $N/2$ and $N/2+1$ largest values, e.g., 2.5 is the median of 1, 2, 3, 4
- *Mode*: the value occurring most frequently (Excel: `MODE`)
- The median is more *robust* to outliers than the mean
- It is meaningful to compute the median of an ordinal variable, but not the mean
- Supplementary video

Means of Dichotomous Variables

- Normally it makes no sense to compute the mean of a nominal-level variable.
- It is valid to compute the mean of a dichotomous variable when it takes values 0 and 1:

  The sum of a 0-1 variable is the number of 1's. The mean of a 0-1 variable is the fraction of 1's.
- For example, consider the data 0, 0, 0, 1, 1.
  - The sum is two, indicating there are two 1's.
  - The mean is $2/5 = 40\%$, indicating that 40% of the values are 1's.
- 0-1 variables are examples of *dummy variables*, which will be covered in chapter 12. They are also called *indicator variables*.

Trimmed Mean Example

Compute the 10% trimmed and Winsorized means[13] of the following 10 numbers:

$$1 \quad 2 \quad 2 \quad 3 \quad 3 \quad 4 \quad 4 \quad 4 \quad 5 \quad 20$$

- The simple mean is

$$\frac{1+2+2+3+3+4+4+4+5+20}{10} = 4.8$$

- To compute the 10% *trimmed mean*, drop 10% of the observations in each tail:

$$\frac{2+2+3+3+4+4+4+5}{8} = 3.375$$

[13] Software packages are not consistent in their definitions of the 10% trimmed mean. Some trim 10% from each tail while other trim 5% from each tail for a total of 10%.

Measures of Position

- *Measures of position*: measured by *quantiles*, which partition a sorted list into (roughly) equally-sized groups.
  - *Quartiles*: 4 groups (denoted by Q_1, Q_2, and Q_3)
  - *Quintiles*: 5 groups
  - *Deciles*: 10 groups
  - *Demideciles*: 20 groups
  - *Percentiles*: 100 groups
  - What measure of central tendency divides the list into two equally-sized groups?
  - Quantiles are sometimes used to form bins
- In addition to being a dividing point between groups, quantile can also refer to the group itself, e.g, "the student scroed in the top quartile."
- Excel: `QUARTILE` and `PERCENTILE`
- SPSS: Analyze / Descriptives / Frequencies or Explore

Estimating Quantiles

- Statisticians have not agreed on the "best" way to estimate quantiles from a sample.[14]
- SAS offers five options, R offers nine, and SPSS offers at least two.
- **The answers in textbook may not match exactly what your software reports.**
- In practice, use a standard statistical package and be consistent, e.g., when comparing the quantiles of two segments, use the same quantile definition.
- Exams may ask you to find a quantile from a frequency distribution. Here is a simple rule:[15] **look for the first cumulative percent that is greater than or equal to the percentile you want to find.**
- The other methods improve on the simple rule by interpolating between numbers.
- Big data issue: computing quantiles (including median) is computationally expensive because sorting is required (means and standard deviations require a single pass)

[14] See Hyndman and Fan (1996), "Sample Quantiles in Statistical Packages," *American Statistician*, 50:4, pp. 361–4.

[15] I like this definition because it emphasizes the basic idea of a quantile without requiring any formulas. The default method in SPSS and definition 4 in SAS are preferred over my simple definition according the Hyndman and Fan.

Measures of Dispersion

- Measure how heterogeneous the population is:
  - Large values indicate a *heterogeneous* population.
  - Small values indicate a *homogeneous* population.

- *Population variance:* "average squared deviation from the mean" (Excel: `VARP`)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- *Population standard deviation*

  $\sigma = \sqrt{\sigma^2}$ (Excel: `STDEVP`)

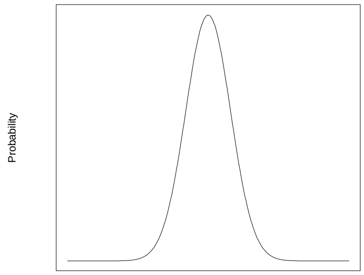
- *Range*

  Maximum − Minimum

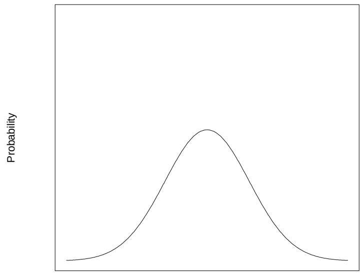
- *Interquartile range*

  $IQR = Q_3 - Q_1$

- IQR robust to outliers



```
Std dev = 7      IQR = 5
Range = 24.5
```



```
Std dev = 16     IQR = 12
Range = 56
```

Example

Total sales (in thousands of dollars) of $N = 6$ stores during a particular week are 10, 8, 14, 20, 11 and 9. Compute the mean and standard deviation (σ).

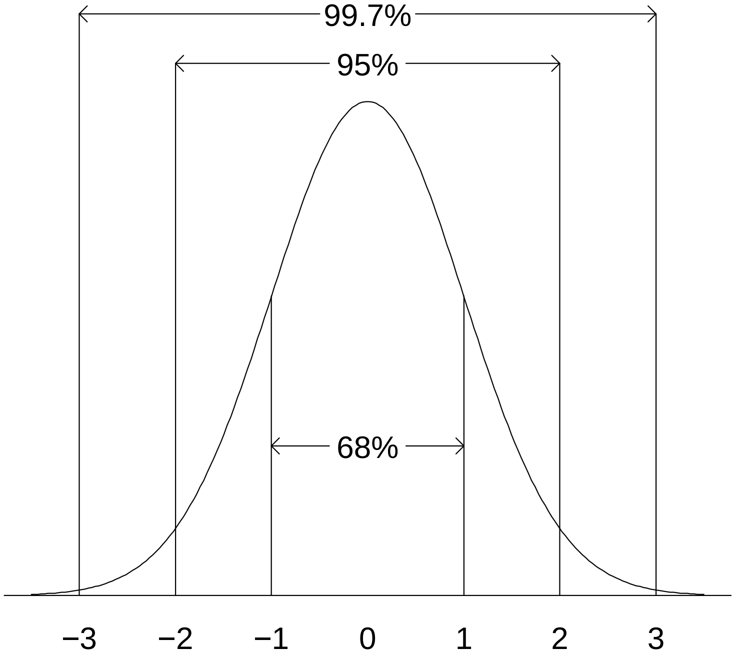
Store number	**Sales (thousands)**	**Squared deviation from mean**
1	$x_1 = 10$	$(10 - 12)^2 = 4$
2	$x_2 = 8$	$(8 - 12)^2 = 16$
3	$x_3 = 14$	$(14 - 12)^2 = 4$
4	$x_4 = 20$	$(20 - 12)^2 = 64$
5	$x_5 = 11$	$(11 - 12)^2 = 1$
6	$x_6 = 9$	$(9 - 12)^2 = 9$
Totals	$\sum x_i = 72$	$\sum (x_i - \mu)^2 = 98$

- The mean is $\mu = 72/6 = 12$
- The *popularion variance* is $\sigma^2 = 98/6 \approx 16.33$
- The standard deviation is $\sigma = \sqrt{98/6} \approx 4.04$
- The *sample* variance is $s^2 = 98/(6 - 1) = 19.6$

Empirical Rule

For data from a **normal distribution**, approximately

- 68% of the observations will fall within 1 standard deviation of the mean
- 95% of the observations will fall within 2 standard deviations of the mean
- 99.7% of the observations will fall within 3 standard deviations of the mean



More Single-Number Summaries

1. *Coefficient of variation* (CV) measures relative variation

   $$\text{CV} = \frac{\sigma}{\mu}$$

   Notes: (1) You must have a ratio scale to compute CV and (2) the units cancel out and CV is a "unitless" measure.

2. *Skewness* (Excel: `SKEW`)

   - Negative values indicate left (negative) skew (see slide 42)
   - 0 indicates symmetric distributions (e.g., normal or uniform)
   - Positive values indicate right skew

Statistics From a Sample

Until now we have talked about summaries from a population:

- N is the number of elements in the population
- μ is the population mean
- σ is the population standard deviation

We use different symbols if the data are from a *sample*:

- n is the number of elements in the sample $x_1, x_2, \ldots, x_n$
- $\bar{x}$ is called the *sample mean* (Excel: `AVERAGE`)

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

- s^2 is called the *sample variance* (denominator slightly different) (Excel: `VAR`)

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

- *Sample standard deviation* (Excel: `STDEV`) $s = \sqrt{s^2}$

Most computer packages assume data are from a sample and report s instead of σ

Computing Means and Standard Deviations from Frequency Distributions Using Weights

Recall from page 34 (assume this is the population):

```
   Frequency Distribution
                                Cum    Cum
# Q1          Freq    Pct     Freq    Pct
------------------------------------
1 Poor          27    0.2       27    0.2
2 Fair         178    1.1      205    1.3
3 Good        1779   11.1     1984   12.4
4 Very Good   7137   44.6     9121   57.0
5 Excellent   6889   43.0    16010  100.0

Mean         4.2919           Variance    0.5159
```

$$\begin{aligned}
\mu &= \frac{1(27)+2(178)+3(1779)+4(7137)+5(6889)}{16010} \\
&= 1(0.2\%)+2(1.1\%)+3(11.1\%)+4(44.6\%)+5(43.0\%) \\
&= 4.2919
\end{aligned}$$

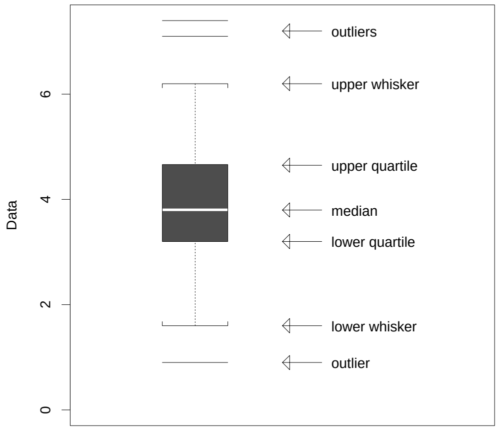
$$\begin{aligned}
\sigma^2 &= \frac{27(1-4.29188)^2+178(2-4.29188)^2+\cdots+6889(5-4.29188)^2}{16010} \\
\sigma^2 &= 0.2\%(1-4.29188)^2+1.1\%(2-4.29188)^2+\cdots+43\%(5-4.29188)^2 \\
&= 0.5159
\end{aligned}$$

$$\mu = \sum_x xP(x) \qquad \text{and} \qquad \sigma^2 = \sum_x P(x)(x-\mu)^2$$

where $P(x)$ is the percentage of cases that have the value x.

Box-and-Whisker Plots

1. Find quartiles (Q_1, Q_2, and Q_3)
2. Draw box—ends at Q_1 and Q_3, middle line at Q_2
3. Plot whiskers:
   - the top whisker is the value of the largest observation that is less than or equal to $Q_3 + 1.5 \times \text{IQR}$.
   - the lower whisker is the value of the smallest observation that is greater than or equal to $Q_1 - 1.5 \times \text{IQR}$.
4. Plot outliers



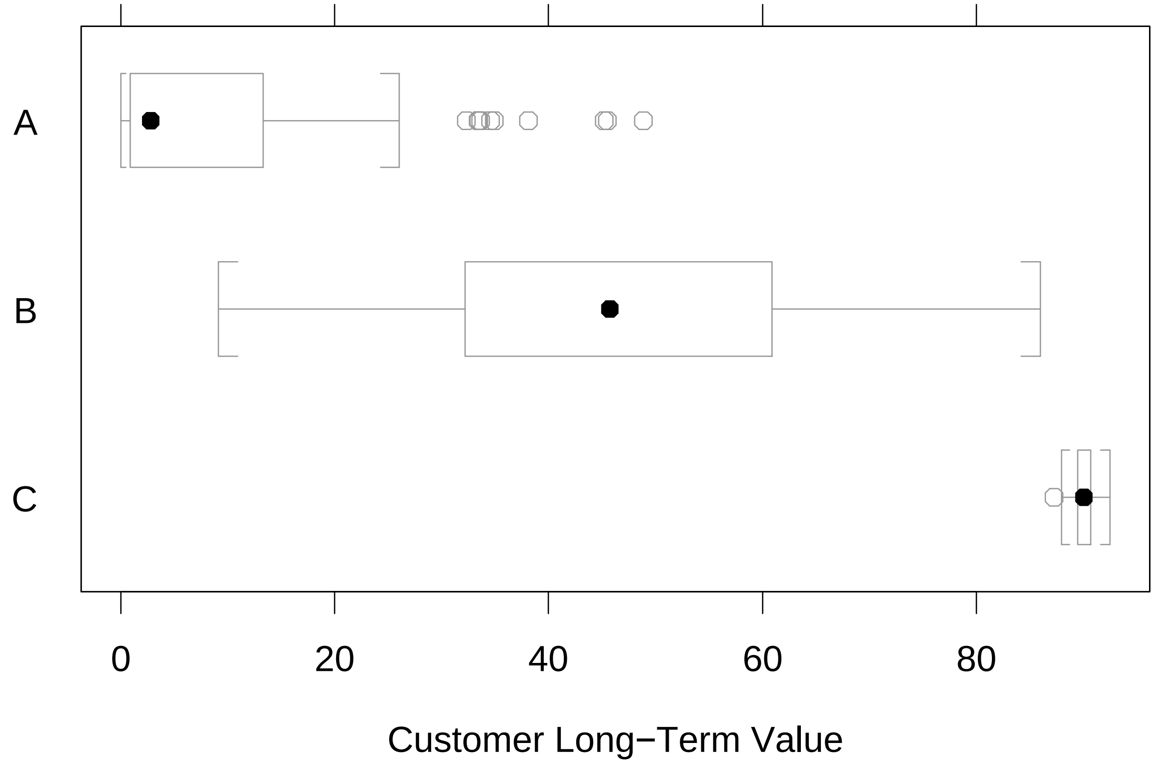
Key points:

- Dot (line) in box shows middle (median) of data (distribution)
- "Middle half" of data in box
- "Most" of data between whiskers
- Dots (lines) show outliers

How to make boxplots in SPSS

Example Boxplots

Consider three segments of customers (A, B and C)



Outliers

- Definition: an *outlier* or *extreme value* is an observation whose value is very large or very small compared with those of the majority of the observations
- Examples:
  - Number-of-times example
  - Donald Trumps' income
- Effects: can dominate analysis
  - Graphs — bulk of data in small region
  - Strong influence on means, standard deviations, regression estimates, cluster analyses, etc.
- Detection:
  - Boxplots show them
  - Compare maximum (minimum) value with upper (lower) quantiles

What should I do about outliers?

First answer this question: **Why is the value an outlier?**

- Erroneous value, e.g., coding error, data-transfer errors, respondent errors (misunderstanding question or deliberate sabotage): if possible, fix the value; otherwise set to missing. (Wendy's example)
- Correct but extreme value:
  - If observation is identifiably different from the rest of the data then form segments, e.g., business-to-business customers versus consumers or high, medium, low profit customers
  - Use robust statistics, e.g., median, IQR, trimmed or Winsorized statistics (maintains original units)
  - Transform variable, e.g., log or square root (different units, but sometimes log units[16] are more "natural," e.g., Richter scales, star intensities. This happens when a unit increase corresponds to a multiplicative, rather than additive, effect on the thing being measured, e.g., an earthquake of magnitude 8 is 10 *times* worse than one of magnitude 7.)[17]
  - Survey researchers often group values on the instrument, e.g., 1–5 times, 6–10 times, 11-15 times, 16–20 times, 21+ times (at extreme, dichotomize variable, e.g., users and non-users)
- Explain what you did and why you did it
- Conclusions shouldn't depend on arbitrary assumptions

[16] From Wikipedia entry of "Logarithmic scale:" Presentation of data on a logarithmic scale can be helpful when the data covers a large range of values; the logarithm reduces this to a more manageable range. Some of our senses operate in a logarithmic fashion (doubling the input strength adds a constant to the subjective signal strength), which makes logarithmic scales for these input quantities especially appropriate. In particular our sense of hearing perceives equal ratios of frequencies as equal differences in pitch.

[17] Logs transform *right*-skewed distributions to be less skewed (ideally normal), while trimming and Winsorizing do not have much effect on skewness (when there are outliers in both tails).

Effects of Rescaling the Data

- Suppose we have numbers $x_1, \dots, x_N$
- Consider the rescaled version

$$y_i = ax_i + b$$

  - if x_i is degrees Celsius on day i, then $y_i = 9/5x_i + 32$ is degrees Fahrenheit
  - if x_i is cents then $y_i = 100 \times x_i$ is dollars

- The following are true:
  - $\mu_y = a\mu_x + b$
  - $\sigma_y = |a|\sigma_x$
  - $\sigma_y^2 = a^2\sigma_x^2$
  - What do you think happens to the median, IQR, range, min, and max values?

- Example: Suppose you produce widgets. Your fixed costs are \$10,000 per month (including all salaries, rent, etc.). The marginal cost of the materials to make a widget is \$10 and the price of a widget it \$20. On average you sell 1200 widgets per month with a standard deviation of 150 widgets. Compute the mean profit per month and its standard deviation.

  Solution: Let x = number of widgets sold per month and $y = (20 - 10)x - 10,000 = 10x - 10,000$. Then $\mu_x = 1200$ and $\sigma_x = 150$. We can compute $\mu_y = 10 \times 1200 - 10000 = 2000$ and $\sigma_y = 150 \times 10 = 1,500$.

Rescaling Examples

1. (Siegel 5.28, 29) Your costs have been forecast has having an average of \$138,000 with a standard deviation of \$35,000. You have just learned that your suppliers are raising prices by 4% across the board. Now what are the average and standard deviation of your costs? Compare the coefficient of variation before and after the price increase. Why does it change in this way?

2. (Siegel 5.30, 31) You are sales manager for a regional division of a beverage company. The sales goals for your representatives have an average of \$768,000 with a standard deviation of \$240,000. You have been instructed to raise the sales goal of each representative by \$85,000. What happens to the mean and standard deviation? Compare the coefficients of variation.

3. Return to the comparative scales case. You coded very ineffective $= -2, \ldots,$ very effective $= 2$. Your colleague coded very inffective=1, ..., very effective $= 5$. You found the average for Restease to be 1.2. Assume further that the standard deviation is 1. What was the mean and standard deviation for your colleague? *Solution: Let x = the rating using your coding and y = your colleagues coding $= 1x + 3$. Then $\mu_x = 1.2$ and $\sigma_x = 1$. Compute $\mu_y = 1.2 + 3 = 4.2$ and $\sigma_x = |1| \times 1 = 1$.*

4. Your other colleague coded very ineffective=5, ..., very effective $= 1$. Find the mean and standard deviation. *Solution: Let $y = -1x + 3$. Compute $\mu_y = -1.2 + 3 = 1.8$ and $\sigma_x = |-1| \times 1 = 1$.*

5. Suppose you code very ineffective=0, ..., very effective $= 100$. Find the mean and standard deviation. *Solution: Let $y = 25(x + 2) = 25x + 50$. Compute $\mu_y = 25 \times 1.2 + 50 = 80$ and $\sigma_x = |25| \times 1 = 25$. Where did the coefficients come from? To many they will be "obvious." If not, one could solve the system of equations, $0 = -2a + b$ and $100 = 2a + b$, since -2 is mapped to 0, 2 to 100, etc.*

Siegel Chapter 4 Homework Solutions

1. (video) Use SPSS (a) 15.6; (b) 14; (c) use SPSS (d) omit; (e) using Tukey's Hinges, $Q_1 = 6$ and $Q_3 = 24.5$ (or $Q_1 = 5$ and $Q_3 = 25$); (f) 0, 34; (g) use SPSS (h) omit; (i) 30 using the simple rule or 31.6 using a weighted average in SPSS; (j) anything[18] from 84th to 90th percentile.

2. do by hand; $\mu = 94$; $Q_1 = 13$; $Q_2 = 32$; $Q_3 = 142$

5. (a) Use SPSS; Sixth edition: (b) 18.18%; (c) 19.6%; (d) It's left skewed and we'd expect the mean to to be less than the median; (e) omit; (f) SPSS reports 15%; (g) about 27%. Fifth edition: (b) 14.71%; (c) 15%; (d) approximately equal; (e) omit; (f) 12%, 18.6%; (g) 17.7%

21. 1.61%; 1.75%; 2%; It might be argued that the mode is the most useful description of the typical loan fee, since half the banks are charging this amount. However, the mode is also the largest value, an extreme value, and so one of the other summaries might be preferable.

25. (a) $\mu = 8.19\%$ and median $= 4.90\%$; (b) omit; (c)[19] 17.2%

26. 24.76 and 24; use SPSS

27. (a) -1.09%; (b) on average the dollar weakened; (c) -2.80%; use SPSS

28. 27 miles per gallon

29. (R video) Answers to fifth and sixth editions only: (a) 5.2175%; (b) 5.205%; (c) Q1=5.1225% and Q3=5.27%; (d) Min=4.91%, Max=5.83% see previous parts for Q1–Q3; (e) Use SPSS; (f) there are two outliers; (g) omit; (h) between 90% and 92.5% from a frequency distribution; (i) approximately 5.23%. Answers to #21 in 4th edition: (a) 5.99%; (b) 6.00%; (c) $Q_1 = 5.92\%$, $Q_3 = 6.08\%$; (d) 5.79%, 5.92%, 6%, 6.08%, 6.15%; (f) omit; (g) 5.90%; (h) 6.1%

30. 11.89%; 13.95%

31. (a) average = 36,483.90; (b) median = 7198; (c) average is much larger because of outliers in right tail; (d) min=317, Q1=1965.75, Q2=7198, Q3=31,960, max=364,299; (e) right skewed with outliers; (f) yes, we expect the average to be larger because right skewness.

32. 102.9; 104; median larger; left skew

39. (a) 4.39, 3; (b) 2.92, 3; (c) average is sensitive to outliers but median is not

[18] You will get different answers depending on the way you choose to compute a percentile. As indicated on page 51 of course packet, statisticians don't agree. If you form a frequency distribution with the cumulative column and use my simple rule, you will see that 28 is the 87 percentile and 30 is the 93 percentile. Therefore, the value 29 must be a percentile between the 87th and 93rd. In fact, it is about one half of the way between them, or the 90th percentile. Siegel's answer is 87th percentile, which he gets from the graph of a cumulative distribution (see the last part of chapter 4 in Siegel). You can get SPSS to compute various percentile under Analyze / Descriptive statistics/ Frequencies / Statistics. If you ask it to find various percentiles, e.g., 84, 85, 86, . . . , 90, you will find that the 84th percentile is 28.88 and the 85th percentile is 29.2. So 29 is about the 84th or 85th percentile. Finding a percentile based on such a tiny sample is going to be unstable. I'd give credit on an exam for anything between 84th and 90th percentile.

[19] Percentile estimates on small data sets are highly dependent on the particular definition of percentile you use. Siegel suggests that the 80th percentile is 17.2%. SPSS (Descriptives / Frequency distribution) gives 16.88. In R, `quantile(share, .8) = 0.1592`. The simple rule gives 15.6. Any of these would be counted correct.

Siegel Chapter 5 Homework Solutions

1. Use SPSS Sixth edition: (a) \$748.41 million; (b) \$253.55 million; (d) \$810.4 million; (f) 33.9% (no units because the cancel out); (g) see fifth edition; (h) 64288 million dollars squared; (i) see answer for fifth edition. Fifth edition: (a) \$49.5 million; (b) \$55.75 million; (d) 185; (f) 1.13; (g) size of budget for these firms typically varies from the average price by 113%; (h) 3,108; (i) typical squared deviation from the mean.

3. Multiply the numbers from question 1 by the exchange rate

6. (video) (a) 10.4; (b) 7.19; (c) 38, more than you would expect; (d) 43, close to what you would expect; (e) 44, close

7. (a) 9.70; (b) 5.53; (c) 29, close; (d) 41, close; (e) 44 close

8. Fifth edition: (a) +3%; (b) +3%; (c) +3%, (d) same. Fourth edition: (a) +4%; (b) +4%; (c) +4%, (d) same. Third edition: +5%, +5%, +5%, same.

9. (a) 186,400; (b) 35,750; (c) .3125 and .1918; (d) .8; (e) 4.8. 4th edition: (a) 126,400; (b) 17,000; (c) 31.3%, 13.4%; (d) .8; (e) 4.8

12. 225,000; 30,000

13. (a) 1.658; (b) .0330; (c) 1.663; (d) .0563; (e) no, variation increased

14. Sixth editon: (a) $\bar{x} = 434.39$; (b) $\sigma = 161.6$. Fifth edition: (a) 475.80; (b) 311.67

15. Sixth edition. (a) $\sigma_{NE} = 155.5$ and $\sigma_{SW} = 174.6$; (b) $\text{CV}_{NE} = 155.5/420 = 37\%$ and $\text{CV}_{SW} = 174.6/431.5 = 40\%$;

21. (a) 179.25; (b) 46.39.

22. (a) 10.9

23. Sixth edition: 238.7. Fifth edition: (a) 503.00

24. (a) 75.99; (b) 6.40; (c) 8.42%; (d) 19.3

25. (a) 23.13, 14.02; (b) 14.88 above the average and is about one standard deviation away from the average; (c) 68.88 above average and is 4.9 standard deviations away from the average; (d) first typical, second atypical

26. (a) 1.21; (b) 1.03; (c) 4.5; (d) 2.75 standard deviations above the mean

27. (a) 7.08%; (b) 15.69%; (c) 2.26 SDs below average

28. (video) 143,520, 36400

29. (video) 25.4% both before and after

30. (video) SD unchanged

31. (video) decrease from 31.3% to 28.1%

32. 5.13%

36. (R video) (a) .1711%; (b) .92%; (c) .1711/5.21.

37. 9.70, 33.7

38. (a) 75,354; (b) 363,982; (c) 75,354/36484; (e) 27 out of 30, or 90%. We expect 68% for a normal distribution, but this is not normal; (f) 29 out of 30, which is higher than the 95% we expect for a normal distribution.

40. (a) 7.08, 41; (b) 2.02, 7; (c) both sensitive

41. 129.08

43. .435%, 1%, .271

45. (a) 5.4; (b) 3.05; (c) 7; (d) 0.5647

46. (a) 13.8; (b) 8.53; (c) 18; (d) 0.6179

Chapter 5 case

Univariate Descriptive Statistics Class Exercises

1. A national survey included the question “How many years of education did you complete (not including kindergarten or pre-school)?” A frequency distribution of the responses is as follows:

```
                                Cumulative  Cumulative
EDUC   Frequency   Percent   Frequency     Percent
-----------------------------------------------------
   0         110       0.6          110         0.6
   1          16       0.1          126         0.7
   2          60       0.3          186         1.0
   3          95       0.5          281         1.5
   4          92       0.5          373         2.0
   5         115       0.6          488         2.6
   6         345       1.8          833         4.5
   7         222       1.2         1055         5.7
   8         661       3.5         1716         9.2
   9         471       2.5         2187        11.7
  10         734       3.9         2921        15.7
  11         758       4.1         3679        19.7
  12        7282      39.0        10961        58.8
  13        1230       6.6        12191        65.4
  14        1902      10.2        14093        75.5
  15         640       3.4        14733        79.0
  16        2410      12.9        17143        91.9
  17         408       2.2        17551        94.1
  18        1103       5.9        18654       100.0
```

   (a) What is the sample size? *Answer: 18654*

   (b) What is the median? *Answer: 12*

   (c) What is the first quartile? *Answer: 12*

   (d) What is the third quartile? *Answer: 14*

   (e) What is the fifth percentile? *Answer: 7*

   (f) What is the range? *Answer: 18*

   (g) What percentage of the sample have 12 or more years of education (i.e., at least high school graduates)? *Answer: 80.3*

   (h) Construct a box-and-whisker plot for the data. *Answer: Use the frequency distribution to find $Q_1 = 12$, $Q_2 = 12$, $Q_3 = 14$, and IQR=2. The lower end of the box is at 12 and the upper end at 14. The top whisker is at $14 + 1.5 \times 2 = 17$. The lower one at $12 - 1.5 \times 2 = 9$. Other values (0,1,2,3,4,5,6,7,8,18) are outliers.*

   (i) Suppose that everyone in the sample returned to school for an additional year of education, e.g., those with 10 years of education would have 11 years after an additional year of education. How would giving every person an additional year of education affect the mean years of education — would the mean increase, decrease, or remain the same? *Answer: Increase. See page .*

(j) How would giving every person an additional year of education affect the standard deviation of the years of education — would the SD increase, decrease, or stay the same? *Answer: Stay the same.*

2. The purpose of this problem is to give you practice using and interpreting descriptive statistics in SPSS. The data are available on Canvas in `defaultsmall.csv`, which is a comma-delimited file.

   The data set default is from a company that provides a service to its customers. The service is targeted at **adults**. Customers join and may cancel at any time. The company would like to understand who is likely to default. Customers pay an initial down payment and then monthly payments over three years. You have a sample of customers who joined over a three-year period of time.

   - `enrolldt` date of enrollment, e.g., 2007/05/01
   - `price` price of the membership
   - `downpmt` down payment
   - `monthdue` monthly dues
   - `pmttype` method of paying monthly dues. The four types of payment that we want to study are Book, Statement, Checking EFT, and Credit Card EFT.
   - `use` number of times the customer used the service during the first month
   - The `default` variable in the data set takes the value 1 if customers has defaulted within the first month of their membership and 0 otherwise
   - `age` and `gender` (1=male, 2=female) of the member

   (a) (video) Read the data into SPSS. Be sure to read the enrollment date variable in as a date. Run descriptives on all variables and examine the minimum and maximum values to see if they are reasonable. Hint: you should have 24,975 cases on all variables except `use`, which has some missing values.

   (b) (video) Enter value labels for the `pmttype` variable, where 1="book," 3="statement," 4="checking EFT," and 5="credit card EFT." Run a frequency distribution and bar plot of `pmttype`. *Answer: Book has 31.2%, statement has 20.9%, checking EFT has 24.2% and creditcard EFT has 23.7%.*

   (c) (video) Consider the `age` variable. Generate a histogram and frequency distribution. Submit a the histogram and a written description of the shape of the distribution (center, dispersion, skewness, number of modes, outliers, etc.) and make a list of features that do not make sense. *Answer: A histogram shows that the distribution is slightly right skewed. Descriptive statistics show that the IQR is from 23 to 36, so that the "middle half" of the customers are quite young. The histogram and frequency distribution show that there is a large group of customers with 0 age, which cannot be. The group of customers who are 99 years old and those between the ages of 1 and 10 are also suspicious. This service is for adults, so it is odd to have anyone younger than 18.*

(d) Set all cases that you consider to be nonsensical values based on your inspection of the age variable. Hint: use Data / Transform / Recode into a new variable, say `age2`, and assign a meaningful label. It is not a good idea to overwrite a variable because you may change your mind on what values to discard. (video)

(e) Find the skewness of age using descriptives or explore. What does it tell you? Is the interpretation consistent with what you learned from the histogram? Submit only skewness and its interpretation. *Answer: Skewness=1.233, indicating a right-skewed distribution, i.e., the right tail is long.*

(f) Generate a histogram of `downpmt` and describe its shape. *Answer: The distribution is highly right skewed with outliers.*

(g) Based on the shape of the `downpmt` distribution, do you expect the mean to be greater than, less than, or equal to the median? *Answer: Since the distribution is right skewed, we would expect the mean to be greater than the median.*

(h) Find the mean, median and 5% trimmed mean of `downpmt` and confirm your answer to the previous part. *Answer: 199.20, 100.00, 146.38.*

(i) (video) Generate a frequency distribution of `downpmt` and sort in descending order of counts (click on "Format" and check off "Descending counts"). Which down payments are more common? Do the values make sense? Do not submit your entire frequency distribution, which will be many hundred lines long. *Answer: The most common values are \$100, 150, 50, 75, 25, 5, 19 and 250. It makes sense that people round to such numbers when deciding on a downpayment.*

(j) Compute the base 10 logarithm of `downpmt`, generate its histogram, and describe its shape. Why are there so many spikes? Hint: to compute the logs use Transform / Compute, type `logdown` in the Target Variable box and `LG10(downpmt+1)` in the Numeric Expression box. *Answer: The shape is roughly normal, but the distribution has many spikes, which represent the common down payment values. The modal spike is at a value of 2, which corresponds to a down payment of* $10^2 = 100$*, indicating that a down payment of \$100 is the most common value.*

(k) What percentage of people have `use` $= 0$? *Answer: From frequency distribution, 39.7%.*

(l) Change the format of `enrolldt` to mm/dd/yyyy by clicking on the variable view tab and then changing the type. To show that you have done this correctly, run descriptives on `enrolldt`. *Answer: The minimum should be 1/01/2007.*

(m) Extract the month from `enrolldt` and submit a frequency distribution. *Answer: Use `month = XDATE.MONTH(enrolldt)` in transform / compute. You should find 3303 in January, 2754 in February, etc.*

(n) **Be sure to save your data set as an SPSS (`.sav`) file.**

3. Enter the following 20 values into column A of Excel.

$$1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 18, 21, 24, 27, 30, 34, 39, 48, 58$$

(a) Find the mean using the `AVERAGE` function. *Answer: 19*

(b) In the second column find the squared deviation from the mean. It is good form to reference the mean instead of typing the number 19.

(c) Find the mean of the squared deviations from the mean. This number is the population variance. *Answer: 247.3*

(d) Find the population standard deviation by computing the square root of the answer in the previous part. *Answer: 15.73*

(e) Confirm your answer part (d) by applying the `STDEV.P` function to the raw data.

(f) Find the sample standard deviation using the column of squared deviations from the mean. *Answer: SQRT(SUM(B1:B20)/19)=16.13*

(g) Confirm your answer using the `STDEV` function.

(h) Find the 5% trimmed mean by dropping the smallest and largest value (1/20=5% from each tail) from your computation. *Answer: AVERAGE(A2:A19)=17.83*

(i) Find the median "by hand." Confirm answer using the `MEDIAN` function. *Answer: 15*

(j) Find the quartiles "by hand." Confirm your answer using the `QUARTILE` function. *Answer: 5.75 and 27.75. Excel interpolates between the 5th and 6th value, and there is more than one acceptable way of doing this interpolation. For example, 5.5 and 28.5 are also acceptable answers using the "Tukey Hinge" approach. SPSS interpolates differently and gives 5.25 and 29.25*

(k) What is the interquartile range? *Answer:* $27.75 - 5.75 = 22$

(l) Create a boxplot "by hand."

4. In **1997** an IMC class was doing a project for a large national company that has a database of customers (consumers). It had a large sample of customers from their database. For this question, assume that the sample is representative of all consumers in the US. The company paid for *demographic overlays* — they paid another company to give them information about the demographics of their customers. One of the overlay variables is `yroccupy`, the "number of years occupancy." Summary statistics for this variable are shown below.

```
N             60622         100% Max  96           99%  96
Mean          87.03          75% Q3   94           95%  96
Std Dev        8.96          50% Med  90           90%  95
N Missing     85868          25% Q1   82           10%  73
Sum         5275872           0% Min  61            5%  67
Variance      80.26                                 1%  62
```

Do the values of the summary statistics make sense considering the description of the variable? Why or why not? If you answer no, suggest an alternative interpretation that makes more sense. Are there other features of the distribution that don't make sense?[20]

[20] It does not make sense that the average length of residence is 87 years and that the minimum is 61 years. The variable is mislabled. It is more likely the *last two digits of the year of the last move*. Under this interpretation, one feature that does not make sense is that the minimum is 1961. Surely some people have lived in their house since before 1961. I suspect that the distribution is truncated because such records were not computerized before the 1960s.

5. I am analyzing a dataset from a company that sells books to consumers with catalogs and a web site. The unit of analysis is a consumer. One of the variables is “age in years.” Summary statistics for this variable are provided below. Discuss problems with this variable based on the summary statistics.[21]

```
N                12612                  0% Min        -7132
Sum Obs         762464                  1%ile            23
Mean           60.4554                  5%ile            34
Std Dev        84.4824                  25% Q1           49
Coeff Var      139.743                  50% Median       61
Skewness       -41.004                  75% Q3           71
Kurtosis       4329.83                  95%ile           82
Std Err Mean  0.75227                   99%ile           89
Variance       7137.27                 100% Max        1994
```

[21] Answer: one cannot have an age of -7132 or 1994, which could be a year of birth. It looks like there are many reasonable values for age (first–99th percentiles), but there are some extreme values that don’t make sense.

The Case of the Suspicious Customer

B.R. Harris arrived at work and found, as expected, the recommendation of H.E. McRorie waiting on the desk. These would form the basis for a quaterly presentation Harris would give this afternoon to top management regarding production leveles for the next three months. The projections would serve as a planning guide, ideally indicating appropriate levels for purchasing, inventory, and human resources in the immediate future. However, customers have a habit of not always behaving as expected, and so these forecasts were always difficult to prepare with considerable judgment (guess-work?) traditionally used in their preparation.

Harris and McRorie wanted to change this and create a more objective foundation for these necessary projections. McRorie had worked late analyzing the customer survey (a new procedure they were experimenting with, based on responses of $n = 30$ representative customers from the entire database of $N = 1,015$ customers) and had produced a draft report that read, in part:

> We anticipate quarterly sales of $1,477,108, with projected sales by region given in the accompanying table. We recommend that production be increased from current levels in anticipation of these increased sales . . . .

2009 Quarter II Projections	2009 Quarter I Actual	2008 Quarter II Actual
1,477,502	1,114,929	1,020,798

Harris was uneasy. They were projecting a large increase both from the previous quarter (32.5%) and from the same quarter last year (44.7%). Historically, the firm had not been growing at near these rates in recent years. Along with that came a recommendation for increased production in order to be prepared for the increased sales. Why the hesitation? Because if these projections turned out to be wrong, and sales did not increase, the firm would be left with expensive inventory (produced at a higher cost than usual due to overtime, hiring of temporary help, and leasing of additional equipment) with its usual carrying costs (including the time value of money: the interest that could have been earned by waiting to spend on the additional production).

Harris asked about this, and McRorie also seemed hesitant. Yet it seemed simple enough: take the average anticipated spending by cusomters as reported in the survey, then multiply by the total number of customers in that region. What could be wrong with that? They decided to take a closer look at the data. Here is their spreadsheet, including background information (the wholesale price the firm receives for each item and the number of active customers by region) and the sampling results. Each of the 30 selected customers reported the number of each item they plan to order during the coming quarter. The Value column indicates the cash to be received by the firm (e.g., Customer 1 plans to buy 3 chairs at $45 and 4 bookshelves at $65 for a total value of $395).

Wholesale	Price
Chairs	$45
Tables	125
Bookshelves	65
Cabinets	350

Active Customers	Number
Northweast	302
Northwest	201
South	103
Midwest	255
Southwest	154
Total	1,015

Discussion Questions

1. Compute the `total` that each customer will purchase by multiplying the quantities of each item by its price.

2. Find the mean of `total` and multiply this by the number of customers in the database. Does this equal the 2009 Quarter II Projections given in the table above?

3. Find the percentage increase over the previous quarter and same-quarter-previous-year figures given in the same table. Do they equal the figures reported in the text?

4. Would the average-based procedure they are currently using ordinarily be a good method? Or is it fundamentally flawed? Justify your answer?

5. Take a closer look at the data using summaries and graphs. What do you find?

6. What would you recommend that Harris and McRorie do to prepare for their presentation this afternoon?

What is Graphical Excellence?[22]

- "Graphical excellence is the well-designed presentation of interesting data — a matter of *substance*, of *statistics*, and of *design*."
- "Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency."
- "Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."
- "Graphical excellence is nearly always multivariate."
- "And graphical excellence requires telling the truth about the data."

Examples:

- Train Schedule by Marey, cover of Tufte's book.
- Napoleon's March on Moscow

[22] From Tufte, *The Visual Display of Quantitative Information*, p. 51.

Types of Findings

- **Descriptive**: describe/measure something, e.g.,
  - Patient's blood pressure is ____.
  - 40% of the population favors a certain policy
- **Casual**: show how dependent variable Y, e.g., sales, depends on independent variables (X). Two types of independent variables:
  - Under your control, e.g., marketing actions including price, advertising, positioning, etc.
  - Out of your control, e.g., competitive actions, weather, the state of the economy, etc.
- "Perception of relationships is the cornerstone of civilization. By understanding how certain phenomena depend on others we learn to predict the consequences of our actions and to manipulate our environment."[23]
- "Relational graphics are essential to competent statistical analysis since they confront statements about cause and effect with evidence, showing how one variable affects another."[24]

[23] Sen and Srivastava, *Regression Analysis*, Springer-Verlag, New York, 1990, page 1.

[24] Tufte *The visual display of quantitative information*, Graphics Press, Cheshire, Connecticut, 1983, page 82.

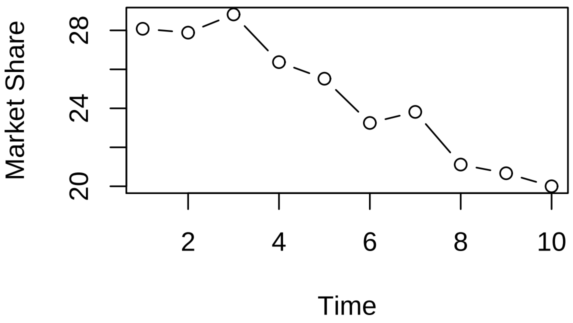
Hierarchy of Good Graphics

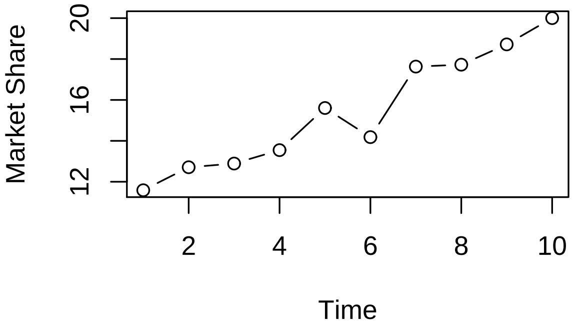
Ultimate question: What should the client do as a result of knowing what is on your graph?

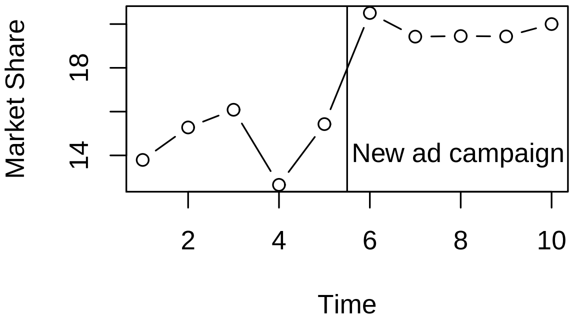
- Seek to show causal relationships.
  - Show how the outcomes you are trying to achieve depend on things under your control (as well as other causal factors out of your control).
  - Answer the question: if we do ____, what happens?
- Purely descriptive graphs over time have some value (showing how something changes over time), but time is usually not a fundamental cause (it is often a proxy for the causal factors).
- Univariate descriptive snapshots are usually not very actionable.

Focus on **outcomes**, not outputs

- Your dashboard tells you that your market share is 20% in period 10. Is this good or bad?
- Adding a time series is more informative
- Indicating causal variables improves it further.







Summarizing the Relationship Between Two or More Variables

- Quantify the *dependence* of one variable on the other by studying a *conditional distribution*
  - *Dependent variable*: variable that we would like to affect, but we do not have direct control over, e.g., sales, response to offer, intentions, attitudes, retention. Also called a *Criterion variable*.
  - *Independent variable*: a variable that affects the dependent variable, ideally under our control, e.g., price, advertising budget, offer

We study the (conditional) distribution of the dependent variable for given values of the independent variable.

Independent Variable(s)	**Dependent Variable**	
	Categorical	**Numerical**
Categorical	Crosstabs Z test proportions Chi-squared test	Compare means, boxplots Independent sample t-test ANOVA
Numerical	Logistic regression Discrete choice Discriminant analysis	Correlations Linear regression Nonparametric regression
Mixed	Generalized linear model	General linear model

Crosstabs in SPSS

			Default Status		Total
			Non-Defaults	Defaults	
Gender	Male	Count	10914	1627	12541
		% within Gender	87.0%	13.0%	100.0%
		% within Default Status	49.5%	55.7%	50.2%
		% of Total	43.7%	6.5%	50.2%
	Female	Count	11138	1296	12434
		% within Gender	89.6%	10.4%	100.0%
		% within Default Status	50.5%	44.3%	49.8%
		% of Total	44.6%	5.2%	49.8%
Total		Count	22052	2923	24975
		% within Gender	88.3%	11.7%	100.0%
		% within Default Status	100.0%	100.0%	100.0%
		% of Total	88.3%	11.7%	100.0%

- Crosstabs give a frequency distribution of all combinations of two variables.
- We divide all counts by the sample size to estimate *joint probabilities*, $P(\texttt{gender} \cap \texttt{default})$, which sum to one.
- Dividing by the row totals gives *row percents*, i.e., $P(\texttt{default}|\texttt{gender})$. Rows sum to one.
- Dividing by the column totals gives *column percents*, i.e., $P(\texttt{gender}|\texttt{default})$. Columns sum to one.

As a general rule, **condition on the causal (independent) variable**.

More Crosstab Examples

- Art museum visits and income (SAS output)

```
          Frequency|     Income
          Percent  |
          Row Pct  |
          Col Pct  |   Low   | Medium |  High  |  Total
Visit     ---------+--------+--------+--------+
Art       No       |    250 |    803 |    858 |   1911
Museum             |   8.33 |  26.77 |  28.60 |  63.70
Last               |  13.08 |  42.02 |  44.90 |
Year               |  70.03 |  65.18 |  60.81 |
          ---------+--------+--------+--------+
          Yes      |    107 |    429 |    553 |   1089
                   |   3.57 |  14.30 |  18.43 |  36.30
                   |   9.83 |  39.39 |  50.78 |
                   |  29.97 |  34.82 |  39.19 |
          ---------+--------+--------+--------+
          Total         357     1232     1411     3000
                      11.90    41.07    47.03   100.00
```

- Your turn: show how each of the percentages in the table is computed from the counts. Write out the interpretations.
- Your turn: How does payment type affect default status?

```
> fit = xtabs(~pmttype+default, default)
> prop.table(fit,1)
            default
pmttype                0           1
  Book       0.756826048 0.243173952
  Statement  0.836335761 0.163664239
  Check EFT  0.990403706 0.009596294
  Credit EFT 0.980717185 0.019282815
```

Click here to see this example in SPSS (**video**)

Cross Tabulations

How does area to improve depend on customer segment?

```
          Frequency|             Area to improve
          Percent  |
          Row Pct  |
          Col Pct  |Decor  |Taste  |Service|New Item|Value  |  Total
          ---------+-------+-------+-------+--------+-------+
Previous  New      |   324 |   107 |    59 |    285 |   488 |   1263
visits      1 time |  3.35 |  1.11 |  0.61 |   2.95 |  5.05 |  13.07
                   | 25.65 |  8.47 |  4.67 |  22.57 | 38.64 |
                   | 14.29 | 21.10 | 12.55 |   9.11 | 14.82 |
          ---------+-------+-------+-------+--------+-------+
          Light    |  1002 |   245 |   198 |   1262 |  1506 |   4213
           2-10    | 10.37 |  2.54 |  2.05 |  13.06 | 15.58 |  43.59
           times   | 23.78 |  5.82 |  4.70 |  29.95 | 35.75 |
                   | 44.20 | 48.32 | 42.13 |  40.35 | 45.75 |
          ---------+-------+-------+-------+--------+-------+
          Heavy    |   941 |   155 |   213 |   1581 |  1298 |   4188
           11+     |  9.74 |  1.60 |  2.20 |  16.36 | 13.43 |  43.34
           times   | 22.47 |  3.70 |  5.09 |  37.75 | 30.99 |
                   | 41.51 | 30.57 | 45.32 |  50.54 | 39.43 |
          ---------+-------+-------+-------+--------+-------+
          Total       2267     507     470     3128    3292     9664
                     23.46    5.25    4.86    32.37   34.06   100.00
```

Your Turn

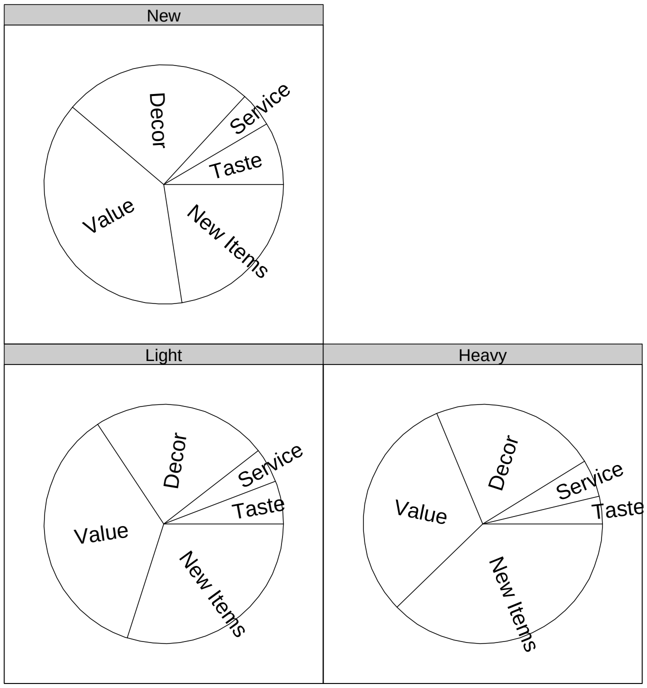
1. What was the sample size?
2. How many checked new items?
3. How many heavy users checked new items?
4. What percentage *of people* are heavy users?
5. What percentage *of people* checked value and are light users?
6. Given you are a heavy user, what is the chance you checked new items?
7. Given you checked new items, what is the chance you are a heavy user?
8. What do best customers want improved?
9. What do new customers want improved?
10. Are there differences between new and heavy users?

Presenting Findings: Restaurant

Area that needs improvement, by previous types of users

User	n	New Items (%)	Value (%)	Decor (%)	Service (%)	Taste (%)
Heavy	4188	37.8	31.0	22.5	5.1	3.7
Light	4213	30.0	35.8	23.8	4.7	5.8
New	1263	22.6	38.6	25.7	4.7	8.5

Note: New users have been to the restaurant 1 time, light users 2–10 times, and heavy users 11+ times. Universe: Adult customers in U.S., 1/1/2005–1/31/2005.



On Pie Charts

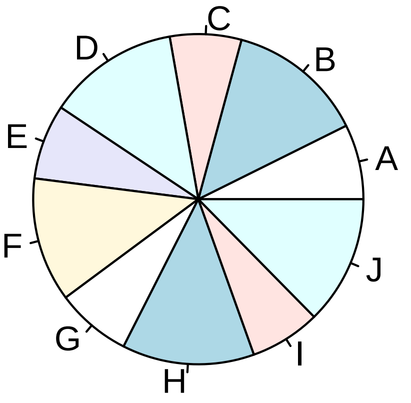
- "A table is nearly always better than a dumb pie chart; the only worse design than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies . . . Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used (Tufte, 1983, p. 178)."
- "Data that can be shown by pie charts always can be shown by a dot chart [or bar graph]. This means that judgments of position along a common scale can be made instead of less accurate angle judgments. Fortunately, pie charts are little used in science and technology, but they are a staple of business and mass media graphics (Cleveland, 1985, p. 264)."
- "Pie charts have severe perceptual problems. Experiments in graphical perception have shown that compared with dot plots, they convey information far less reliably. But if you want to display some data, and perceiving information is not so important, then a pie chart is fine (*S-Plus Trellis Graphics User's Manual*, p. 48)."

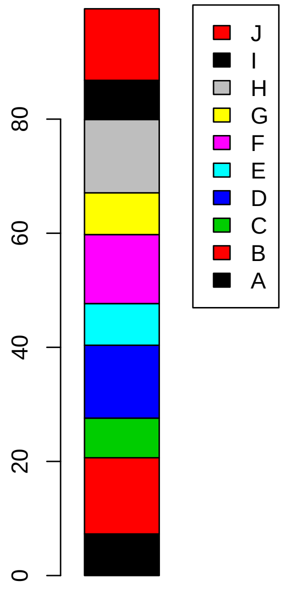
Graphical Perception

- The goal of a graph should be to convey the intended information in a way that can be understood quickly and without ambiguity by the intended audience. ("complex ideas communicated with clarity, precision, and efficiency")
  - We begin with an "idea" and data that support it
  - Data are encoded using graphical encryption devices
  - Audience decodes graphical device to understand idea
- Hierarchy of graphical encryption devices (See Cleveland's *The Elements of Graphing Data*, chapter 4)
  1. Position along a common scale is usually best
  2. Lengths (not with a common starting point as with a stacked bar chart) are more difficult to decode than (1), but easier than (3)
  3. Angles, area, color intensity and volumes are usually difficult to decode
- Erase unnecessary ink (see Tufte)
- Eliminate *chartjunk* (see Tufte)
- Proximity increases the accuracy of comparisons

Graphical Perception Example

What's the story?

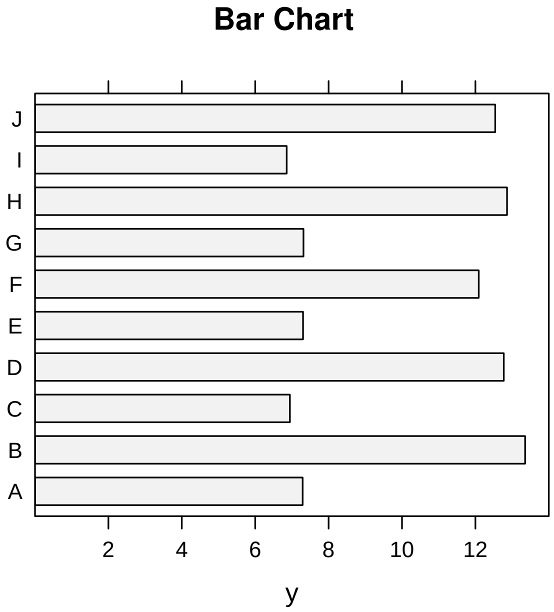




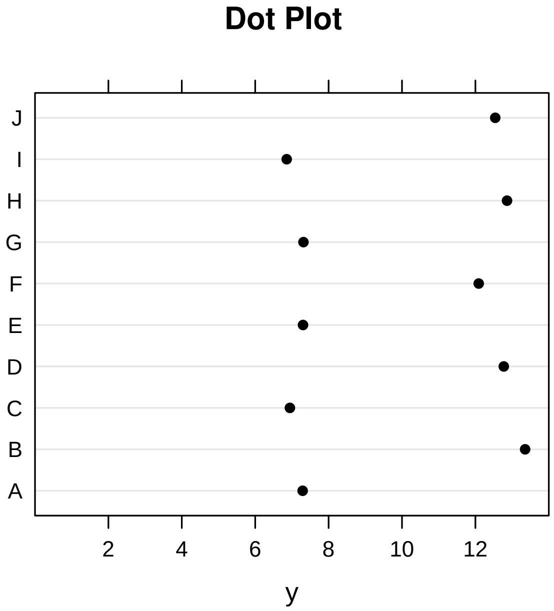
This example is motivated by Figures 4.19 and 4.20 in Cleveland, *The Elements of Graphing Data*

Graphical Perception Example

- Bar charts use position along a common scale, making the story obvious

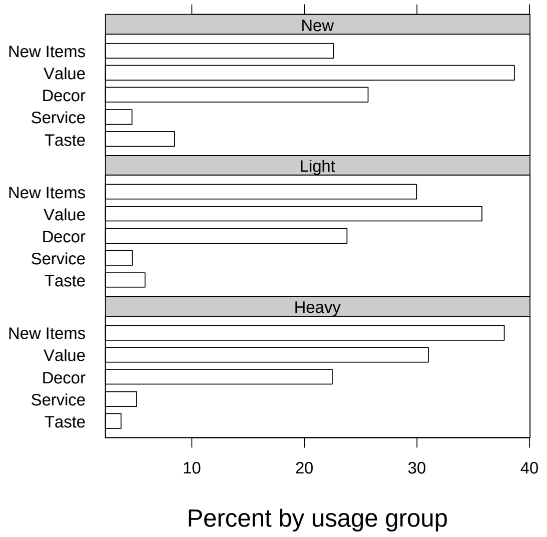


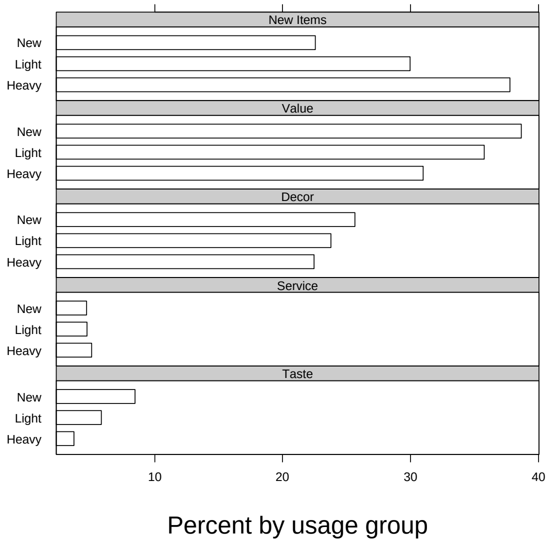
- Erase unnecessary ink to free up space



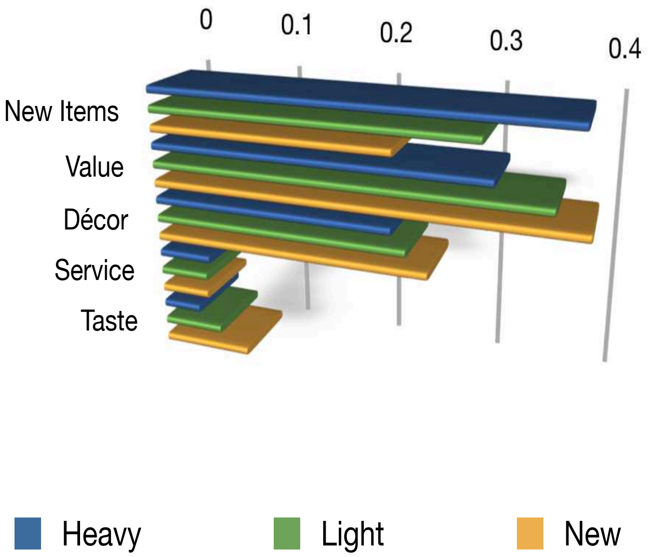
Presenting Findings: Restaurant

Proximity increases the accuracy of comparisons



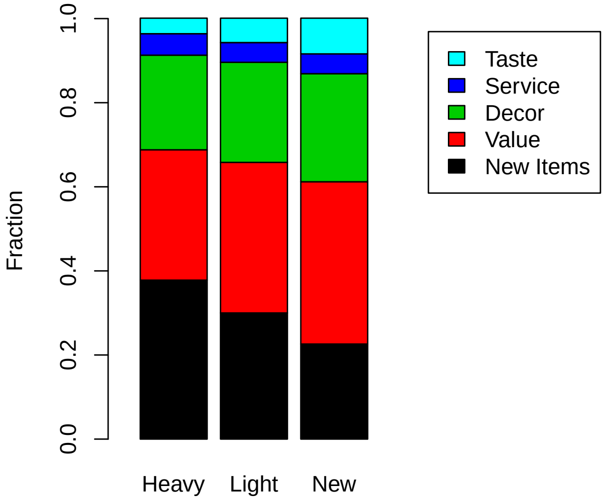


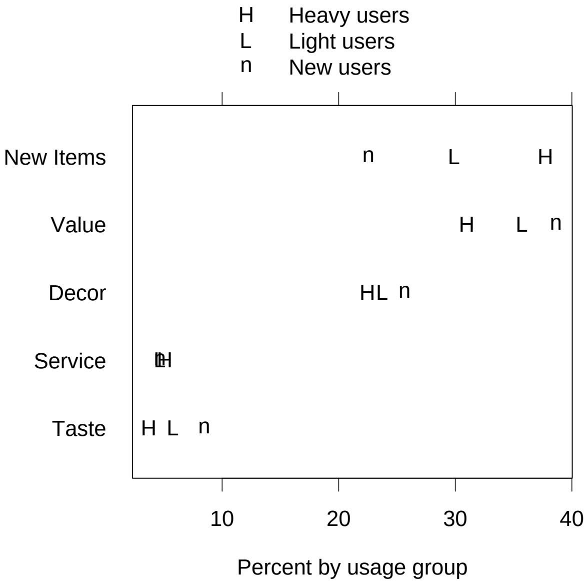
On Chartjunk



"Chartjunk does not achieve the goals of its propagators. The overwhelming fact of data graphics is that they stand or fall on their content, gracefully displayed. Graphics do not become attractive and interesting through the addition of ornamental hatching and false perspective to a few bars. Chartjunk can turn bores into disasters, but it can never rescue a thin data set. The best designs . . . are *intriguing and curiosity-provoking*, drawing the viewer into the wonder of the data, sometimes by narrative power, sometimes by immense detail, and sometimes by elegant presentation of simple but interesting data. But no information, no sense of discovery, no wonder, no substance is generated by chartjunk. (Tufte, p. 121, 1983)"

Presenting Findings: Restaurant





Studying Dependence

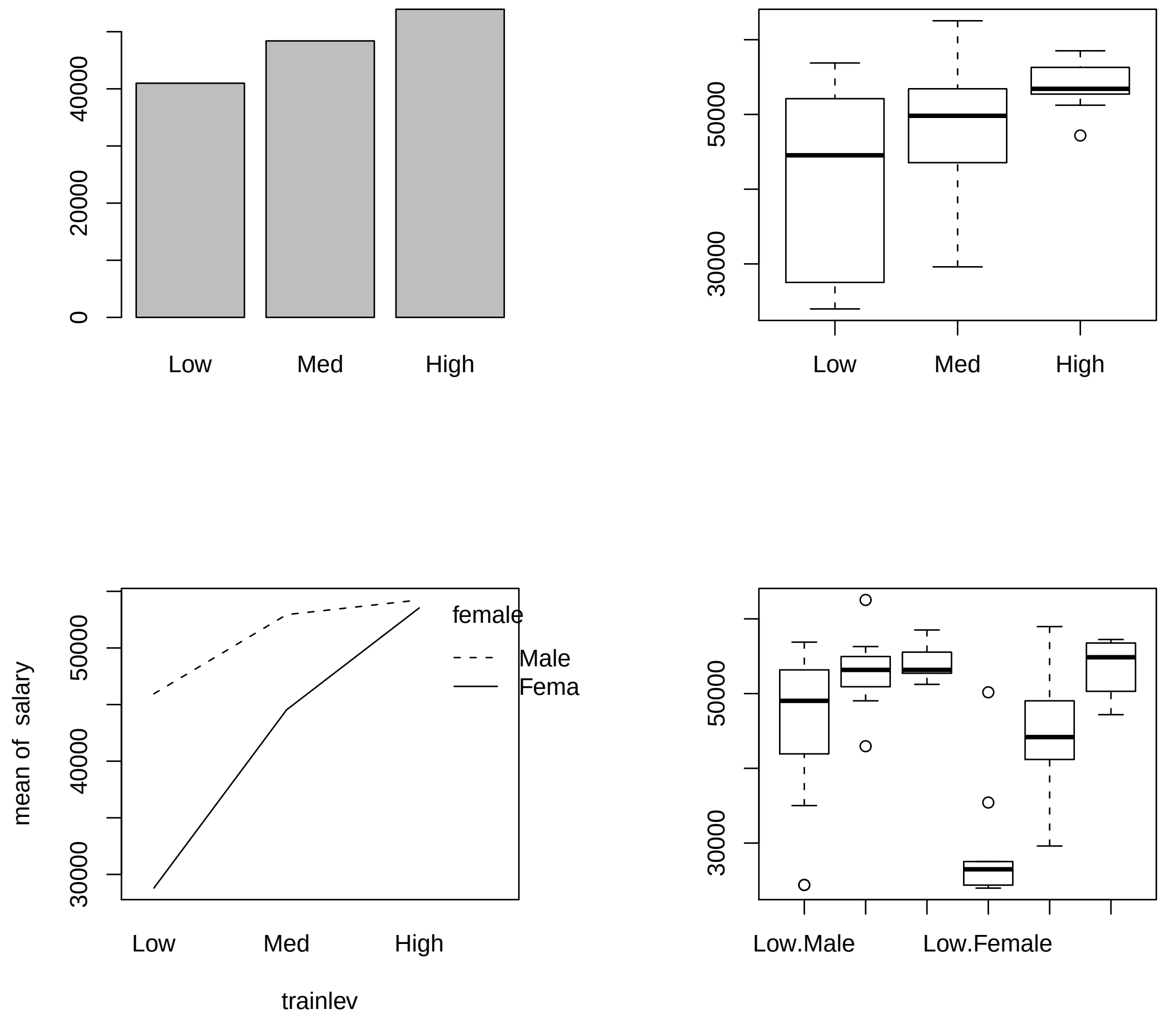
- How does some *dependent variable* Y depend on an *independent variable* X?
- Answer: study the distribution of Y for given values of X (i.e., "the conditional distribution of Y given X")
- Sometimes we'll be interested in the entire conditional distribution of Y, but we'll usually be interested in just the mean: $E(Y|X)$. Use Compare Means in SPSS or Pivot Tables in Excel. Average salary for low training is 40984.55, ...

trainlev	Mean	N	Std. Deviation	Median
A	40984.55	38	11550.861	44533.50
B	48387.17	24	8220.615	50209.75
C	53926.89	9	3451.407	53429.00
Total	45127.42	71	10816.877	48847.70

- We can condition on more than one variable at a time using "Layers" in SPSS' Compare Means

trainlev	sex	Mean	N	Std. Deviation
A	male	45952.15	27	8829.645
	female	28791.36	11	7796.847
	Total	40984.55	38	11550.861
B	male	52940.82	11	4883.575
	female	44534.08	13	8634.801
	Total	48387.17	24	8220.615
C	male	54239.40	5	2848.148
	female	53536.25	4	4536.929
	Total	53926.89	9	3451.407
Total	male	48703.58	43	8242.012
	female	39635.46	28	12084.268
	Total	45127.42	71	10816.877

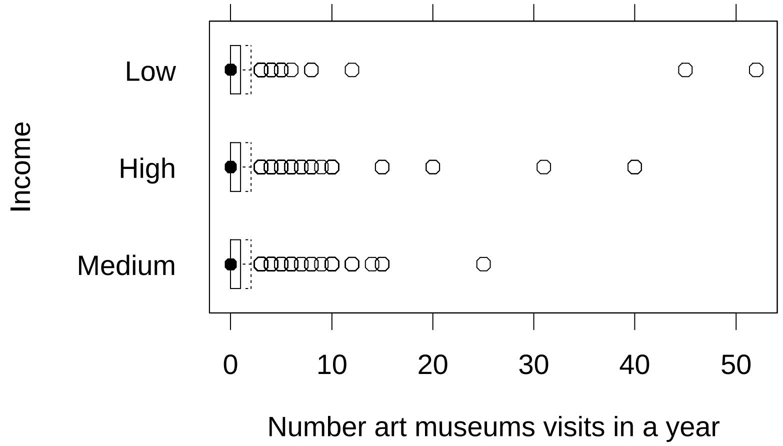
Visual Displays



Art Museums and Income

```
Income        N       Mean   Std Dev     Min   Max
--------------------------------------------
Low         370     0.8907     3.83       0     52
Medium     1248     0.8369     1.94       0     25
High       1424     0.8972     2.38       0     40
--------------------------------------------
```

Means suggest that high and low income go roughly equally often and middle income goes less often. Boxplot shows outliers:



Income	1% trimmed mean
Low	0.5904
Medium	0.6510
High	0.7199

After trimming, the conclusions are consistent with crosstab. Do we want 1% of the data to have such a strong influence on our conclusions?

Summary of Key Points

- Use relational graphs to show how what your client cares about (dependent variable) is affected by factors under your control (your marketing) and other causal factors (e.g., competitor, weather). Purely descriptive graphs are usually not as actionable.
- Plot dependent variable on vertical axis and (the first) independent variable on the horizontal axis. Encode other independent variables using multiple/clustered lines/bars (color) and/or small multiples (panels).
- Use graphical devices that can be easily decoded: position along a common scale usually best, followed by lengths. Avoid using angles, areas, color intensity, and volumes.
- Erase unnecessary ink, eliminate chartjunk, and keep proximity in mind.

Multivariate Descriptive Statistics Class Exercises

1. A manufacturer of floor waxes has recently developed a new wax. The company is considering designs for two different containers for the wax, one plastic and one metal. The company decides to make the final determination on the basis of a limited sales test in which the plastic containers are introduced in a random sample of 10 stores and the metal containers are introduced in an independent random sample of 10 stores. The sales results are as follows: (**video**)

Store	Plastic	Store	Metal
1	432	11	365
2	360	12	405
3	397	13	396
4	408	14	390
5	417	15	404
6	380	16	372
7	422	17	378
8	406	18	410
9	400	19	383
10	408	20	400

   (a) Type the data into SPSS. Have one column labeled "sales" and another labeled "material". In the sales column enter the 20 numbers. In the material column enter the value 1 for plastic and 2 for metal. Set the values labels so that 1="Plastic" and 2="Metal".

   (b) Construct box plots and error bar plots comparing the distribution of sales for metal and plastic.

   (c) Use the Analyze / Compare means to compute the means and standard deviations for plastic and metal.

   (d) Which material seems to have higher sales? (Note: we really should do a significance test here. You will do that for this problem in chapter 10.)

2. This problem continues using the `default.csv` data from last week. The pupose is to give you practice using compare means and crosstabs. **Be sure to use the data set you saved from last week**. (**video**)

   (a) The `gender` variable gives the gender of the member, where 1=male and 2=female. Assign value labels so that male and female show up in the output rather than 1 and 2. Are males or females more likely to default? Submit a cross tab in support of your answer and indicate which is more likely to default. Hint: click on analyze / descriptives / cross tab; copy `gender` in to the Rows box and `default` into the Columns box. Click on Cells and check off the row percent box. *Answer: females have a 10.4% chance of defaulting while males have a 13.0% chance. Males are slightly more likely than females to default.*

   (b) Answer the previous question using "compare means" instead of cross tabs. Hint: click on analyze / compare means / means. Copy `default` into the dependent list and gender into the independent list. Submit the output. Note that sometimes it is easier to use compare means while other times it will be easier to use crosstabs; you should therefore know both methods. *Answer: The mean of default for males should be the same as the percents in the previous part.*

(c) Which age groups are most likely to default? To answer this question, you will implement a *bin smoother*, with is also known as a *regressogram* (extending the idea of a histogram). First group the ages into bins by using Transform/Compute. Type "agebin" in the target variable box and type the following into the Numeric Expression box: `10*RND(age2/10)`. Make sure you understand this transformation—it is very useful! Note that you created `age2` last week by removing the "dirty" data from the `age` variable. Support your answer with appropriate descriptive statistics (using either cross tabs or compare means) and a charts (try a bar chart, line chart and error bar chart) plotting the mean of `default` against `agebin`. All three graphs are often useful and you should know them all. Discuss the nature of the relationship and which groups are more likely to default. *Answer: A good way to answer this is with a bar chart that plots the default rate for each age group. The probability of defaulting decreases linearly as age increases.*

(d) This part illustrates another application of compare means. Which ages are assigned to each level of `agebin`? To answer this question, click on Analyze / Compare Means / Means. Copy `age2` into the Dependent list box and `agebin` into the Independent list box. Click on Options and request the minimum and maximum values as well as the number of cases. (Be sure to understand exactly how the formula in the previous part works. This is a good trick to know. For you to think about but not submit: how could you form 5-year bins?) *Answer: agebin=20 (18 – 24); 30 (25 – 34); 40 (35 – 44); 50 (45 – 54); 60 (55 – 64); 70 (65 – 74); 80 (75 – 79); 90 (88 – 88). One way to form 5-year bins is `5*RND(age/5)`.*

(e) Generate and submit a bar chart plotting mean default status against the `age` variable. Compare this bar chart with the one that used `agebin` on the horizontal axis. In particular, why is the `age` version so jagged? *Answer: It has the same basic shape as before. The jags are due to the small sample sizes. This is an indication that we are showing an artifact of this particular sample. Having fewer bars is preferred for this reason. This modivates the need for a bin smoother.*

(f) How does the month during which a customer joins affect the likelihood of defaulting? Support your answer with appropriate descriptive statistics / graphs. Which months, if any, have higher default rates? *Answer: Default rates are slightly lower in winter (November – March) at around 9% than during the rest of the year, around 14%.*

(g) Form quartiles of `downpmt` and `use`. Hint: click on Transform / Rank cases. Copy `downpmt` and `use` into the variables box. Uncheck "Display summary tables." Click on Rank types, uncheck Rank, and check Ntiles. Make sure that Ntiles is set to 4 so that you get quartiles. Notice that SPSS has added two new variables to your dataset. Submit the minimum and maximum down payment for each level of `Ndownpmt`, e.g, what is the range of down payments in the bottom quarter. Note that this is an alternative way of forming bins and is very useful. *Answer: Use compare means as you did in part* 2d. *For down payment, the first quartile ranges from 0–45, second quartile 45–100, third quartile 101–175, and fourth quartile 176–9371. For use, the first and second quartiles are 0, the third quartile ranges from 1–3, and the fourth quartile ranges from 4–8.*

(h) Run a frequency distribution of `Nuse`. Note that it only takes three values, instead of four. You would expect it to take four possible values because it is the quartile number. Explain why it only takes three values. Hint: run a frequency distribution of `use`. *Answer: Note that the value 0 constitutes 39.7% of the data and the value 1 an additional 23.1% of the data. Clearly $Q_1 = 0$ and $Q_2 = 1$ (which is what Explore tells us, along with $Q_3 = 4$). The question is how does "Rank Cases" partition a distribution? In particular, what does SPSS do with the boundary values of 1? Do the 1s go into the second or third quartile? They seem to be assigned to the thrid.*

(i) How does payment type affect the likelihood of defaulting? Support your answer with appropriate descriptive statistics / graphs. Which payment types, if any, have higher default rates? *Answer: Payment type has a strong effect on the default rates. The rates are: book 24%, statement 16%, checking EFT 1%, and credit card EFT 2%. Notice the large differences in default rates across types.*

(j) How does the quartile of down payment affect the likelihood of defaulting? Support your answer with appropriate descriptive statistics / graphs. Which down payments, if any, have higher default rates? Why is it necessary to use the quartiles rather than raw down-payment values? Note that you are doing another regressogram. *Answer: Down payment also has a strong effect on defaulting. The bottom quartile (0–45) has a default rate of 25%, the second quartile (45–100) has 13%, the third quartile (101–175) has 5%, and the fourth quartile (175+) has 1%. The higher the down payment, the lower the default rate. Without the bins the plot would be too jagged, showing artifacts of this sample.*

(k) How does the quartile of usage affect the likelihood of defaulting? Support your answer with appropriate descriptive statistics / graphs. Which levels of usage have higher default rates? *Answer: Use has a strong effect on default rates. The bottom two quartiles (use=0) has a default rate of 21%, the third quartile has 8%, and the fourth quartile has 0.4%. The more that a person uses the service, the less likely the person is to default.*

(l) Generate the following line plot and turn it in. Click on Graph / Legacy dialog / Line chart. Click on Multiple then Define. Under Lines Represent check off Other statistics (e.g., mean) and copy `default` into the Variable box; copy `Ndownpmt` onto the Category Axis; copy `Nuse` into the Define Lines by box; copy `pmttype` into the Panel by Columns box. Write (only) one or two sentences describing what this graph tells you. *Answer: Three strikes and you are out. A person who has book or statement, with a low down payment and who is not using the service, is highly likely to default. Having any two of these problems still has a low default rate. This is call a three-way interaction.*

(m) Memberships are for 3 years. What relationship do you expect between `price`, `downpmt`, and `monthdue`? Does this relationship hold (approximately)? *Answer: Compute a new variable* ***`diff`*** *as* ***`price`*** $-$ ***`downpmt`*** $- 36\times$***`monthdue`***. *One would expect this variable to be close to 0. We can evaluate how often it is close to 0 by constructing either a box plot or looking at the percentiles. The tenth percentile is -288 and the 25th percentile is 275. If we define "close to 0" as being within \$250, then we are close to 0 less than 15% of the time. Often we are way off. The median is 583, indicating that for half the cases the price is at least \$583 higher than the down payment-monthly due combination. Something is wrong with at least one of these variables. We should seek clarification.*

(n) For you to think about but not turn in: which of the variables are "important" predictors? How can you quantify variable importance? *Answer: The concepts of sums of* *squares, R-square, and partial sums of squares/R-square will quantify variable importance. You will study these when you get to multiple regression in chapter 12.*

(o) For you to think about but not turn in: to reduce attrition at this company, what other research do you need in addition to this database analysis? *Answer: You need to understand why people join and default. This probably requires some exploratory research.*

3. (video) Three of the most important variables in database marketing and CRM are *RFM*:

- *Recency*: the length of time since the most recent purchase.
- *Frequency*: the number of previous purchases.
- *Monetary*: the amount of previous purchases.

In practice, analysts must compute RFM from a database "table" of all transactions using compare-means functionality. This process is called "rolling up a transaction file." This exercise will show you how this is done for a tiny database of 3 customers and 6 transactions. (In real life you might have millions of customers and tens of millions of transactions!)

`custid`	`transdate`	`amount`
1	12/30/2011	100
2	12/01/2011	50
2	11/01/2011	20
3	12/15/2011	20
3	11/30/2011	30
3	11/15/2011	40

(a) Type the data into SPSS (you will have to change the type of trandate to being a date). For example, customer 1 had one transaction on 12/30/2011, spending $100.

(b) Suppose that "today" is January 1, 2012. Create a new variable called `diff`, giving the number of days since a transaction. Hint: `DATEDIFF(DATE.MDY(1,1,2012),transdate,"days")`.

(c) Find RFM for each customer using compare means. Hint: recency is the minimum value of `diff`, frequency is the number of cases, and monetary is the sum of `amount`. *Answer: For customer 1 RFM = (2 days, 1 purchase, $100); for customer 2 RFM = (31, 2, $70); for customer 3 RFM = (17, 3, $90). Next quarter you will learn about a different function in SPSS for computing these.*

(d) Another important variable is the *average order amount*, which is the mean of `amount` by customer. Find it for all three customers. *Answer: For customer 1: 100; 2: 35; 3:30*

(e) Another important variable is *customer tenure*, which is the length of time that a customer has been a customer. It is the maximum value of `diff`. Find tenure for all three customers. *Answer: For customer 1: 2 days; 2: 61; 3:47*

4. A national survey included the questions: "What is your age?" and "How many times did you attend an art fairs during the past year?" Crosstab:

Art Fair	Age				
Attendance	< 30	30–44	45–64	65+	Total
None	1245	2260	2386	1625	7516
Light	617	1257	1153	415	3442
Heavy	296	656	603	205	1760
Total	2158	4173	4142	2245	12718

(a) What percentage of people under age 30 are heavy attenders? *Answer:* 296/2158 = .1372

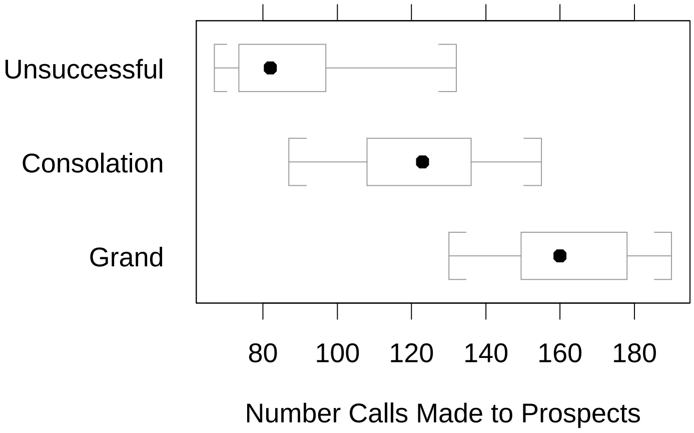
(b) What percentage of heavy attenders are in the 65+ age group? *Answer:* 205/1760 = .1165

(c) What percentage of people are in the 30–44 age group and are light art fair attenders? *Answer:* 1257/12718 = .0988

(d) Does art fair attendance depend on age? If so, how? Justify your conclusion by giving appropriate numbers computed from those in the table in support of your answer. You will be partially evaluated by the clarify and conciseness of the presentation of evidence. *Answer: The 65+ group is least likely to attend. This is a very good problem because it asks you to make sense of a cross tab and determine what the story should be. You must judge what is more important, and filter out small differeces that are not so important. Column percents are summarized in the table below. Some of the best answers made a bar chart.*

Art Fair	*Age*			
Attendance	< 30	*30–44*	*45–64*	*65+*
None	*58%*	*54%*	*58%*	*72%*
Light	*29%*	*30%*	*28%*	*18%*
Heavy	*14%*	*16%*	*15%*	*9.1%*

5. A company rewards the best members of its salesforce with a “grand” prize. The good members of the salesforce receive “consolation” prizes. The other members of the salesforce are “unsuccessful” and receive no prize. The figure below shows boxplots of the distribution of the number of calls made to new prospects that each of the three types of sales people made.



(a) Of the “unsuccessful” sales people, is the skewness of the distribution of calls left, right, symmetric, or unknown, based on these boxplots? *Answer: right*

(b) Approximately what percentage of "grand" prize winners made less than 150 calls to new prospects? *Answer: 25%*

(c) Does it appear that the number of calls to new prospects was the only variable used in determining the three groups? Or do you think that other variables were also used? Explain briefly. *Answer: overlap of boxplots implies others used*

6. A store wants to test whether giving customers an in-store coupon will increase the likelihood that they will purchase a certain product. The store decides to execute a test. Some customers are given the coupon while others do not receive any coupon. The purchase behavior of both groups is tracked during their visit. The table below summarizes the results, e.g., 170 males received the coupon and bought.

	Males		Females	
	Buy	No Buy	Buy	No Buy
Coupon	170	30	100	300
No Coupon	160	40	15	85

(a) What is the conditional probability of buying given that the person is male and received the coupon? *Answer:* $170/(170+30) = 85\%$*.*

(b) Considering only the population of males, should the coupon be recommended? Support your recommendation with appropriate statistics. *Answer: The chance that a male buys without the coupon is* $160/(160+40) = 80\%$*, which is less that 85% from the previous part. The coupon increases the likelihood that a male will buy. Many of you said that since a high percentage of males would buy anyway, the coupon should not be used. If the store gives the coupon to all males then it will be giving the perk to the 80% who would have bought anyway. These are valid concerns, but the problem does not give sufficient information to resolve it. For example, we would need know know what the perk offered by the coupon costs so that we could compare what it costs to give the perk to all males with the incremental profit from the additional 5% buyers. I've not deducted any points for raising this concern. In my defense, the problem states that the store wants to determine whether the coupon increases the likelihood of buying. I had intended for profit to be evaluated later.*

(c) Considering only the population of females, should the coupon be recommended? Support your recommendation with appropriate statistics. *Answer: The chance that a female buys with the coupon is* $100/(100+300) = 25\%$*, and without the coupon is* $15/(15+85) = 15\%$*. The coupon increases the likelihood that a female will buy and should be used.*

(d) Considering the entire population of males and females combined, find the conditional probability of buying given the coupon, and the conditional probability of buying given no coupon. Based only on these computations, would you recommend using the coupon? *Answer: The chance that a person buys with the coupon is* $(100+170)/(200+400) = 45\%$*, and without the coupon is* $(160+15)/(200+100) = 58\%$*. Now the coupon seems to deter purchases.*

(e) Considering all parts and any other computations you may wish to do, should the coupon be used? *Answer: The coupon increases the likelihood of buying. This is an example of Simpson's paradox. For some reason a much higher percentage of females received coupons, yet females as a group are less responsive. Giving a higher percentage of females coupons is a flaw in the design of the test because it introduces another causal factor — gender — that confounds the coupon variable. Females are less likely to buy with or without the coupon, and females are more likely to receive the coupon.*

IMC 451

Whodunit?

Uh-oh. The defect rate has ben rising lately and the responsibility has fallen on your shoulders to identify the problem so that it can be fixed. Two of the three managers (Jones, Wallace, and Lundvall) who supervise the production line have already been in to see you (as have some of the workers), and their stories are fascinating.

Some accuse Jones of being the problem, using words like "careless" and "still learning the ropes" based on anecdotal evidence of performance. Some of this is ordinary office politics to be discounted, of course, but you feel that the possibility should certainly be investigated nonetheless. Jones has countered by telling you that defects are actually produced at a higher rate when others are in change and that, in fact, Wallace has a much higher error rate.

Soon after, Wallace (who is not exactly know for tact) comes into your office, yelling that Jones is an (unprintable) . . . and is not to be believed. After calming down somewhat, there is mumbling: something about being given difficult assignments by the upper-level management. However, even when asked directly, the high error rate is not denied. You are suspicious: It certainly looks like you've found the problem. However, you are also aware that Wallace (although clearly no relation to Miss Manners) has a good reputation among technical experts and should not be accused without first considering possible explanations and alternatives.

While you are at it, you decide that it would be prudent to also look at Lundvall's error rates, as well as the two different types of production: one for domestic clients and one for overseas clients (who are much more demanding as to the specifications). The data are as follows:

	Domestic Clients		Overseas Clients	
Supervisor	Defective	Nondefective	Defective	Nondefective
Wallace	3	293	255	1,247
Lundvall	12	307	75	359
Jones	131	2,368	81	123

1. Is Jones, correct? That is, using the more complete data set, it is true that Jones has the lowest defect rate overall? Find the percentages.

2. Is Wallace correct? That is, what percent of Wallace's production was the more demanding? How does this compare to the other two managers?

3. Look carefully at conditional defect rates given various combinations of manager and production client. What do you find?

4. Should you recommend that Wallace start looking for a new job? If not, what do you suggest?

Solution

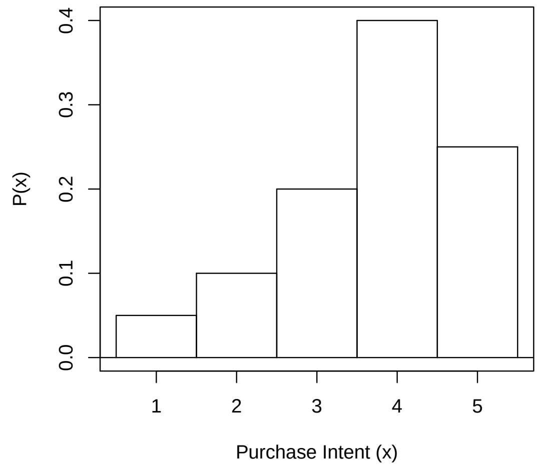
Probability Distributions

- Definition — a *random variable* assigns some number to the outcome of a random experiment
- *Probability mass functions* give the probability that a random variable will assume a particular value
- Experiment: measure purchase intent (e.g., 5-point scale) of randomly selected member of target market
- Possible mass function (notation: $P(x = 1) = .05$):

x	1	2	3	4	5
$P(x)$	.05	.1	.2	.4	.25

Note that the probabilities sum to 1 (.05+.1+.2+.4+.25=1)

- Probability mass function:



Descriptions of Distributions

- Measure of center: *Population mean* (or *Expectation*)

  $$\mu = \mu_x = E(x) = \sum_x P(x)x$$

  Purchase intent example on previous page:

  $$E(x) = .05(1) + .1(2) + .2(3) + .4(4) + .25(5) = 3.7$$

- Measures of spread
  - *Population variance*

    $$\sigma^2 = \sigma_x^2 = V(x) = \sum_x P(x)(x - \mu_x)^2$$

    $$\begin{aligned} \sigma_x^2 &= .05(1-3.7)^2 + .1(2-3.7)^2 + .2(3-3.7)^2 + \\ & \quad .4(4-3.7)^2 + .25(5-3.7)^2 \\ &= 1.21 \end{aligned}$$

  - *Population standard deviation*

    $$\sigma = \sigma_x = \sqrt{\sigma_x^2}$$

    $$\sigma_x = \sqrt{1.21} = 1.1$$

- Use *Ed's table* for fool-proof way of computing the mean and variance of a discrete random variable:

x	$P(x)$	$xP(x)$	$P(x)(x-\mu_x)^2$
1	.05	0.05	$.05(1-3.7)^2 = .3645$
2	.10	0.20	$.1(2-3.7)^2 = .2890$
3	.20	0.60	$.2(3-3.7)^2 = .0980$
4	.40	1.60	$.4(4-3.7)^2 = .0360$
5	.25	1.25	$.25(5-3.7)^2 = .4225$
Sum	1	$\mu_x = 3.7$	$\sigma_x^2 = 1.21$

Your Turn

1. (Siegel 7.4) On a given day, assume there is a 30% chance you will receive no orders, a 50% chance you will receive one order, a 15% chance of two orders, and a 5% chance of three orders. Find the expected number of orders and the standard deviation. *Answer:* $\mu = .95$; $\sigma = 0.8047$

2. (video) (Siegel 7.10) You have identified the four major problems, the extent to which each one occurs (i.e., the probability that this problem occurs per item produced), and the cost of reworking to fix each one. Assume that only one problem can occur at a time. *Answer: (a) .275, .246, .045, .029, broken case most serious; (b)* $\mu = .5954$; *(c)* $\sigma = 2.1642$

Problem	$P(X)$	X
Broken Case	0.04	\$6.88
Faulty electronics	0.02	\$12.30
Missing connector	0.06	\$0.75
Blemish	0.01	\$2.92

   (a) Compute the expected rework cost for each problem separately. Compare the results and indicate the most serious problem in terms of expected dollar costs.

   (b) Find the expected rework cost due to all four problems together.

   (c) Find the standard deviation of rework cost.

   (d) Write a brief memo, as if to your supervisor, describing and analyzing the situation.

3. (Siegel 7.2) (video) The length of time a system is down (that is, broken), is described (approximately) by the probability distribution as follows: Minor problems take 5 minutes to fix and occur 60% of the time when there is a problem; Substantial problems require 30 minutes to fix and occur 30% of the time; Catastrophic problems require 120 minutes fix and occur 10% of the time. These are the only three types of problems and assume that they always take this long (5, 30, or 120 minutes) to fix. *Answer: (a) discrete; (b) 24; (c) 33.90; (d) .4; (e) 90%*

   (a) What type of probability distribution is this?

   (b) Find the mean downtime.

   (c) Find the standard deviation of the downtime.

   (d) What is the probability that the downtime will be greater than 10 minutes?

   (e) What is the probability that the downtime is literally within one standard deviation of its mean? Is this about what you would expect for a normal distribution?

4. An investment will pay \$105 with probability 0.7 and \$125 with probability 0.3. Find the risk (as measured by the standard deviation) for this investment. *Answer:* $\mu = 111$; $\sigma = 9.17$

Your Turn: Siegel Chapter 7

7.1. (a) 8.5; (c) 10.1366; (e) .25; (f) .9

7.2. (video) (b) 24; (c) 33.90; (d) .4; (e) 90%

7.3. $\mu = 111$; $\sigma = 9.17$

7.4. $\mu = .95$; $\sigma = 0.8047$

7.7. $\mu = 37,500$; $\sigma = 16,771$

7.8. $\mu = 32.5\%$; $\sigma = 48.15\%$; $P(X > 40) = .6$; σ

7.9. *Apartment:* $\mu = \$106,000$ and $\sigma = \$29,394$. *House:* $\mu = \$84,000$ and $\sigma = \$19,596$. *Sell land:* $\mu = \$60,000$ and $\sigma = \$0$. *Casino:* $\mu = \$50,000$ and $\sigma = \$150,000$. Casino has lower expected payoff and higher risk, so eliminate it. With others, high expected payoffs are associated with higher risks.

7.10. (video) (a) .275, .246, .045, .029, broken case most serious; (b) $\mu = .5954$; (c) $\sigma = 2.1642$

Additional Problem

- Despite all safety measures, accidents do happen at Sam's Manufacturing Corporation. Let x denote the number of accidents that occur during a month at this company. The following table lists the probability distribution of x. *Answer: (2a) 0.45; (2b) 0.75; (3) 1.55; (4) 1.2836.*

x	0	1	2	3	4
$P(x)$	.25	.3	.2	.15	.1

1. Draw a graph of the probability distribution (be sure to label your axes).
2. Determine the probability that the number of accidents that will occur during a given month at this company is
   (a) at least 2
   (b) less than 3
3. Calculate and interpret the mean of this distribution.
4. Calculate the standard deviation of this distribution.

Binomial Distribution

- Definition: A *Bernoulli trial* is a random experiment that has two possible outcomes, labeled “success” and “failure” (dichotomous)
  - Flip a coin
  - Did you get exam question 5 correct?
  - Did you respond to a particular offer?
  - Did you recall seeing advertisement?
  - Do you subscribe to the *Tribune*?

  Let π be the probability of “success”
- Definition: The *binomial distribution* tells us the probability of getting
  - exactly X_B successes
  - out of n identical and independent Bernoulli trials
- Excel: `BINOMDIST`

Example

Suppose that the chance a viewer recalls seeing your ad is $\pi = .6$. If we ask $n = 4$ viewers if they recall it, what is probability that exactly $X_B = 3$ do?

Responses	Probability
YYYN	$.6 \times .6 \times .6 \times (1 - .6) = 0.0864$
YYNY	$.6 \times .6 \times (1 - .6) \times .6 = 0.0864$
YNYY	$.6 \times (1 - .6) \times .6 \times .6 = 0.0864$
NYYY	$(1 - .6) \times .6 \times .6 \times .6 = 0.0864$
Sum	0.3456

This is really just

$$4 \times .6^3 \times (1 - .6)^1$$

More generally

$$(\#\text{ ways}) \times (\text{Pr success})^{\#\text{ successes}} \times (\text{Pr failure})^{\#\text{ failures}}$$

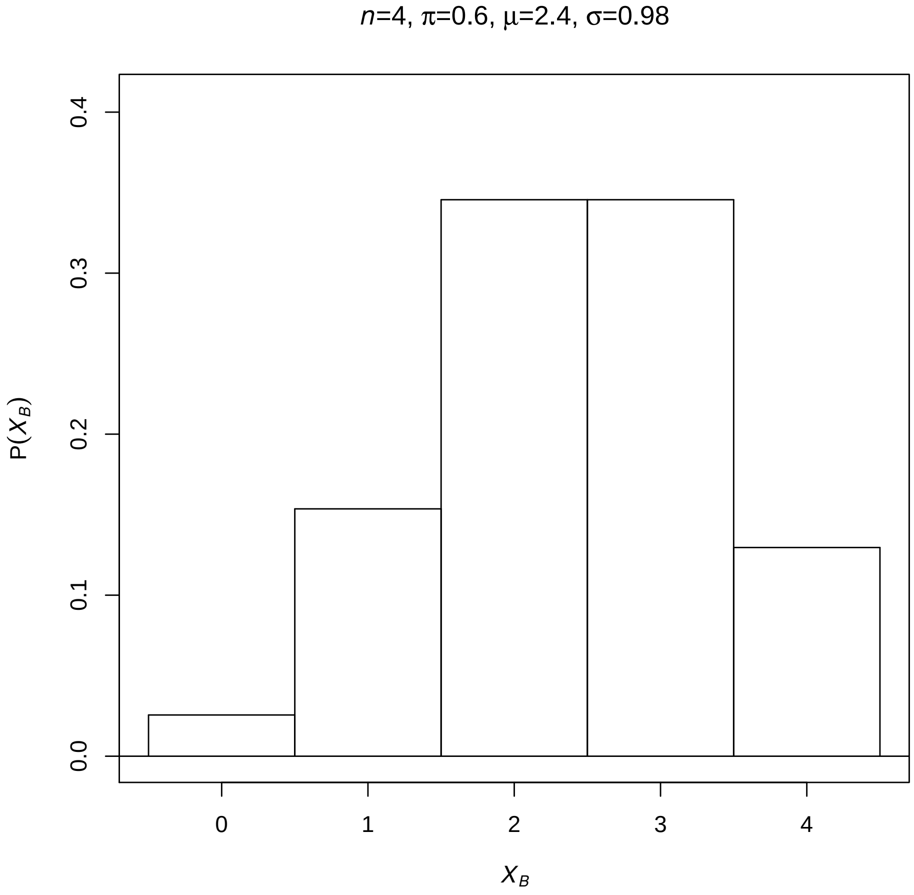
```
=BINOMDIST(3,4,0.6,FALSE)
```

Example

(**video**) Let X_B be the number of people who recall seeing the ad out of 4. The *distribution* of X_B is binomial:

X_B	$P(X_B)$	$X_B P(X)$	$(X_B - \mu_{X_B})^2 P(X_B)$
0	$1 \times .6^0 \times (1 - .6)^4 = .0256$	0	0.1475
1	$4 \times .6^1 \times (1 - .6)^3 = .1536$	0.1536	0.3011
2	$6 \times .6^2 \times (1 - .6)^2 = .3456$	0.6912	0.0553
3	$4 \times .6^3 \times (1 - .6)^1 = .3456$	1.0368	0.1244
4	$1 \times .6^4 \times (1 - .6)^0 = .1296$	0.5184	0.3318
Sum	1	$\mu_{X_B} = 2.4$	$\sigma^2_{X_B} = 0.96$

Note: (1) The probabilities add to 1, (2) all are between 0 and 1, and (3) we can compute the mean and standard deviation



The Binomial Formula and Properties

- Binomial formula: the probability of getting X_B successes in n tries where the probability success is π

  $$\mathcal{B}(x; n, \pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

  See `BINOMDIST` in Excel

- The mean (expected) number of "yes" responses is

  $$E(X_B) = \mu_x = n\pi$$

- The standard deviation of the number of "yes" responses is

  $$\sigma_x = \sqrt{n\pi(1-\pi)}$$

We are often more interested in a linear combination of X_B:

$$p = \frac{X_B}{n}$$

- Using the formulas on 62 we can show

  $$E(p) = \mu_p = \pi \quad \text{and} \quad \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- We will see next time that X_B and p are approximately normal under certain conditions (page 118)

Your Turn

1. (Siegel 7.11) (**video**) Suppose that 8% of the loans you authorize as vice president of the consumer loan division of a neighborhood bank will never be repaid. Assume further that you authorized 284 loans last year and that loans go sour independently of one another. *Answer: 22.72; 4.5719; 0.016098*

   (a) *How many* of these loans, authorized by you, do you *expect* will never be repaid? What percentage do you expect?

   (b) Find the *usual measure of the level of uncertainty* in the *number of loans* you authorized that will never be repaid. Briefly interpret this number.

   (c) Find the *usual measure of the level of uncertainty* in the *percentage of loans* you authorized that will never be repaid. Briefly interpret this number.

2. (Variation on Siegel 7.14, 15) (**video**) An election coming up next week promises to be very close. In fact, assume that 55% are in favor and 45% are against. *Answer: (a) 0.01769, [.51, .59]; (b) 0.03518, [.48, .62]; (c) 0.03464, [.53, .67].*

   (a) Suppose you conduct a poll of 791 randomly selected likely voters. Find the standard deviation of the percent in favor from this poll. Compute $\pi \pm 2\sigma_p$.

   (b) Repeat the previous question, but assume $n = 200$

   (c) Repeat the first question, but assume $n = 200$ and $\pi = .6$

3. Your company is planning to market a new reading lamp and has segmented the market into three groups, avid readers, regular readers and occasional readers, and currently assumes that 25% of avid readers, 15% of regular readers, and 10% of occasional reader will want to buy the new product. As part of a marketing survey, 400 individuals will be randomly selected from the population of regular readers. Using the current assumptions, find the mean and standard deviation of the percentage among those surveyed who will want to buy the new product. *Answer: 0.15, 0.01785*

4. A company is conducting a survey of 235 people to measure the level of interest in a new product. Assume that the probability of a randomly selected person being "very interested" is 0.88 and that people are selected independently of one another. *Answer: (a) 0.02120; (b) 4.982; (c) 206.8; (d) 0.88*

   (a) Find the standard deviation of the percentage who will be found by the survey to be very interested.

   (b) How much uncertainty is there in the number of people who will be found to be very interested?

   (c) Find the expected number of people in the sample who will say that they are very interested.

(d) Find the expected percentage that the survey will identify as being very interested.

5. You have just performed a survey interviewing 358 randomly selected people. You found that 94 of them are interested in possibly purchasing a new cable TV service. How much uncertainty is there in this number 94 as compared to the average number you would expect to find in such a survey? (You may assume that exactly 25% of all people you might have interviewed would have been interested.) *Answer: 8.193*

6. (**video**) Your firm has decided to interview a random sample of 10 customers in order to determine whether or not to change a consumer product. Your main competitor has already done a similar but much larger study and has concluded that exactly 86% of consumers approve of the change. Unfortunately, your firm does not have access to this information (but you may use this figure in your computations here). *Answer: (a) binomial; (b) 8.6; (c) 1.0973; (d) .86; (e) .1097; (f) .2639; (g) 0.4184; (h) .8455; (i) 0.7387*

   (a) What is the name of the probability distribution of the number of consumers who will approve of the change in your study?

   (b) What is the expected *number* of people, out of the 10 you will interview, who will approve of the change?

   (c) What is the standard deviation of the *number* of people, out of the 10 you will interview, who will approve of the change?

   (d) What is the expected *percentage* of people, out of the 10 you will interview, who will approve of the change?

   (e) What is the standard deviation of the *percentage* of people, out of the 10 you will interview, who will approve of the change?

   (f) (**video**) What is the probability that exactly eight of your interviewed customers will approve of the change?

   (g) What is the probability that eight or less of your interviewed customers will approve of the change?

   (h) What is the probability that eight or more of your interviewed customers will approve of the change?

   (i) What is the probability that between seven and nine of your interviewed customers will approve of the change?

Your Turn: Siegel Chapter 7

7.11. (video) (a) 22.72; (b) 4.5719; (c) .016098

7.14. (video) .0178

7.15. .01270

7.16. 8.19

7.17. .1678

7.18. (a) securities independent with constant probability; (b) $E(x) = 12$; (c) $\sigma_x = 1.55$; (d) $P(x = 15) = .035$; (e) $P(x = 10) = .1032$; (f) $P(x \geq 13) = .231 + .132 + .035 = .398$

7.19. (video), part 2 (a) binomial; (b) 8.6; (c) 1.0973; (d) .86; (e) .1097; (f) .2639; (g) .2639 + .3602 + .2213 = .8455

Additional Problem

1. Every week a resort interviews six randomly chosen vacationers on the island about their experience. In general, each vacationer's comments can be classified as mainly positive or mainly negative. Suppose that only 5% of all visitors are dissatisfied with their visit.

   (a) *How many* customers during a particular week do you expect to indicate that they are "satisfied."

   (b) What is the standard deviation of the number of customers who indicate they are satisfied?

   (c) What is the standard deviation of the percent of people who indicate they are satisfied?

   (d) What is the chance that exactly 4 customers indicate they are satisfied during a particular week?

   (e) What is the chance that 4 or more customers are satisfied?

   (f) What is the chance that 5 or fewer customers are satisfied?

   (g) What is the chance that between 4 and 5 customers are satisfied?

   Answer: (a) 5.7; (b) 0.5339; (c) 0.0890; (d) 0.0305; (e) .9978; (f) .2649; (g) .2627.

The Normal Distribution

- Definition: a random variable X has a *normal* or *Gaussian* distribution if it has the familiar "bell shape"[25]
- We need to find three types of normal probabilities:
  - Left tail: $P(X < x) =$ `NORMDIST(x, mu, sigma, TRUE)` or convert to Z and use the normal table or `NORMSDIST(z)`.
  - Right tail: $P(X > x) = 1 - P(X \leq x) = P(Z \leq -z)$
  - Slices: $P(x_1 < X < x_2) = P(X < x_2) - P(X \leq x_1)$
- Definition: Z has a *standard normal* distribution if it is normal with mean 0 and variance 1. By convention, Z is reserved for standard normal variables
- Important theorem: if X is normal with mean μ_x and variance σ_x^2, then

$$Z = \frac{X - \mu_x}{\sigma_x}$$

has a *standard normal distribution.* Z is measured in *standard units*, indicating the number of standard deviations from the mean of X (also called Z-scores).

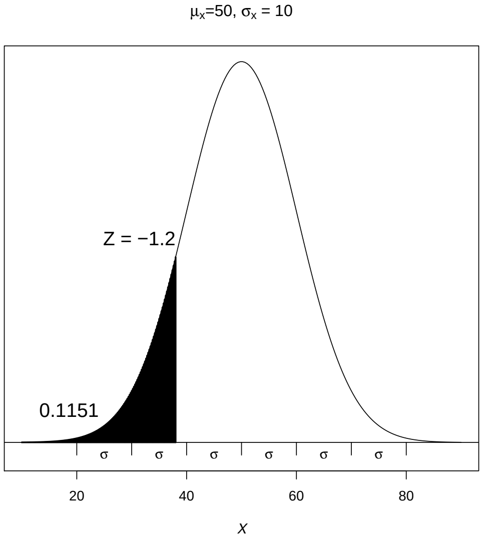
[25] More technically, if x has a normal distribution with mean μ_x and variance σ_x^2 then

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right].$$

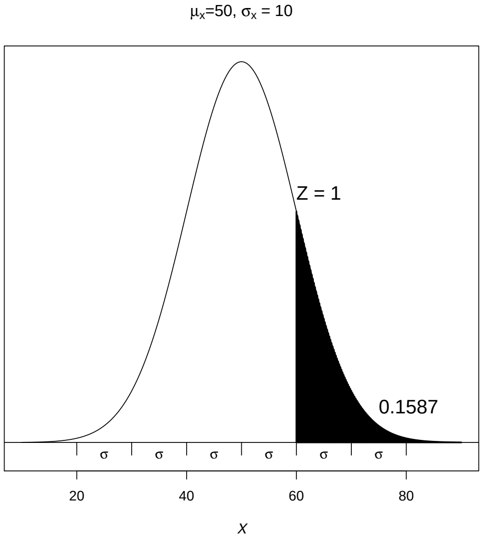
Normal Distribution Problem

A standardized test has a mean of $\mu = 50$ and a standard deviation of $\sigma = 10$. Let X be the test score of a random student. Find the probability that a test is less than 38

$$\begin{aligned} P(X < 38) &= P\left(Z < \underbrace{\frac{38-50}{10}}_{-1.2}\right) \\ &= \texttt{NORMSDIST(-1.2)} \\ &= \texttt{NORMDIST(38, 50, 10, TRUE)} \\ &\approx 0.1151 \end{aligned}$$



Note that 38 is 1.2 standard deviations to the left of the mean. Find the probability that a test is greater than 60 (note 60 is one standard deviation to the right of the mean)

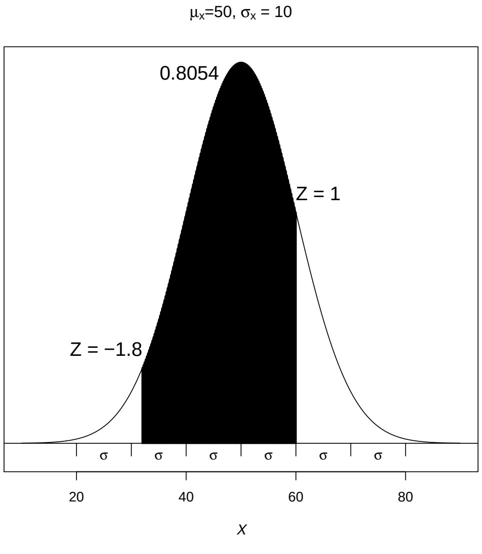


$$P(X > 60) = P\left(Z > \underbrace{\frac{60-50}{10}}_{1}\right) = \underbrace{1 - P(Z < 1)}_{\text{complement rule}} = \underbrace{P(Z < -1)}_{\text{symmetry}} = .1587$$

Normal Slices

Find the probability that a test score is between 32 and 60.

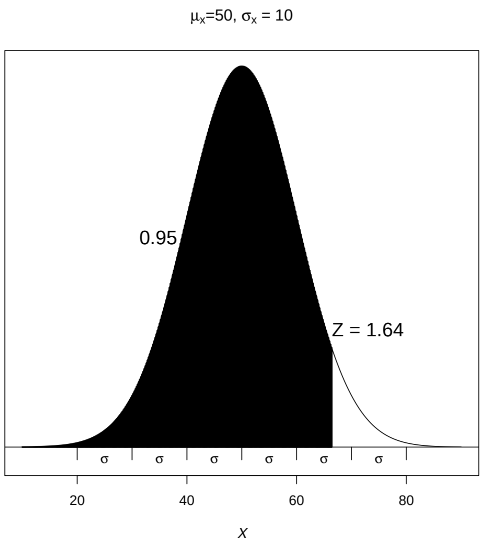
$$\begin{aligned}
P(32 < X < 60) &= P(X < 60) - P(X < 32) \\
&= P\left(Z < \underbrace{\frac{60-50}{10}}_{1}\right) - P\left(Z < \underbrace{\frac{32-50}{10}}_{-1.8}\right) \\
&= \texttt{NORMSDIST}(1) - \texttt{NORMSDIST}(-1.8) \\
&= .8413 - .0359 \\
&= .8054
\end{aligned}$$



Normal Percentiles

What is the 95th percentile? That is, what is the score such that 95% of all students score less than this value?

$$P(X <?) = .95$$



To find the answer in Excel use `NORMINV(.95,50,10)` = 66.45.

Or use either the normal table or `NORMSINV(.95)` to find $z = 1.645$

$$.95 = P(X < x) = P\left(Z < \underbrace{\frac{x-50}{10}}_{1.645}\right)$$

Solve for x

$$\frac{x-50}{10} = 1.645 \Longrightarrow x = 66.45$$

Your Turn

1. (7.26) (video) (video) The amount of ore (in tons) in a segment of a mine is assumed to follow a normal distribution with mean 185 and standard deviation 40. Find the probability that the amount of ore is less than 175 tons. *Answer: .4013*

2. (7.27) You are a farmer about to harvest the crop. To describe the uncertainty in the size of the harvest, you feel that it may be described as a normal distribution with mean value 80,000 bushels and a standard deviation of 2,500 bushels.

   (a) Find the probability that your harvest will be less than 76,000 bushels. [.0548]

   (b) Find the probability that your harvest will exceed 84,000 bushels. [.0548]

   (c) What is the probability of a "normal" year, where the harvest is between 76,000 and 84,000 bushels? [.8904]

3. (7.22) Under usual conditions, a distillation unit in a refinery can process a mean of 135,000 barrels per day of crude petroleum, with a standard deviation of 6,000 barrels per day. You may a assume a normal distribution. *Answer: (a) .5; (b) .0478*

   (a) Find the probability that more than 135,000 barrels are produced in a day.

   (b) Find the probability that less than 125,000 barrels are produced in a day.

4. (7.29) Although you don't know the exact total amount of payments you will receive next month, based on past experience it will be approximately \$2,500 more or less than \$13,000, and will follow a normal distribution. Find the probability that you will receive between \$10,000 and \$15,000 next month. *Answer: .6731*

5. (7.30) A new project will be declared "successful" if you achieve a market share of 10% or more in the next two years. Your marketing department has considered all possibilities and decided that it expects the product to attain a market share of 12% in this time. However, this number is not certain. The standard deviation is forecast to be 3%, indicating the uncertainty in the 12% forecast as 3 percentage points. You may assume a normal distribution. *Answer: .75; .25; .16; .26*

   (a) Find the probability that the new project is successful.

   (b) Find the probability that the new project fails.

   (c) Find the probability that the new project is wildly successful, defined as achieving at least a 15% market share.

   (d) To assess the precision of the marketing projections, find the probability that the attained market share falls close to the projected value of 12%, that is, between 11% and 13%.

Your Turn: Siegel Chapter 7

7.21. .3820

7.22. .5; .7967; .0062; .0475; .0000

7.23. .84

7.24. .49; .34; .20; .09; .08

7.25. .58; .31; .11

7.26. (video) .4013

7.27. .0548

7.28. .1056

7.29. .67

7.30. .75; .25; .16; .26

Additional Problem

1. Assume that electronic microchip operating speeds are normally distributed with a mean of 1.5 gigahertz and a standard deviation of 0.4 gigahertz.

   (a) What percentage of your production would you expect to be "superchips" with operating speeds of 2 gigahertz or more?

   (b) What percentage of your project will have an operating speed of less than 1.3 gigahertz?

   (c) What percentage of your project will have an operating speed between 1.3 and 1.9 gigahertz?

   (d) What is the skewness of operating speeds?

   (e) What is the speed of a chip that is slower than 95% of all chips?

   (f) What is the interquartile range of operating speeds?

   Answer: *(a)* $P(X > 2) = P(Z < -1.25) = 0.1056 = 1 - \texttt{NORMDIST}(2, 1.5, .4, \texttt{TRUE})$; *(b)* $P(X < 1.3) = P(Z < -0.5) = .3085 = \texttt{NORMDIST}(1.3, 1.5, .4, \texttt{TRUE}$; *(c)* $P(1.3 < X < 1.9) = .5328 = \texttt{NORMDIST}(1.9, 1.5, .4, \texttt{TRUE} - \texttt{NORMDIST}(1.3, 1.5, .4, \texttt{TRUE}$; *(d) 0; (e)* $-1.645 \times 0.4 + 1.5 = 0.842 = \texttt{NORMINV}(.05, 1.5, .4)$; *(f)* $0.4 \times 2 \times 0.6745 = 0.5396 = \texttt{NORMINV}(.75, 1.5, .4) - \texttt{NORMINV}(.25, 1.5, .4)$.

The Normal Approximation to the Binomial

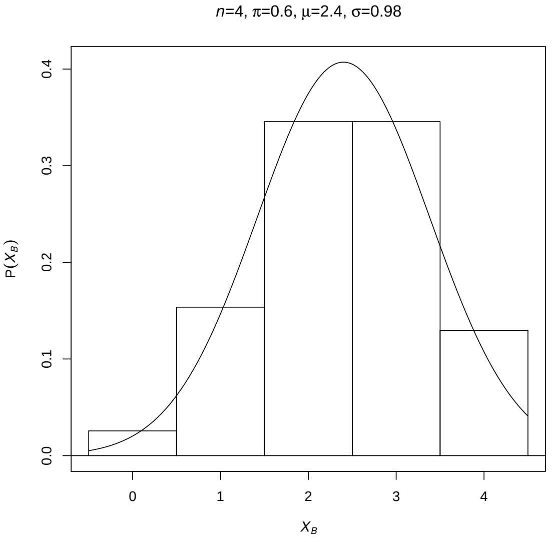
Normal approximation to the binomial: if X_B has a binomial distribution (probability of success π and number of trial n) and $p = X_B/n$ is the proportion of "successes," then it turns out that

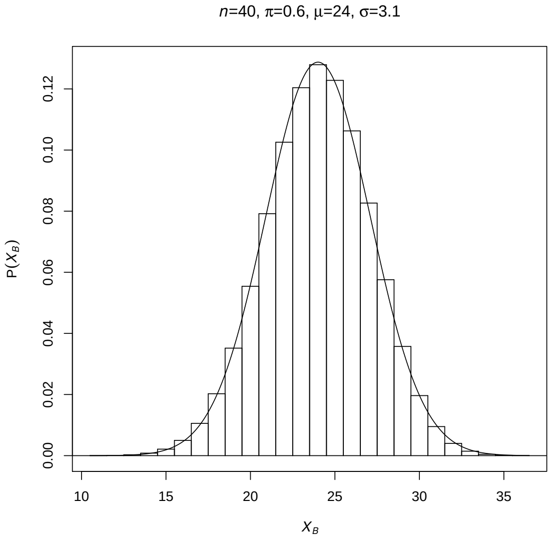
$$Z = \frac{X_B - n\pi}{\sqrt{n\pi(1-\pi)}}$$

$$Z = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}}$$

have *approximate* standard normal distributions, provided that $n\pi > 5$ and $n(1-\pi) > 5$

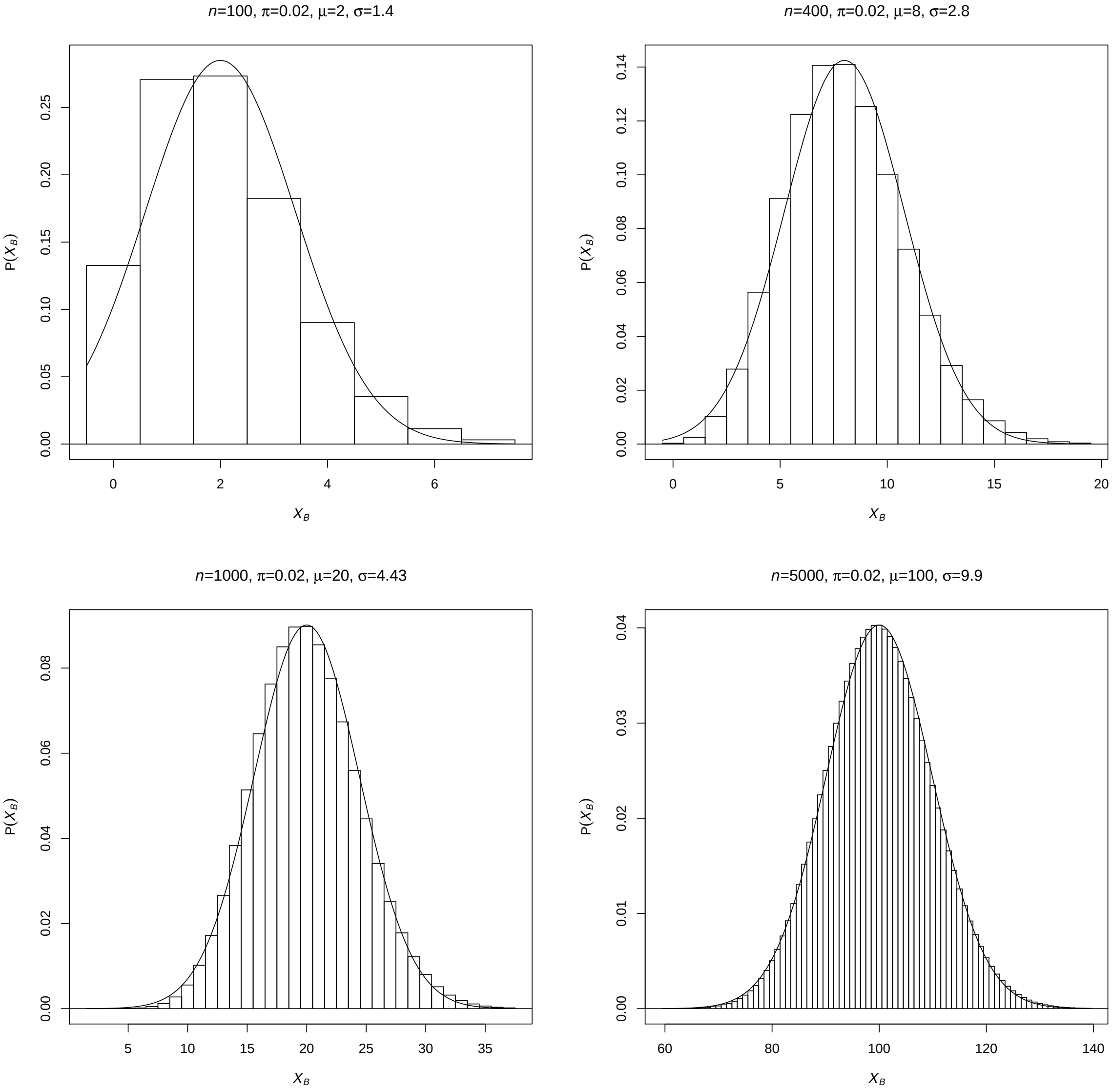
Consider example from page 107:





Beware of Skewed Distributions

The condition $n\pi > 5$ guards against right-skewed distributions and $n(1-\pi) > 5$ against left-skewed distributions. For example, let $\pi = .02$.



n=100, π=0.02, μ=2, σ=1.4

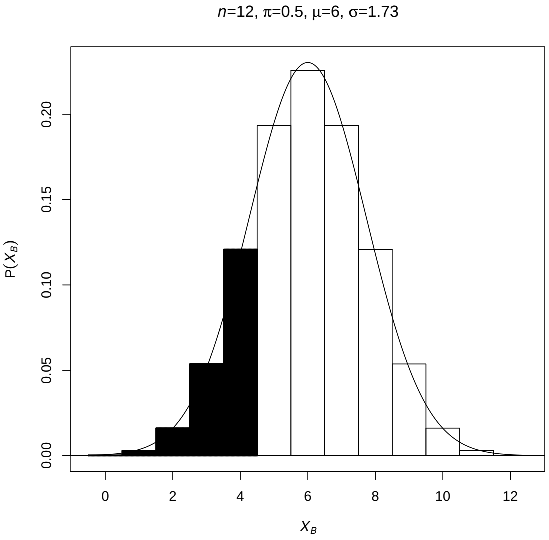
n=400, π=0.02, μ=8, σ=2.8

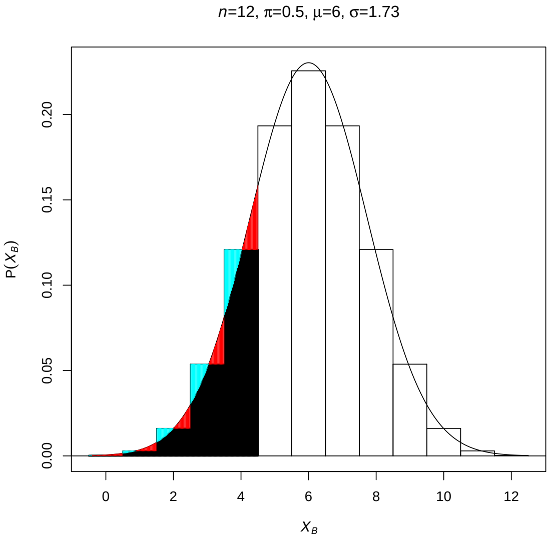
n=1000, π=0.02, μ=20, σ=4.43

n=5000, π=0.02, μ=100, σ=9.9

Continuity Correction

Suppose X_B has a binomial distribution with $n = 12$ and $\pi = .5$. Find $P(X_B \leq 4)$.





The mean of X_B is $\mu = 12 \times .5 = 6$ and standard deviation is $\sigma = \sqrt{12 \times .5 \times .5} = 1.73$. Visualize bars of width 1 centered over each possible value of X_B.

$$P(X_B \leq 4) \approx P\left(Z < \underbrace{\frac{4.5 - 6}{1.73}}_{-0.87}\right) = 0.1932$$

Exact: $\texttt{BINOMDIST}(4, 12, .5, \texttt{TRUE}) \approx 0.1938$

Comments on Continuity Correction

1. Using the correction will usually, but not always, produce more accurate results.
2. The big idea is that we can use the normal distribution to make probability statements about counts (X_B) (also called *totals*), and proportions (p) (which are special *means*).
3. On an exam, you will full credit if you use it or if you choose not to use it. It is your choice. You may also use `BINOMDIST`.
4. The rationale for my policy is as follows: In real life, if you need an extremely accurate binomial probability then use Excel or R. Using the continuity correction (e.g., adding .5) is still only an approximation.

Your Turn

1. (Siegel 7.32, 33) (video) A vote is scheduled tomorrow, and it looks to be close. Assume that the number of yes votes follows a binomial distribution. You expect 300 people to vote. Assume a probability of 0.53 that a typical individual will vote yes.

   (a) Identify n and π for this binomial random variable.

   (b) Find the mean and standard deviation of the *number* who will vote yes.

   (c) Find the (approximate) probability that a majority will vote yes. Find the exact probability.

   (d) Now assume that 1,000 people will vote and repeat the previous parts, continuing to assume $\pi = 0.53$.

   Answer: (a) $n = 300$, $\pi = .53$; (b) $\mu = 159$, $\sigma = 8.645$; (c) .837262, exact .8372696; (d) $n = 1000$, $\pi = .53$, $\mu = 530$, $\sigma = 15.78$, 0.9692, exact .9691136

2. (Siegel 7.37) (video) You have just sent out a test mailing of a catalog to 1,000 people randomly selected from a large database. You will go ahead with the mass mailing to the remaining names provided you receive orders from 2.7% or more from the test mailing within two weeks. Find the (approximate) probability that you will do the mass mailing under each of the following scenarios that in reality,

   (a) exactly 2% of the population would send in an order within two weeks.

   (b) exactly 3% of the population would send in an order within two weeks.

   (c) exactly 4% of the population would send in an order within two weeks.

   Answer: .0708; .7422; .9854. Without correction: .0571; .7123, .9821. Exact: .0758; .7363; .9890

3. (Siegel 7.36) You are planning to interview 350 consumers randomly selected from a large list of likely sales prospects to assess the value of this list and whether you should assign sales people the task of contacting them all. Assuming that 13% of the large list will respond favorably, find (approximate) probabilities for the following

   (a) Less than 10% of randomly selected consumers will respond favorably.

   (b) More than 15% of randomly selected consumers will respond favorably.

   (c) Between 10% and 15% of randomly selected consumers will respond favorably.

   Answer: Using p: .0476, .1329, .8195. Exact: .0363, .1337, .8299. With CC: $P(X_b < 35) = .0402$; $P(X_b > 52.5) = .1329$; $P(35 \leq X_b \leq 52.5) = .8269$

4. (Siegel 7.34) Assume that if you were to interview the entire population of Detroit, exactly 18.6% would say that they were ready to buy your product. You plan to interview a representative random sample of 250 people. Find the (approximate) probability that your observed sample percentage is overoptimistic, where this is defined as the observed percentage exceeding 22.5%. *Answer: .0565; exact .05472819*

5. (Siegel 7.35) Suppose that 15% of the items in a large warehouse are defective. You have chosen a random sample of 250 items to examine in detail. Find the (approximate) probability that more than 20% of the sample is defective. *Answer: 1-normdist(.2, .15, sqrt(.15*.85/250),1) = 0.01341311 or 1-binomdist(50, 250, .15, 1) = 0.01298611*

(7.31) (a) binomial; (b) 126; (c) 10.86563; (d) .63; (e) .69; (f) .9090, or exact is .9074; (g) ≈ 1.

(7.32) (**video**) $n = 300$ and $\pi = .53$; $\mu = 159$ and $\sigma = 8.644652$; .84.

(7.33) $n = 1000$ and $\pi = .53$; $\mu = 530$ and $\sigma = 15.7829$; .9692, or exact .9691.

(7.34) .0565. (7.35) .0134. (7.36) (a) .94; (b) .5; (c) .13; (d) .81.

(7.37) (**video**) (a) .0708; (b) .7422; (c) .9854. Without correction: (a) .0571; (b) .7123, (c) .9821. Exact: (a) .0758; (b) .7363; (c) .9890

Sampling — Key Concepts

- *Population* or *universe* — collection of all units we wish to study, e.g., individuals who are 18 years or older living in the United States
- *Parameter* — a numerical fact about a population, e.g., average age, average intention to buy product x
- *Statistic* — a numerical fact about a sample that approximates a parameter, e.g., average age of a sample
- *Simple random sampling* (SRS or random sampling *without* replacement) — a sample where every possible collection of n units has the same chance of being drawn included (there are $\binom{N}{n}$ possible samples)
- Random sampling *with* replacement — repeat the following n items: (1) pick a unit at random; (2) replace the unit so that it can be selected again
  - For large[26] populations there is little difference between SRS and random sampling with replacement.
  - Formulae in computer packages usually use random sampling with replacement
  - For small populations you must adjust output

[26] The key is actually the *sampling fraction*, $f = n/N$. If the fraction sampled is large, e.g., $> 5\%$, you should worry about the difference.

Sampling Distribution Illustration

Consider the population consisting of the $N = 3$ elements $\{0, 2, 4\}$

- The population mean and variance are

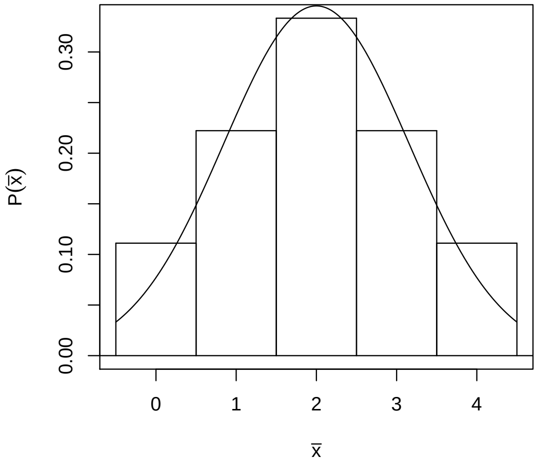
$$\mu_x = \frac{0+2+4}{3} = 2 \qquad \sigma_x^2 = \frac{(0-2)^2 + (2-2)^2 + (4-2)^2}{3} = \frac{8}{3}$$

- We wish to draw a sample of $n = 2$. The following samples (with replacement) are possible:

Sample	$\bar{x}$	**Sample**	$\bar{x}$	**Sample**	$\bar{x}$
0 0	0	2 0	1	4 0	2
0 2	1	2 2	2	4 2	3
0 4	2	2 4	3	4 4	4

- Sampling distribution of $\bar{x}$

$\bar{x}$	$P(\bar{x})$	$\bar{x}P(\bar{x})$	$P(\bar{x})(\bar{x}-\mu_{\bar{x}})^2$
0	1/9	0	4/9
1	2/9	2/9	2/9
2	3/9	6/9	0
3	2/9	6/9	2/9
4	1/9	4/9	4/9
Sum	1	$\mu_{\bar{x}} = 2$	$\sigma_{\bar{x}}^2 = 4/3$



- Note that $E(\bar{x}) = \mu_{\bar{x}} = \mu_x$

- The variance of the sampling distribution is $(\sigma_{\bar{x}}^2)$

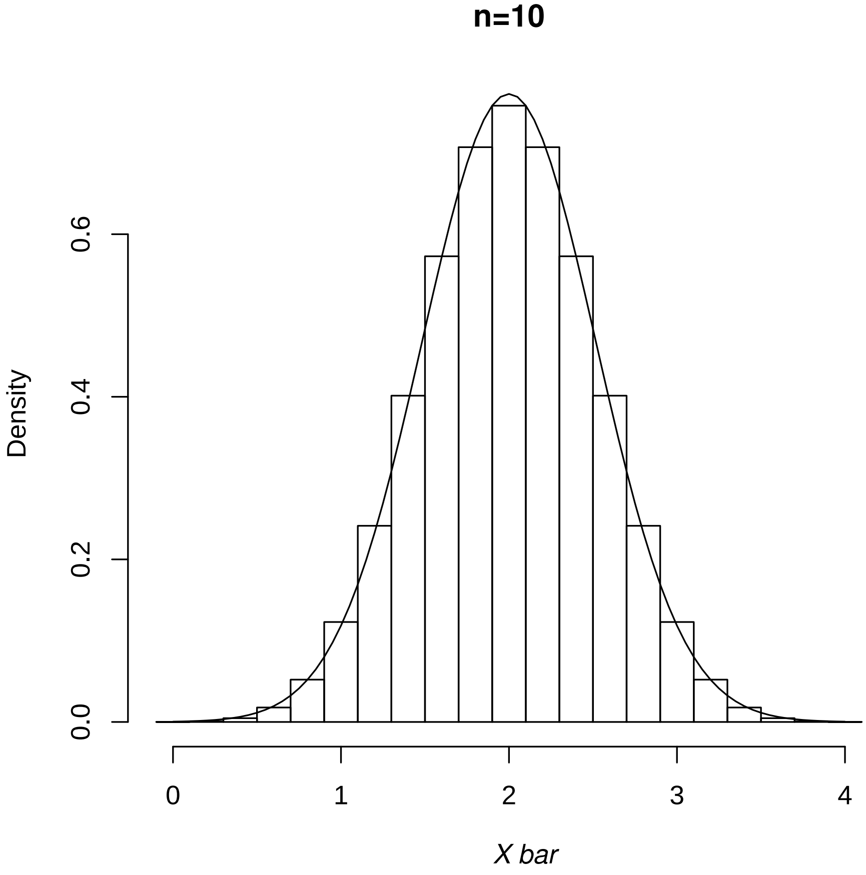
$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n} = \frac{8/3}{2} = \frac{4}{3}$$

Your Turn

Enumerate all $3^3 = 27$ possible samples of size $n = 3$ (with replacement) from the population $\{0, 2, 4\}$, find the sample means of each sample, and tabulate the sampling distribution. Show that the mean of the sampling distribution is still $E(\bar{x}) = 2$, but the variance is $V(\bar{x}) = (8/3)/3 = 8/9$. Plot the distribution and comment on the shape. Hint: you should find

$\bar{X}$	0	2/3	4/3	6/3=2	8/3	10/3	12/3=4
$P(\bar{X})$	1/27	3/27	6/27	7/27	6/27	3/27	1/27

Here is the sampling distribution for $n = 10$:



Sampling Distribution of the Total

Again, consider the population of $N = 3$ elements $\{0, 2, 4\}$

- The population mean and variance are

$$\mu_x = \frac{0+2+4}{3} = 2 \qquad \sigma_x^2 = \frac{(0-2)^2+(2-2)^2+(4-2)^2}{3} = \frac{8}{3}$$

- We wish to draw a sample of $n = 2$. The following samples (with replacement) are possible:

Sample	$\bar{x}$	$\hat{T}$	**Sample**	$\bar{x}$	$\hat{T}$	**Sample**	$\bar{x}$	$\hat{T}$
0 0	0	0	2 0	1	2	4 0	2	4
0 2	1	2	2 2	2	4	4 2	3	6
0 4	2	4	2 4	3	6	4 4	4	8

- Sampling distribution of $\hat{T}$ (and $\bar{x}$)

$\bar{x}$	$\hat{T}$	$P(\hat{T}) = P(\bar{x})$	$\hat{T}P(\hat{T})$	$P(\hat{T})(\hat{T}-\mu_{\hat{T}})^2$
0	0	1/9	0	16/9
1	2	2/9	4/9	8/9
2	4	3/9	12/9	0
3	6	2/9	12/9	8/9
4	8	1/9	8/9	16/9
Sum		1	$\mu_{\hat{T}} = 4$	$\sigma^2_{\hat{T}} = 16/3$

- Note that $E(\hat{T}) = \mu_{\hat{T}} = n\mu_x$
- The variance of the sampling distribution is $(\sigma^2_{\hat{T}})$

$$\sigma^2_{\hat{T}} = n\sigma_x^2 = 2 \times \frac{8}{3} = \frac{16}{3}$$

Sampling Distributions

Assume a *random* sample of size n from a population of size N, mean μ_x, and variance σ_x^2. Denote the measurements by $x_1, \ldots, x_n$. Then the following are true about the *sampling distributions* of the *sample total* $\hat{T} = x_1 + \cdots + x_n$ and *sample mean* $\bar{x} = \hat{T}/n$

- Theorem: On average we get the "true" value
$$E(\bar{x}) = \mu_{\bar{x}} = \mu_x \qquad \text{and} \qquad E(\hat{T}) = n\mu_x$$
- Theorem: If we sampled *with* replacement, then the *standard deviation of the mean* ($\sigma_{\bar{x}}$) and the *standard deviation of the total* ($\sigma_{\hat{T}}$) are
$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \qquad \text{and} \qquad \sigma_{\hat{T}} = \sigma_x\sqrt{n}$$
- Theorem: If we sampled *without* replacement, then[27]
$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \approx \frac{\sigma_x}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}}$$
- *Normal-normal theorem*: if the population distribution is normal, then the sampling distributions of $\bar{x}$ and $\hat{T}$ are normal. In particular, these have standard normal distributions:
$$Z = \frac{\bar{x} - \mu_x}{\sigma_x/\sqrt{n}} \qquad \text{and} \qquad Z = \frac{\hat{T} - n\mu_x}{\sigma_x\sqrt{n}}$$
- *Central limit theorem* (CLT): for all population distributions, the sampling distributions of $\bar{x}$ and $\hat{T}$ are *approximately* normal (for sufficiently large n)

[27] The *finite population correction* is $1 - n/N = 1 - f$.

Sampling Distribution Illustration

Suppose we have a population consisting of the $N = 5$ elements $\{0, 3, 9, 12, 15\}$ and we wish to draw a sample of $n = 3$. The following samples (without replacement) are possible $(5!/(3!2!) = 10$ samples):

Sample	$\bar{x}$	**Sample**	$\bar{x}$
0 3 9	4	0 12 15	9
0 3 12	5	3 9 12	8
0 3 15	6	3 9 15	9
0 9 12	7	3 12 15	10
0 9 15	8	9 12 15	12

- The population mean $\mu_x = (0 + 3 + 9 + 12 + 15)/5 = 7.8$
- The population variance σ_x^2 is

$$\sigma_x^2 = [(0 - 7.8)^2 + \cdots + (15 - 7.8)^2]/5 = 30.96$$

- Sampling distribution of $\bar{x}$

$\bar{x}$	$P(\bar{x})$	$\bar{x}P(\bar{x})$	$P(\bar{x})(\bar{x} - \mu_{\bar{x}})^2$
4	.1	0.4	1.444
5	.1	0.5	0.784
6	.1	0.6	0.324
7	.1	0.7	0.064
8	.2	1.6	0.008
9	.2	1.8	0.288
10	.1	1.0	0.484
12	.1	1.2	1.764
Sum	1	$\mu_{\bar{x}} = 7.8$	$\sigma_{\bar{x}}^2 = 5.16$

- Note that $E(\bar{x}) = \mu_{\bar{x}} = \mu_x = 7.8$
- The variance of the sampling distribution is $(\sigma_{\bar{x}}^2)$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \times \frac{N - n}{N - 1} = \frac{30.96}{3} \times \frac{5 - 3}{5 - 1} = 5.16$$

Your Turn

1. (Siegel 8.12) The mean account balance is \$500 and the standard deviation is \$120 for a large population of bank accounts. Find the standard deviation of the average balance of groups of eight accounts (chosen independently of one another). *Answer: 42.43*

2. (Siegel 8.14) (**video**) You have eight machines operating independently. The mean production rate for each machine is 20.3 tons per day, and the standard deviation is 1.4 tons per day. The distribution of production from a single machine is normal. How much do you expect to produce in total from the 8 machines and approximately how much uncertainty is there in the average daily production for the eight machines? What is the shape of the distribution? What is the mean and standard deviation of the the average production per machine? *Answer: $\sigma_{\bar{x}} = .495$ and $\sigma_{\hat{T}} = 3.9598$. Normal.*

3. (Siegel 8.20) (**video**) Breakfast cereal is packed into packages labeled "net weight 20 ounces, packed by weight not by volume; some settling may occur during shipment." However, weights of individual packages are not really all exactly equal to 20 ounces — although they are close, they do have some randomness. Based on past observation, assume that the mean weight is 20.04 ounces, the standard deviation is 0.15 ounce, and the distribution is approximately *normal*. Consider the average weight of 30 packages selected independently and at random. [Answers: (a) 20.04; (b) .0274; (c) .07]

   (a) What is the mean weight of this random variable?
   (b) What is the standard deviation of this variable?
   (c) What is the shape of its distribution?
   (d) What is the probability that the average weight is less than 20 ounces?

4. (Siegel 8.24) (**video**) A typical incoming telephone call to your catalog sales force results in a mean order of \$28.63 with a standard deviation of \$13.91. You may assume that orders are received independently of one another. *Answer: (a) No; (b) 3,149.30, 145.89; (c) approx normal; (d) .15; (e) .50*

   (a) Based only on this information, can you find the probability that a single incoming call will result in an order of more than \$40? Why or why not?
   (b) An operator is expected to handle 110 incoming calls tomorrow. Find the mean and standard deviation of the resulting orders ($\hat{T}$ = total dollars). You may assume that these 110 calls represent a random sample of calls.
   (c) What is the approximate shape of the probability distribution of the total order to be received by the operator in part b tomorrow? How do you know?
   (d) Find the (approximate) probability that the operator in part b will generate a total order more than \$3,300 tomorrow.

(e) Find the (approximate) probability that the operator in part b will generate an average order between \$27 and \$29 tomorrow.

5. (Siegel 8.25) (video) Your restaurant will serve 50 dinner groups tonight. Assume that the mean check size of dinner groups in general is \$60, the standard deviation is \$40, and the distribution is *slightly skewed with a longer tail towards high values.* [Answers: (a) 3000, 282.84; (b) 60, 5.6569; (c) normal; (d) .36; (e) .45]

   (a) Find the mean and standard deviation for the total of all 50 checks.

   (b) Find the mean and standard deviation for the average of all 50 checks.

   (c) What is the approximate shape of the distributions of the mean and total?

   (d) Find the probability that the total of all 50 checks is more than \$3,100, assuming a normal distribution.

   (e) Find the probability that the average of all 50 checks is between \$58 and \$65, assuming a normal distribution.

6. (Siegel 8.23) (video) You have analyzed a project using four scenarios, with the results shown in the table below. Suppose you actually have 40 projects just like this one and that they pay off independently of one another. [Answers: (a) $\mu_x = 5.55$ and $\sigma_x = 6.968$, not normal; (b) $E(\hat{T}) = 222$ and $\sigma_{\hat{T}} = 44.07$, normal from the CLT; (c) $E(\bar{x}) = 5.55$ and $\sigma_{\bar{x}} = 1.102$; (d) 0.0384; (e) .35]

Scenario	$P(X)$	X	$P(X) \times X$	$P(X)(X-\mu_x)^2$
Really bad	0.10	-10		
So-so	0.15	2		
Pretty good	0.50	5		
Great	0.25	15		

   (a) Find the mean and standard deviation of profit for a single project. Is the distribution normal?

   (b) Find the expected total profit and its standard deviation from all 40 projects. What is the shape of the distribution of total profit?

   (c) Find the mean profit per project and its standard deviation from all 40 projects.

   (d) Find the (approximate) probability that your total profit will exceed 300.

   (e) Find the (approximate) probability that your average profit per project will be between \$5 million and \$6 million.

Summary of Formulas

Statistic	"Expected" Mean	"Uncertainty" Std. Dev.
X_B Binomial count "How Many?" Number	$E(X_b) = n\pi$	$\sigma_{X_B} = \sqrt{n\pi(1-\pi)}$
p Sample percent	$E(p) = \pi$	$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$
$\bar{x}$ Sample mean	$E(\bar{x}) = \mu_x$	$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$
$\hat{T}$ Sample total	$E(\hat{T}) = n\mu_x$	$\sigma_{\hat{T}} = \sigma_x\sqrt{n}$

Your Turn: Siegel Chapter 8

1. (a) Bad; (b) bad; (c) systematic sample, acceptable, but not good; (d) SRS, good; (e) stratified random sample, good. The sample in (c) will contain transmissions from throughout the day and may even be more "representative" than the SRS in (d), but the main problem is that you cannot assess sampling error (you have a random sample of size 1, the starting point of your every-20th sample).

2. Sample c

3. Sample d

5. (a) statistic; (b) parameter; (c) parameter; (d) statistic; (e) parameter; (f) statistic

8.12. 42.43

8.13. 38.73

8.14. (video) $\sigma_{\bar{x}} = .495$ and $\sigma_{\hat{T}} = 3.9598$. The wording is not completely clear on this problem.

8.15. $-32,000$

8.16. (a) .5398; (b) .0668

8.17. $\sigma_{\bar{x}} = 10.39$

8.18. $s_p = 0.02601$

8.19. (a) 90; (b) 5.534; (c) approx normal; (d) .1477.

8.20. (video) (a) 20.04; (b) .0274; (c) .07

8.21. .1170

8.22. .14

8.23. (video) .35 Also find the (approximate) probability that your total profit will exceed 300 (answer: 0.0384):

8.24. (video) (a) No; (b) 3,149.30, 145.89; (c) approx normal; (d) .15; (e) .50

8.25. (video) (a) 3000, 282.84; (b) 60, 5.6569; (c) independence, n large enough for CLT to apply (no outliers, etc.) (d) .36; (e) .45

8.26. (a) 117,045; (b) 8,552.96; (c) CLT; (d) .92; (e) .02; (f) .62; (g) .04; (h) .57

8.27. (a) 100; (b) 2.3717; (c) approx normal, CLT; (d) .20; (e) .79

8.29. (a) see me if you have problems; (c) 5.5%

Standard Errors

- We usually need to know the *standard deviations of the mean* ($\sigma_{\bar{x}} = \sigma_x/\sqrt{n}$) and *total* ($\sigma_{\hat{T}} = \sigma_x\sqrt{n}$). Likewise for the *standard deviation of a percent* ($\sigma_p = \sqrt{\pi(1-\pi)/n}$).
- In practice, we rarely know σ_x or π, so we cannot compute $\sigma_{\bar{x}}$, $\sigma_{\hat{T}}$, or σ_p.
- Our solution is to estimate them with *standard errors*

	Standard	Standard
Statistic	Deviation	Error
X_B	$\sigma_{X_B} = \sqrt{n\pi(1-\pi)}$	$S_{X_B} = \sqrt{np(1-p)}$
p	$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$	$S_p = \sqrt{\frac{p(1-p)}{n}}$
$\bar{x}$	$\sigma_{\bar{x}} = \sigma_x/\sqrt{n}$	$S_{\bar{x}} = S_x/\sqrt{n}$
$\hat{T}$	$\sigma_{\hat{T}} = \sigma_x\sqrt{n}$	$S_{\hat{T}} = S_x\sqrt{n}$

where, from slide 56, the *sample standard deviation* is

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

Names are formed by prepending "standard error" to the statistic, e.g., the *standard error of the mean* is $S_{\bar{x}}$.

Your Turn: Siegel Chapter 8

8.31. (**video**) (a) 499.38; (b) 280.70; (c) 77.85; (d) 249,692

8.32. No, standard deviation and standard error confused

8.33. 1.15

8.34. (a) 2.33; (b) 1.17

8.35. 94, 40.82

8.36. 115.35, 9.35. Note that there is a typo in the fifth edition of Siegel — this is problem 5.4.23.

8.37. 1.658, .00738

8.38. .0168

8.39. (a) 19.92%; (b) .0145

8.40. (**video**) (a) .0063; (b) no, 4% is only 0.63 standard errors higher than 3.6%; (c) yes, because 10% is 10.12 standard errors higher than 3.6%

8.41. (a) 84%; 7.33%

Your Turn

1. (8.31) (video) Here is a list of the dollar amounts of recent billings: 994, 307, 533, 443, 646, 148, 307, 524, 71, 973, 710, 342, 494. [Answers: (a) 499.38; (b) 280.70; (c) 77.85; (d) 249,692]

   (a) Find the average sale. What does this number represent?

   (b) Find the standard deviation. What does this number represent?

   (c) Find the standard error. What does this number represent?

   (d) You are anticipating sending another 500 billings similar to these next month. What total amount should you forecast for these additional billings?

2. (8.40) (video) Based on careful examination of a sample of size 868 taken from 11,013 inventory items in a warehouse, you learn that 3.6% are not ready to be shipped. [Answers: (a) .0063; (b) No, 4% is only .63 standard error higher than 3.6%; (c) yes, because 10% is 10.12 standard error higher than 3.6%]

   (a) Find the standard error associated with this estimated percentage and indicate its meaning.

   (b) Would you be surprised to learn that in fact 4% of the 11,013 inventory items are not ready to be shipped? Why or why not?

   (c) Would you be surprised to learn that in fact 10% of the 11,013 inventory items are not ready to be shipped? Why or why not?

3. (8.41) From a list of 729 people who went on a cruise, 25 were randomly selected for interview. Of these, 21 said that they were "very happy" with the accommodations. [Answers: (a) 84%; 7.33%]

   (a) What percent of the sample said that they were "very happy?"

   (b) If you had been able to interview all 729 people, approximately how different a percentage would you expect to find as compared to your answer in part a? To answer this, please indicate which statistical quantity you are using, and compute its value.

Sampling Controversy

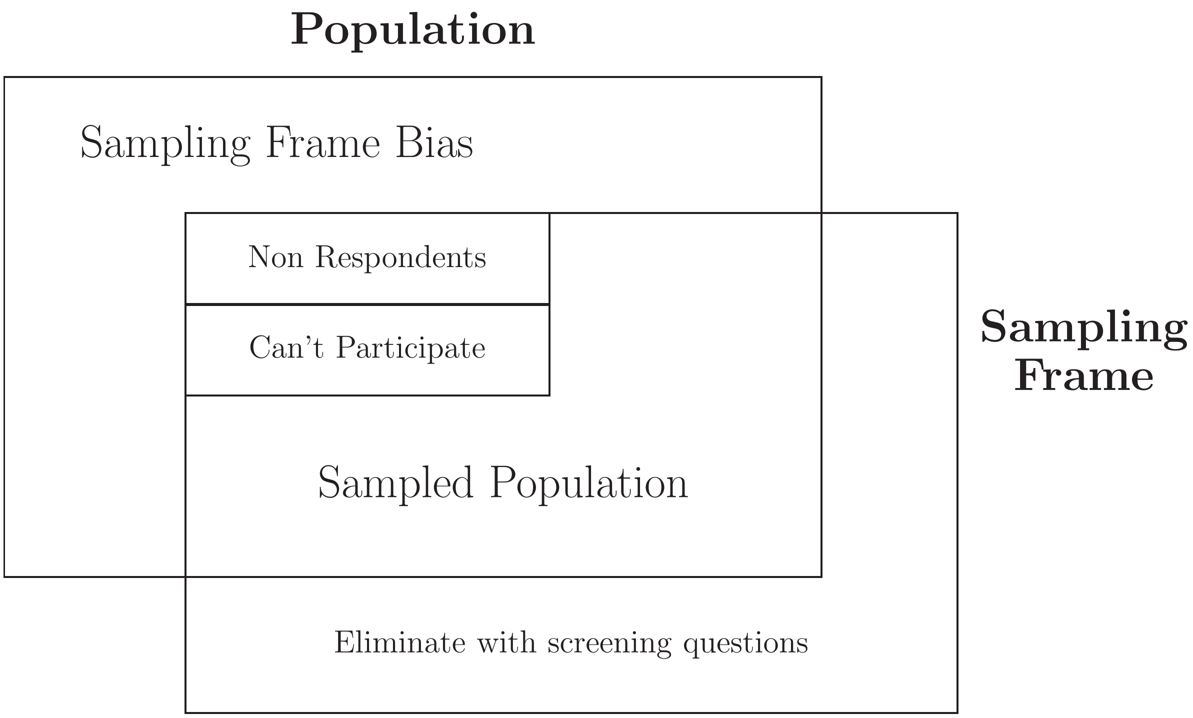
- Based on a sample of size $n = 4,500$, Shere Hite's book *Women and Love: A Cultural Revolution in Progress* (1987) reports that
  - 84% of women are "not satisfied emotionally with their relationships" (p. 804)
  - 70% of all women "married five or more years are having sex outside of their marriages" (p. 856)
  - 95% of women "report forms of emotional and psychological harassment from men with whom they are in love relationships" (p. 810)
  - 84% of women report forms of condescension from the men in their love relationships (p. 809)
- *Self-selection bias*: 100,000 questionnaires mailed, 4.5% responded
- *Sampling frame bias*: Questionnaires mailed to organizations such as professional women's group, counseling centers, church societies, and senior citizens' centers.
- *Measurement bias*: Questionnaire consisted of 127 essay questions, most with several parts. Who will tend to return such a survey?
- *Measurement bias*:
  - Many questions used vague words such as *love* (which is a latent variable)
  - Many *leading questions*, e.g., "Does your husband/lover see you as an equal? Or are there times when he seems to treat you as an inferior? Leave you out of the decisions? Act superior?" (p. 795)

Sampling: Eight-Step Procedure

1. Define the population
2. Identify a sampling frame
3. Select a sampling plan
4. Determine sample size
5. Select sampling units
6. Execute survey, e.g., see (in alphabetical order) Comscore, Ipsos, MySurvey, Simmons, SurveyLot, SurveyMonkey, Zoomerang
7. Second attempt to contact nonrespondents if necessary, e.g., second mailing.
8. Summarize results and do statistical inference

Sampling Frames

- *Sampling frame* — A listing of population units (sample is chosen from these units).
- *Sampling frame bias* — A bias that occurs when the population as implied by the sampling frame differs from the (ideal) population in a systematic way.



Your Turn

(Hypothetical) The Chicago Transit Authority (CTA) commissioned a market research agency to conduct a survey measuring attitudes toward the CTA, satisfaction and dissatisfaction with the services, and services that people would like to see offered. They drew 1000 names at random from the Chicago phone book (white pages), called each name, and administered the survey.

1. What is population?
2. What is sampling frame?
3. Give examples of groups that omitted because of the sampling frame (sampling frame bias).
4. Discuss how the omitted groups may have systematically different responses than the sampled population.

Probability Versus Non-probability Sampling

Probability samples — Every element of the population has a *known, non-zero* probability of inclusion in the sample.

Non-probability samples — Population elements are selected in a non-random manner.

Probability samples imply:

- One can quantify the amount of sampling error that is introduced because a sample is used instead of a census.
- Usually a higher cost.

With many samples, we must identify factors that introduce biases and take them into consideration when forming our conclusions

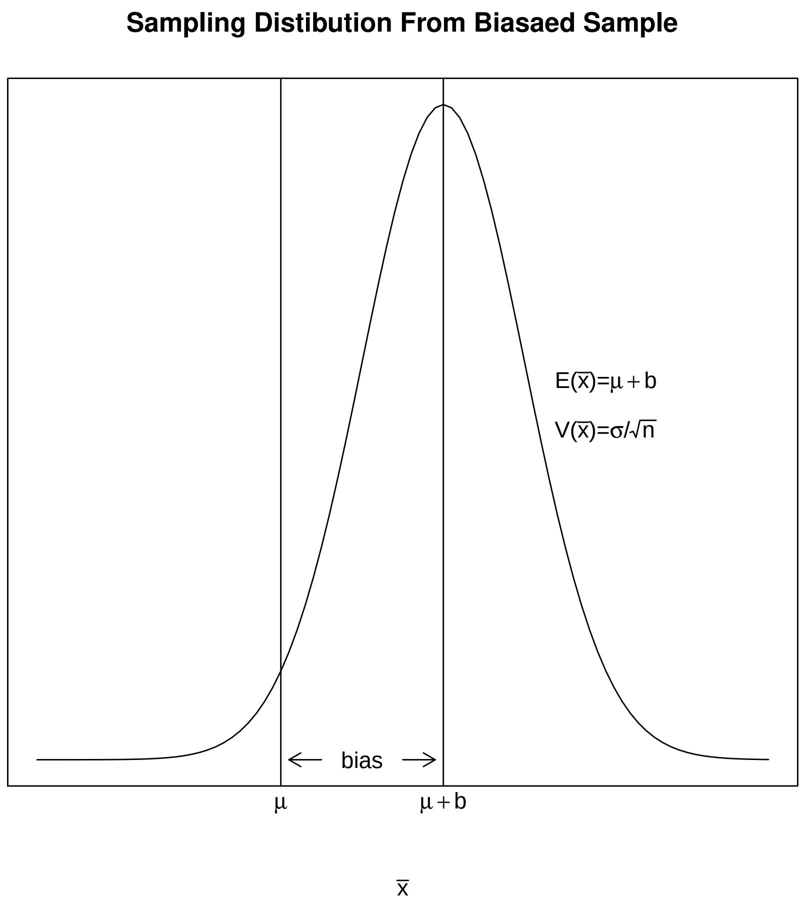
Bias Versus Sampling Error

$$\text{statistic} = \text{parameter} + \text{bias} + \text{sampling error}$$

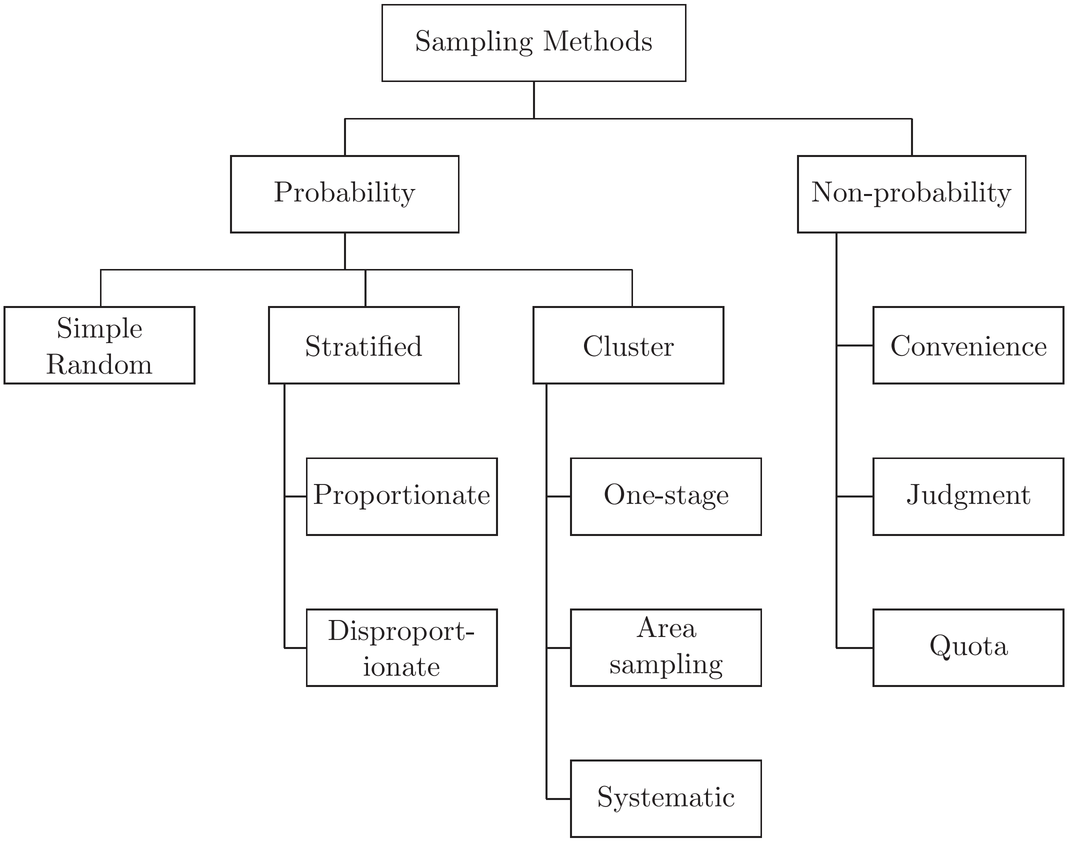
- The "sampling error" has mean 0
- A statistic (estimate) is said to be *unbiased* if the "bias" term is 0 — on average the estimate equals the parameter
- If the statistic is the sample mean ($\bar{x}$) then

$$\bar{x} = \mu + b + e$$

and the standard deviation of e is $\sigma/\sqrt{n}$



Types of Sampling Procedures



Sampling Methods

Probability

Non-probability

Simple Random

Stratified

Cluster

Proportionate

Disproportionate

One-stage

Area sampling

Systematic

Convenience

Judgment

Quota

Non-sampling Errors (Biases)

Non-sampling Errors: the difference between the mean of the statistic and the parameter

- Respondent errors
  - *Self-selection bias*: when sampled respondents choose whether or not to participate in a survey.
  - *Deliberate falsification*: when respondents deliberately give false answers on a survey
  - *Unconscious misrepresentation*: when respondents misread questions.
- Administrative errors
  - *Interviewer error*: interview biases respondent, cheats, etc.
  - *Administrative error*: mistakes made in keying, coding, recording data.

Non-probability Samples

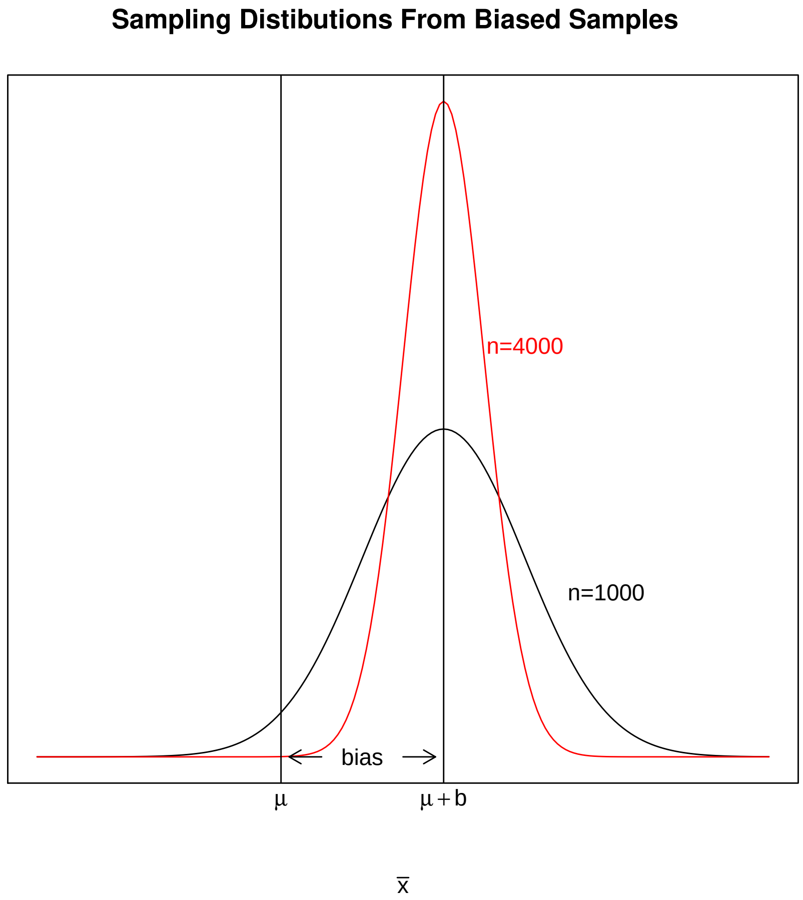
- *Convenience sample* — those included enter by accident in that they just happen to be where the study is being conducted.
- *Judgmental sample* — respondents handpicked by researcher because they are expected to serve research purpose.

Circumstances under which a non-probability sample may be used:

- Exploratory research, e.g., focus groups
- When probability samples are too costly (e.g., poor sampling frame)
- When sample size is too small (e.g., test marketing)

Large Biased Samples are Still Biased

When a selection procedure is biased, taking large samples does not help. This just repeats the basic mistake on a larger scale.



Larger samples produce statistics that are, on the average, closer to $\mu + b$, but we want to know μ, not $\mu + b$. **Larger biased samples give you a more precise measure of the wrong thing.**

Example Convenience Sample: Landon-Roosevelt Election

- In 1936 first-term President Roosevelt was running against Alfred Landon
- Observers thought Roosevelt would be an easy winner
- *Literary Digest* magazine, which had correctly predicted elections since 1916, predicted overwhelming victory for Landon 57%–43%
- Actual result: Roosevelt 62% and Landon 38%
- *The Literary Digest* prediction based on sample of size 2.4 million
  - Response rate: mailed to 10 million people, 24% rate
  - Sampling frame: telephone books (25% of households had phones), club membership lists, automobile registration lists
- **Non-respondents can be very different from respondents. When there is a high non-response rate, look out for non-response bias.**

Can This Survey Be Saved?

	Pilot Study	First Sample	Second Sample	Both Samples	All Combined
Initial Mailing					
Mailed	12	400	200	600	612
Responses	12	216	120	336	348
Average	$39,275	$3,949	$3,796	$3,894	$5,114
Std Dev	$9,062	$849	$868	$858	$6,716
Follow-up Mailing					
Mailed	0	184	80	264	264
Responses	0	64	18	82	82
Average		$1,238	$1,262	$1,244	$1,244
Std Dev		$153	$157	$153	$153
Initial+Follow-up					
Mailed	12	400	200	600	612
Responses	12	280	138	418	430
Average	$39,275	$3,330	$3,465	$3,374	$4,376
Std Dev	$9,062	$1,364	$1,180	$1,306	$6,230

1. Which data sets should be used? Should you include the pilot study? Should you use the first sample? Second sample? Both?
2. Are the follow-ups a good idea (ignore $1 incentive for now)? Should you use the initial, follow-up, or both?
3. Is the $1 incentive a good idea?
4. Why are there differences in the estimated averages?
5. Using the sample of your choice, do you think that the average under-estimates the true mean, over-estimates it, or is about right (on average)? Can you think of a better estimate of the mean?
6. Does it make sense to compute a 95% confidence interval? Discuss.

Your Turn

1. For each of the following situations, identify the appropriate target population and sampling frame.[28]

   (a) The national Head Injury Foundation Inc. wants to test the effectiveness of a brochure soliciting volunteers for its local chapter in Indianapolis, Indiana.

   (b) A regional manufacturer of yogurt selling primarily in the Pacific Northwest wants to test market three new flavors of yogurt.

   (c) A national manufacturer wants to assess whether adequate inventories are being held by wholesalers in order to prevent shortages by retailers.

   (d) A large wholesaler dealing in electronic office products in Chicago wants to evaluate dealer reaction to a new discount policy.

   (e) Your school cafeteria system wants to test a carbonated milk product manufactured in the Food Science department and sold by the cafeteria system.

   (f) A local cheese manufacturer wants to assess the satisfaction with a new credit policy offered to mail-order customers.

   (g) A regional manufacturer of hog feed wants to conduct an on-farm test of a new type of hog feed in Carroll County, Indiana.

2. A leisure-wear manufacturer wants to determine consumer preference for several varieties of t-shirts. Respondents will participate in a touch test; that is, they will touch several different t-shirts and then state their preferences. Some aspects of the t-shirts that will be compared are armbands, neckbands, and shirt material. The marketing

[28] (a) The population could be defined as all persons 16 years of age or older currently residing in Indianapolis. The sampling frame could be the city directory (for example R.L. Polk's city directory) which would provide a list of addresses from which samples of dwelling units could be developed. (b) The population could be defined as all cities in the Pacific Northwest. The sampling frame would be a list of the cities that are representative in terms of certain external criteria. The Statistical Abstract of the United States could be used to compare cities on characteristics deemed important to consumption of the yogurt products. (c) The population could be defined as all wholesalers that distribute the manufacturers' products. The sampling frame could be the list of such wholesalers which should be available in the company's records. (d) The population could be defined as all retailers that sell electronic office products in Chicago. The sampling frame might be the telephone book yellow pages. Alternatively, there might be an association of office product suppliers. If so, the membership roster could serve as the sampling frame. (e) The population could be defined as all students enrolled at your university. The sampling frame would be the listing of students currently enrolled, which would be available at the Registrar's office. Alternatively, a student directory could be used. (f) The population could be defined as all persons 18 years of age or oder in the particular town or city or alternatively it could be defined simply as current mail order customers. The sampling frame could b the city or county directory which would provide a list of addresses from which samples of dwellings can be developed. If the focus was current mail order customers, then a list of the manufacturer's mail order customers would serve as the sampling frame. (g) The population could be defined as all farms with hogs in Carroll County, In. The county directory might be used as the sampling frame. However, we would only want to include those farms in the sample with hogs. Since this might be hard to determine a priori, some screening questions would have to be used. Alternatively, there might be a county or state hog producer's organization. If so, the membership list or a portion thereof could serve as the sampling frame.

researcher conducting the study has recommended that the touch tests be conducted by mall intercepts.[29]

(a) What problems, if any, do you see in trying to use the results of this study to estimate the population of consumers who buy t-shirts?

(b) Suggest an alternative to a mall intercept study for determining consumer preferences for t-shirts. Why is this a better method?

3. A leading film-processing company wishes to investigate the market potential for a new line of digital-processing equipment. Because this is a completely new technology, the company believes that industry leaders might offer insights into the desires of customers and consumers. However, a list of industry leaders does not exist. Discuss possible methods of generating a sampling frame for the influential people in the digital-processing industry.[30]

4. The My-Size Company, a manufacturer of clothing for large-sized consumers, was in the process of evaluating its product and advertising strategy. Initial efforts consisted of a number of focus-group interviews. Each focus group consisted of 10 to 12 large men and women of different demographic characteristics who were selected by the company's research department using on-the-street observations of physical characteristics.[31] (a) What type of sampling method was used? (b) Critically evaluate the method used.

5. The owners of a popular bed and breakfast inn in Door County, Wisconsin, had noticed a decline in the number of tourists and length of stay during the past three years. An overview of industry trends indicated that the overall tourist trade was expanding and growing rapidly. The managers decided to conduct a study to determine people's attitudes toward the particular activities that were available at the inn. Because they wanted to cause the minimum amount of inconvenience to their guests, the owners devised the following plan. Interview request cards, which were available at the chamber of commerce office, the visitor information center, and three of the more popular restaurants in Sturgeon Bay, indicated the nature of the study and encouraged visitors to participate. Visitors were asked to report to a separate room at either the chamber

[29] (a) In a mall intercept only people at the mall can be included in the study, so t-shirt consumers who only shop at discount stores (e.g., Walmart) will not be included in the study. This would bias the study by not sampling the entire target population of t-shirt consumers. A mall intercept study usually consists of a convenience sample, which may or may not project to the entire population. (b) One potential alternative to a mall intercept is to recruit a probability sample by phone to a central location. The central location may be a shopping center; the distinction lies in that by recruiting probabilistically by phone the sampling method is improved twofold. First by using probability sampling the study can be projected to the target population. Second, the target can be defined as more than people who shop at the mall.

[30] Some possible sources of names to develop the sampling frame include: (1) the names of authors of journal articles on digital processing; (2) names listed in footnotes of journal articles on digital processing; (3) trade shows for new technologies; (4) word of mouth from suppliers, distributors, and resellers of image processing equipment. After an initial list is generated, ask the people on the list to name other influential leaders in the digital processing industry. This is called a *snowball sample*.

[31] (a) The method is a judgment sample because the respondents were selected on certain prescribed physical characteristics. (b) Here again a judgment sample is appropriate as a convenience or random sample might not generate subjects that possess the required characteristics.

of commerce office or the visitor information center. Personal interviews, lasting 20 minutes, were conducted at these locations.[32] (a) What type of sampling method was used? (b) Critically evaluate the method used.

6. A national manufacturer of processed meats was planning to enter the Japanese market. Before the final decision about launching its product, management decided to test market the products in two cities. After reviewing the various cities in terms of external criteria, such as demographics, shopping characteristics and so on, the research department settled on the cities of Yokohama and Hiroshima.[33] (a) What type of sampling method was used? (b) Critically evaluate the method used.

[32] (a) The method is a convenience sample, as volunteers were asked to come forward. (b) Convenience samples may or may not be representative. In this case, volunteers might possess the characteristics of being helpful and kind and might provide a biased or favorable view of the actual situation. Alternatively, a random sample of visitor attractions or eating establishments could have been selected and questionnaires could have been left for visitors to complete. This, of course, could produce a problem in non-response bias because some guests would probably not complete the questionnaire. An alternative strategy would be to select the local attractions or eating establishments randomly and attempt to conduct personal interviews with the visitors at them.

[33] (a) The method used is a judgment sample, as cities that were deemed most representative were selected. (b) The judgment sample often exhibits selection bias. However, in this situation the approach is probably satisfactory since only two cities are to be used. Further, since the cities were evaluated on various external characteristics as being representative, selection bias will probably be negligible, although Yokohama is fairly unusual.

Steps in Questionnaire Design

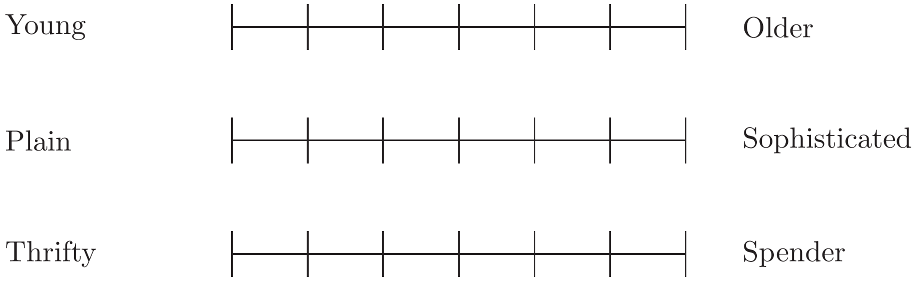
1. Specify information sought (e.g., scenario tables)
2. Decide on administration method
3. Write questions
4. Order sections and questions
5. Format survey
6. Pretest, pretest, pretest
7. Make sure survey will meet objectives (when results come back, will you be able to make actionable suggestions?)
8. Execute survey

As a general rule of thumb, seek to understand relationships: How do things under your control relate to outcomes you want to achieve?

Types of Questions

- *Semantic Differential (antonyms)*

The typical client of Jacques' Hair Care is



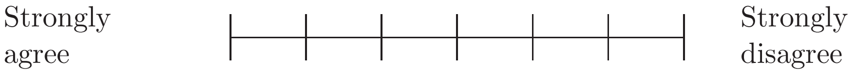
Young — Older

Plain — Sophisticated

Thrifty — Spender

- *Likert (strongly agree . . . strongly disagree)*

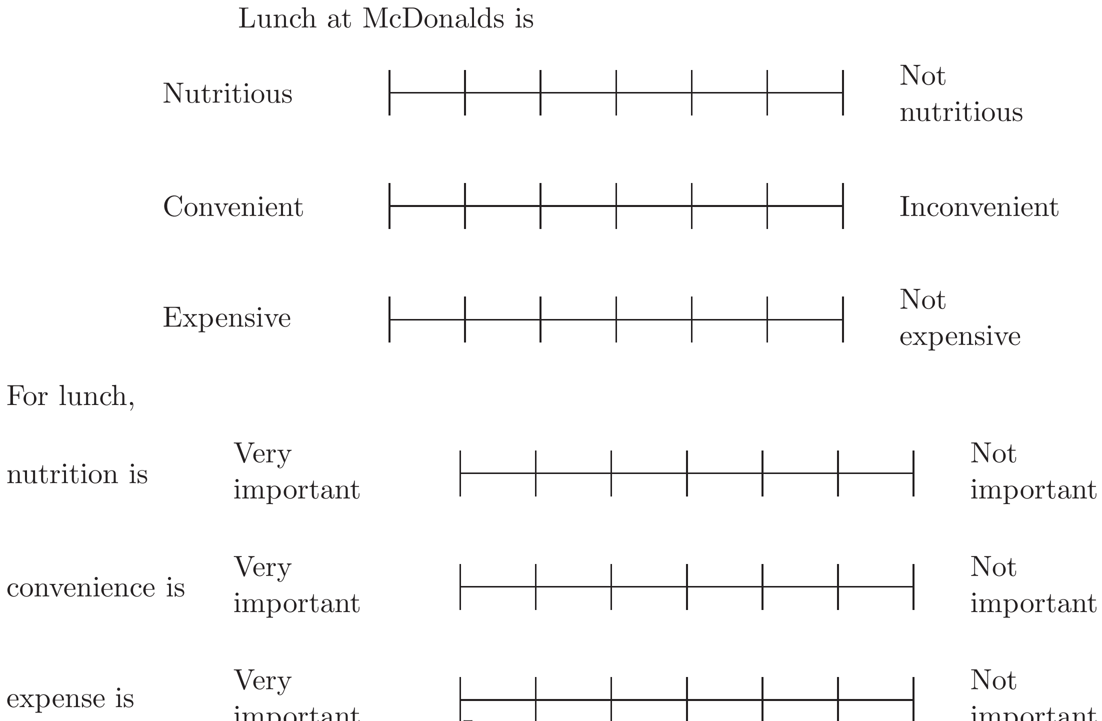
Cancun is my favorite vactions spot



Strongly agree — Strongly disagree

- Guidelines
  - Allow for midpoint
  - 3 or 5 points for phone
  - 5, 7, 9 or 11 (0-10) points for mail
  - Provide anchors on each end
- These scales are easily analyzed. Use them!

Example Attributes and Their Importance



Lunch at McDonalds is

Nutritious — Not nutritious

Convenient — Inconvenient

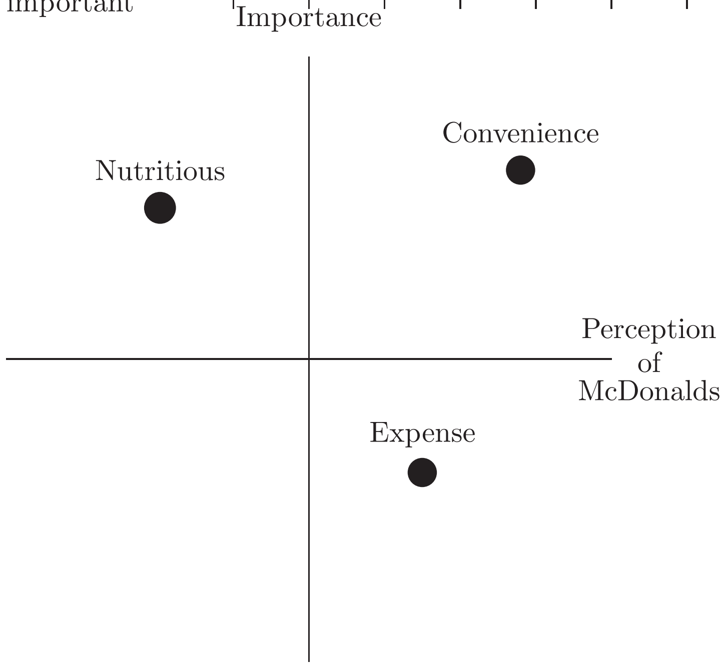
Expensive — Not expensive

For lunch,

nutrition is Very important — Not important

convenience is Very important — Not important

expense is Very important — Not important



Ranks versus Ratings

1. Please rank the following movies (1=most liked, 3=least liked)

    ____ Gone with the Wind

    ____ Raiders of the Lost Ark

    ____ It’s a Wonderful Life

2. Rate these movies

Gone with the Wind	Really Liked	├──┼──┼──┼──┼──┼──┤	Didn’t like
Raiders of the Lost Ark	Really Liked	├──┼──┼──┼──┼──┼──┤	Didn’t like
It’s a Wonderful Life	Really Liked	├──┼──┼──┼──┼──┼──┤	Didn’t like

3. In general, ratings are much easier to analyze (can use correlation, regression, factor analysis, ...)

4. Rankings are difficult to analyze, but force respondent to make a choice.

Important Issues in Questionnaire Design

- *Can the research objective(s) be fulfilled without asking this question?*
  - Exactly how am I going to use the data generated by this question?
  - What decision will it help me make?
- *Use the language of the population to be surveyed and avoid "internal marketing speak."*

  Q: How is Target positioned relative to Wal-Mart?
- *Meaning of words should be clear.*

  Q: How many members are there in your family?

  Q: What is your income?

  Q: Do you use product X regularly?
- *Avoid double barrelled questions.*

  Q: Do you believe that McDonalds has *fast* and *courteous* service?
- *Do respondents have the necessary information?*

  In a survey of lawyers and the general public, 50% of the lawyers and 75% of the general public expressed an opinion on the performance of the "National Bureau of Consumer Complaints."

Important Issues in Questionnaire Design

- *Avoid negatives and double negatives*

  Q: Students should not be required to take a comprehensive exam to graduate.

  Q: It is not a good idea to not turn in homework on time.

- *Avoid leading/loaded questions.*

  Q: Do you think the US should *allow* public speeches against democracy?

  Q: Do you think the US should *forbid* public speeches against democracy?

  44% of a sample said "no" to the first question and 28% of a similar sample said "yes" to the second question.

  Q: What did you dislike about the product you just tried?

  Q: Did you dislike any aspect of the product you just tried?

  Q: If yes, then what did you dislike about the product you just tried?

  Q: Do you think Johnson and Johnson did *everything possible* in its handling of the Tylenol poisoning case?

- *Avoid Complex Questions.*

  Q: Of the total number of miles you drove during the past month, approximately what percentage was for driving to and from work?

Important Issues in Questionnaire Design

- *Response categories should not overlap.*

  Q: What is your age?

  □ under 20

  □ 20 – 30

  □ 30 – 40

  □ 40+

- *Avoid open-ended questions.*

  Q: What do you like best about our product?

  - Require more time and effort to answer than closed-ended questions.
  - Difficult and subjective to code.
  - Different respondents will give different levels of detail, e.g., some list 5 things, others only 1 or 2.
  - When a respondent doesn't mention something, is it because the respondent forgot about it, or because it wasn't important?
  - It is better to use in-depth interviews or focus groups to develop closed-ended questions.

- *When appropriate, allow for "Not applicable."*

  Q: How satisfied or disatisfied are you with the gynecological services offered by your health provider?

Important Issues in Questionnaire Design

- *Allow for "Don't know" responses when appropriate.* When respondents don't know anything about a question and are forced to answer, their responses add only noise.
  - *Standard format.* "Don't know" response category not included.

    Q: "Do you agree or disagree that the Russian leaders are basically trying to get along with America?"
  - *Quasi filter.* Offer "don't know" response category

    Q: "Do you agree, disagree or have no opion the the Russian …"
  - *Full filter.* First ask if respondent has opinion, and, if yes, ask question.

    Q: "Here is a statement about another country. Not everyone has an opinion on this. If you do not have an opinion then say no. Do you have an opinion on whether the Russian …" If yes, "Do you agree or disagree that the Russian …"

	Standard	Quasi Filter	Full Filter
Agree	48.2%	27.7%	22.9%
Disagree	38.2%	29.5%	20.9%
No opinion	13.6%	42.8%	56.3%

Important Issues in Questionnaire Design

- *Avoid Implied Assumptions.*

  Q: Are you in favor of requiring all new refrigerators to be built with the most effective insulation available as an energy conservation measure?

  "Even though it will mean a 25% increase in the retail price of the refrigerator"

- *Order Bias.* Does the order of questions matter? Six studies were conducted where respondents rated the effectiveness of three analgesics. The order of the questions were different in each study. Scores were normed to sum to 100.

Order 1		**Order 2**		**Order 3**	
Tylenol	10	Tylenol	15	Bufferin	30
Bufferin	60	Excedrin	55	Excedrin	10
Excedrin	30	Bufferin	30	Tylenol	60

Order 4		**Order 5**		**Order 6**	
Bufferin	35	Excedrin	25	Excedrin	35
Tylenol	30	Tylenol	5	Bufferin	35
Excedrin	35	Bufferin	70	Tylenol	30

Solution: use software to randomize response categories.

Your Turn: Are these good questions?

- You don't smoke, do you?
- Does your employer provide health insurance and a pension?
- How do you spend your free time? reading, watching TV, or what?
- My grocery store purchases exhibit brand loyalty (strongly agree · · · strongly disagree).
- How did you feel about your brother when you were 6 years old?
- I eat out frequently (strongly agree · · · strongly disagree).
- Most Americans believe social security will be defunct when they retire; do you?
- Miller Lite is (less filling · · · tastes great)
- How often do you drink too much?
- If a new grocery store were to open down the street, would you shop there?
- How many gallons of gasoline did you buy last year?
- What should be done about murderous terrorists who threaten the freedom of good citizens and the safety of our children?

Question Sequencing

- Screening questions (make sure subject is in defined population)
- Start zippy — grab their attention
- Within section, start general, get specific
- Don’t ask difficult or personal questions at the beginning. Put them towards the end.
  - Allow respondents to get comfortable first and “warmed up” so that they are more likely to answer.
  - If they get mad and quit, you have at least some answers from them.
- Put demographics and sensitive questions (which may cause them to terminate early) at end

Practice Midterm Questions

1. On a given day, assume that there is a 30% chance that you will receive no orders, a 50% chance that you will receive one order, and a 20% chance of two orders.

   (a) Find the expected number of orders.

   (b) Find the standard deviations of the number of orders.

   (c) (A–) For this part, suppose that your fixed costs are \$500 per day and that you make \$1000 in profit from each order. Compute the expected *profit* from a single day. Hint: think of profit as a *linear transformation* of orders.

   (d) Continuing part (c) find the standard deviation of *profit* from a single day.

   (e) A "bad day" is when you receive no orders. Assume that the number of orders you receive on a particular day is independent of the number of orders you receive on other days. Out of the next 5 days, find the probability that you have exactly 3 bad days.

   (f) Continuing the previous part, find the probability of having at least one bad day out of the next 5.

   Answer: (a) 0.9; (b) 0.7; (c) $y = -500 + 1000x$ *so* $\mu_y = -500 + .9(1000) = 400$*; (d)* $\sigma_y = .7|1000| = 700$*; (e)* $P(X_b = 3) = \binom{5}{3}0.3^3(1 - 0.3)^2 =$ *0.1323; (f)* $P(X_b \geq 1) = 1 - P(X_b = 0) = 1 - (1 - .3)^5 =$ *.83193.*

2. (B+) You are planning to interview 400 randomly selected consumers from a large list of likely sales prospects, to assess the value of this list and whether you should assign sales people to the task of contacting them all. Assume that 13% of the large list will respond favorably, as opposed to not responding favorably. Assume also that the consumers respond independently of each other.

   (a) Out of your sample of 400, how many consumers do you expect to respond favorably?

   (b) Find the standard deviation of the *number* of consumers who will respond favorably.

   (c) Find the standard deviation of the *percent* of consumers who will respond favorably.

   (d) Find the probability that more than 15% will respond favorably.

   (e) Find the probability that between 50 and 60 will respond favorably (i.e., $50 \leq X \leq 60$).

   (f) Find the probability that less than 50 or more than 60 will respond favorably (i.e., P($X < 50$ or $X > 60$)).

   (g) Find the 99th percentile of the random variable "number of consumers who will respond favorably out of 400."

   (h) For this part only, consider the first 15 people you call in your sample. Find the probability that exactly 3 consumers respond favorably out of the first 15 you call.

   Answer: Note that $n = 400$ *and* $\pi = .13$*. (a)* $E(X_b) = 400 \times 0.13 = 52$*. (b)* $\sigma_x = \sqrt{400 \times 0.13 \times 0.87} = 6.7261$*. (c)* $\sigma_p = \sqrt{0.13 \times 0.87/400} = 0.016815$*. (d)* $P(p > 0.15) = P(Z > (0.15 - 0.13)/0.01685) =$ *0.1171407. For exact, note* $.15(400) = 60$*, so use* $P(p >$

.15) $= P(X_b > 60) = 1 - P(X_b \leq 60) =$ *0.1048979. (e) Exact:* $P(50 <= X_b <= 60) = P(X_b \leq 60) - P(X_b \leq 49) =$ *0.5340634. Using normal approximation,* $P(49.5 < X_n < 60.5) \approx P(-0.37 < Z < 1.26) = 0.8962 - 0.3557 = 0.5405$. *(f)* $1 - 0.5405 = 0.4595$. *(g)* $2.33 = (x - 52)/6.7261$ *implies* $x = 67.7$. *(h)* $455 \times 0.13^3 \times 0.87^{12} = 0.1879631$.

3. The arrival time of the 7 p.m. flight from Los Angeles is known to be normally distributed with a mean arrival time of 7:00 p.m. and a standard deviation of 15 minutes.

   (a) What time should you arrive at the gate so that there is only a 5% chance that the flight arrives at the gate before you?

   (b) Suppose a friend of yours is on the flight. What time should you arrive at the gate so that there is only a 10% chance that your friend will have to wait more than 20 minutes to see you?

   Answer: (a) about 6:35; (b) 0.77 minutes after 7:00. Hints: (a) The 5th percentile of a standard normal distribution is -1.645 *(from the table).* $Z = (x - \mu)/\sigma$ *so that* $-1.645 = (x - \mu)/15$*. Solve for* x *and get* $x = \mu - 25$ *so you should arrive 25 minutes before the mean (7:00). (b) similar except you use* $Z = -1.282$ *and you have to add 20 minutes at the end.*

4. The Nelson Company makes the machines that automatically dispense soft drinks into cups. Many national fast food chains such as McDonald's and Burger King use these machines. A study by the company shows that the actual volume of soft drink that goes into a 16-ounce cup per fill is normally distributed, with a mean of 16 ounces and a standard deviation of 0.35 ounces. A new 16-ounce cup that is being considered actually holds 16.7 ounces of drink.

   (a) Calculate the proportion of cups that will be "overfilled" by the machine, i.e., those that will be given more than 16.7 ounces of drink.

   (b) They wish to adjust the machine so that the overfill percentage is no greater than 0.5% Determine the mean required to fulfill this wish.

   (c) If the mean is set at 16 ounces, calculate the standard deviation that would be required to meet this stipulation in the previous part.

   (d) Which of the two procedures described in parts (b) and (c) do you prefer? Why?

   (e) What is the interquartile range of the amount of beverage filled into a cup?

   (f) What is the skewness of the amount of beverage filled into a cup?

   Answer: (a) $P(X > 16.7) = P(Z > 2) = 0.0228$. *(b)* $(16.7 - \mu_x)/0.35 = 2.576$ *implies that* $\mu_x = 15.7984$. *(c)* $\sigma_x = 0.7/2.576 = .2717$. *(d) c. (e) Note that the IQR of a standard normal distribution is 1.35 (look up the* Z *value for the 75th percentile and double it). The IQR is thus* $1.35 \times 0.35 = 0.4725$. *(f) 0.*

5. Suppose we wish to construct a boxplot for a variable that has a normal distribution.

   (a) What percentage of the distribution will lie in the box?

   (b) What percentage of the distribution will lie between the whiskers?

Answer: (a) 50%; (b) 99.3%. Hint: (b) Using the normal table the 75th percentile of the standard normal distribution is .6745, so IQR = 1.35. The upper whisker would be at .6745 + 1.5 × 1.35 = 2.70. Look this up in the normal table to find that $P(Z < 2.70)$ = .9965 so that .0035 lies outside this. Therefore .9965−.0035 = 99.3% of the distribution falls between the whiskers.

6. Gossage's Beverages recently sent a special advertisement to a large number of people in its marketing area. It offered a special price on root beer for purchases of one to four packages of six bottles or cans. In planning for the special promotion, the brand manager assess the probability distribution of the number of packages that each customer would buy during the promotion. There is a 60% chance that a customer will buy 0 packages, a 15% chance that the customer will buy 1 package, an 10% chance of 2 packages, 5% chance of 3 packages, and a 10% chance of 4 packages.

   (a) What is the mean (expected) number of packages that a customer will buy?

   (b) Find the standard deviations of the number of packages.

   (c) (B+) Suppose that the *profit* per package is \$0.20. Compute the expected *profit* from a single customer. Hint: think of profit as a *linear transformation* of orders.

   (d) Continuing the previous part, find the standard deviation of *profit* from a single customer.

   Answer: (a) 0.9; (b) 1.34; (c) 0.18; (d) 0.2676.

7. (video solution) The Bank of Connecticut issues Visa and Mastercard credit cards. It is known that the balances on all Visa credit cards issued by the Bank of Connecticut have a mean of \$845 and a standard deviation of \$270. Assume that the balances on all these Visa cards follow a normal distribution.

   (a) What is the probability that a randomly selected Visa card issued by this bank has a balance between \$1000 and \$1400?

   (b) What percentage of Visa cards issued by this bank have a balance of \$750 or more?

   (c) For parts (c) – (g) assume that we have drawn a sample of 100 customers and computed the sample mean balance, $\bar{x}$. What is the distribution of $\bar{x}$, i.e., what is its shape (name of distribution), mean, and standard deviation?

   (d) What is the probability that $\bar{x}$ is greater than \$875?

   (e) The bank defines a *small* balance to be one that is less than \$750. Suppose we count the number of customers in our sample of 100 that have small balances. What is the name of the distribution of this count. Also find its mean and standard deviation? Hint: use your answer from part (b). (I want the name of the *exact* distribution, not an approximation of the distribution)

   (f) What is the chance that 40 or more of the people in our sample will have small balances?

   Answer: (a) .2630 or .2646 with rounding; (b) .6375 or .6368 with rounding; (c) normal, $\mu_{\bar{x}} = 845$, $\sigma_{\bar{x}} = 27$; (d) .1333 or .1335 with rounding; (e) binomial, $E(X_b) = 36.25$, $\sigma_{X_b} = 4.8075$; (f) $P(X_b \geq 40) = .2495$.

8. The hours worked per week by the current population of all workers employed in the private industrial sector has a normal distribution with $\mu = 34.2$ and a $\sigma = 4$.

   (a) What percentage of workers in this sector work overtime (defined as someone who works more than 40 hours per week)?

   (b) What is the probability that the number of people who work overtime in the sample of 80 is less than or equal to 10?

   (c) For the remaining parts of this question, suppose we draw a random sample of 80 such workers and compute the sample mean number of hours worked, $\bar{x}$. What is the mean, standard deviation, and shape of the distribution of $\bar{x}$?

   (d) What is the probability that the sample mean $\bar{x}$ will be greater than 35?

   Answer: *(a)* $P(X > 40) = P(Z > (40 - 34.2)/4) = .07353$. *(b)* $E(x) = 5.882$, $\sigma_x = 2.3343$ *and* $P(x \leq 10) = .9761$ $P(X_b \leq 10) = P(X_n < 10.5) = P(Z < (10.5 - 5.882)/2.3343)$. *Note that* $\sqrt{80 \times .07353 \times (1 - .07353)} = 2.3343$. *(c) use normal-normal theorem to get normal,* $\mu_{\bar{x}} = \mu_x = 34.2$, $\sigma_{\bar{x}} = 4/\sqrt{80} = .4472$*]. (d) .0367.*

9. The June 1994 issue of *PC World* included a report on reliability and service support for personal computers (PCs). One of the conclusions, "25% of new PCs have problems," formed the top headline of the May 23, 1994, issue of USA Today. Every issue of PC World since October, 1993 had included a survey form asking questions about users' hardware troubles. Survey respondents for each month were entered in a drawing to win a new PC, and over 45,000 responses were received.

   (a) What target population is implied by the USA Today article?

   (b) What is the sampling unit?

   (c) Briefly critique the sampling methodology. In particular, identify any biases.

   Answer: (a) New PCs. (b) A computer user — note this doesn't match part (a). (c) There are at least two problems: sampling frame bias (PC World readers may not represent all computer users) and self-selection bias (those who choose to respond may be different than those who don't). The drawing also creates and incentive to reply more than once and there could be problems with leading questions, etc.

Introduction to Statistical Inference

After we have made an estimate, two key questions:

- How stable is our estimate? (Chapter 9)
  - Margin of error
  - Confidence intervals
- What can we say about the parameter in comparison with other numbers? (Chapter 10)
  - Is it larger (smaller) than some constant?
  - Is it larger (smaller) than some other parameter(s)?

If we had drawn a different sample, would our conclusions have been different?
Another question is about the value of another observation. The answer will be prediction and/or tolerance intervals.

Introduction to Confidence Intervals

- For now, assume …
  - we have a random sample from a large population
  - the sampling distribution is approximately normal (the sample size is at least 40 or so[34] *and* sufficiently large for the CLT to apply if the population distribution is not normal—see page 228)
- Definition: the **margin of error** of the estimate is

$$E = \pm 1.96 s_{\bar{x}} = \pm 1.96 \frac{s_x}{\sqrt{n}}$$

- Definition: a 95% **confidence interval** (CI) for μ_x is

$$\bar{x} \pm 1.96 s_{\bar{x}} \quad \text{or} \quad \bar{x} \pm 1.96 \frac{s_x}{\sqrt{n}}$$

- Literal interpretation

$$\begin{aligned} .95 &= P(\mu_x - 1.96 s_{\bar{x}} < \bar{x} < \mu_x + 1.96 s_{\bar{x}}) \\ &= P(\bar{x} - 1.96 s_{\bar{x}} < \mu_x < \bar{x} + 1.96 s_{\bar{x}}) \end{aligned}$$

  The CI shows the likely size of the amount the estimate differs from the parameter. Think of it as a "give-or-take" number.

- 95% is called the **confidence level**, denoted by $1-\alpha$, so $\alpha = 5\%$

[34] If the sample is not at least 40 then you need to use a value from the t distribution instead of 1.96 from the normal table.

Example

The owner of a local store wants to know the average age of his customers. He drew a random sample of 87 customers and asked them their ages. The average was 20 years and the standard deviation was five years. Compute the margin of error and a 95 percent CI for the mean age:
Solution:

$$n = 87 \qquad \bar{x} = 20 \qquad s_x = 5$$

Margin of Error: $\pm 1.96 \frac{5}{\sqrt{87}} = \pm 1.05$

$$20 \pm 1.96 \frac{5}{\sqrt{87}} \quad \text{or} \quad (18.95, 21.05)$$

Possible interpretation? There is a 95% chance that the true average age of his customers is between 18.95 and 21.05.
Excel:

```
=20+NORMSINV(.975)*5/sqrt(87)
```

Your Turn

9.1. (**video**)
$103.6 \pm 1.96 \times 1.194 = [101.26, 105.94]$

9.3. $[316.42, 328.46]$

9.31. $2.34 \pm 1.96 \times .036$

9.33. (b) .2769; (c) $101.37 \pm 1.96 \times .2769$. (d)–(g) omit.

Interpretation of Confidence Intervals

"The latest New York Times/CBS News Poll is based on telephone interview conducted from last Wednesday through Saturday with 1,190 adults around the United States, excluding Alaska and Hawaii.

The sample of telephone exchanges called was selected by a computer from a complete list of exchanges in the country. The exchanges were chosen to assure that each region of the country was represented in proportion to its population. For each exchange, the telephone numbers were formed by random digits, thus permitting access to both listed and unlisted numbers. Within each household, one adult was designated by a random procedures to be the respondent for the survey. The results have been weighted to take account of household size and number of telephone lines into the residence and to adjust for variations in the sample relating to region, race, sex, age, and education.

In theory, in 19 cases out of 20 the results based based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all American adults.

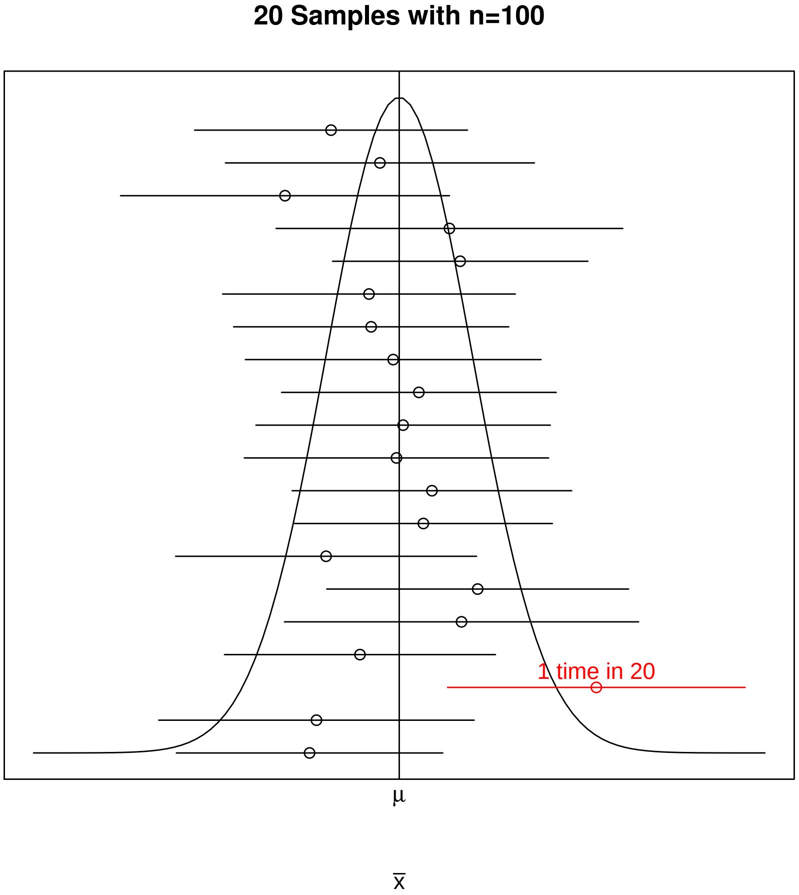
The potential sampling error for smaller subgroups is larger. For example, it is plus or minus five percentage points for Republicans, Democrats or independents taken as a group.

In addition to sampling error, the practical difficulties of conducting any survey of public opinions may introduce other sources of error into the poll. Variations in question wording or the order of questions, for example can lead to somewhat different results."

- "19 cases out of 20" = 95%
- "results based on such samples" = statistic p
- "differ by no more than three percentage points" = margin of error
- "from what would have been obtained by seeking out all American adults" = parameter π

Interpretation of Confidence Intervals

Suppose we draw 20 samples, each of size $n = 100$, and construct a 95% confidence for each. The dots show the sample mean and the lines show the confidence interval.



Notice how in 19 of the samples the confidence interval covers the true mean μ, but one time in 20 (red line) it does not.

Confidence Intervals for Proportions

- Assume that the sample size is sufficiently large for the CLT to apply to the sampling distribution ($n\pi > 5$ and $n(1-\pi) > 5$)
- The "population" variance is estimated by $s^2 = p(1-p)$
- Inserting our estimate of s into the formula for means, a 95% confidence interval for π is

$$p \pm 1.96 s_p \quad \text{or} \quad p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

- The *margin of error* of the estimate is

$$\pm 1.96 s_p = \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

Your Turn

9.14. (video) [.9981, 1.0453]

9.17. $[0.11, 0.18]$

9.20. (a) .1992; (b) .01446; (c) $.1992 \pm 1.96(.01446)$; (e) CLT.

9.21. $.2796 \pm 1.96 \times .01350$

- Find the margin of error and a 95% confidence interval if $p = .7$ and $n = 100$. Answer: $\pm .0898$, $[.61, .79]$
- Find the margin of error and a 95% confidence interval if $X_B = 150$ and $n = 400$. Answer: $\pm .04744$, $[.33, .42]$

Other Confidence Intervals

- So far we have talked only about 95% CIs. We can compute other CIs as well, e.g., 90% and 99%. The following confidence levels are commonly used:

Confidence level	Z value
90%	1.645
95%	1.96
98%	2.326
99%	2.576
99.9%	3.29

- To construct a 99% CI (assuming large sample), use

$$\bar{x} \pm 2.576 \frac{s}{\sqrt{n}}$$

- The interval becomes wider as the confidence level increases
- **Your Turn:** What would be the Z value to find an 80% confidence interval?

  [Answer: `NORMSINV(.9)` ≈ 1.28]

Example for Proportions

On December 16, 1998 ABC News conducted a nationwide poll of 510 adults. Respondents were asked "As you may know, the House of Representatives is expected to vote soon on whether or not to impeach Bill Clinton. If the House impeaches him, the Senate will hold a trial to decide wether or not Clinton should be removed from office. Based on what you know, do you think the House should or should not impeach Clinton?" The number of "shoulds" was 204. Compute the estimated proportion, 90% CI, 95% CI and 99% CI.

$$p = 204/510 = 40\%$$

$$.4 \pm 1.645\sqrt{\frac{.4(1-.4)}{510}} \quad \text{or} \quad 40\% \pm 3.57\%$$

$$.4 \pm 1.96\sqrt{\frac{.4(1-.4)}{510}} \quad \text{or} \quad 40\% \pm 4.25\%$$

$$.4 \pm 2.576\sqrt{\frac{.4(1-.4)}{510}} \quad \text{or} \quad 40\% \pm 5.59\%$$

Your Turn (now)

9.18. (video) [.7638, .8870]

9.19. [.5412, .7379]

1. Find 90, 95, 98 and 99% confidence intervals if $n = 80$, $\bar{x} = 10$ and $s = 2$. Answers: [9.63, 10.37], [9.56, 10.44], [9.48, 10,52], [9.42, 10.58]

Your Turn

1. (video) (Siegel 9.15) A market survey has shown that people will spend an average of \$15.48 each for your product next year, based on a sample survey of 483 people. The standard deviation of the sample was \$2.52. Find the two-sided 95% confidence interval for next year's mean expenditure per person in the larger population. *Answer:* $15.48 \pm 1.96 \times .1147$

2. (video) (Siegel 9.18) A recent survey of 252 customers, selected at random from a database with 12,861 customers, found that 208 are satisfied with the service they are receiving. Find the 90%, 95%, 99% and 99.7% confidence interval for the percentage satisfied for all customers in the database. *Answer:* $p = 208/252; 99\% : p \pm 2.576\sqrt{p(1-p)/252}$

3. (video) (video) (Siegel 9.14) Your bakery produces loaves of bread with "1 pound" written on the label. Here are weights of randomly sampled loaves from today's production: 1.02, 0.97, 0.98, 1.10, 1.00, 1.02, 0.98, 1.03, 1.03, 1.05, 1.02, 1.06. Find the 95% confidence interval for the mean weight of all loaves produced today.

4. Out of 763 people chosen at random, 152 were unable to identify your product.

   (a) Estimate the percentage of the population (from which this sample was taken) who would be unable to identify your product.

   (b) Find the standard error of the estimate found in part a.

   (c) Find the two-sided 95% confidence interval for the population percentage.

   (d) Why is this statistical inference approximately valid even though the population distribution is not normal?

The t Distribution

- The normal-normal theorem tells us that *if we sample from a normal population* the following has a standard normal distribution

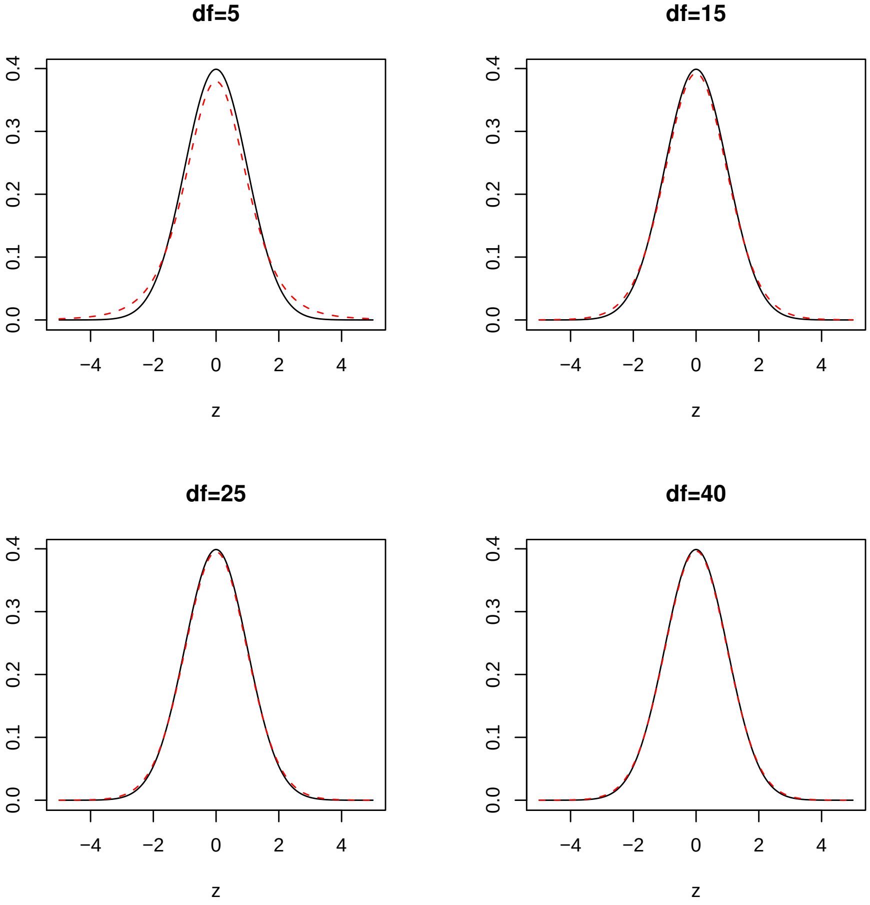
$$z = \frac{\bar{x} - \mu_x}{\sigma_x/\sqrt{n}} = \frac{\bar{x} - \mu_x}{\sigma_{\bar{x}}}$$

- In practice, we do not know σ_x so we estimate $\sigma_{\bar{x}}$ with its standard error $S_{\bar{x}}$, but then the normal-normal theorem no longer applies
- If a sample is drawn from a normal distribution and the standard deviation is *unknown* (and thus must be estimated with S_x), then the following has a t *distribution* (with $n-1$ "degrees of freedom")

$$t = \frac{\bar{x} - \mu_x}{S_x/\sqrt{n}} = \frac{\bar{x} - \mu_x}{S_{\bar{x}}}$$

- The t distribution looks like the normal distribution, but it has "thicker" tails, with the thickness determined by the sample size
- For sample sizes greater than, e.g., 40, the t distribution is essentially the same as the normal distribution
- By default SAS, R, SPSS, Minitab and Excel use the t distribution when constructing CIs and testing hypotheses about means and regression coefficients
- If the sample is from a non-normal population and the sample size is not "sufficiently large" for the central limit theorem to apply, then you *cannot* use t or normal distributions to compute confidence intervals — call a statistician for help!

The t Distribution



Your Turn (now)

9.5. (video) (a) 2.36; (b) 3.50; (c) 5.41; (d) 1.89

9.6. (a) 2.09; (b) 2.85; (c) 3.85; (d) 1.72

9.7. (a) 1.96; (b) 2.574; (c) 3.291; (d) 1.645

9.8. (a) 2.01; (b) 2.68; (c) 3.51; (d) 1.68

Details For Means

1. Examine summary statistics for outliers, anomalies, and the shape of the population distribution (is it normal?)
2. Compute $\bar{x}$ and s^2
3. Determine distribution of estimate ($\bar{x}$), i.e., the sampling distribution
   (a) If population is normal and variance (σ^2) is known, use normal distribution (normal-normal theorem)
   (b) If population is normal and variance is *unknown*, use t distribution with $n - 1$ "degrees of freedom" (for $n > 40$ this will be nearly normal)
   (c) If population is *not* normal and sample is sufficiently large, use normal distribution (CLT)
   (d) If population is *not* normal and sample is *not* sufficiently large, call a statistician

   Note: computer packages usually assume (b). What should you do if you are using a computer package but have (a), (c) or (d)?
4. A 95% CI for μ is

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm t_{n-1,.975}\frac{s}{\sqrt{n}}$$

Your Turn: Siegel Chapter 9
Homework Solutions

9.2 [24.27, 24.55]

9.4 $41.93 \pm 2.093 \times 0.04 = [41.85, 42.01]$, where TINV(.05,19) = 2.093.

9.11. SPSS (a) $s = 15.43$; (b) about one standard error, 5.46; (c) [23.0, 48.8]; (d) [16.8, 55.0]

9.12. (a) [15.0369, 16.4940]; (b) [14.8073, 16.7237]

9.13. (a) [0.0464, 0.0525]; (b) [0.0455, 0.0534]

9.14. (**video**) [.9981, 1.0453]

9.15. (**video**) $15.48 \pm 1.96 \times .1147$

9.16. (a) 3.1152; (b) [127.6435, 141.3565]; (c) [124.8248, 144.1752]

9.22. [.4185, .5015]

9.24. If we assume a normal population and $25,000 = S$ (instead of σ, then use $T_{.975,7} = 2.365$ to find [153912, 195714] in part (a) and $T_{.975,7} = 3.499$ to find [143882, 205744] in part (b). Fifth edition and earlier: (a) [52728, 81938]; (b) [46082, 88584]

9.26. From SPSS: Sixth: (57.88, 66.08) Fifth: -1.10 to 5.40

9.28. (a) [407.6, 476.1], note use `TINV(.05,18-1)=2.1098` instead of 1.96.

9.29. (a) -0.0494; (b) 0.2053; (c) .0530; (d) $[-0.1631, 0.0643]$; (e) 90%: $[-14.27, 4.39]$ and 95%: $[-16.31, 6.43]$; (f) omit)

9.30a. $.524 \pm 1.96 \times .01646$

9.31. [2.2692, 2.4108]

9.32. (a) $n = 200$, $p = .21$, $S_p = .02880$; (b) $.21 \pm 1.96 \times .02880$; (c) $5,924 \pm 1592 = [4,331, 7,516]$

9.33. (a) a coil; (b) 0.2769, mean from sample of 93; (c) [100.82, 101.92] (d–g) omit

9.34a. [6.04, 8.82]

9.37. (a) typical customer who place orders with you under current circumstances; (b) not using the normal distribution; (c) use $t = 2.306$ (you're not responsible for getting this number) to get $3,782 \pm 2.306 \times 430$; (d) use $t = 1.860$ to get $3,782 \pm 1.860 \times 430$;

9.38. (a) $53.01 \pm 1.96 \times 2.33$; (b) $53.01 \pm 2.576 \times 2.33$

9.39. SPSS [1.67, 186.33]

9.40. SPSS [88.1, 142.7]

9.41. SPSS [1.629, 1.687]

Formula for Sample Size

- Theorem. Let

  E acceptable margin of error

  z standard normal variate value for a specified confidence level

  σ_x^2 population variance

  Then

$$n = \frac{z^2\sigma_x^2}{E^2}$$

- Proof. Margin of error is defined as

$$E = z\frac{\sigma_x}{\sqrt{n}}$$

  Solve this equation for n to get the result.

Sample Size Problems

The transit system of a major metropolitan area was interested in determining the average number of miles a commuter drives to work. Past studies have shown that the variation (σ) in commuting was five miles. The managers of the transit system want to be 95 percent confident in the result and do not want the error to exceed 0.75 miles.

1. What sample size should they use? Solution:

$$n = \frac{z^2\sigma_x^2}{E^2} = \frac{1.96^2 \times 5^2}{0.75^2} \approx 170$$

2. The results shows that commuters drive on average 20 miles to work with a standard deviation of 10. Compute a 95 percent confidence interval. Solution:

$$20 \pm 1.96\frac{10}{\sqrt{170}} \quad \text{or} \quad 20 \pm 1.5$$

3. Compute the 95 percent confidence interval if the average is 20 miles and the standard deviation is five miles. Solution:

$$20 \pm 1.96\frac{5}{\sqrt{170}} \quad \text{or} \quad 20 \pm 0.75$$

Methods of Estimating σ_x^2

- For means
  1. Results from previous studies
  2. Results from a pilot survey
  3. Secondary data
  4. Judgement
  5. For ratings scale means, see example
  6. One-sixth of the range of normal variables
- For proportions, use 1–4 above or $\pi = 0.5$ for conservative estimate (recall $\sigma^2 = \pi(1-\pi)$)
- A pollster wants to estimate the fraction of people who will vote for a particular candidate. The maximum error should not be greater than 5% at the 95-percent level. How large should the sample be? Solution:

$$n = \frac{1.96^2 \times .5 \times (1 - .5)}{.05^2} \approx 384 \approx 400$$

- Worst-case sample sizes (assuming $\pi = 0.5$)

Sample size	Margin of error
$1067 \approx 1000$	3%
$384 \approx 400$	5%
$96 \approx 100$	10%

Your Turn

1. (**video**) A survey was being designed by the marketing research department of Conner Peripherals, Inc., a large manufacturer of disk drives. The goal was to assess customer satisfaction with the company's disk drives. Management wanted to measure the average maintenance expenditure per year per computer, the average number of malfunctions or breakdowns per year and the length of service contracts purchased with new portable computers. Management wanted to be 95% confident in the results. Further the magnitude of the error was not to exceed ±$20 for maintenance expenditures, ±1 malfunctions and ±3 months. The research department noted that although some individuals and businesses would spend nothing on maintenance expenditures per year, others might spend as much as $400. Also, although some portable computers would experience no breakdowns within a year, the maximum expected would be no worse than three. Finally, although some portable computers might not be purchased with service contracts, others might be purchased with up to a 36-month contract. *Answer: Note: use 1/6 the ranges to estimate the standard deviations. (a) 43, 1, 16; (b) 43; (c) (82, 118)*

   (a) What sample size would you recommend for the three variables considered separately?

   (b) What size would you recommend *overall* given that management thought that accurate knowledge of the expenditure on repairs was most important and the service contract length least important?

   (c) The survey indicated that the average maintenance expenditure is $100 and the standard deviation is $60. Estimate the confidence interval for the population parameter μ. What can you say about the degree of precision?

2. The management of a brewery wanted to determine the average number of ounces of beer consumed per resident in the state of Washington. Past trends indicated that the variation in beer consumption (σ) was 4 ounces. A 95% confidence level is required, and the error is not to exceed ± 0.5 ounces. (a) What sample size would you recommend? (b) Management wanted an estimate twice as precise as the initial precision level and an increase in the confidence to 99 percent. What sample size would you recommend? *Answer: (a) 246; (b) 1699.*

3. The director of a state park recreational center wanted to determine the average amount of money that each customer spent traveling to and from the park. On the basis of the findings, the director was planning on raising the entrance fee. The park director noted that visitors living near the center had no travel expenses but that visitors living in other parts of the state or out of state traveled upwards of 250 miles and spend about 24 cents per mile. The director wanted to be 95 percent confident of the findings and did not want the error to exceed ±50 cents. *Answer: (a) 1537; (b)* 12 ± 0.37

(a) What sample size would you recommend to determine the average travel expenditure?

(b) After the survey was conducted, the park director found that the average expenditure was $12.00 and the standard deviation was $7.50. Construct a 95% confidence interval. What can you say about the level of precision?

4. (video) A large manufacturer of corrugated paper products recently came under severe criticism from various environmentalists for its disposal of waste. In response, management launched a campaign to counter the bad publicity. A study of the effectiveness of the campaign indicated that about 20 percent of the residents of the city were aware of the campaign, based on a sample of 480. Six months later, it was believed that 40% of the residents were aware of the campaign. However, management decided to do another survey and specified a 99% confidence level and a margin of error of ±3%. *Answer: (a) 1770; (b) [.32, .38]*

   (a) What sample size would you recommend?

   (b) The survey indicated that 35% of the population was aware of the campaign. Construct a 99% confidence interval for the population parameter.

5. Pac-Trac, Inc., manufactures video games. The marketing research department is designing a survey to determine attitudes toward the products, the percentage of households owning video games and the average usage rate per week are to be determined. They want to be 95% confident of the results and do not want the error to exceed ±3% for video ownership and ±1 hour for average usage rate. Previous reports indicated that about 20% of the households own video games and that the average usage rate is 15 hours with a standard deviation of 5 hours. *Answer: (a) 683; (b) 96; (c) 683; (d) [.266, .334]; (e) [12.7, 13.3]*

   (a) What sample size would you recommend, assuming that only the percentage of households owning video games is to be determined?

   (b) What sample size would you recommend, assuming that only the average usage rate per week is to be determined?

   (c) What sample size would you recommend, assuming that both of the preceding variables are to be determined?

   (d) After the survey was conducted, the results indicated that 30% of the households own video games and that the average usage rate is 13 hours with a standard deviation of 4. Find the 95% confidence interval for both variables. Comment on the degree of precision.

Confidence Intervals and Sample Sizes

By Edward Malthouse

Introduction

Consider the following problems:

- The manager of a muffler store wants to know the want to know the average rating on a seven-point semantic differential scale measuring customer satisfaction of all customers served during a specified month. A one on the scale denotes "very dissatisfied" and a seven denotes "very satisfied."

- An entrepreneur is considering opening a restaurant in downtown Evanston. He wants to know how many people eat out for lunch and how much they spend. He also wants to know the fraction of people who spend between \$5 and \$10, the expected cost of a lunch in his planned restaurant.

- A pollster wants to estimate the fraction of people who will vote for a particular candidate.

- The government wants to estimate the fraction of households in the United States having a colored television.

In each case there is a quantifiable attribute called a *parameter* that we would like to know, e.g., average rating, fraction of households with a colored television. If we had measurements on every member of the population, we could simply compute the value of the parameter. Unfortunately, it is usually not possible to survey every member of the population for the following reasons:

- *Cost.* The cost of talking to each member can be prohibitively expensive.

- *Time.* There may be insufficient time to make contact with each member.

- *Difficulty.* It many cases it would be very difficult to seek out each member of a population. For example, finding all lunch diners in Evanston would be a difficult task.

Instead of talking to each member of the population, organizations often take measurements on a subset of the population called a *sample*. Using the measurements on the sample, they compute a guess for the parameter called a *statistic*. Examples of parameters and statistics are given in the table below. In the customer satisfaction example, the parameter μ is the average score on the seven-point scale of *all* customers served during the specified month; the statistic $\bar{x}$ is the average score for a sample of the customers.

Suppose there are N subjects in the *population* and let n be the number of subjects in the *sample*. Denote the measurements of the n subjects in the sample by $x_1, \ldots, x_n$. We show how each of the "guesses" or statistics are computed in the table below.

Name **(Parameter)**	**Statistic**
Mean (μ)	$\overline{x} = (x_1 + \cdots + x_n)/n$
Proportion (π)	$p = (x_1 + \cdots + x_n)/n$
Total (T)	$\hat{T} = N\bar{x}$
Std dev (σ)	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Correlation (ρ)	r (use SPSS/Excel)
Intercept (α)	a (use SPSS/Excel)
Slope (β)	b (use SPSS/Excel)

Because we are unable to measure every subject in the population, our guess (statistic) will most likely not equal the parameter. If we had drawn a sample with different subjects, we most likely would have computed a different statistic value. The relationship between the two is as follows:

statistic = parameter+sampling error+bias error

The value of the statistic is polluted by two types of error.

- *Sampling error.* During the sampling lecture, I defined a probability sample as a sample where *every member of the population has a known, nonzero, probability of being selected into the sample.* Sampling error refers to error introduced by using a probability sampling procedure. If a different random sample had been drawn, the estimate would most likely be different.

- *Bias error.* This includes errors due to: (1) self-selection or non-response — respondents choose whether or not they are included in the sample by choosing to respond or not respond, e.g., mail surveys; (2) selection — the interviewer chooses who is included in the sample rather than a random procedure; (3) data entry error.

All of the discussion and formulas in this handout assume that *there is no bias error.* Technically, the formulas do not apply to samples when there is bias error and should not be used in these cases. In practice, it is hardly ever possible to draw a truly random sample without biases, and market researchers follow a more pragmatic approach. They make every effort that is practically possible to eliminate sources of bias, they use the formulas, and they discuss possible sources of bias and their effects on the estimate. Random samples are usually more expensive than non-random samples and the market researcher must balance these costs with the importance of the decision and the objectives of the research.

In this reading I will address the following questions:

- How good is a given guess for a parameter?
- How large a sample is needed to make a guess with a specified quality?

Confidence Intervals

Means

Let $x_1, \ldots, x_n$ be observations from a normal distribution. This means that a histogram of the measurements of all subjects in the population has the familiar "bell shape." Then a $100 \times (1-\alpha)\%$ confidence interval for μ, the population mean, is given by

$$\bar{x} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}},$$

where $t_{1-\alpha/2,n-1}$ is the $(1-\alpha/2)^{\text{th}}$ percentile of a t distribution with $n-1$ degrees of freedom. For example, to find an $\alpha = 95\%$ confidence interval with $n = 400$, $t_{1-0.05/2,400-1} = t_{0.975,399} \approx 1.96$.

The formula given above requires that the distribution of the population be normal. There are many cases where a distribution is not normally distributed. The distribution for income is usually not normal because it usually has a long right tail. The distribution of the number of people with a certain property, e.g., owning a color television, has a binomial distribution. When the observations do not come from a normal population, the formula given above can still be used if the sample size is sufficiently large. The meaning of "sufficiently large" depends on how much the population distribution deviates from normality. I give some rules of thumb for proportions later in this section.

Example Daily demand for coffee at the Coffee Corner is approximately normally distributed. A random sample of the demand for five days out of the last year showed the demand in pounds per day to be 29, 22, 32, 19, and 28. A 95% confidence interval for the mean daily demand is computed as follows.

$$\bar{x} = 26 \qquad s = 5.34 \qquad n = 5 \qquad t_{0.975,4} = 2.776$$

$$26 \pm 2.776 \frac{5.34}{\sqrt{5}} = (19.37, 32.63)$$

The correct interpretation of the resulting interval is in 19 cases out of 20 (19/20 = 95%), the results based on such samples will differ by no more than $2.776 \times 5.34/\sqrt{5}$ in either direction from what would have been obtained by computing the mean daily demand using data from all 365 days. It is tempting to interpret the confidence interval as "there is a 95% chance that the population mean falls in the interval (19.37, 32.63)." This interpretation is incorrect because the population mean is not a random quantity. It makes no sense to make chance statements about a fixed quantity. The randomness is in the selection of sampling units, which is brought out in the correct interpretation given earlier.

Totals

Let $x_1, \ldots, x_n$ be observations from a normal distribution. A $100\times(1-\alpha)\%$ confidence interval for the population total is

$$N\bar{x} \pm t_{1-\alpha/2,n-1}\frac{Ns}{\sqrt{n}}.$$

The comments made in the means section for non-normal distributions also apply for totals.

Example (Coffee Corner example continued) A 95% confidence interval for the total yearly demand (assuming 365 days in a year) is

$$365 \times 26 \pm 2.776\frac{365 \times 5.34}{5} = (7070, 11910).$$

Proportions

Suppose $x_1, \ldots, x_n$ are responses to a dichotomous question — for example, responses to the question "Do you own a color television, yes or no?" The easiest way to find a confidence interval for the true proportion of households owning a color television is to transform this problem so that the formula for means can be applied. If we assign the value 1 to yes responses and 0 to no responses, the sample mean, $\bar{x}$, is the proportion we want. If we interviewed $n = 100$ households and 60 answered yes, we would add 60 1's to 40 0's and divide by 100 giving 60%. We can also save some effort in computing s by remembering that $s^2 = p(1-p)$. With this in mind, a $100 \times (1-\alpha)\%$ confidence interval for π is

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{\text{th}}$ percentile of a standard normal distribution. For example, to compute a 95% confidence interval, $z_{1-\alpha/2} = z_{0.975} = 1.96$.

The only problem with using the formula for means is that it only applies for "sufficiently large n." A rule of thumb that is often given in elementary statistics books is (1) $n \geq 30$, (2) $n\pi \geq 5$, and (3) $n(1-\pi) \geq 5$. The following example shows how the formula above can give misleading results when n is very large but these conditions are not satisfied.

Example In 1980 a survey was conducted to determine the fraction of households owning a computer. One thousand households were surveyed and only 2 had a computer. A 95% confidence interval for π is computed

$$0.2\% \pm 1.96\sqrt{\frac{0.2\% \times 99.8\%}{1000}} = (-0.1\%, 0.5\%)$$

Clearly something is wrong, because percentages cannot be negative. The problem is that the conditions for using normal approximation are not satisfied ($0.2\% \times 1000 = 2 < 5$). In cases where the conditions are not satisfied, one must use another method to compute confidence intervals.

Sampling Without Replacement

All of the formulas given in this section so far assume that sampling is done *with* replacement. After selecting a person to be included in our sample, we replace the person into our pool of possible subjects before drawing another subject. Under these rules, it is thus possible to include the same person in a sample more than once. In practice, we would not want to measure a subject more than once. This assumption is made to simplify the calculations and when the size of the sample is small relative to the population, i.e., the sampling fraction $f = n/N$ is small, the chance of including the same subject in a sample more than once is negligible. But when the sampling fraction is large, this chance can be substantial.

When sampling *with* replacement, the variance of the sample mean is $V(\bar{x}) = \sigma^2/n$. When sampling *without* replacement, this variance is shrunk with the following *correction factor*[35]:

$$V(\bar{x}) = \frac{\sigma^2}{n}\Big(1 - \frac{n}{N}\Big) = \frac{\sigma^2}{n}\Big(1 - f\Big).$$

The intuition behind the variance of $\bar{x}$ getting smaller is that removing items from the sampling frame *tends* to reduce the variability slightly. So the variance of $\bar{x}$ for drawing without replacement is a little less than for drawing with replacement. The correction factor can be applied to the confidence interval formulas for means, totals, and proportions given above:

$$\bar{x} \pm t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\sqrt{1-f}$$

$$N\bar{x} \pm t_{1-\alpha/2,n-1}\frac{Ns}{\sqrt{n}}\sqrt{1-f}$$

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{1-f}.$$

A rule of thumb for deciding when to use the correction factor is to use it whenever the sampling fraction is greater than 0.05. When the sampling fraction is less than 0.05, the effect of the correction factor is small.

Example One public opinion poll uses a simple random sample of size 1,000 drawn from a town with population 50,000. Another poll uses a simple random sample of size 1,000 from a town with a population of 500,000. Other things being equal, which of the following is correct?

1. the first poll is likely to be quite a bit more accurate than the second.
2. the second poll is likely to be quite a bit more accurate than the first.
3. there is not likely to be much difference in accuracy between the two polls.

The correct answer is (3). The correction factor for the smaller town is $\sqrt{1 - 1000/50,000} \approx 0.9999$. The correction factor for the larger town is $\sqrt{1 - 1000/500,000} \approx 1.0000$. The effect on the confidence interval is thus negligible.

Computing Sample Sizes

Market researchers often need to know how many subjects to include in their sample to achieve a certain accuracy before they draw the sample. In this section we discuss some formulas to answer this question. We need to know following information to estimate sample size:

- Magnitude of acceptable error (E) — the radius of the confidence interval.
- Variability in the population (σ^2) — a measure of how homogeneous the population is.
- Confidence level (α) — the chance that you get a sample where the resulting confidence interval does not cover the parameter.

Estimates for the required sample size can be made as follows:

$$n = \frac{z^2_{1-\alpha/2}\sigma^2}{E^2},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{\text{th}}$ percentile of a standard normal distribution.

Example Suppose we want to estimate the mean of responses on a seven-point scale within $E = 0.5$ unit of the true value with 95% confidence. Previous studies suggest that the standard deviation is around 1.2. The sample size required is

$$n = \frac{1.96^2 \times 1.2^2}{0.5^2} \approx 22.$$

[35] The exact correction factor is $(N-n)/(N-1)$. The -1 in the denominator is usually ignored to simplify calculations giving $(N-n)/N = 1-f$.

To use this formula for proportions, recall that for dichotomous random variables, $\sigma^2 = \pi(1-\pi)$.

Example Suppose we want to estimate the fraction of households in the United States with color televisions. We want the estimate to be within 3% of the true fraction with 95% confidence. When no prior information is given about the variance, a conservative estimate is to take $\pi = 0.5$ and estimate $\sigma^2 = 0.5(1-0.5) = 0.25$ (see next section). The sample size is

$$n = \frac{1.96^2 \times 0.25}{0.03^2} = 1067.$$

When sampling is to be done *without* replacement, the sample size required is:

$$n^* = \frac{n}{1+n/N} = \frac{n}{1+f},$$

where n is given by the formula for sampling *with* replacement. Notice that n and n^* are nearly equal when f is small.

Example We want to estimate the fraction of households with color television sets in a town with 500 households. We want the estimate to be within 3% of the true value with 95% confidence. In the previous example we found $n = 1067$. This n applies to sampling with replacement and clearly we would not want to take 1067 measurements on 500 households. Using the formula given above,

$$n^* = \frac{1067}{1+1067/500} \approx 340.$$

Estimating σ^2

To use the sample size formula, one must know the standard deviation, σ, before drawing the sample. Here are some ways to estimate σ.

- *Previous study.* If similar studies have been done in the past, the σ from the previous study can be used.
- *Pilot study.* Do a pilot study to make an estimate of σ.
- *Rating scales.* When a rating scale (e.g., Likert or semantic differential) is used, tables can be used to estimate σ. The table below is based on a table given in Churchill.

A	B	C	D	E
4	2.5	0.7–1.3	320	80
5	3	1.2–2.0	342	85
6	3.5	2.0–3.0	376	94
7	4	2.5–4.0	384	96
10	5.5	3.0–7.0	355	89

A = Number of points on scale

B = Mean

C = Variance

D = E 5% of the mean

E = E 10% of the mean

The values given in the table assume the highest variance and $\alpha = 0.05$ ($z = 1.96$). For example, if a seven-point scale is used, the largest sample required to estimate μ within 5% of 4 = 0.2 at the 95% level is 384.

- *Proportions.* For proportions, the most conservative estimate for σ is 0.5. The reason for this is that the variance, $\sigma^2 = \pi(1-\pi)$, has a maximum value when $\pi = 0.5$.

Non-Response Rates

Not all subjects who have been drawn into a sample will respond. Many will not be at home or will refuse to participate in phone surveys. Subjects will forget or refuse to respond to mail surveys. The implication of these non-response problems is that the researcher must attempt to survey many more people than the n from the formulae in the previous sections to get n completed surveys. The researcher must estimate the non-response rate and adjust the sampling procedure accordingly. For example, if the response rate of a mail survey is estimated at 10%, i.e., for every ten questionnaires mailed, one is returned, and if $n = 100$ responses are necessary to achieve a certain accuracy, then 1000 surveys should be mailed out. Response rates vary from survey to survey depending on factors like the survey method (phone, mail, email, etc.), incentives, nature of the topic you are studying, complexity of the questions, and characteristics of the population.

Subjects who are selected to be in the sample but do not participate for one reason or another technically have a *zero* probability of being included in the sample. Therefore the sample is no longer a random sample and the formulas for confidence intervals and hypothesis tests do not apply. In situations where it is important to have a random sample, researchers will try to eliminate non-response problems. When phone interviews are used and a random sample is important, the caller will make several attempts to reach a person selected into the sample. When using mail surveys, there is often a "second mailing," where all those who didn't respond to the first survey are sent a second copy.

Sample Size for Hypothesis Tests

In many cases the market researcher conducts a survey to test a hypothesis instead of computing a confidence interval. For example, if customer satisfaction measured on a seven-point scale falls below a specified level, management might want to take corrective action. This hypothesis takes the form:

$$H_0: \quad \mu = \mu_0$$

$$H_a: \quad \mu < \mu_0.$$

To estimate n, the market researcher must know the following:

- Chance of a type I error (α) — the probability of rejecting a true H_0. This is the chance of taking corrective action when nothing is wrong. It is also called the "false-positive" or "false-alarm" rate.
- Population variability (σ).
- Chance of a type II error (β) — the probability of accepting a false H_0. This is the chance of *not* detecting a situation where something is wrong.
- Magnitude of difference to detect (δ). The precise meaning of δ is that the type II error rate applies when $\mu \le \mu_0 - \delta$. Equivalently, if $\mu = \mu_0 - \delta$, the probability of drawing a sample where we can reject H_0 is $1 - \beta$.

After picking these four values, the formula for n is

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{\delta^2}.$$

Market researchers often need to test a hypothesis about two means. For example, the average number of visits to art museums is higher for females than for males. In this case the hypothesis takes the following form:

$$H_0: \quad \mu_f = \mu_m$$

$$H_1: \quad \mu_f > \mu_m,$$

where μ_f is the average number of visits per year for females and μ_m is the average for males. Assuming the sample sizes for males and females are equal, the sample sizes are estimated:

$$n_m = n_f = \frac{(z_{1-\alpha} + z_{1-\beta})^2 (\sigma_f^2 + \sigma_m^2)}{\delta^2}.$$

The precise meaning of δ in this case is that the type II error rate β applies when $\mu_f \ge \mu_m + \delta$. Equivalently, the chance of drawing a sample so that H_0 can be rejected when $\mu_f = \mu_m + \delta$ is $1 - \beta$. One can also think of δ as being a "substantial" difference; for example if $\mu_m = 1$ and $\mu_f = 1.05$, then $\mu_f > \mu_m$, but the difference between the two is probably not large enough to take managerial action on the difference.

Example We want to know if females attend art museums more frequently than males. A previous study suggests that the standard deviation for the number of times females attend per year is $\sigma_f = 2.5$ and that the standard deviation for males is $\sigma_m = 2.3$. We want the type I and type II error rates to be $\alpha = \beta = 5\%$. The type II rate is to apply to differences greater than or equal to $\delta = 0.5$ visits per year. We need

$$n_f = n_m = \frac{(1.645 + 1.645)^2(2.3^2 + 2.5^2)}{0.5^2} \approx 500$$

of each gender in our sample.

Additional Reading

1. (Readable discussion) Freedman, D., Pisani, R. Purves, and R. Purves (2007), *Statistics*, W.H. Norton & Company.

2. (Advanced reference on sampling) Cochran, W.G. (1977), *Sampling Techniques*, John Wiley.

A Problem

McDonald's has been conducting consumer surveys for many years. Over this time span, the ratings of McDonald's on various attributes have been averaged to form attribute norms. For example, the norm for "service" is $\mu_x = 2.6$. Suppose McDonald's wishes to test whether or not a new training program will improve the service score. Employees of a particular store (which has a mean of 2.6) are sent to the new program. After they return to work, McDonald's selects a random sample of 100 customers and administers its satisfaction survey. The sample mean on the service scale is $\bar{x} = 2.74$ with a sample standard deviation of $s_x = 0.8$. Does the new program improve satisfaction with service?

Solution: Let μ be the true mean for the new program. We would like to evaluate the two hypotheses

$$H_0 : \mu = 2.6 \quad \text{business as usual}$$

$$H_1 : \mu > 2.6 \quad \text{new program works better}$$

- H_0 is called the *null hypothesis* (pronounced "H naught")
- H_1 is called the *alternative* or *research hypothesis*
- The *reference*, *test* or *null* value is 2.6
- Usually we want to reject H_0 in favor of H_1
- Analysis strategy: evaluate whether the observed data could have arisen if H_0 were true.

Four Possible Outcomes

When testing hypotheses there are four possible outcomes:

	The Truth	
	H_0 True	H_0 False
Decision	(does not work)	(new prog better)
Accept H_0 (Don't Roll Out)	Correct Decision	*Type II Error* *False negative* Prob $= \beta$
Reject H_0 (Roll Out Program)	*Type I Error* *False positive* Prob $= \alpha$	Correct Decision $Power = 1 - \beta$

We must manage a trade-off between type I and II errors

Solution to McDonald's Example

The usual solution is to consider the following question: if H_0 were true, how likely is it that we would draw a sample giving $\bar{x} \geq 2.74$?

- How strong is our evidence against H_0?
- How plausible is H_0? Note: under H_0, $\mu_{\bar{x}} = 2.6$ and $s_{\bar{x}} = .8/\sqrt{100}$

$$P(\bar{x} \geq 2.74) \approx P\left(Z \geq \underbrace{\frac{2.74 - 2.6}{.8/\sqrt{100}}}_{1.75}\right) \approx 4\%$$

```
=1-NORMDIST(2.74,2.6,0.8/SQRT(100), TRUE)
```

- If H_0 were true, there would be only a 4% chance of drawing a sample producing a sample mean as extreme or more extreme than 2.74. This means that H_0 is not so plausible in view of the evidence at hand ($\bar{x} = 2.74$)
- The number 4% is call the *P-value* or *observed significance*
  - If $P < 5\%$, the result is said to be *statistically significant* (often denoted by, e.g., 2.74^*)
  - If $P < 1\%$, the result is *highly significant* (denoted with 2 stars, e.g., 2.74^{**})

Two Types of Tests

There are two types of tests. To find out which to use, look at the alternative hypothesis.

- *One-sided tests*, e.g.,

$$H_1 : \mu > 2.6$$

- *Two-sided tests*, e.g.,

$$H_1 : \mu \neq 2.6$$

Key point: the P-value of a two-sided test for means is twice as large as the P-value of the corresponding one-sided test.

Question: Which should I use?

Answer:

- As long as you state your hypotheses it doesn't matter, since the reader can always divide or multiply by 2
- The conservative approach is to use two-tailed tests all the time
- For means, correlations and regression coefficients, R, SAS, SPSS, Minitab and Excel report two-sided P-value by default. If you want a one-sided P-value, divide by 2.

Procedure for Testing Hypotheses

1. State null and alternative hypotheses (data snooping not allowed!), decide on significance level α
2. (Gather data) Study exploratory statistics for outliers, anomalies and the shape of the population distribution
3. Compute P-value and compare with α (you're not responsible for the "critical-value" method discussed in text)
   - If $P \leq \alpha$, reject H_0
   - If $P > \alpha$, you cannot reject H_0 (Note: you are *not* concluding H_0 is true, only that it is plausible)

   The P-value gives probability of drawing a sample producing an estimate as extreme or more extreme than the current one, assuming H_0 is true

   Alternatively, reject H_0 if reference value falls outside a $100(1-\alpha)\%$ confidence interval (*confidence interval approach to hypothesis testing*)
4. When reporting results, state ...
   - Null and alternative hypotheses
   - Name of the test used, e.g., t test
   - P-value
   - Conclusion

Your Turn

1. (Siegel 10.20) (video) You are considering a new delivery system and wish to test whether delivery times are significantly different, on average, than your current system. It is well established that the mean delivery time of the current system is 2.38 days. A test of the new delivery system shows that, with 48 observations, the average delivery time is 1.91 days with a standard deviation of 0.43 days.

   (a) Identify the null and research hypotheses for a two-sided test using both words and mathematical symbols.

   (b) Perform a two-sided test at the 5% significance level and describe the result. Give the P-value. Also use the confidence interval approach.

2. (Variation of Siegel 10.1) (video) To help your restaurant marketing campaign target the right age levels, you want to find out if there is a statistically significant difference, on the average, between the age of your customers and the age of the general population in town, 43.1 years. A random sample of 50 customers shows an average age of 39.6 years with a standard deviation of 16.2 years.

   (a) Identify the null and research hypotheses for a two-sided test using both words and mathematical symbols.

   (b) Perform a two-sided test at the 5% significance level and describe the result. Do so by finding the P value and also use the confidence interval approach.

3. (Siegel 10.13) At a recent meeting, it was decided to go ahead with the introduction of a new product if "interested consumers would be willing to pay, on average, \$20.00 for the product." A study was conducted, with 315 randomly interested consumers indicating that they would pay an average of \$18.14 for the product. The standard deviation was \$2.98.

   (a) Identify the reference value for testing the means for all interested consumers. Identify the null and research hypotheses for a two-sided test using both words and mathematical symbols.

   (b) Perform a two-sided test at the 5% significance level and describe the result. Give the P-value and use the confidence-interval approach.

4. (Siegel 10.29) A recent poll of 1,235 randomly sampled likely voters shows your favorite candidate ahead, with 52.1% in favor. There are two candidates. Use hypothesis testing to infer to the larger group of all likely voters. (video)

   (a) Carefully identify the two-sided hypotheses.

   (b) Perform the hypothesis test at the 0.05 level and give the result. [$P = .1388$ using $\sigma_p = .01423$ so do not reject]

(c) Make a careful, exact statement summarizing the result of the test and what it means.

(d) Repeat parts b and c assuming that the percentage is 58.3% instead of 52.1%.

(e) Explain why a one-sided test would be inappropriate here by showing that each of the three possible outcomes of a two-sided test would be of interest.

10.1. (video) $H_0 : \mu = 43.1$ versus $H_1 : \mu \neq 43.1$; P-value ≈ 0 so reject. The video might use different numbers.

10.2. still reject

10.7. (a) .1559; (b) .01347; (c) 2169; (d) $.1559 \pm 1.96 \times .01347$; (e) $H_0 : \pi = .1$ versus $H_1 : \pi > .1$; P-value ≈ 0 so reject at .05 level

10.8. $\bar{x} = 466.17$; inventory at all instants of time last year; 296.06 to 636.27; P-value .67, so do not reject $H_0 : \mu = 500$

10.9. (video) (video) .998 to 1.045; $H_0 : \mu = 1$ versus $H_1 : \mu \neq 1$; do not reject H_0 at .05 level

10.10. Yes, $P = .000 < .05$ so we reject $H_0 : \mu = .38$.

10.12. Sixth: $P = .000$, reject $H_0 : \mu = 71.991$. do not reject $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. The one-sided P-value is .09.

10.13. $H_0 : \mu = 20$ versus $H_1 : \mu \neq 20$; P-value ≈ 0 so reject H_0 at .05 level and .01 level

10.14. $H_0 : \mu = 20$ versus $H_1 : \mu < 20$; P-value ≈ 0 so reject H_0 at .05 level

10.17. (video) $p = .5266$; P-value=.1310, so do not reject at .05 level

10.18. P-value = .3734, so do not reject at .05 level[36]

10.19. $H_0 : \pi = .1$ versus $H_1 : \pi > .1$; P-value = .0228 so reject at .05 level but not at .01 level

10.20. (video) $H_0 : \mu = 2.38$ versus $H_1 : \mu \neq 2.38$; P-value ≈ 0 so reject at .05 and .01 levels[37]

10.21. (video) (a) $\bar{x} = 14.34$; (b) yesterday's $s = .31$; (c) 14.23 to 14.45; (d) $H_0 : \mu = 14.5$ versus $H_1 : \mu \neq 14.5$; reject H_0 since $P = .009$

10.22. (video) $\bar{x} = 11.058$; $s = 1.66756$; 9.87 to 12.25; $H_0 : \mu = 11.396$ versus $H_1 : \mu \neq 11.396$; do not reject H_0

[36] $H_0 : \pi = .2$ versus $H_1 : \pi \neq .2$ note that $\sigma_p = \sqrt{.2 \times .8/500} = 0.01789$. $P(p < .184) = P(Z < (.184 - .2)/.01789) = P(Z < -.89) = .1867$, so $P = 2 \times .1867 = .3734$.

[37] $H_0 : \mu = 2.38$ versus $H_1 : \mu \neq 2.38$, $s_{\bar{x}} = .43/\sqrt{48}) = .0621$. $P(\bar{x} < 1.91) = P(Z < (1.91 - 2.38)/.0621) = P(Z < -7.57) = 0$ so $P = 2 \times 0 = 0$

10.23. (**video**) $\bar{x} = 208.25$; $s = 3.17453$; 198.39 to 218.11; $H_0 : \mu = 200$ versus $H_1 : \mu \neq 200$; reject H_0 at .05 level

10.24. $s = 10.91788$; $s_{\bar{x}} = 4.45720$; 5,541 to 28,459; Do not reject H_0

10.25. do not reject

10.26. (a) reject, (b) $P < .001$, (c) $H_0 : \mu \geq 25$, $H_1 : \mu < 25$, random sample, normal; (d) one-sided tests will give more rejections

10.27. Out of control

10.29. (**video**) (a) $H_0 : \pi = .5$, $H_1 : \pi \neq .5$; (b) $P = .1388$ using $\sigma_\pi = .01423$ so do not reject; (d) $P \approx 0$ so reject; (e) The three outcomes are that you are ahead, you are behind, and you are tied. If you are behind you might want to change what you are doing. If you are ahead you would want to keep doing what you are currently doing. If the race is tied then you need to find some way of differentiating yourself from your opponent.

10.30. (a) $.12 \pm 1.96 \times .11$; (b) $P = .2754$ so do not reject; (c) $H_0 : \mu = 0$, $H_1 : \mu \neq 0$; (d) H_0 is plausible, but not necessarily true; (e) no, type II possible

10.32. (a) 3000; (b) $H_0 : \mu \geq .416667$, H_1: $\mu < .416667$; (c) omit; (d) Find[38] P-value =.4960; (e) do not reject H_0. Hints: You are allowed a total error of \$5,000 across all 12,000 accounts. Therefore the average error per account can be \$5,000/12,000=\$0.4167. If have found the average error across your sample to be \$0.25. Multiply this by 12,000 to get \$3,000. This number is less than \$5,000, so you have evidence that the errors are not material. But can you rule out the possibility that you drew a "lucky" sample?

10.33. (a) $P \approx 0$ so reject $H_0 : \mu = 24$ against a one-sided alternative.

[38] $P(\bar{x} < .25) = P(Z < (0.25 - 0.416666)/12.21) = P(Z < -0.01) = .4960$

Practical Significance

- P-values depend on sample size
- For large samples, even small differences will be "highly significant"
- This doesn't mean they are important
- Conversely, important differences may not be statistically significant if the sample is too small
- Hypothesis testing requires that you combine information about sampling variation (e.g., using P-values) with sound business judgement and good common sense

Suppose we compare the results of a vocabulary test for big-city and rural children. We have a simple random sample of 20,000 big-city children and a simple random sample of 20,000 rural children. The big city children have $\bar{x} = 26$ with $S_x = 10$ and the rural children have $\bar{x} = 25.5$ with $S_x = 10$. The result is very highly significant.

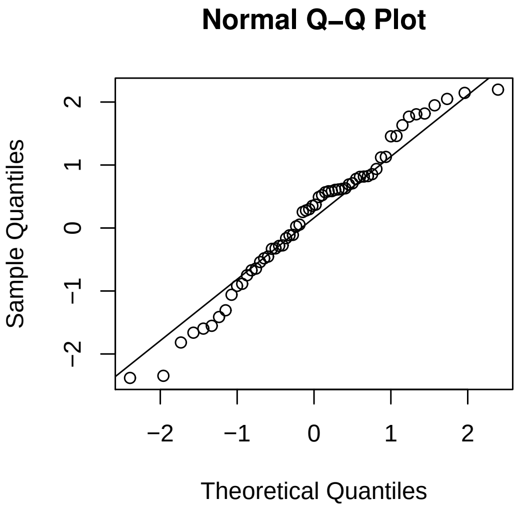
Evaluating Normality

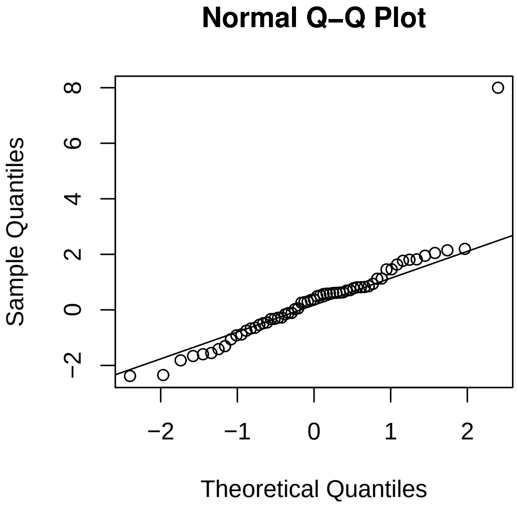
- *When the CLT has not "converged"* then the population distribution must be normal for you to use the t distribution.

- Evaluate normality using a *normal probability plots*, which plot the observed quantile against normal quantiles ("Q-Q plot").

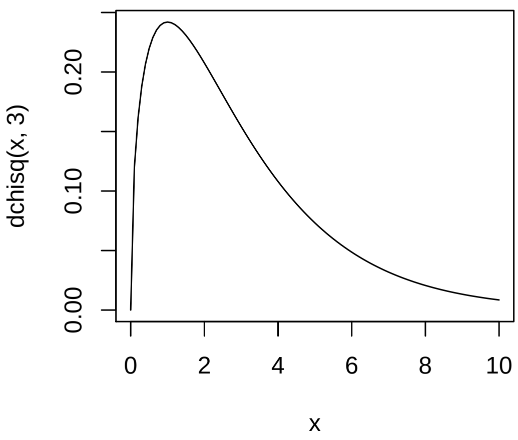
- Points falling on a line indicate normality.

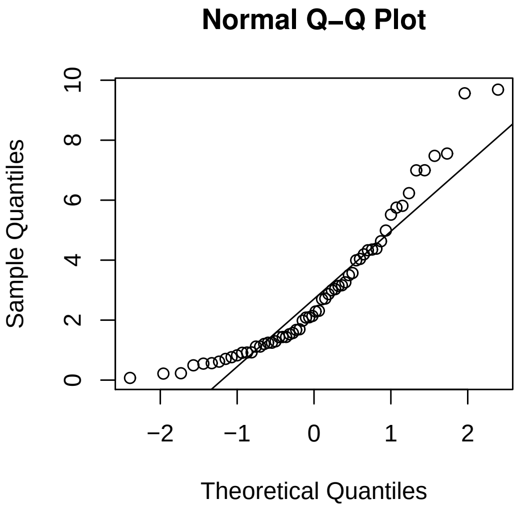
- SPSS: use Analyze / Descriptives / Explore, check normal plots

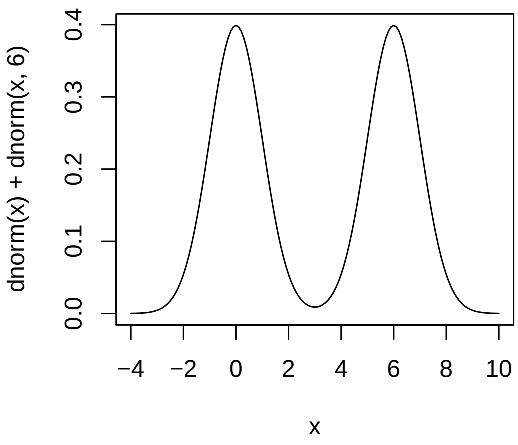


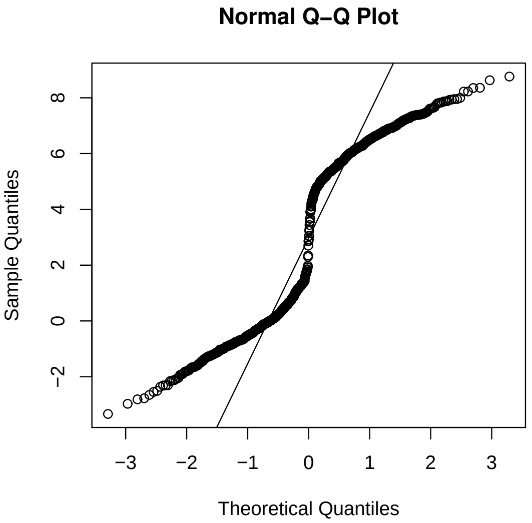


Evaluating Normality









Testing Normality

We can perform an explicit test of normality with the Shapiro-Wilk test. In SPSS use Descriptives / Explore and check Normality Tests
H_0: Population distribution normal
H_1: Population distribution not normal
Using SPSS, $P = .9876 > .05$, indicating that we cannot reject the null. It is plausible that the population distribution is normal.
This test does not replace looking at your data with Q-Q plots, boxplots and/or histograms.

Your Turn

(**video**) (Siegel 10.22) Although your product, a word game, has a list price of $12.95, each store is free to set the price as it wishes. Your marketing department believes that games generally sell at a mean discount of 12% from the list price. You have just completed a quick survey, and the marked prices at a random sample of stores that sell the product were as follows: 12.95, 9.95, 8.95, 12.95, 12.95, 9.95, 9.95, 9.98, 13.00, 9.95.

Two-Sample Tests for Means

- *One-group pretest-posttest design (before-after)*

  $$O_1 \qquad X \qquad O_2$$

  - O_t is a measurement, X indicates a *treatment*, time occurs from left to right.
  - Use *paired-sample t test*. Key point: two outcome measurements (*pre-* and *post-measures*, O_1 and O_2) are taken on each sampling unit.

- *After-only with control group* (ideally we have random assignment to treatment and control groups, indicated by (R), TG means treatment group and CG means control group)

TG:	(R)	X	O_1
CG:	(R)		O_2

  Use *independent-sample t test*. Key point: a single outcome measure is taken on each sampling unit from two independent samples.

Dependent Samples

- Definition — Two samples are said to be *paired*, *matched* or *dependent* samples when for each data value collected from one sample there is a corresponding data value collected from the second sample.
- For each *sampling unit* we have two measurements that we wish to compare. These are often, but not always, *pre-* and *post-measures* (before/after).
- Testing Procedure — compute differences between post- and pre-measures and use single-sample formulae.
- (Siegel 10.5) (**video**) Some of your advertisements seem to get no reaction, as though they are being ignored by the public. You have arranged for a study to measure the public's awareness of your brand before and after viewing a TV show that includes the ad in question. You wish to see if the ad has a statistically significant effect as compared with zero, representing no effect. Your brand awareness, measured on a scale from 1 to 5, was found to have increased an average of 0.22 points when 200 people were shown an ad and questioned before and after. The standard deviation of the increase was 1.39 points.

  1. Identify the null and alternative hypotheses for a two-sided test, using both words and mathematical symbols.
  2. Perform a two-sided test by finding the P-value.

Example: Lie Detector Problem

(**video**) (Siegel 10.40) Stress levels were recorded during a true answer and a false answer given by each of six people in a study of lie-detecting equipment, based on the idea that the stress involved in telling a lie can be measured. Test at the 5% level whether the stress levels are different.

Subject	True	False	$D = F - T$
1	12.8	13.1	0.3
2	8.5	9.6	1.1
3	3.4	4.8	1.4
4	5.0	4.6	−0.4
5	10.1	11.0	0.9
6	11.2	12.1	0.9
Mean	8.5	9.2	$\overline{D} = 0.7$
Std Dev			$S_D = 0.6481$

1. Let μ be the true mean difference between the False and True responses. $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$.

2. Compute difference column and check for normality (since sample size small)

3. Compute standard error:

$$S_{\overline{D}} = \frac{S_D}{\sqrt{n}} = \frac{0.6481}{\sqrt{6}} = 0.2646$$

4. Compute P-value:

$$P(\overline{D} > 0.7) = P\left(T_5 > \underbrace{\frac{0.7 - 0}{0.2646}}_{2.66}\right) = 0.02283.$$

The P-value is $2 \times 0.02283 = 0.0457$, where `TDIST(2.66, 5, 1)` $= 0.02283$

5. We reject H_0 because $0.0457 < 5\%$. This P-value is close to 0.05, so we have weak evidence against the null. Conclude that the lie detector works.

Your Turn

1. (Siegel 10.6b) P-value = .0125

2. (Siegel 10.41) A group of experts has rated your winery's two best varietals. Ratings are on a scale from 1 to 20, with higher numbers being better. The results are as follows:

   (a) Is this a paired or unpaired situation?

   (b) Test to see if the average ratings are significantly different.

   (c) Find the average rating for each varietal and the average difference in ratings (Chardonnay − Cabernet Sauvignon). Find the standard error of the difference.

   (d) Find the 95% two-sided confidence interval for the mean difference in rating.

   (e) Is the difference significant? Use both confidence intervals and P-values.

3. A company sent seven of its employees to attend a course in building self-confidence. These employees were evaluated for their self-confidence before and after attending this course. Here are the scores of the employees before and after attending the course:

Before	8	5	4	9	6	8	5
After	10	7	5	11	6	7	9

   Test at the 1% significance level if attending this course changes the mean score of employees. To receive full credit you must state the null and two-sided alternative hypotheses, compute the P-value, and state your decision. *Answer:* $P(\bar{D} > 1.4286) = P(T_6 > 2.3355) = .0291$*, so* $P = 0.058$ *and so we cannot reject* H_0*, although it is close to the cut-off. We need a bigger study.*

Independent-Sample Test for Large Samples

Do those who have high levels of utilitarian engagement with the editorial content of a magazine rate the ads in the magazine higher than those who are not so engaged? Study design:

TG: X O_1
CG: O_2

- Let μ_1 and σ_1 be the mean and standard deviation of those with low levels
- Let μ_2 and σ_2 be the mean and standard deviation of people who have high levels of utilitarian engagement
- We would like to test the hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2$$

- Let $D = \mu_2 - \mu_1$. We can equivalently test

$$H_0 : D = 0 \quad \text{versus} \quad H_1 : D \neq 0$$

	Utilitarian			Standard	Std. Error
	Engagement	n	Mean	Deviation	Mean
Attitude	≥ 3.5	787	4.3043	1.30159	0.04640
Towards Ad	< 3.5	2806	3.8025	1.19825	0.02262

- We can estimate D with $\hat{D} = 4.3043 - 3.8025 = 0.50183$

Independent-Sample Test for Large Samples

1. When both samples are large (e.g., $n_H \geq 40$ and $n_L \geq 40$) and the CLT applies, the distribution of $\hat{D}$ is approximately normal and you can use the Z distribution to find probabilities. For small samples use T distribution, but degrees of freedom are complicated.

2. Variation: it can be shown that

$$\sigma_{\hat{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad \text{and} \qquad s_{\hat{D}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$s_{\hat{D}} = \sqrt{\frac{1.30159^2}{787} + \frac{1.19825^2}{2806}} = 0.05162$$

3. Tail probability and P-value:

$$P(\hat{D} > 0.50183) \approx P\left(Z > \underbrace{\frac{0.50183 - 0}{0.05162}}_{9.722} \right) \approx 0$$

The P-value is also approximately 0. We reject H_0 and conclude that engagement affects the copy testing measures.

4. 95% confidence interval for difference

$$0.50183 \pm 1.96 \times 0.05162 = (0.401, 0.603)$$

5. SPSS output: Use row for “Equal variances not assumed”

	t-test for Equality of Means						
Equal Variances	*t*	df	Sig. (2-tail)	Mean Diff.	Std. Err Diff.	95% CI for Difference	
Assumed	10.184	3591	0.000	0.50183	0.04928	0.40522	0.59844
Not Assumed	9.722	1185.3	0.000	0.50183	0.05162	0.40056	0.60310

Small-Sample Considerations

- When the sample sizes are small, it is desirable to assume that the variances are equal in the two groups. Don't make this assumption unless it's plausibly true (otherwise use the "unequal variance" row in SPSS). With equal variances, we can estimate a single *pooled* estimate rather than making the two separate estimates s_1^2 and s_2^2:

$$S^2_{\text{pool}} = \frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- For the engagement problem

$$s^2_{\text{pool}} = \frac{(787 - 1)1.30159^2 + (2806 - 1)1.19825^2}{787 + 2806 - 2} = 1.492347,$$

$$s_{\hat{D}} = \sqrt{\frac{1.492347}{787} + \frac{1.492347}{2806}} = 0.04928, \text{ and}$$

$$P(\hat{D} > 0.50183) \approx P\left(T_{3591} > \underbrace{\frac{0.50183 - 0}{0.04928}}_{10.184}\right) \approx 0$$

- In general, use a T distribution with $n_1 + n_2 - 2$ degrees of freedom.

- For small samples both population distributions must be normal, otherwise call a statistician.

Procedure for Testing the Equality of Two Means

1. Pick α, state null and alternative hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2$$

2. Examine data for outliers, miscodings, and normality[39]

3. Compute the P-value using your favorite statistics package.

   (a) Large sample: assume unequal variances (unless you're sure the variances are equal)

   (b) Small sample: decide if the two population variances are equal (look at boxplots, standard deviations, test for equality of variances)

   (c) Reject H_0 if P-value $< \alpha$. Alternatively, reject if reference value 0 falls outside $100(1-\alpha)\%$ CI

4. State your decision, null and alternative, test used and conclusion

[39] If the sample sizes are large, it is not so important for the distributions to be normal. If the samples are small and the distributions are not normal, consult a statistician.

Your Turn

1. (10.47) (video) Consider the weights for two samples of candy bars, before and after intervention given in the data set `candybar.sav`.
   (a) Is this a paired or unpaired situation? Why?
   (b) Find the mean and standard deviation before and after. What are your conclusions?
   (c) Find the 95% confidence interval for the population mean difference in weight per candy bar (after minus before).
   (d) Did intervention produce a significant change in weight? How do you know?
   (e) Discuss the variances? Are they different?

2. (video) Test the significance of the difference in problem 1 on page 93.

3. (10.42) (video) To understand your competitive position, you have examined the reliability of your product as well as the reliability of your closest competitor's product. You have subjected each product to abuse that represents about a year's worth of wear-and-tear per day. See reliability.sav for the data.
   (a) Find the average time to failure for your and your competitor's products. Find the average difference (yours minus your competitor's).
   (b) Find the appropriate standard error for this average difference. In particular, is this a paired or unpaired situation? Why?
   (c) Find the two-sided 95% confidence interval for the mean difference in reliability.
   (d) Test at the 5% level if there is a difference in the reliability between your products and your competitor's.

4. (10.45) There are two manufacturing processes, old and new, that produce the same product. The defect rate has been measured for a sample of days for each process, resulting in the following summaries. Old: defect rate = 0.047, standard deviation = 0.068, sample size = 50. New: defect rate 0.023, standard deviation = 0.050, sample size = 44.[40]

[40] (a) .024; (b) .01246 (pooled); (c) $H_0 : \mu_{\text{new}} \geq \mu_{\text{old}}$, $H_1 : \mu_{\text{new}} < \mu_{\text{old}}$; (d) Find $S^2_{\text{pooled}} = (.068^2(50-1)+.05^2(44-1))/(50+44-2) = .003631$, Using the pooled standard deviation, $se(\hat{D}) = \sqrt{.003631(1/50+1/44)} = .01246$ and the two-sided 95% confidence interval is $0.024 \pm 1.96 \times 0.01246$. (e) $P(\hat{D} > .024) = P(Z > .024/.01246) = P(Z > 1.93) = .0270 < 5\%$ so reject. Following Siegel's advice to use the unpooled standard deviations you get $se(\hat{D}) = 0.012219$ and a P-value of 0.02475, computed with the normal distribution. In this case the variances are approximately equal, so use the pooled variance. Using an F test not covered in this class, the variances are significantly different, but not highly significantly different. It's good to look at both P-values for the difference in means. In this case they both give the same conclusion.

(a) By how much would we estimate that the defect rate would improve if we switched from the old to the new process?

(b) What is the standard error of your answer to part a?

(c) Your firm is interested in switching to the new process only if it can be demonstrated convincingly that the new process improves quality. State the null and alternative hypotheses for this situation.

(d) Find the 95% confidence interval for the long-term reduction in defect rate.

(e) Is the improvement statistically significant?

5. You are considering a new machine. Given below are the times required to assemble a part by the old and new machine. Is there a significance difference in the times? The times of the new machine are: 42.1, 41.3, 42.4, 43.2, 41.8, 41.0, 41.8, 42.8, 42.3, 42.7. The times of the old machine are: 42.7, 43.8, 42.5, 43.1, 44.0, 43.6, 43.3, 43.5, 41.7, 44.1.

Your Turn: Siegel Chapter 10

10.42. (**video**) SPSS (a) 4.475, 2.360, 2.115; (b) 1.0581, unpaired; (c) $[-0.9367, 5.1667]$, (d) using unequal variances $P = .061$, do not reject; (e) $P < .05$ assuming equal variances. Repeat this using logs and you will have a borderline significant result.

10.43. SPSS $P = .01326$ is the one-sided value, reject H_0 at the .05 level

10.44. SPSS (a) unpaired; (b) 2, 3.66667; (c) .557773; (d) [.424, 2.909]; (e) reject

10.45. (a) .024; (b) .01246 (pooled); (c) $H_0 : \mu_{\text{new}} \geq \mu_{\text{old}}$, $H_1 : \mu_{\text{new}} < \mu_{\text{old}}$; (d) Find $S^2_{\text{pooled}} = (.068^2(50-1) + .05^2(44-1))/(50+44-2) = .003631$, Using the pooled standard deviation, $se(\hat{D}) = \sqrt{.003631(1/50+1/44)} = .01246$ and the two-sided 95% confidence interval is $0.024 \pm 1.96 \times 0.01246$. (e) $P(\hat{D} > .024) = P(Z > .024/.01246) = P(Z > 1.93) = .0270 < 5\%$ so reject. Following Siegel's advice to use the unpooled standard deviations you get $se(\hat{D}) = 0.012219$ and a P-value of 0.02475, computed with the normal distribution. In this case the variances are approximately equal, so use the pooled variance.

10.46. Use SPSS (a) 67.417, 85.84; (b) 12.9187, 10.4959; (c) 18.423, 7.2076 assuming equal variances (note that the standard error is 7.0603 without the assumption); (d) [2.12, 34.73]; (e) $P = .031$, reject. To do this by hand, note that $S_{\text{pooled}} = 11.90294$, $\hat{D} = 18.4233$, $S_{\hat{D}} = 7.20759$ and $P(\hat{D} > 18.4223) = P(T_9 > (18.4233-0)/7.20759) = .03088$ from Excel.

10.47. (**video**) Use SPSS (a) unpaired; (b) $[-0.0245, 0.0348]$; (c) no. Why is this unpaired? The candy bars are different. Paired is when you have some sampling unit and take two measurements on it, e.g., we have a sample of people and we take before and after measurements on them. Here, the before candy bars are one sample and the after candy bars are different.

10.48. (a) Detroit 13.733%, Kansas 9.585%; (b) $\hat{D} = 4.15\%$; (c) 1.408%; (d) [1.4%, 7.0%]; (e) P=.003213, reject. Note: $p_{\text{pool}} = .1156$, $\text{se}(\hat{D}) = \sqrt{.1156(1-.1156)(1/983+1/1085)} = 1.408\%$, $P(\hat{D} > 4.15\%) = P(Z > 4.15\%/1.408\%) = P(Z > 2.95) = .001606$ so $P = 2 \times .001606 = .0032$.

10.49. You should assume unequal variances and find $\hat{D} = .42$, $S_{\hat{D}} = .1711$ and $P = 2 \times .00705282 = .0141$. Reject at the 5% level, but not at 1% level.

10.50. $S_{\hat{D}} = .1715$, $P = .0000$, reject at the 5% and 1% level. With unequal variances $\hat{D} = .18$, $S_{\hat{D}} = 2 \times .1994$ and $P = .0000$.

10.51. $S_{\hat{D}} = .1833$, $P = .3267$, do not reject at the 5% or 1% level. With unequal variances $S_{\hat{D}} = .1919$ and $P = .3481$.

Testing Equality of Two Proportions

A random sample of 500 was selected in a large metropolitan area to determine various information concerning consumer behavior. Among the questions was "Do you intend buy product A?" Of 240 males, 136 answered yes. Of 260 females, 174 answered yes. Is there evidence of a significant difference between males and females?

- $H_0 : \pi_M = \pi_F$ versus $H_1 : \pi_M \neq \pi_F$
- Consider a new parameter $D = \pi_F - \pi_M$ and estimate $\hat{D} = 174/260 - 136/240 = 0.1026$. We can equivalently test $H_0 : D = 0$ versus $D \neq 0$
- What is the shape and variance of the distribution of $\hat{D}$?
  - If sample sizes are large, then $\hat{D}$ will have an approximate normal distribution
  - Under H_0 the proportions, and thus the variances, are equal. It is customary to compute a pooled estimate of $\pi_M = \pi_F$ (call it p) and use it to find the variance of the difference

    $$p = \frac{136 + 174}{240 + 260} = 0.62$$

    Using the pooled estimate,

    $$s_{\hat{D}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.62(1-0.62)\left(\frac{1}{240} + \frac{1}{260}\right)} = 0.04345$$

- Now we can compute the P-value (Click here for SPSS solution)

  $$P(\hat{D} > 0.1026 | H_0) \approx P\left(Z > \underbrace{\frac{0.1026}{0.04345}}_{2.3606}\right) \approx 0.0091$$

  The two-sided P-value is thus $2 \times 0.0091 = 0.0182$

- A 95% confidence interval for the difference is given by

  $$0.1026 \pm 1.96 \times 0.04345 = (0.017, 0.19)$$

- This difference is significant ($0.0182 < 5\%$) but not highly significant ($0.0182 > 1\%$).

Your Turn

1. (17.11) A mail-order company is interested in whether or not "order rates" (the percent of catalogs mailed that result in an order) vary from one region of the country to another. In the east, there were 926 mailings that resulted in an order and 22,113 that did not. In the west there were 352 mailings that resulted in an order and 10,617 that did not.

   (a) Find the order rate for each region. Which region appears to be more responsive?

   (b) Are the order rates significantly different from each other?

2. (17.14) The eastern factory had 28 accidents last year out of a work force of 673. The western factory had 31 accidents during this time period out of 1,306 workers.

   (a) Which factory had more accidents? Which factory had a greater accident rate?

   (b) Is there a significant difference between the accident rates of these two factories?

3. (17.13) Does it really matter how you ask a question? A study was conducted that asked whether or not people would pay \$30 to eat at a particular restaurant. One group was told "there is a 50% chance that you will be satisfied," while the other was told "there is a 50% chance that you will be dissatisfied." The only difference in the wording is the use of dissatisfied in place of satisfied. The results were that 25.83% of the 240 people who were asked the "satisfied" version said that they would eat there, as compared with only 11.16% of the 215 people who were asked the "dissatisfied" question. Is the difference (between 25.83% and 11.16%) significant, or is it possible to have differences this large due to randomness alone? How do you know?

Computer Homework 2
IMC 451: Market Research and Statistics

You may work in self-selected groups of at most six students. The data set is available on Canvas.

1. (video) (36 points) A European book company mails catalogs to customers and also operates a web site. The company wants to test the effect of including a letter with each catalog that has customized recommendations (e.g., "We know that you like history books and this month we have ________, which is on page xx, ..."). In particular, the company wants to determine if the customized recommendations increase response rates and/or order sizes. It randomly assigns customers to receive letter with either the customized recommendations (treatment group) or with generic recommendations (control group). The file book.sav has data resulting from this test. The variable `tx` indicates whether the customer received the treatment (customized letter) or control (generic letter). The `totitem` variable gives the total number of items that the customer ordered and the `toteuro` variable gives the total amount of the order in Euros.

   (a) (2 point) Is this paired or unpaired experiment? Explain

   (b) (2 point) Define a new variable `response` that takes the value 1 if the customer responded (bought anything) and 0 otherwise. Hint: use Transform / Compute, type `response` into the target box, and "`toteuro`> 0" into the Numeric Expression box. Run a crosstab with `response` in the columns and `tx` in the row. Include percents to show the response rates (percent who respond) for the treatment and control groups. Submit only the crosstab.

   (c) (4 point) Test at the 5% level whether there is a difference in response rates between the treatment and control groups? To receive full credit, define the parameters, state the null and alternative hypotheses, the appropriate P-value, and your conclusion.

   (d) (2 points) Continuing the previous part, give a 95% confidence interval for the difference between the parameters in the previous part.

   (e) (3 points) In addition to studying response rates, the company also wants to determine if the customization increases the amount of the order. Restrict your attention to responders. Hint: use Data / Select cases and select all cases where `response`=1. Generate a boxplot of `toteuro` by `tx` and comment on its shape. In particular, comment on whether the assumptions of a z-test are satisfied.

   (f) (3 points) Compute the logarithm of `toteuro` as `logeuro = LN(toteuro+1)`. For responders only, generate boxplots of `logeuro` for the treatment and control group separately (copy `tx` into the Category Axis). Discuss whether the assumptions of the t-test are more closely satisfied.

   (g) (3 points) For responders only, test at the 5% level whether there is a difference in the log average order amount (`logeuro`) between the treatment and control groups? To receive full credit, state the null and alternative hypotheses, the appropriate P-value, and your conclusion.

   (h) (3 points) The company wants to understand whether the customization has a different effect on various customer segments. The `RFseg` variable gives four segments (recent one-time buyers, recent multi-time buyers, one-year-lapsed customers, and 2+ year lapsed customers). Turn off the filter under Data / Select cases so that you are using all observations again. Generate the following graph. Go to Graph / Legacy / Error Bar and select a Clustered graph. Copy `response` into the "Variable" box, `RFseg` into the "Category Axis" box, and `tx` into the "Define Clusters By" box. Submit the resulting graph and comment on whether the segments are equally responsive to this customization treatment.

(i) (3 points) For each of the four segments, test at the 5% level whether there is a difference in response rates between the treatment and control groups? Hint: click on Data / Split File, click on "organize output by groups," then copy `RFseg` into the "Groups Based On" box. Then run the t-test. Complete the table below and comment on the RF segments where the customization is effective.

	Response Rate		P-Value for
RF Segment	Treatment	Control	Difference
$R < 1, F = 1$			
$R < 1, F > 1$			
$1 < R < 2$			
$R > 2$			

(j) (3 points) For responders only, generate separate boxplots of `logeuro` for treatment and control groups within each of the four RF segments. Hint: Turn off the Split File option (Data / Split File / Analyze all cases). Now go into Data / Weight Cases and Weight Cases By the `response` variable (this assigns 0 weight to non-responders and is an alternative to select cases). Click on Graph / Boxplot / Simple. Copy `logeuro` into the variable box, `tx` into the Category Axis box, and `RFseg` into the Panel By Columns Box. Submit the resulting graph and comment on the shape of the boxplots.

(k) (5 points) For responders only, run t-tests comparing order amounts for the treatment and control groups for each RF segment separately. You do not need to include output. Instead, briefly discuss whether the order amount is increased for any RF segment. Use the 5% level of significance. Hint: Split the file by `RFseg` again and make sure that cases are weighted by response. Use the independent-sample t-test, copy `toteuro` and `logeuro` into the Test Variable Box and `tx` into the Group Variable box.

2. (**video**) (20 Points) The file `orders.sav` gives a complete list of orders amounts (in Euros) that a catalog company received during a one-month period. You may assume that the order amounts from these two periods are representative of all orders, i.e., while people may be more likely to place orders during some periods of the year than others, the distribution of order amounts is the same. Thus, you can regard this as a very large random sample of all orders.

(a) (4 points) Estimate the probability of a single order being under 40 Euros. Briefly describe how you arrived at this estimate.

(b) (2 points) Briefly describe the shape of the distribution of (single) orders, supported with appropriate descriptive statistics, e.g., give the mean, standard deviation, and discuss other aspects such as skewness, and number of modes.

(c) (4 points) The company expects to receive 1000 orders tomorrow, following the same distribution as those in orders.sav. Consider the random variable, “the total amount of revenue from the 1000 orders tomorrow.” (Do not worry about returns, bad debt, etc. for this problem.) Describe the shape of the distribution of the total amount, i.e., give its mean, standard deviation, and name.

(d) (2 points) Estimate the probability of having a “great day” tomorrow, where the total revenue exceeds 40,000 Euros.

(e) (2 points) Estimate the probability of generating between 38,000 and 40,000 Euros in total orders tomorrow.

(f) (3 points) Estimate the probability that the average order amount across the 1000 orders tomorrow will be between 35 and 37 Euros.

(g) (3 points) Suppose that 10% of all orders will be returned and that 5% of the non-returned orders will result in bad dept. Suppose further that the company makes 40% profit on every Euro it receives. Find the mean and standard deviation of the total amount of profit (after accounting for returned orders, bad debt and the margin) from the 1000 orders tomorrow. Hint: think of profit as a linear transformation of the total number of orders.

(h) (0 points) For you to think about but not turn in: how could you account for the fact that the number of returned and bad debt orders is itself a random variable? Hint: start by thinking only about returns, ignoring bad debt, and assume a binomial distribution for the number of paid orders.

Crosstabs Revisited

(17.14) The eastern factory had 28 accidents last year, out of a work force of 673 (so 645 not accidents). The western factory had 31 accidents during this period, out of 1,306 workers (so 1,275 not accidents).

			Accident		Total
			No	Yes	
Region	East	Count	645	28	673
		Expected Count	652.9	20.1	673
		% within Region	95.8%	4.2%	100%
		% within Accident	33.6%	47.5%	34%
		Residual	−7.9	7.9	
		Std. Residual	−0.3	1.8	
		Adjusted Residual	−2.2	2.2	
	West	Count	1275	31	1306
		Expected Count	1267.1	38.9	1306
		% within Region	97.6%	2.4%	100%
		% within Accident	66.4%	52.5%	66%
		Residual	7.9	−7.9	
		Std. Residual	0.2	−1.3	
		Adjusted Residual	2.2	−2.2	
Total		Count	1920	59	1979
		% within Region	97.0%	3.0%	100%
		% within Accident	100%	100%	100%

Independent Sample Z-test for Proportions

- Is there a relationship between region and accident. The row percents indicate that accidents appear to occur more commonly in the east ($p_E = 4.2\%$) than the west ($p_W = 2.4\%$), but could this difference be an artifact of the sample?
- Earlier we would answer this question with

$$H_0 : \pi_E = \pi_W \quad \text{versus} \quad H_1 : \pi_E \neq \pi_W,$$

  where π_W is the probability that a worker in the west has an accident and π_E is the probability for the east.
- We observe a difference in percents of

$$\hat{D} = .04160 - .02374 = 0.01787$$

- The standard error of this difference is

$$S_{\hat{D}} = \sqrt{\frac{59}{1979}\left(1 - \frac{59}{1979}\right)\left(\frac{1}{673} + \frac{1}{1306}\right)} = 0.008070$$

- The tail probability is

$$P(\hat{D} > 0.01787) \approx P\left(Z > \underbrace{\frac{0.01787}{0.0008070}}_{2.2141}\right) \approx 0.01341$$

- $P = 2 \times 0.01341 = 0.02682 < 5\%$, so reject at the 5% level. Accidents are more common in the east.

Statistical Independence

- Equivalently, we can evaluate whether the rows are *independent* of the columns. Recall that A and B are *independent* if and only if

  1. $P(A \cap B) = P(A) \times P(B)$
  2. $P(A|B) = P(A)$
  3. $P(B|A) = P(B)$

- This implies the following compound H_0:

$$\begin{aligned} H_0: \quad & P(Y|E) = P(Y|W) = P(Y) \quad \text{and} \\ & P(N|E) = P(N|W) = P(N) \end{aligned}$$

  against the alternative that any one of the above is not equal.

- Under independence, the *expected cell counts* are

$$\frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

- Wikipedia page on chi-square test

Chi-square Test of Independence

- To test the hypotheses with type I error rate α:

  H_0: rows and columns are independent

  H_a: rows and columns are dependent

  assume H_0 is true and evaluate the likelihood of observing data as "extreme" or more extreme than what have using the random variable

  $$\text{Test Stat} = \sum_{\text{all cells}} \frac{(\text{obs} - \text{exp})^2}{\text{exp}},$$

  which is a random variable with an asymptotic χ^2 (Chi-squared)[41] distribution with $(\#\text{row} - 1)(\#\text{col} - 1)$ degrees of freedom

- Note: if H_0 is true, $(\text{obs} - \text{exp}) \approx 0$ and the test statistic will be approximately 0. Positive test statistics indicate dependence.

- Reject H_0 if the P-value is less than α (e.g., .05).

- This test may be used only if none of the *expected* cell counts are less than 5.

- For this problem,

  $$\frac{(645 - 652.9)^2}{652.9} + \frac{(28 - 20.1)^2}{20.1} + \frac{(1275 - 1267.1)^2}{1267.1} + \frac{(31 - 38.9)^2}{38.9} = 4.902$$

- The P-value is $P(\chi^2_1 > 4.902) = \texttt{CHIDIST}(4.902, 1) = 0.02682 < 5\%$, so reject H_0 and conclude that accidents are dependent on region. Note that the P-value matches the one from the independent-sample Z test.

[41] A chi-squared distribution is the square of a standard normal (Z) variable.

Chi-Square Cell Values

- If you cannot reject independence, stop. Otherwise, you may then look at row/column percents and chi-square residuals to understand the nature of the relationship.
- *Expected Count* (or *fitted value under* H_0): if region and accidents were independent, the number of people we would expect to find in a particular cell, e.g., if they were independent, we'd expect to find 20.1 in the east-Yes cell
$$\frac{673 \times 59}{1979} = 20.1$$
- *Unstandardized Residuals* or *Deviation*: difference between observed and expected counts, e.g., $28 - 20.1 = 7.9$.
- *Standardized residuals*: (observed − expected)/ $\sqrt{\text{expected}}$, e.g., $7.9/\sqrt{20.1} = 1.8$. R: `residuals`
- *Adj Std residuals* have an asymptotic standard normal distribution. R: `stdres`
$$\frac{\text{observed} - \text{expected}}{\sqrt{\text{expected}(1 - \text{row }\%)(1 - \text{col }\%)}}$$
e.g.,
$$7.9/\sqrt{20.1(1 - 0.03)(1 - 0.34)} = 2.21$$
Values greater than 2 or less than -2 show deviations from independence

Your Turn

1. A sample of 1000 adults was shown a new product and asked about their purchase intentions. Here are the results:

	Purchase Intent	
	Yes	**No**
Male	500	0
Female	0	500

   Is gender related to purchase intent? Does the χ^2 test apply?

2. Perform a χ^2 test of independence on problem 2 on page 283.

3. Perform a χ^2 test of independence on problem 3 on page 283.

4. Two drugs were administered to two groups of randomly assigned patients to cure the same disease. One group of 60 patients and another group of 40 patients were selected. The following table gives information about the number of patients who were cured and not cured by each of the two drugs:

	Cured	**Not Cured**
Drug I	46	14
Drug II	18	22

   Test at the 1% significance level whether or not the two drugs are similar in curing and not curing the patients.

5. (17.12) A commercial bank is reviewing the status of recent real estate mortgage applications. Some applications have been accepted, some rejected, and some are pending while waiting for further information. The data are shown below. Are the differences between residential and commercial customers significant? Perform a chi-square test and analyze the residuals.

	Residential	**Commercial**
Accepted	78	57
Information Requested	30	6
Rejected	44	13

6. (Siegel 17.15 modified) (video) Peterson and Wilson (1992) conducted a study relating to the manner in which a satisfaction question is posed. They asked a closed-end question about satisfaction with a household's primary vehicle in a telephone interview. An identical question was asked in each of two versions of the

questionnaire. The only difference was that in one version participants were asked how "satisfied" they were, whereas in the other they were asked how "dissatisfied" they were; response categories were identical in the two question versions. The sample consisted of randomly selected middle-class automobile owners in two large metropolitan areas and respondents were assigned randomly to one of the two question-wording groups. The cross tabulation below gives the counts.

	Satisfied	**Dissatisfied**
Very satisfied	139	128
Somewhat satisfied	82	69
Somewhat dissatisfied	12	20
Very dissatisfied	10	23

(a) Perform a chi-square test at the 5% level and examine the residuals.

(b) Assume that the four-point satisfaction scale is interval and perform an independent-sample t-test. Discuss the differences between what you can conclude from this test compared with the chi-square test.

(c) Discuss threats to internal and external validity.

Siegel Chapter 17
Homework Solutions

17.1 (a) There is no problem since the *expected* cell count is greater than 5. (b) There is a problem because the *expected* cell count is less than 5.

17.4. (a) .2247; (b) .0187; (c) .0449; (d) .0008416; (e) .2247; (f) Should be equal.

17.5. (a) Add them up. (b) You should do the whole table. The overall percent for manager and better is 2.7%. (c) You should do them all. For example, there is a 26.7% chance that a manager says better. (d) You should do them all. For example, there is an 11.1% chance that someone who said better is a manager. (e) You should note that, for example, all of the percents within an attitude are about equal (90% for other, 10% for manager). This is an indication that the variables are independent. Information about one variable does *not* tell us very much about the other variable.

17.6. (a) Information about whether or not a person is a manager does not tell us anything about their response. (b) 17.3; (c) Use SPSS; (d) 5.224; (e) 3.

17.7. Use the P-Value, .156 > .05 so you cannot reject the null hypothesis of independence.

17.8. (a) Add them up. (b) Do them all. For example, there is a 5.3% chance that someone says "very likely" and is a stockholder. (c) Do them all. For example, there is a 13% chance that a stockholder says "very likely." (d) Do them all. For example, there is a 40.9% chance that someone who says "very likely" is a stockholder. (e) No. The row percents do not differ from the marginals.

17.9. (a) Information about whether the person is a stockholder does not change what we think their opinion will be. (b) 17.8; (c) Use SPSS. (d) .729; (e) 4.

17.10. Use the P-value $= .948 > .05$ so do not reject H_0.

17.11. (a) 4% in the east versus 3.2% in the west. (b) Test statistic $= 13.488$ with $P = .000 < .05$ so reject H_0.

17.12. (a) Commercial more likely to be "accepted" while residential more likely to be in "information requested" or "rejected." (b) Yes. The test statistic is 12.142 with $P = .002 < .05$, so reject H_0 of independence.

17.13. Perform a chi-squared test of independence. The test statistics is 15.923 with $P = .000 < .05$ so reject H_0 of independence.

17.14. (video) (a) The west has more accidents but also more workers. The east has the higher accident rate. (b) The test statistic is 4.902 with $P = .02682 < .05$, so reject H_0.

17.15. (video) (a) Those asked if they were satisfied. (b) Those asked if they were dissatisfied. (c) Yes, since the test statistic is 8.675 with $P = .034 < .05$, so reject H_0 of independence.

17.17. The test statistics is 4.444 on 2 degrees of freedom. $P = .108 > .05$ so do not reject H_0.

How Large is “Sufficiently Large”?

- “How large is large enough? If the distribution of individuals is not too skewed, $n = 30$ is generally sufficient. However, if the distribution is extremely skewed or has large outliers, n may have to be much larger. If the distribution is fairly close to normal already, then n can be much smaller than 30, say, 20, 10, or even 5. Of course, if the distribution was normal to begin with, then $n = 1$ is enough.” (Siegel, 2012, page 198, footnote 8)
- “If the distribution of the parent variable is normal, the means of samples of size $n = 1$ will be normally distributed. If the distribution of the variable is symmetrical but not normal, samples of very small size will produce a distribution in which the means are normally distributed. If the distribution of the variable is highly skewed in the parent population, samples of a larger size will be needed. (Churchill, 1994, page 594)”

Example Questions

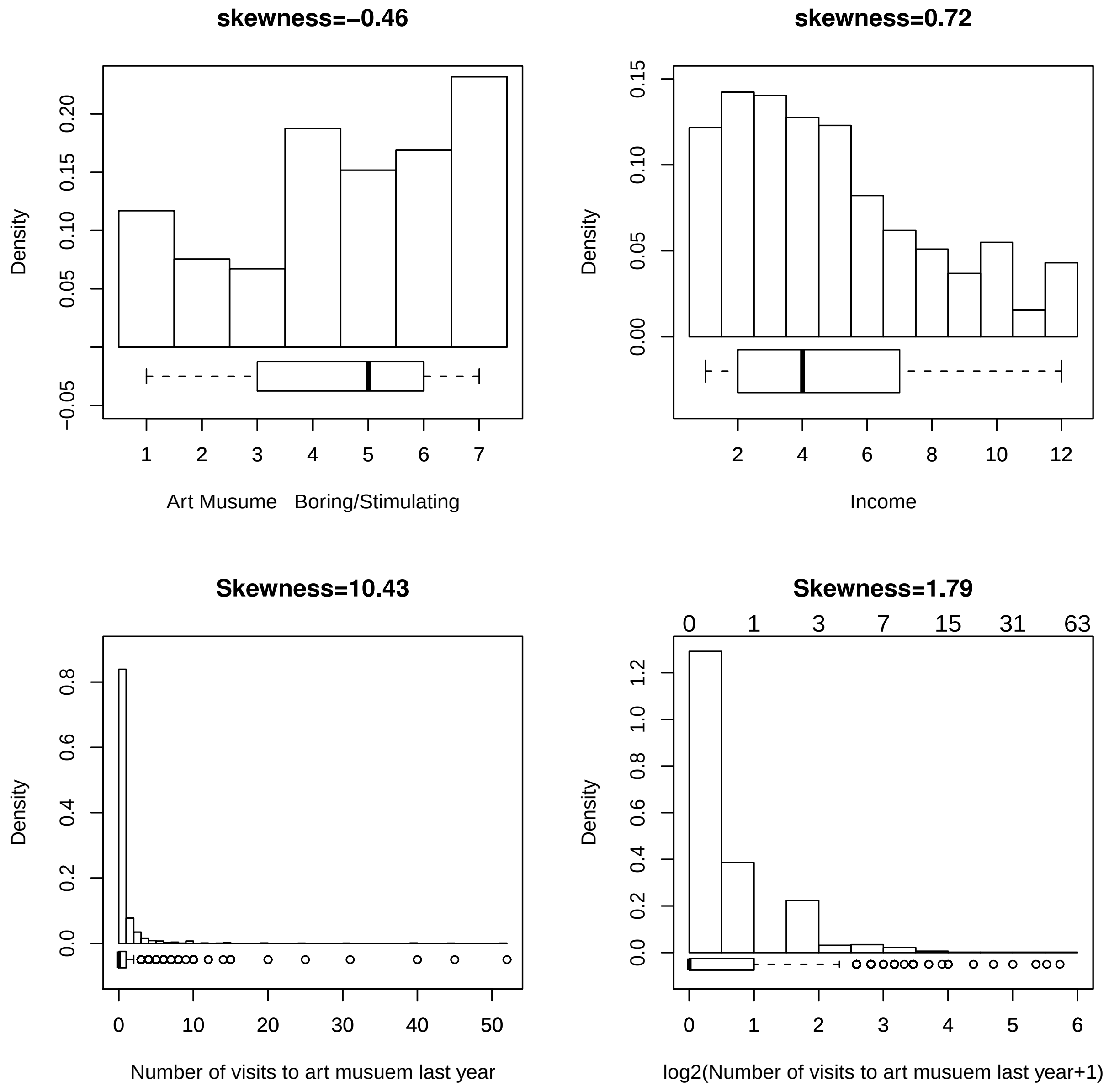
Can I use the t-test option in R/SAS/SPSS/Minitab to test hypotheses (or do I need to call a statistician)?

1. Suppose I have a sample of size 12 from the customers of Lands End catalog. I want to test a hypothesis about the mean dollar amount spent during the most recent six-month season.
2. Same as previous part, but sample size is 50,000.
3. Suppose I have a sample of size 100 from a population that I can assume is normal. I want to test a hypothesis about the mean of this distribution.
4. Same as part (a), but sample size is 12.
5. Suppose I have a sample of size 100. I want to test a hypothesis about the mean of a five-point satisfaction scale that has a left-skewed distribution.
6. Same as previous part, but sample size is 10.
7. Suppose I have a sample of size 20. I want to test a hypothesis about the average age of the population. The distribution is fairly symmetrical.

Simulation Study

- Evaluate type I error rates for the four distributions on the next slide
- Repeat the following 100,000 times
  1. Draw a sample of size n *with replacement* from the available data.
  2. Test $H_0 : \mu =$ true mean against a two-sided alternative using $\alpha = 5\%$. I estimate σ with s and use the t distribution. Note that we know the true mean by computing it from our "population"
  3. Count the number of rejections (which are false positives)
- We expect 5000/100,000 = 5% type I errors
- Type I error rates not equal to 5% indicate that the CLT has not yet converged

Four Distributions



How do we do?

		Number	Type I	95% Confidence	
		Type I	Error	Interval	
Variable	n	Errors	Rate	Lower	Upper
normal	2	5,102	5.102	4.97	5.24
uniform	10	5,535	5.535	5.39	5.68
uniform	15	5,178	5.178	5.04	5.32
uniform	20	5,070	5.070	4.93	5.21
artbore	5	6,468	6.468%	6.32	6.62
artbore	10	6,034	6.034	5.89	6.18
artbore	15	5,560	5.560	5.42	5.70
artbore	20	5,434	5.543	5.29	5.57
artbore	25	5,256	5.256	5.12	5.39
artbore	30	5,170	5.170	5.03	5.31
artbore	35	5,122	5.122	4.99	5.26
artbore	40	5,065	5.065	4.93	5.20
income	5	8,111	8.111	7.94	8.28
income	10	6,691	6.691	6.54	6.85
income	20	5,746	5.746	5.60	5.89
income	30	5,462	5.462	5.32	5.60
income	40	5,438	5.438	5.30	5.58
income	50	5,189	5.189	5.05	5.33
income	60	5,222	5.222	5.08	5.36
income	70	5,192	5.192	5.05	5.33
income	80	5,179	5.179	5.04	5.32
income	90	5,150	5.150	5.01	5.29
income	100	5,089	5.089	4.95	5.23
artmus	5	28,338	28.338	28.06	28.62
artmus	50	16,744	16.744	16.51	16.98
artmus	100	13,939	13.939	13.72	14.15
artmus	500	9,157	9.157	8.98	9.34
artmus	1000	7,139	7.139	6.98	7.30
artmus	3000	5,676	5.676	5.53	5.82
artmus	5000	5,377	5.377	5.24	5.52
artmus	10000	5,202	5.202	5.06	5.34
logartmus	40	6,521	6.521	6.37	6.67
logartmus	80	5,765	5.765	5.62	5.91
logartmus	160	5,468	5.468	5.33	5.61
logartmus	320	5,234	5.234	5.10	5.37

Assessing Normality

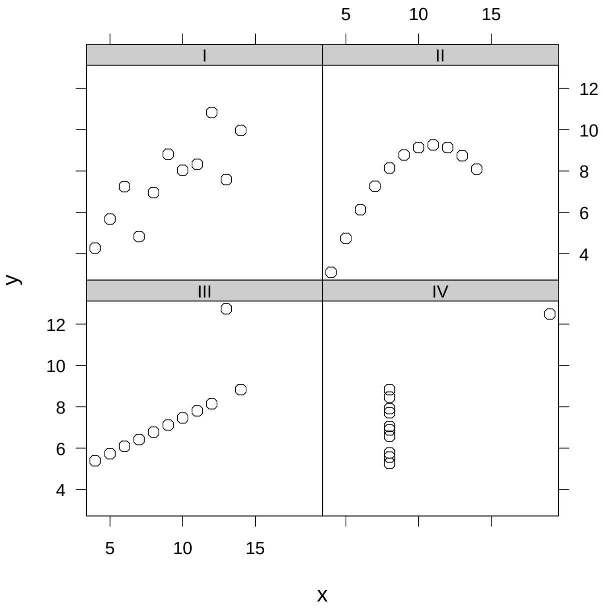
- For **large data sets**, watch out for outliers. If there are outliers, handle them appropriately. It is not as important to have a normal distribution because the CLT tells us the sampling distribution will be approximately normal.
- For **small samples** (e.g., < 100) there are two approaches:
  - Examine normal probability plot, look for a straight line.
  - Perform significance tests. The null hypothesis is that the data come from a normal distribution
  - SPSS: Analyze / Descriptive Statistics / Explore
  - SAS: `PROC UNIVARIATE` with `NORMAL` option
  - R: `qqnorm`, `qqline`, `shapiro.test`

  If the distribution is right skewed and you need to do significance tests, log or square root transformations often help.
- Don't try to generate a normal plot with thousands of observations — SPSS will crash

Anscombe's Example

```
dat = data.frame(
  dataset = factor(c(rep(1,11),rep(2,11),rep(3,11),rep(4,11)),
    labels=c("I","II","III","IV")),
  x=c(10,8,13,9,11,14,6,4,12,7,5, 10,8,13,9,11,14,6,4,12,7,5,
      10,8,13,9,11,14,6,4,12,7,5, rep(8,10),19),
  y=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68,
      9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74,
      7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73,
      6.58,5.76,7.71,8.84,8.47,7.04,5.25,5.56,7.91,6.69,12.5)
)
```

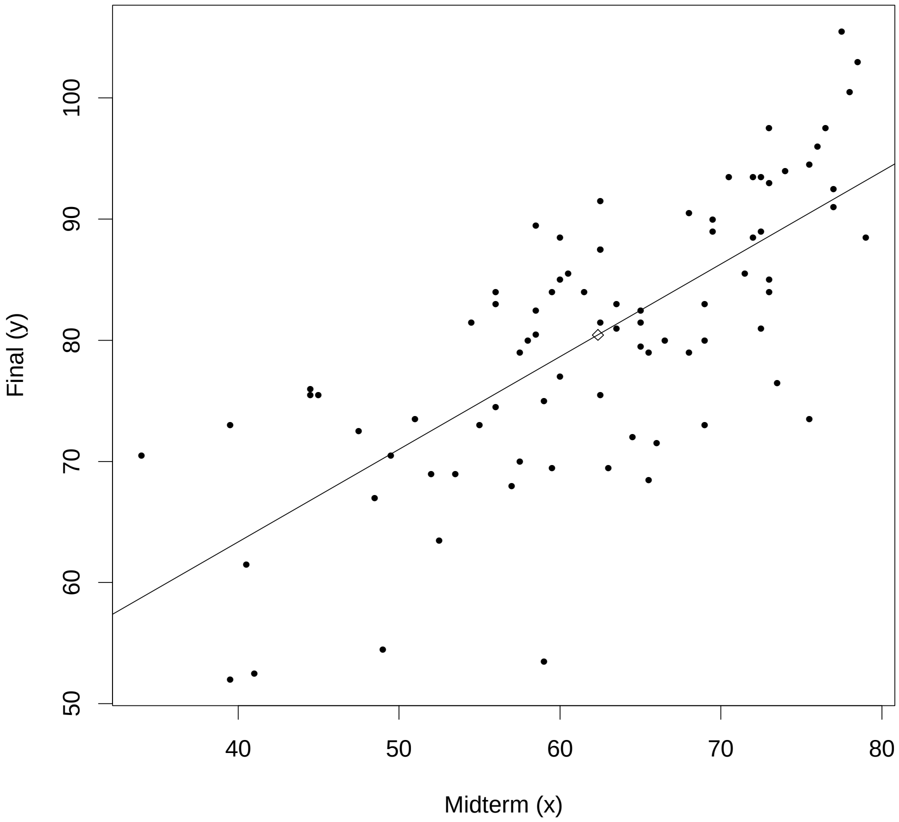
```
> library(lattice)
> xyplot(y~x | dataset, dat)
> tapply(dat$x, dat$dataset, mean)
  I  II III  IV
  9   9   9   9

> tapply(dat$y, dat$dataset, mean)
       I        II       III        IV
7.500909 7.500909 7.500000 7.482727

> tapply(dat$x, dat$dataset, var)
  I  II III  IV
 11  11  11  11

> tapply(dat$y, dat$dataset, var)
       I        II       III        IV
4.127269 4.127629 4.122620 4.151322

> summary(lm(y ~ dataset*x, dat))
```



```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.000e+00  1.125e+00   2.666 0.011431 *
datasetII     8.182e-04  1.592e+00   0.001 0.999593
datasetIII    2.364e-03  1.592e+00   0.001 0.998823
datasetIV    -3.291e-02  1.592e+00  -0.021 0.983618
x             5.001e-01  1.180e-01   4.239 0.000149 ***
datasetII:x  -9.091e-05  1.668e-01  -0.001 0.999568
datasetIII:x -3.636e-04  1.668e-01  -0.002 0.998273
datasetIV:x   1.636e-03  1.668e-01   0.010 0.992229
```

Always look at your data!

Grades Example

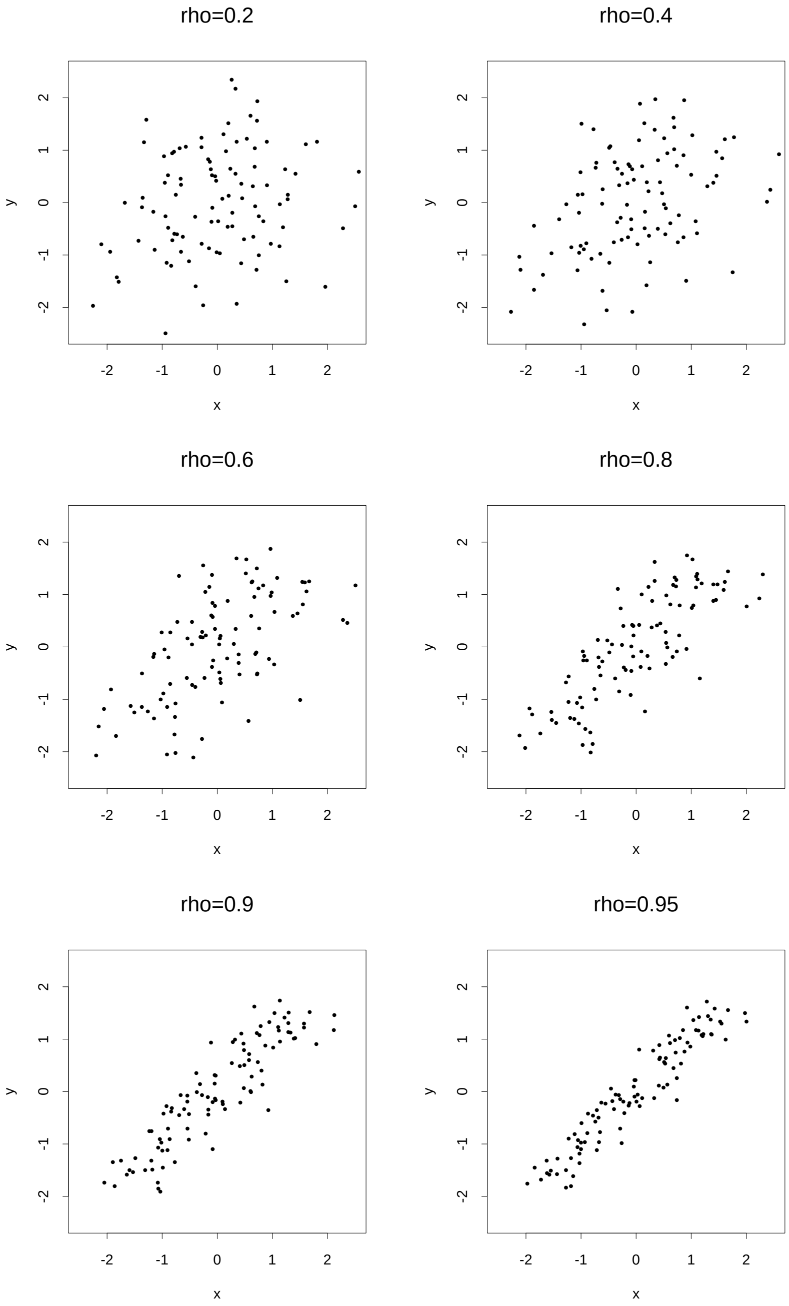


- Midterm: $\bar{x} = 62.35$, $S_x = 10.61$, $n = 83$
- Final: $\bar{y} = 80.44$, $S_y = 11.19$
- Correlation: $r = .7243$

Measures of Association: Correlations

- Commonly used measures:
  - Pearson product moment correlation: Population value ρ, sample statistic r, $-1 \leq \rho \leq 1$.
- Objective: measure *direction* and *strength* of the *association* between two variables.
  - Positive values ($\rho > 0$) indicate positive association — when one variable increases, so does the other.
  - Negative values ($\rho < 0$) indicate negative association — when one variable increases, the other decreases.
  - Zero values ($\rho \approx 0$) indicate no *linear* association.

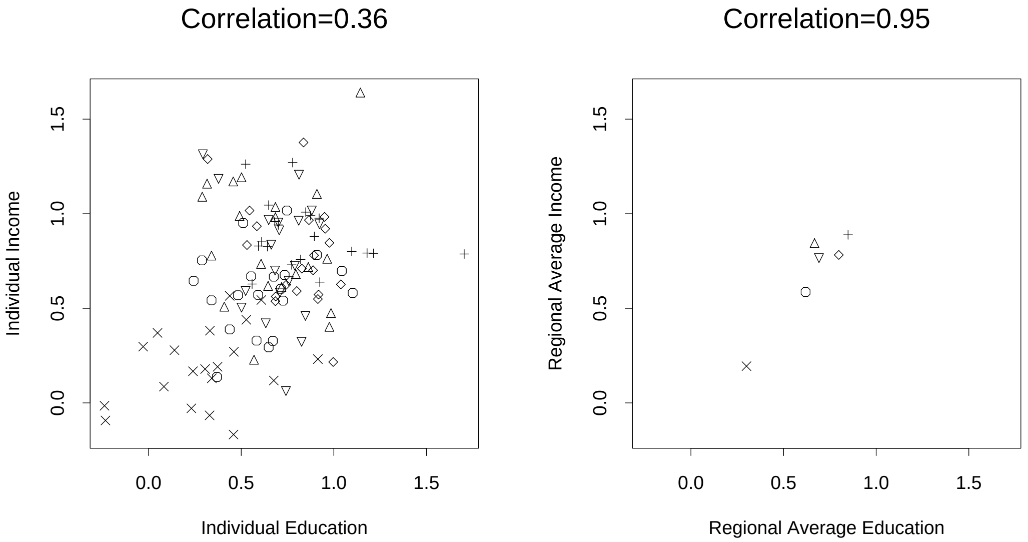
Measures of Association: Correlations



“Association is not Causation”

- Correlation measures association. But association is not the same as causation.
- See “Food Store and Restaurant Spending” example in Siegel
- In the Great Depression, better-educated people tended to have shorter spells of unemployment. Does education protect you against unemployment? Answer: age is a confounding variable
  - Younger people were better educated
  - Employers preferred younger job-seekers
  - Controlling for age makes the effect of education on unemployment much weaker
- Bahry and Silver (1987) showed that subjects who viewed the KGB as effective were less likely to describe themselves as activists. Which way does the causality go?
  - If you think the KGB is efficient, then you don’t demonstrate.
  - If you are an activist, you find out that the KGB is inefficient.
  - People of certain personality types are more likely to describe themselves as activists and are also more likely to describe the KGB as inefficient.
- Breast cancer example

Ecological Correlations



- *Ecological correlations* are based on rates or averages and tend to overstate the strength of an association. Beware whenever rates or averages are correlated!
- For example, beware of a correlation involving average store sales

Notation

- Suppose x has mean μ_x and standard deviation σ_x. The *standard units* (also called *Z-scores*) tell how many standard deviations each observation is from the mean:

  $$Z_x = \frac{x - \mu_x}{\sigma_x}$$

  To compute Z-scores in SPSS, use Analyze / Descriptive Statistics / Descriptives and check box labeled "Save standardized values as variables."

- Let $x_1, \ldots, x_n$ be observations from a distribution with mean μ_x and standard deviation σ_x. The *sample mean* and *sample standard deviation* are

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Let $y_1, \ldots, y_n$ be observations from a distribution with mean μ_y and standard deviation σ_y. Likewise let $\bar{y}$ be the sample mean and S_y be the sample standard deviation.

Interpretation and Computation of Correlations

- *Pearson correlation*

$$\begin{aligned}\rho &= \frac{\sum_{i=1}^N (x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^N (x_i-\mu_x)^2 \sum_{i=1}^N (y_i-\mu_y)^2}} \\ &= \sum_{i=1}^N \left[\frac{(x_i-\mu_x)}{\sqrt{\sum (x_i-\mu_x)^2}} \times \frac{(y_i-\mu_y)}{\sqrt{\sum (y_i-\mu_y)^2}}\right] \\ &= \frac{1}{N}\sum_{i=1}^N \left[\frac{(x_i-\mu_x)}{\sigma_x} \times \frac{(y_i-\mu_y)}{\sigma_y}\right]\end{aligned}$$

  Thus, the Pearson correlation coefficient is the *average of the products of the standardized versions of x and y.*

- We estimate ρ with r

$$r = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i-\bar{y})^2}}$$

- Note that Pearson correlations are unaffected by
  - interchanging the two variables
  - replacing a variable x with $ax+b$, where $a>0$

- There are formulas for hypothesis tests and CIs, e.g., under $H_0 : \rho = 0$, $t = r\sqrt{(n-2)/(1-r^2)}$ is distributed T_{n-2}.

Example: Maintenance Costs

(Siegel 11.1) The age in years and maintenance cost in thousands of dollars for 5 machines are provided in the table below.

	Age	Cost	Standard Units		Product
	x	y	Z_x	Z_y	$Z_x \cdot Z_y$
	2	6	-1.247	-1.023	1.275
	5	13	-0.147	-0.105	0.015
	9	23	1.320	1.206	1.592
	3	5	-0.880	-1.154	1.015
	8	22	0.953	1.075	1.025
Mean	5.4	13.8	0	0	0.985
σ	2.728	7.626	1	1	

$$r = \frac{102.40}{\sqrt{37.2 \times 290.8}} = 0.9845$$

Note: I use σ rather than S

Introduction to Regression

- *Objective:* to quantify the relationship between an interval-level *response* variable and one or more *predictor* variables.
- *Response variable* (also called *dependent*, *criterion*, or *output* (Y) variable): a variable that we wish to study and is causally dependent on other variables
- Note: we will also study categorical dependent variables, but this will be called the *classification problem*
- *Predictor variables* (also called *independent* (X) variables, *covariates*, or *inputs*): variables that are (causally) related to the response variable
  - *Factors* refer to categorical variables
  - *Covariates* refer to numerical variables
- *Why?*
  - *Prediction*: primarily interested in estimating response variable, i.e., $\hat{y}$ (pronounced "y-hat") values
  - Causal attribution: primarily interested in how predictor variables affect response variable, e.g., "β_j" values. This is more *prescriptive*

  Causal attribution is much more difficult, with many more considerations.

Linear Regression

Assume that

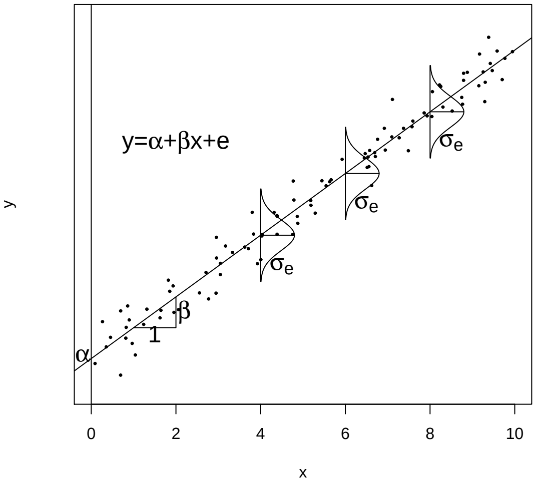
$$y_i = \alpha + \beta x_i + e_i,$$

where **error** e_i has a normal distribution with mean 0, standard deviation σ_e (for all i), and e_i independent of e_j for $i \neq j$. This implies that

$$\mu_{Y|x} = E(Y|x) = \alpha + x\beta.$$

This is called the **regression of y on x**.

- α: the **intercept** or **constant**. On average $Y = \alpha$ when $x = 0$.
- β: the **slope**. Every unit increase in x is associated with an increase in Y of β, *on the average*
- σ_e: **standard deviation of the errors** (σ_e^2 called **error variance**). If errors are normal, empirical rule tells us for any x, 68% of points will fall within one σ_e of the mean ($\alpha + \beta x$).



Estimating the Regression Model

- In practice, we do not know value of *parameters* α, β, and σ_e
- Estimate α, β, and σ_e with a, b, and S_e
- The estimate of $\mu_{Y_i|x_i}$ is denoted by **fitted**, **predicted** or **y-hat value** $\hat{y}_i = a + bx_i$
- The **residual** for observation i is

$$\hat{e}_i = y_i - \hat{y}_i = y_i - a - bx_i$$

- We choose a and b so that they minimize the **least-squares criterion** (SSE means **sum of squared errors**):

$$\text{SSE} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2,$$

- The **ordinary least-squares estimates** (OLS) are

$$b = r\frac{S_y}{S_x} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

  Note, for every SD change in x, there is a change of r SD's in y, on average.

- Estimate σ_e with the **residual standard error**

$$S_e = \sqrt{\frac{\text{SSE}}{n-2}} = S_y\sqrt{(1-r^2)\frac{n-1}{n-2}}$$

Sampling Distribution of Parameters

- Theorem: *standard errors* are given by

$$S_b = \frac{S_e}{S_x\sqrt{n-1}} \quad \text{and} \quad S_a = S_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2(n-1)}}$$

- Theorem:
  - $E(b) = \beta$ and $E(a) = \alpha$
  - b and a have normal distributions and the following have t distributions with $n-2$ degrees of freedom:

$$\frac{b-\beta}{S_b} \quad \text{and} \quad \frac{a-\alpha}{S_a}$$

- Definition/Theorem: The **coefficient of determination** (aka R^2) is the square of the correlation and tells you the percentage of the variability of y that is "explained" by x

$$R^2 = r^2 = 1 - \frac{\text{SSE}}{\text{SST}},$$

where $\text{SST} = \sum(y_i - \bar{y})^2 = (n-1)S_y^2$ is the **total sum of squares**.

- Gauss-Markov Theorem: b has a variance that is "smaller" than that of any other linear, unbiased estimator and is called the *best linear unbiased estimator* ("BLUE").

Grades Example (continued)

- The regression coefficient is

$$b = r\frac{S_y}{S_x} = .7243 \times \frac{11.19}{10.61} = .76$$

- The intercept is

$$a = \bar{y} - b\bar{x} = 80.44 - .76(62.35) = 32.80$$

- The standard error if the estimate is

$$S_e = 11.19\sqrt{(1 - .7243^2)\frac{83-1}{83-2}} = 7.765$$

- The standard error of b is

$$S_b = \frac{7.765}{10.61\sqrt{83-1}} = .0808$$

- A 95% confidence interval for β is roughly

$$.76 \pm 1.96 \times .0808$$

- We can test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$

$$P(b > .76) \approx P(Z > \frac{.76 - 0}{.0808}) = P(Z > 9.45) \approx 0$$

- The fraction of variation in the final scores "explained" by the midterm is

$$R^2 = .7243^2 = 52\%$$

Click Ball Point Pens Example

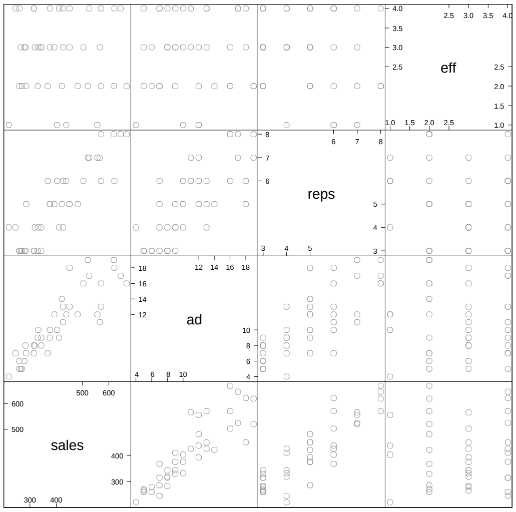
y_i Sales in territory $i = 1, \ldots, 40$

x_{i1} Advertising (number of TV spots) in territory i

x_{i2} Number of sales reps in territory i

x_{i3} Wholesaler efficiency index in territory i (4=outstanding, 3=good, 2=average, 1=poor)

```
> round(cor(click), 4)
       sales     ad    reps     eff
sales 1.0000 0.8802  0.8818  0.0019
ad    0.8802 1.0000  0.7763  0.0321
reps  0.8818 0.7763  1.0000 -0.1896
eff   0.0019 0.0321 -0.1896  1.0000

> plot(click)
```



This is a **scatterplot matrix**

Estimates From Click Ball Point Pens

```
                      Analysis of Variance

                     Sum of         Mean
Source        DF    Squares       Square    F Value    Prob>F

Model          1     463451       463451    130.644    0.0001
Error         38     134802         3547
C Total       39     598253

    Root MSE      59.56023      R-square       0.7747
    Dep Mean     411.28750      Adj R-sq       0.7687

                 Parameter Estimates

              Parameter    Standard    T for H0:
Variable  DF   Estimate       Error   Parameter=0   Prob>|T|

INTERCEP   1     135.43       25.91        5.228      0.0001
AD         1      25.31        2.21       11.430      0.0001
```

- Root MSE: $S_e = 59.56$, standard error of estimate
- R-square $= 0.7747$: fraction of variation explained by model
- Adj R-sq: 0.7687 adjusted for number of parameters

Interpretation of Output

- The estimated regression model is

$$\hat{y} = 135 + 25.3\texttt{ad}$$

What does 25.3 tell us? What about 135?

```
                    Parameter    Standard    T for H0:
Variable  DF    Estimate      Error   Parameter=0    Prob>|T|
INTERCEP   1      135.43      25.91         5.228      0.0001
AD         1       25.31       2.21        11.430      0.0001
```

- Standard errors are $S_a = 25.91$ and $S_b = 2.21$
- A 95% CI for β: $25.3 \pm 2.02 \times 2.21 \approx (20.8, 29.8)$
- To test the hypotheses (with Type I error rate .05)

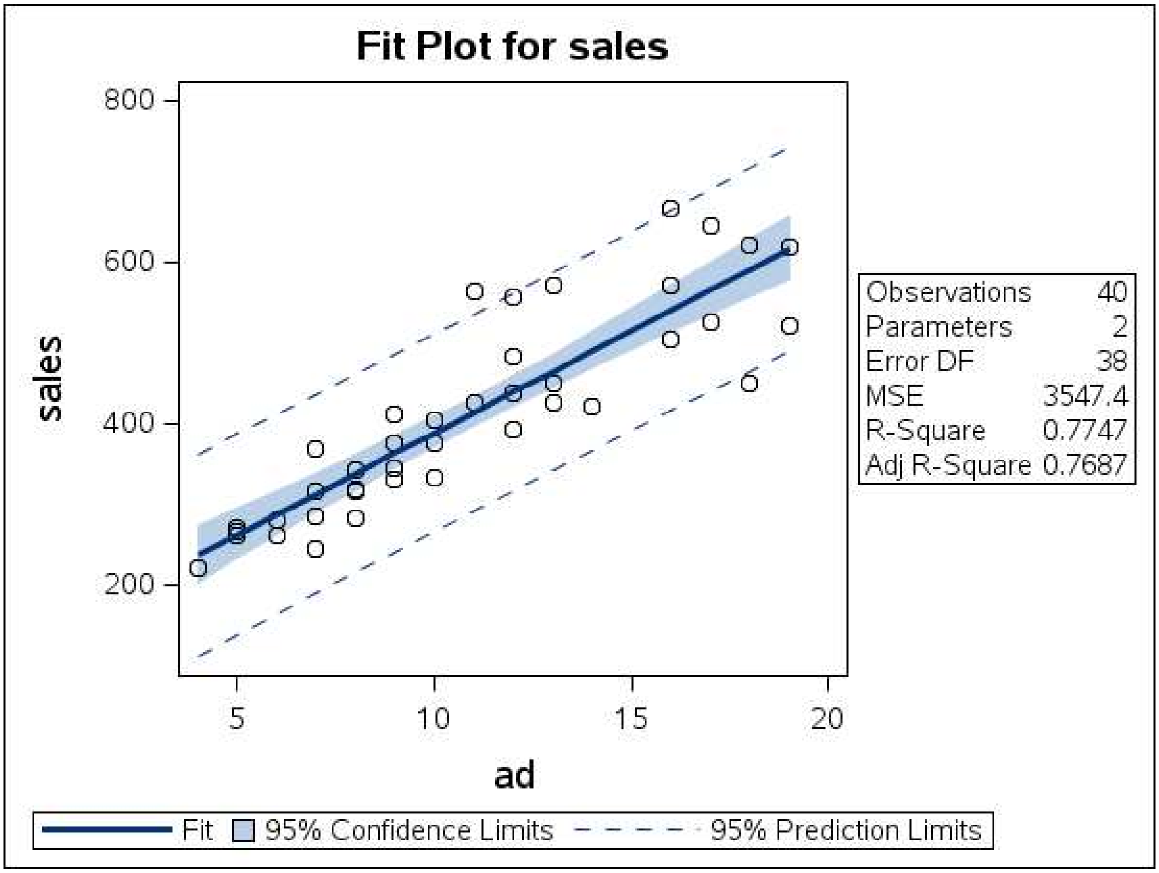
$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0,$$

P-value$= 0.0001 < .05$, so reject H_0 and conclude $\beta \neq 0$.

- The expected sales when advertising is 5 (spots) is

$$\hat{y} = 135 + 25.3 \times 5 = 261.97$$

Prediction and Confidence Intervals



- *Confidence interval for mean prediction* (shaded blue area): 95% confidence interval for $\hat{y} = a + bx$, i.e., indicates sampling variation of predicted values $E(Y|x)$.
- *Prediction interval* (dashed lines): indicates where the middle 95% of the distibution of Y for a given x_0 falls. If we knew parameters, it would be $(\alpha + \beta x_0) \pm 1.96\sigma_e$.

Additional Results

- The standard error of a new observation Y given x_0:

$$S_{Y|x_0} = \sqrt{S_e^2\left(1+\frac{1}{n}\right) + S_b^2(x_0-\bar{x})^2}$$

Example: find a 95% prediction interval for the mean sales when there are 5 ads. Hint: the 97.5 percentile of a t distribution with $40-2=38$ degrees of freedom is 2.024.

$$S_{Y|x_0} = \sqrt{3547\left(1+\frac{1}{40}\right) + 2.214^2(5-10.90)^2} = 61.70$$

$$262.0 \pm 2.024 \times 61.699 = (137.1, 386.9)$$

- The standard error of a predicted value $\hat{Y}$ given x_0

$$S_{\hat{Y}|x_0} = \sqrt{\frac{S_e^2}{n} + S_b^2(x_0-\bar{x})^2}$$

- Example: find a confidence interval for the mean sales when there are 5 ads.

$$S_{\hat{Y}|x_0} = \sqrt{\frac{3547}{40} + 2.214^2(5-10.90)^2} = 16.104$$

$$262.0 \pm 2.024 \times 16.104 = (229.4, 294.6)$$

Your Turn

The sales (y) were observed for various levels of advertising (x):

x	0	0	0	15	15	15	30	30	30	45	45	45	60	60	60	75	75	75
y	8	6	8	12	10	14	25	21	24	31	33	28	44	39	42	48	51	44

1. Find the equation of the regression line.
2. Graph the line on a scatterplot.
3. Compute and interpret the standard error of the estimate.
4. Compute and interpret the coefficient of determination.
5. Find a 99% CI for α
6. Find a 99% CI for β
7. Test at the 1% level if the slope differs from 0.
8. Estimate the mean
9. sales for ad=50.
10. Find a 99% CI for the mean amount when ad=50.
11. Find a 99% PI for the mean amount when ad=50.

```
dat=data.frame(
circ = c(2081995,1374858,1284613,1057536,970051,963069,828236,779259,768288,
  691771,663693,657015,645623,533384,528777,514702,492002,486426,
  443592,349182),
linerate=c(37.65,18.48,14.50,14.61,16.47,16.07,13.82,13.05,13.78,12.25,10.53,
  14.18,12.83,7.81,5.17,11.08,6.58,8.77,6.03,6.77),
row.names=c("WSJ","NY Daily News","USA Today","LA Times","NYT", "NY Post",
  "Philadelphia","Chi Tribune","Wash Post","SF Chronicle","Chi Sun Times",
  "Detroit News","Detroit Free Press","Long Island Newsday","KC Times",
  "Miami Herald","Cleveland","Milwaukee","Houston","Baltimore"))
dat$milline=1000000*dat$linerate/dat$circ
```

Does the Regression Make Sense?

- You can fit a regression line for any two variables, but should you?
  - Is the relationship linear?
  - Do other assumptions hold (when necessary)? We will discuss this in detail next quarter, e.g., homoscedasticity, errors independent of each other and of x, x fixed.
  - Does it make sense? Does y depend on x (linearly)?
- Rectangle example $r = .98$

$$\text{area} = 0.74 \times \text{perimeter} - 2.21$$

- Regression is a powerful tool, but it is not a substitute for understanding.

```
rectangle = data.frame(base=runif(100)+1, height=1+runif(100))
rectangle$p = 2*(rectangle$base+rectangle$height)
rectangle$a = rectangle$base*rectangle$height
fit = lm(a~p, rectangle)
summary(fit)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.20676    0.07278  -30.32   <2e-16 ***
p            0.74201    0.01199   61.89   <2e-16 ***

plot(rectangle$p, rectangle$a); abline(fit)
```

Does Education Pay?

- In 1993 a random sample of $n = 555$ California men age 25–29 were asked their education level (x) and income (y). Summary statistics are as follows
  - Education: $\bar{x} \approx 12.5$ years, $S_x \approx 4$ years
  - Income: $\bar{y} \approx \$21,500$, $S_y \approx \$16,000$
  - Correlation: $r \approx 0.35$
- Estimate the regression equation. Solution:

$$b = r \times \frac{S_y}{S_x} = 0.35 \times \frac{16,000}{4} = 1,400$$

$$a = 21,500 - 1,400 \times 12.5 = 4,000$$

$$\hat{y} = 1,400x + 4,000$$

- Interpret b. Solution: Associated with each extra year of education, there is an increase of \$1,400 in income, on the average.
- Find the coefficient of determination. Solution: $R^2 = 0.35^2 = 0.1225$
- Find the standard error of the estimate. Solution:

$$S_e = 16,000\sqrt{(1 - 0.35^2)\frac{555 - 1}{555 - 2}} = 15,002$$

- Find the standard error of the slope. Solution:

$$S_b = \frac{15,002}{4\sqrt{555 - 1}} = 159$$

Does Education Pay?

- Find a 95% confidence interval for β. Solution:

$$1,400 \pm 1.96 \times 159$$

- Test whether the slope is "significant." Solution: $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

$$P(b > 1,400) = P\left(Z > \frac{1,400 - 0}{159}\right) = P(Z > 8.79) \approx 0$$

  So $P = 2 \times 0 = 0$, we reject H_0, and conclude that education is associated with income.

- Suppose a random sample of men in this age range with 12 years of education were sent to college (for 4 years) to earn a degree. The slope suggests that their income would increase by $4 \times \$1,400 = \$5,600$, on the average. Would this be the case?

  Solution: not necessarily. We do not have a true experimental design here, but rather an observational study. In the original study those with 12 years of education and those with 16 years were likely different with respect to many different factors besides education, e.g., intelligence, ambition, family background. These factor confound the effect of education.

- Key point: *If you run an observational study*, the regression line only describes the data that you have. The line cannot be relied on for predicting the results of interventions. Maybe predictions are accurate, but maybe not.

Siegel Chapter 11
Homework Solutions

1. SPSS (b) $r = .985$; (c) Predicted Cost $= -1.0645 + 2.7527$Age; (d) 18,204; (e) 1,725; (f) .969; (g) yes, reject $H_0 : \beta = 0$; (h) find 95% CI and note $20 \notin [1.852721, 3.652655]$

2. (a) -957; (b) $10 \times 85 = 850$; (c) 11.26 hours; (d) 12.32%; (e) 87.68%.

3. SPSS Sixth/seventh edition: (b) $r = .6376$; (c) $R^2 = .4065$; (d) $\hat{y} = .1041 + .8853$ (Jan-Apr performance); (e) $\hat{y} = .7963$, $\hat{e} = -0.03049$; (f) 0.03961; (g) 0.4043; (h) $[-0.07080419, 1.8414498]$, (i) $P = .0647 > .05$ so we cannot reject $H_0 : \beta = 0$. Fifth edition: (b) $r = .465$; (c) .216; (d) $\hat{y} = 26.3746 + .7063$ (one-year performance); (e) 80.3; (f) 3.375; (g) .4751; (h) $[-0.39, 1.80]$

4. SPSS (a) $r = .961$; (b) $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, P=.000 so reject H_0; (c) $\hat{y} = 0.474 + 0.934$asset; (d) $H_0 : \beta = 1$ versus $H_1 : \beta \neq 1$, $P(b < 0.934|\beta = 1) = P(T_{11} < (0.934 - 1)/0.08088) = 0.2151$ which implies that P=.43 so do not reject H_0;

6. SPSS (b) no; (e) log failures $= -1.912 + 1.0719$ (log population; (f) [0.96, 1.18]; (g) reject; (h) do not reject

7. Fifth Edition: (b) Positive, linear, homoscedastic; (c) $r = .682$; (d) $R^2 = .465$; (e) $1 - .465 = .535$; (f) $\hat{y} = -0.302 + 1.146$dowjones, beta $= 1.146$; (g) $(.758, 1.534)$; (h) (d) $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, $P = .000$ so reject H_0; (i) (d) $H_0 : \beta = 1$ versus $H_1 : \beta \neq 1$, the confidence interval covers 1 so we cannot reject H_0. Fourth Edition: (c) $r = .4610$; (d) .21249; (e) 78.8%; (f) McDonald's $= 0.0008924 + 0.2115$Dow Jones; (g) [0.085, 0.338]; (h) reject; (i) reject

8. omit

9. (b) .7967; (c) yes, reject; (d) $P < .01$

12. $\hat{y} = 565.2$ and $\hat{e} = 34.8$.

14. Fifth edition: (a) $R^2 = .337$; (b) The rate of return for Core EqC is 2.38 above what we would expect for a fund with its expense ratio. This is the residual value for Core EqC. (c) $\hat{y} = 164.34 - 83.60$(expenseratio). (d) The regression coefficient is not significant, with $P = .131$.

16. Fifth edition: (a) $r = .994$ and the coefficient of *determination* is $R^2 = .988$; (c) $\hat{y} = 46.801 + .976$(Y1999); (d) $\hat{y} = 531.78$ and $\hat{e} = -52.78$.

17. Fifth edition: (a) 0.059 million dollars; (b) 0.010; (c) (0.035, 0.082); (d) Yes, $P = .000$.

18. (a) $r = 1$; (b) $\hat{y} = 16.24 + 383.697$wt; (d) $S_e = 5.2826$; (e) (370.870, 396.59); (f) $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, $P = 0$ so reject.

19. (a) $r = .5246$; (b) Predicted capacity = 3,296.124 + 0.18115(Existing Units); (c) $-1,118$; (d) $[-0.1125, 0.4748]$; (e) No, do not reject.

20. (a) figure 11.3.1 has $r \approx 0.9$; (b) figure 11.3.2 has $r \approx 0$; (c) figure 11.3.3 has $r \approx -0.5$; (d) figure 11.3.4 has $r \approx 0$.

21. (a) $r = 0.503$; (b) $\hat{y} = -21.621 + .977$price; (c) $\hat{y} = 112.28$ and $\hat{e} = -1.98$; (d) $(-2.109, 4.062)$; (e) $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, $P = .388$ so we cannot reject.

22. (a) $R^2 = .004$; (b) $\hat{y} = 152657 + 237.71$dols; (c) $\hat{y} = 158359$ and $\hat{e} = -15334$; (d) $S_e = 29,726$; (e) $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, $P = .871$ so we cannot reject.

23. (a) $r = .9987$; (b) Predicted sales = 1,393.83 + 24.1121 (list size); (c) 1,514; (d) .9973; (e) Yes, reject. It would be better to take logs of both sides before the regression.

24. (a) The correlation, $r = -0.487$, suggests a moderate negative association with longer maturities associated with lower yields. (b) Predicted Return = 0.0604 − 0.0054(Maturity), where a 10% return is expressed as 0.1. If returns are expressed in percentage points (so that 10 represents 10%) then multiply these coefficients by 100. (c) Predicted Return = 0.055, or 5.5%. (d) Omit. (e) No, the relationship is not significant.

25. (a) Predicted Production = 232.61 + 40.54 Workers. (b) $b = 40.54$. (d) Expected Production = 516.39, indicating that we expect production to be 516.39, on average, when 7 workers are assigned. Residual = −33.39, indicating that on this day (corresponding to the first pair in the data set) the actual production was 33.39 lower than expected. (e) $S_b = 7.3220$, indicating the uncertainty in the estimated slope b. (f) The 95% confidence interval extends from 24.2 to 56.9. (g) Yes there is a significant relationship.

26. Predicted value is −41.3.

27. Predicted value is 510.6.

28. Predicted value is $414.41, after estimating the least-squares prediction equation for these three data pairs, finding $a = -41.7553$ and $b = 3.0411$.

29. $r = 0.73$.

30. Predicted value = $21.89 based on $a = 10.4959$ and $b = 1.1393$.

31. You would expect $73,952, computed as the predicted value from the least squares equation which has $a = 43,095$ and $b = 308,571$ (alternatively, divide b by 100 if you express 10% as 10 instead of 0.1; the predicted value will still be the same).

32. nonlinear

33. unequal variability

34. (b) $r = 0.423$. There appears to be a weak positive relationship between earnings per share and stock price. (c) The apparent link between earnings and price is not statistically significant; we cannot reject the null hypothesis that there is no link. The estimated slope is $b = 7.5016$, with a standard error $S_b = 3.7907$. It is not significant. (d) Testing at the 10% level, we reject the null hypothesis of no link, and accept the research hypothesis that earning and price are related. There is a significant link at the 10% level ($P < 0.1$). (e) Although earnings and price appear to be positively related ($r = 0.4227$), this relationship is not statistically significant when tested at the conventional 5% level. We would expect to see such a large correlation more than 5% of the time even if there were no relation. However,

the relationship is significant when tested at the weaker 10% level. (f) Predicted Price = 14.5551 + (7.5016)(0.05) = \$14.93. (g) Omit. (h) Omit.

35. (b) $r = .930$; (c) $\hat{y} = 0.28218 + 0.00001581$ Circulation; (d) $P = 0$ reject; (e) 15.62, 0.85. The predicted value says that, for a newspaper with circulation 970,051, we expect an open-line rate of 15.62, on average. The residual values says that *The New York Times* has an open-line rate that exceeds this expected value by 0.85. In particular, the open-line rate for *The New York Times* is higher than you would expect for a newspaper with this circulation size.

36. (a) predicted open-line rate = 3.1097 + 0.00001166 circulation; (b) [20.58, 34.22], (c) *WSJ* an outlier

37. (b) $r = -0.125$; (c) .0156; (d) do not reject

38. (a) $r = 0.7535$; (b) $r = -0.3090$; (c) Temperature variability accounts for more of the variation in the defect rate because its R^2 is higher (56.77% compared to 9.55%). (d) The relationship between temperature variability and defect rate is significant. In fact, it is very highly significant ($P < 0.001$), as may be seen by testing the slope of a regression with X = temperature variability and Y = defect rate. The estimated slope is $b = 0.5881$ with standard error $S_b = 0.1372$. The relationship between stoppages and defect rate is not significant ($P > 0.05$), as may be seen by testing the slope of a regression with X = stoppages and Y = defect rate. The estimated slope is $b = -0.3911$ with standard error $S_b = 0.3217$. (e) The scatterplot, with its correlation $r = -0.3090$, shows a weak negative association, suggesting that when there are many stoppages fewer defects are produced. From the above testing results, we know that the relationship is not significant. There is considerable scatter and, although there is a slight downward tilt, it could be due to randomness alone. (f) The scatterplot, with its correlation $r = 0.7535$, shows a moderately strong positive association with some scatter, suggesting that more defects are produced when temperature variability is high. From the previous testing results, we know that this relationship is very highly significant. The upward tilt to the right is unlikely to be due to randomness alone. (g) The analyses carried out here suggest that the defect rate is largely due to the temperature fluctuations in the production process. We may attribute 56.77% of the variation in the defect rate to variation in temperature. Only 9.55% of this variability is due to daily assembly line stoppages. While correlation does not by itself prove causation, these conclusions appear reasonable for our production process. To remedy the problem it will be necessary to study the system controlling the temperature to find out the cause of these fluctuations, and then to correct them. 33.7% of the determining factors have not yet been accounted for. If the changes made in the temperature control system do not bring the defect rate to an acceptable level, it will be necessary to search for additional causes of this problem.

Outline of Lecture

- What is causality?
- Experimental designs
- Is the design good? (Internal and external validity)
- Analyzing the results

"To this day, I believe that the most important tool available to the direct marketer is the ability to test, a form of pragmatic research that doesn't just tell you how a consumer *might* react to an ad but how consumers really *do* react. Direct marketing makes advertising totally measurable and accountable" (Wunderman, 1996, p. 56).

The Concept of Causality; Inferring Causality

1. *Causality* — A change in one variable will produce a change in another variable.
2. Three types of evidence to *infer* causality
   (a) Concomitant variation/association between cause and effect: the extent to which the cause and effect occur together.
   (b) Time order of occurrence of variables: cause *must* precede effect.
   (c) Elimination of other possible causal factors, e.g., spurious association.

Nature of Experimentation

- *Experimentation* — The manipulation of one or more variables (cause) by the experimenter in such a way that its effect on one or more other variables can be measured
- *Independent variable* — (X's) Variables being manipulated
- *Dependent (or criterion) variables* — (Y's) Variables that will reflect the impact of the independent variable (e.g., response to offer, sales, purchase intent)
- *Treatment group* (TG) — Portion of sample (or population) exposed to the manipulation of the independent variable
- *Control group* (CG) — Independent variable is unchanged
- *Pre-Experimental (observational) study*: subjects assign themselves to different groups (if any) — researcher as little control over when and to whom treatment is given
- *True experimental study* (*randomized controlled experiment*): researcher assigns subjects to treatment and control groups at random

Notation for Representing Experimental Designs

- X — exposure of an individual, group, or other entity to the experimental treatment.
- O — observation or measurement of the test unit.
- R — randomized assignment

Movement through time is represented by a horizontal arrangement of X's and O's from left to right. Simultaneous exposure or measurement by a vertical arrangement.

Your Turn: (Churchill, page 172, # 9) A new manufacturer of women's cosmetics was planning to retail the firm's products through mail order. The firm's management was considering the use of direct mail ads to stimulate sales of their products. Prior to committing themselves to advertising through direct mail, management conducted an experiment. A random sample of 1000 housewives was selected from Memphis, Tenn. The sample was divided into two groups with each prospect being assigned randomly to one of the two groups. Direct mail ads were sent twice over a period of one month to prospects of one of the groups. Two weeks later, both the groups were mailed the company's catalog of cosmetics. Sales to each group were monitored.

What makes a good test?

- Factors other than the experimental variable that affect the dependent variable are called *nuisance*, *confounding*, or *extraneous* factors/variables.
- When nuisance factors are not properly controlled, they are said to *confound* the effects of the experimental variable.
- *Internal validity* — The ability of the experiment to *unambiguously show a cause and effect relationship*, i.e., to what extent can we attribute the effect that was observed to the experimental variable and not other (confounding) factors? See THIS MESS.
- *External validity* — The extent to which the results of the experiment can be *generalized* to other people, settings, and time, i.e., will we get similar results in other settings?

There is a managerial trade-off between internal and external validity

Pre-Experimental Designs

- *After-only design (one-shot case study)*

$$X \qquad\qquad O$$

  - Provides no basis for comparing what happened in the presence of X with what happened when X was absent. Cannot infer causality.
  - Serious threats to IV: history, maturation, selection, mortality
  - Example: *recall study*. Suppose that telephone interviews are conducted until 200 respondents are identified who watched a particular TV show last night. Respondents are asked if they remember seeing a commercial for a product in the category of interest (*unaided recall*). If they can recall the commercial, they are asked what specific copy execution points of the advertisement they remember. If the commercial they remember is not for the test brand, the respondent is asked whether he or she recalls the commercial for the brand being tested (*aided recall*). Results are compared with Burke category norm scores with a single-sample t test ($H_0 : \mu = \text{norm}$).

Pre-Experimental Designs

- *One-group Pretest-Posttest Design (Before-after)*

  $$O_1 \qquad X \qquad O_2$$

  - The result of interest: $\hat{D} = O_2 - O_1$
  - Analysis: paired-sample t test or GLM (general linear model)
  - Threats to IV: history, maturation, pre-measurement, placebo effect.
  - Example: *copy testing*. Subjects are recruited to come to, e.g., a hotel, under the premise that they will evaluate a new TV program. Before seeing the TV program, they complete a survey (*pre-measure*) about their attitudes, purchase intent and behavior of various brands. Next, they watch the TV program with advertisements (*treatment*). After the program and advertisements, they complete a second survey (*post measure*).
  - Example: *"Heavy-up" advertising*. Sales of a product are monitored for one week (*pre-measure*). The advertising budget is doubled (*treatment*). Sales are monitored for one week after the advertising increase (*post-measure*).

Threats to Internal Validity

- *History* — Refers to any variables or events, other than the one(s) manipulated by the experimenter, that occur between the pre- and post-measures and affect the value of the dependent variable.
  - Heavy-up ad: economy declines between O_1 and O_2, e.g., plant closing, layoffs, recession
  - Heavy-up ad: competitor did something, e.g., cut price, increased promotion
- *Maturation* — Represents the biological or psychological processes that systematically vary with passage of time, independent of the experimental variable.
  - Copy test: people become tired, hungry, grumpy
  - Heavy-up ad: stores change, e.g., physical layout or decor
- *Experimental mortality* — Differential loss of respondents from different groups.

Threats to Internal Validity

- *Pre-measurement (or main testing) effect* — The taking of a measurement has a direct effect on performance in a subsequent measurement (O_1 affects O_2).
  - Suppose that copy testing showed no difference between O_1 and O_2. We might conclude that the advertising was ineffective. An alternative explanation is that subjects sought to maintain consistency between pre- and post-measures.
- *Interaction (or interactive testing) effect* — When a pre-measure changes the respondent's sensitivity or responsiveness to the independent variable(s). This is only a threat to external validity.
  - In copy testing, perhaps people pay more attention to an advertisement because they were asked about the brand during the pre-measure.
- *Instrument Variation* — Refers to changes in the measuring instrument that might account for differences in the measurement, e.g., change in scales or interviewer.

Threats to Internal Validity

- *Placebo effect* — respondents acts differently because they know that they are being exposed to the treatment
  - See Tufoil example in Churchill, page 141
  - In the early 1960s a study was carried out to investigate the efficacy of a technique known as gastric freezing to treat ulcer patients. The treatment required patients to swallow a balloon, which was positioned in the stomach. A coolant was then passed through the balloon. It was reported that the patients given this treatment experience relief of the symptoms, and the treatment was recommended for ulcer patients. Later, a randomized controlled experiment was conducted, where the control group also swallowed a balloon without the coolant. There was no difference between the treatment and control groups.
  - If both the subject and the investigator are kept unaware of the treatment, the study is called a *double blind* study.

Pre-Experimental Designs

- *Static-Group Comparison*

TG:	X	O_1
CG:		O_2

  - The result of interest: $\hat{D} = O_1 - O_2$
  - Analysis: independent-sample t test or GLM/ANCOVA
  - Threats to IV: **selection**, maturation / mortality (if treatment unpleasant),
  - Example: *Department store patronage.* Two groups of respondents are recruited *on the basis of convenience*, e.g., one group in the morning and the other in the afternoon. One group is shown a TV commercial about a department store. Both groups are asked about their attitudes toward the store.
  - Example: *Magazine experiences.* Experiences such as “it’s my personal timeout” were measured for readers of 100 magazines. Respondents were shown an ad for bottled water (*treatment*) and asked standard measures of their attitude towards the ad (*post-measure*). After *controlling* for bottled water consumption and attitude towards ads in general (through a regression analysis), we found a highly significant positive relationship between “it’s my personal timeout” and the attitude towards the water ad across magazines.

Threats to Internal Validity

- *Selection bias* — When the groups formed for the purposes of the experiment are initially unequal with respect to the dependent variable or in the propensity to respond to the independent variable.

  Remedies:

  - *Randomization* — assign subjects to treatment and control group using a random procedure.
  - *Matching* — match treatment and control groups with respect to variables which you suspect influence response (used with small sample sizes, e.g., stores).
    - Form *blocks* of units that are similar on key nuisance factors
    - *Randomly* assign units within a block to treatment and control groups
    - Blocking is closely related to stratification. Blocking is used in experiments and stratification in surveys.
  - Control for other causal factors with regression models.
  - *Propensity score models* for observational studies. Find matched "twin(s)" for each treated case that is as similar as possible prior to self-selection into treatment

Threats to Internal Validity

- *Statistical regression* — When individuals are assigned to groups because of their scores on some measurement, such as initial attitude towards a brand. (Also called the *regression effect*)
  - "In direct marketing, you never do as well in roll-out as you do in test."
  - Every year, baseball's major leagues honor their outstanding first-year players with the title "Rookie of the Year." From 1949 to 1987, the overall batting average for the Rookies of the Year was 0.285, far above the major league average of 0.257. However, Rookies of the Year don't do so well in their second year: their overall second-season batting average was on 0.272. Baseball writers call this "sophomore slump," their explanation being that star players get distracted by outside activities like product endorsements and television appearances. Do you agree?

Threats to External Validity

All the previous threats to internal validity are also threats to external validity. In addition, there are the following:

- *Surrogate situation* — Occurs when the environment, the population sampled, and/or the treatments are different from those that will be encountered in the actual situation, e.g., copy testing . . . forced attention.
- *Measurement timing* — When pre- or post-measurements are made at an inappropriate time to indicate the effect of the experimental treatment, e.g., effect of temporary price cut on forward buying.

True Experimental Designs

- *After-only with control group* sometimes called a *completely randomized design* or *randomized controlled experiment*

TG:	(R)	X	O_1
CG:	(R)		O_2

  - The result of interest: $\hat{D} = O_1 - O_2$
  - Analysis: independent-sample t test or GLM/ANCOVA
  - Threats to IV: none
  - Large between-unit variation implies that larger samples will be required to detect differences compared with the before-after with control design
  - Example: Memphis direct marketing example from Churchill
  - Example: department store (cont). Suppose the department store recruits subjects and *randomly* assigns them to treatment and control groups. Both groups are shown a TV program. The treatment group sees an advertisement for the store (*treatment*) while the control group doesn't (*placebo*). Both groups are asked about their attitude towards the store (*post-measure*).

True Experimental Designs

- *Before-after with control group*

TG:	(R)	O_1	X	O_2
CG:	(R)	O_3		O_4

  - The result of interest: $\hat{D} = (O_2 - O_1) - (O_4 - O_3)$
  - Analysis: Compute differences between post- and pre-measures and compare with independent sample t-test or GLM/ANCOVA. More generally, use repeated measures models.
  - Threats to IV: Interactive testing effect
  - Example: department store (cont). Suppose the department store recruits subjects and *randomly* assigns them to treatment and control groups. Both groups are given a questionnaire and shown a TV program. The treatment group sees an advertisement for the store (*treatment*) and the control group sees other unrelated advertisements (*placebo*). Both groups are asked about their attitude towards the store (*post-measure*).

True Experimental Designs

- *Randomized Block Design*:
  - Form *blocks* of units that are similar
  - Randomly assign treatments within each block
- When between-subject variation is large, blocks can reduce sample size requirements
- Example: department store (cont). Suppose that the researcher believes that store patronage will also affect the dependent variable (attitude towards store). The researcher forms three blocks: non-users, occasional users and frequent users. Subject from each block are recruited. The subjects from each block are *randomly* assigned to treatment and control groups. The rest of the experiment is the same as the after-only with control or before-after with control experiments.
- Analysis: GLM/ANOVA/ANCOVA

True Experimental Designs

- *Separate-Sample Before-After Design*

  In many circumstances — such as a national introduction of a new product, service, or implementation of federal legislation — it is impossible to find an unaffected subgroup to use for a control. In such cases a separate random sample is sometimes interviewed for the before measure:

  $$\begin{array}{lllll} \text{CG:} & \text{(R)} & O_1 & (X) & \\ \text{TG:} & \text{(R)} & & X & O_2 \end{array}$$

  The bracket (X) indicates that all groups are exposed to the treatment but only some are interviewed after the treatment.

  - The result of interest: $\hat{D} = O_2 - O_1$
  - Analysis: independent-sample t test on differences between pre and post.
  - Threats to IV: history, maturation
  - Example problem: An outdoor advertising firm wanted to test the effectiveness of a billboard campaign to promote awareness about metric conversions in Quincy, Illinois. The campaign consisted of 10 billboards with the message "1 ounce = 28.3 grams." The boards were located at various locations within the Quincy area. The firm wanted to know whether over a 3-month period the billboard message would have any significant impact on the public's awareness of this particular metric conversion.

Quasi-Experimental Designs

- *Quasi-Experimental Designs*: the researcher is unable to achieve complete control over the scheduling of the treatments or cannot randomly assign respondents to experimental treatment conditions.
- *Before-After with Control Quasi Design*

TG:	O_1	X	O_2
CG:	O_3		O_4

  - The result of interest: $\hat{D} = (O_2 - O_1) - (O_4 - O_3)$
  - Analysis: Indepdendent-sample t test on differences
  - Somtimes, treatment and control *matched*: units assigned to treatment and control based on key factors.
  - Threats of validity: selection, interactive testing effects
  - Example: smoking study

Quasi-Experimental Design

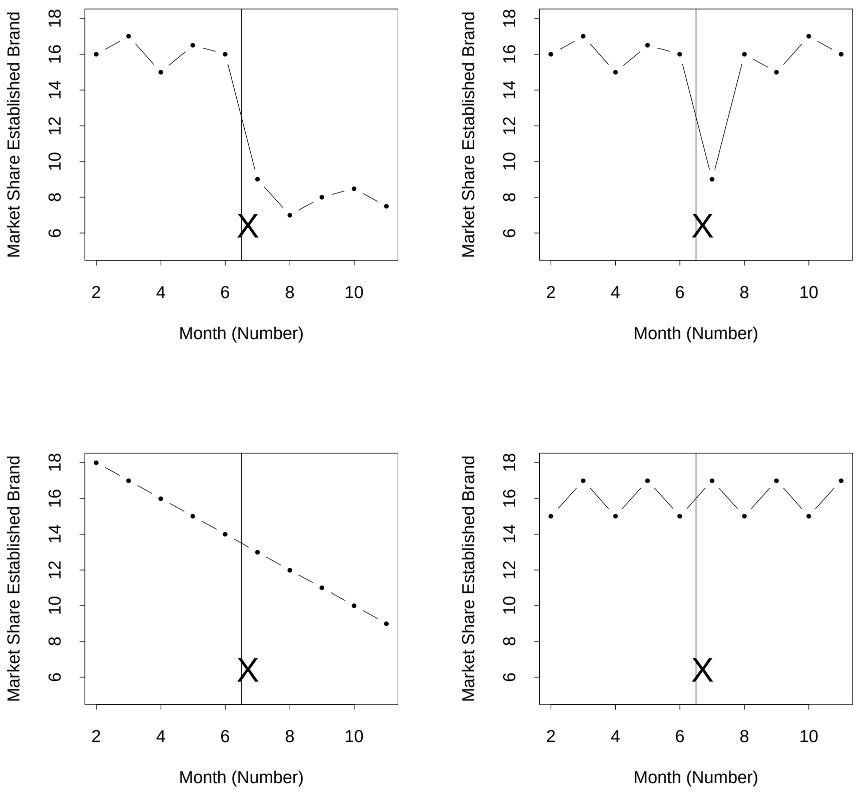
- *Time series*

$$O_1 \quad O_2 \quad O_3 \quad O_4 \quad X \quad O_5 \quad O_6 \quad O_7 \quad O_8$$

  – Threats to IV: history, instrumentation, interactive test, mortality

  – Analysis: Regression and time series models

  – Example: chewing gum. Suppose we do monthly probability surveys of chewing gum users measuring brand awareness, consumer perceptions and usage. In June we launch a new "diet" chewing gum. We continue doing the monthly surveys. Consider four possible outcomes:



Experimental Environments: Laboratory versus Field

	Laboratory	**Field**
Validity	High internal Low external	High external Low internal
Time and Cost	Low	High
Exposure to Competition	Low	High
Nature of the Manipulation	Realistic in a contrived setting	Difficult to duplicate in the lab

Hybrid Approaches: Simulated test markets

Short Cases

1. To test the effect of the first Obama-McCain Presidential debate in 2008, researchers picked 1000 names at random from the Chicago telephone directory. An attempt was made to telephone all 1000 one day before the debate and ask their voting intentions. Only 512 were actually questioned; 488 either were not at home or refused to participate. One day after the debate, 387 of the 512 were reached and again asked their voting intentions. In addition, they were asked if they had seen the broadcast. The before-after change of the viewers and non-viewers were compared.

2. The Schwerin commercial testing technique used the following procedure. About 1,000 individuals were each sent 4 to 8 invitations to participate in a session to improve TV programs. They were told that they had a chance to win a prize if they came. Usually 300–400 responded and saw a pilot television program in the Schwerin movie theater in New York City (or, sometimes, other cities). Before the program each respondent was given a list of brands in each product class involved in the test. They were asked to check the brand they wished to receive if they were the winner of a drawing which was then held. During the program the audience saw the commercial being tested. Immediately after the program the audience members wrote down what they recalled from the commercials they saw. Then they were given another list of the brands and were asked to check the brand they now wanted; another drawing was held.

Short Cases

3 In 1995 one of the largest firms of stock brokers in the United States noted that its option trading business was growing extremely rapidly. This growth had occurred even though the firm had never had an active training program in options for its account executives. The top management of the company believed that by instituting a comprehensive training program in options trading for its account executives, their sales performance could be further improved. With this goal in mind, the firm's training department put together a comprehensive training program in options utilizing a number of leading business and academic experts. The 50 account executives in the New York area who had done the greatest volume of options trading in the first six months of 1995 were selected to participate in the course. Before exposing further large groups of account executives to the new program, the firm decided to monitor the performance of the employees who had participated in the course. In the first six months of 1996 the average options volume of these account executives rose by 5% over the comparable period in 1995. The average option volume of the account executives in the New York area who had not participated in the program was 29% higher in the first six months of 1996 as compared to the comparable period in 1995.

Experimental Research Practice Questions

1. Charlie Sharp is the national sales manager of Hitech Inc. Charlie recently hypothesized that "Hitech's increase in sales is due to the new sales personnel that we recruited from the vocational school over the last several years. Sales of the new salespeople are up substantially, whereas sales for longer-term salespeople have not increased." Identify the casual factor X and the effect factor Y in the preceding statement.

2. To gather support for his hypothesis, Charlie asked the research department of Hitech to investigate the sales of each of the company's salespeople. Using criteria supplied by management, the department categorized territory sales changes as "increased substantially," "increased marginally," or "no increase." Consider the following table, in which 260 sales personnel have been classified as old or new:

	Territory Sales Change			
Salesperson Assigned	**Increased Substantially**	**Increased Marginally**	**No Increase**	**Total**
New	75	30	5	110
Old	50	40	60	150

   (a) Does this table provide evidence of concomitant variation? Justify your answer.

   (b) What conclusions can be drawn about the relationship between X and Y on the basis of the preceding table?

   (c) Draw a diagram for this experiment.

3. Consider the following statement: "The increase in repeat-purchase frequency is due to retailers' decisions to stock our product in the gourmet food section of supermarkets during the last nine months. Repeat purchases from the gourmet section are up as much as 50 percent from our previous store location." Identify the causal factor X and the effect factor Y in the preceding statement.

4. The research department in the preceding question investigated the change in repeat purchase for each store location. Using criteria supplied by management, the department categorized repeat-purchase frequency changes as *increased substantially*, *increased marginally*, or *no increase*. Consider the following table, in which 624 store locations have been classified as *old* or *gourmet*:

	Repeat-Purchase Frequency			
Store Location	**Increased Substantially**	**Increased Marginally**	**No Increase**	**Total**
Gourmet	180	72	12	264
Old	120	96	144	360

   (a) Does this table provide evidence of concomitant variation? Justify your answer.

   (b) What conclusions can be drawn about the relationship between X and Y on the basis of the preceding table?

(c) Draw a diagram for this experiment.

5. A leading manufacturer of frozen food products decided to test the effectiveness of an in-store display. Four large supermarkets, located near the company's main office, were selected for the experiment. This display was set up in two of the stores, and sales were monitored for a period of two weeks. The sales volume for the frozen food products increased 2 percent more in the stores that used the in-store displays than in the stores that did not use the displays.

6. A branch of Alcoholics Anonymous wanted to test consumer attitudes toward an anti-drinking advertisement they plan to run in their area. Two random samples of respondents in Piscataway, New Jersey, were selected for the experiment. Personal interviews relating to consumer attitudes toward alcoholism were conducted with both samples. One of the samples was shown the anti-drinking advertisement, and following this, personal interviews were conducted with both samples to examine consumer attitudes toward alcoholism. Identify the design and diagram it. Discuss the threats to internal and external validity for the design. Can the design be improved? If so, how?

7. A manufacturer of a line of office equipment, based in Houston, Texas, marketed its products in the southwest United States. The region consisted of 30 geographic divisions, each headed by a divisional manger who had a staff of salespeople. The firm's management wanted to test the effectiveness of a new sales training program in which the sales personnel in five of the divisions typically participated. The divisional mangers of these five divisions were instructed to monitor sales for each salesperson for each of the five months before and after the training program. The results were to be sent to the vice-president of sales in Houston. Identify the design and diagram it. Discuss the threats to internal and external validity for the design. Can the design be improved? If so, how?

Experimental Research
Homework Solutions

1. Causal factor (X): new sales personnel recruited from the vocational school Effect factor (Y): change in sales volume.

2. (a) Do a Chi-square test of independence and find that $P = 0.000$ so that you can reject the null hypothesis of independence. About 68% of the new sales personnel have substantially increased sales, while only 33% of the old sales personnel have substantially increased sales. (b) Although there is support for concomitant variation and time order, all that can be said is that the association makes the hypothesis more tenable. It certainly does not prove it. There may be a number of other factors accounting for the sales increase, e.g., the new sales personnel were assigned to more lucrative territories in terms of untapped potential. For example, suppose that the new people were assigned to territories of retiring salespeople. Suppose further that those retiring had not pursued new leads as aggressively as others. Therefore, the new people would have more untapped leads than the "old" salespeople. This is an example of the regression effect, where territories are assigned based on a pre-measure, with those territories that are not doing well getting a new sales person. (c) This is a *before-after with control quasi design* as shown on page 278.

3. Causal factor (X): store location. Effect factor (Y): increase in repeat-purchase frequency.

4. Do a Chi-square test of independence. The table provides evidence of concomitant variation. Substantial increases in repeat purchase frequency occur more often when the product is located in the gourmet food section. About 68% of the gourmet locations have substantially increased repeat purchases, while only 33% of the old store locations have substantially increased repeat purchases of the product. Although there is supporting evidence of concomitant variation, all that can be said is that the association makes the hypothesis more tenable. It certainly does not prove it. There may be a number of other factors accounting for the repeat purchase increase, e.g., the new store locations are also provided with smaller packages or higher quality labels.

5. This is a *before-after with control quasi design* as shown on page 278. Notice the word "increase," indicating that the post sales are compared with the pre sales. The stores are not assigned randomly to treatment and control, which could cause a selection threat to internal validity. External validity is affected by the fact that the experimental stores might not be representative of the total population of stores in the chain (they are close to the headquarters).

6. This is an example of the before-after with control group true experimental design. Extraneous sources of error such as history, maturation, main-testing effect, statistical regression, and instrument variation would affect both groups equally. However, external validity could be affected by experimental mortality (e.g., if all the young people drop out of the study, the conclusions don't necessarily apply to young people any more). In addition, internal validity is affected by the interactive effect of testing. The subjects might, after giving their attitudes about alcoholism, react quite differently to an anti-drinking advertisement than if they had not previously been interviewed. This possibility calls into question the generalizability of any observed result.

7. This is an example of a quasi-experimental design, in particular a time-series experiment. The group of salespeople is observed over a period of time, then the training program is introduced, and then the test units are again observed. See page 279. Maturation, instrument variation, statistical regression, and the main-testing effect are not problems. Selection affects external validity, but not internal validity. History can affect internal validity; for example, the competitor could do something, the economy could change, etc.

Measuring the Effects of Marketing and Showing ROI

By Edward Malthouse

Measuring the effects of marketing efforts and their ROI is one of the most important tasks that marketing managers face. In the past, marketing and advertising have been viewed as an expense by the financial officers of many firms, but there is a shift towards viewing them as an investment in customers with a quantifiable financial return. Consequently, managers are increasingly being required to show a return on marketing activities. At the same time, the quality of data and the ability to measure the outcomes is improving. This whitepaper will discuss good ways of measuring and proving such outcomes.

Use Controlled Tests with Matching/Blocking

The best way to measure the effect of some marketing effort is usually to use some sort of controlled test. The idea is to give some customers—the *treatment group*—the new marketing and give others—the *control group*—the existing marketing. We want the treatment and control groups to be identical in all ways except that one gets the *treatment* (the new marketing) and the other does not. In this way we can isolate the effects of the new marketing and get a true reading of its effectiveness. If the groups are not identical then the results from the test are questionable. For example, suppose that those assigned to the treatment group were better customers to begin with than those in the control group. Then any observed difference in sales between the two groups could be due to the fact that those receiving the new marketing were better to begin with, rather than the new marketing being more effective. This is called a *selection bias*, which is said to *confound* the effects of the new marketing. In this example customer quality is called a *confounding variable* because it is related to both the treatment (better customers received the treatment) and to the outcome (sales).

The problem of designing a good test is even more complicated because not all customers are the same. Some customers are better than others. Marketers call this *heterogeneity*. There are several ways to address heterogeneity. The first approach is to *randomly* assign customers to the treatment and control groups. While randomization is a good basic strategy, one problem with it is that the treatment and control groups may not be equivalent because of bad luck. An example will make this clear. Suppose that customers vary in purchase levels, with some great customers and some weaker customers. If we randomly assign customers to treatment and control groups, there is some chance that too many great customers will be assigned to one group and too many weaker ones to the other. This is just like flipping a coin 10 times—we will probably not get exactly 5 heads and 5 tails. When the groups are not equal we have the same problem that was described earlier, where the effects of the treatment confound customer quality. Just as the percentage of heads will get closer to 50% as we flip the coin more times, the chances of having a disproportionate fraction of great customers in one of the two groups will decrease as the sample size increases, but larger samples are more expensive.

A second way to address heterogeneity is called *blocking*. The idea is to form groups (blocks) using the confounding variables. Continuing the example, suppose that we could identify great, OK and weaker customers before we run the test, for example by looking at their previous purchase history. We want to avoid having a disproportionate number of great customers getting the treatment, and so we could randomly assign half of the great customers to receive the new marketing and the other half to the control group. Likewise, we would randomly assign half the OK customers to treatment and the other half to control, and do the same for weaker customers. In doing so we will have insured that customer quality cannot confound the treatment. We also reduce the cost of our test because we remove differences in customer quality from the test. This process is summarized by those who design experiments with saying, "**block what you can, randomize what you cant**." In other words, we want to block on confounding variables that we can control, and then randomly assign customers within each block to treatment and control.

The last question we will address is how to form blocks. We are worried about customer characteristics that relate to the criterion we are testing (usually sales) and confound the treatment (the marketing program we are testing). Across many businesses in many different categories there are certain variables that are almost universally related to sales. These variables begin with "RFM:" recency, frequency and monetary value. Recency is the time since the most recent purchase, frequency is the number of times that a customer has purchased in the past, and monetary value is the total amount that a customer has spent in some past period (like one or two years, depending on the purchase cycle). When physical stores are involved then distance to the store is also an important variable. Good blocks can be formed by crossing these variables.

The Problem with Historical Controls

The procedure discussed above involves having two groups of customers, with the treatment group getting the new marketing and the control group getting the "business-as-usual" existing marketing. Managers will often object to withholding the new marketing from the control group. "If the new marketing really works better, then why would I want to lose money by not giving it to all customers?" They will often suggest comparing sales under the new marketing programs with sales under the old programs from a previous period. In other words, monitor sales with the old marketing during the *pre-period*, implement the new programs, then monitor sales on the same customers during the *post-period.* The difference in sales between the post- and pre-periods is supposed to give the effect of the new marketing.

There are many potential problems with this approach, usually due to "something else" happening between the pre- and post-periods. Suppose, for example, that the main competitor was running a price promotion during the pre period, and changed the regular price during the post period. When the competitor is charging a lower price, our sales probably go down, and when the competitors price increases, our sales go up. Notice that the competitors price is a confound because the change in sales between the two periods is affected by it rather than only our new marketing program. Likewise, changes to the competitors marketing, the way products are displayed in stores, etc. could also be confounds. In addition to competitive effects, there sales can also be highly seasonal (e.g., the sale of steak sauce spikes around the 4th of July and in summer) or affected by other external factors such as weather (e.g., beer sales go up when it is hot out and hot chocolate sells better in winter).

Sometimes we can adjust for these confounding variables with sophisticated models, but such models must make assumptions that may not be true. A failsafe way of measuring effectiveness is to use the controlled tests described above.

Measure Incremental Spend

The last question we will address is what to measure. It is always important to have the right basis of comparison and quantify how the new marketing improves over the status quo. For example, suppose we send out a promotion that drives people into physical stores. What is the proper way to quantify the effects of such a promotion? Monitoring the sales of those who received the promotion is not enough because some of them would have made purchases anyway, e.g., due to their previous habits of shopping with you, exposure to mass advertising, word of mouth, etc. A better way to isolate the effect of the promotion is to compare the sales to those receiving it to the sales of a control group. It is the difference in sales (or percentage increase) between the groups that gives the true effect of the promotion. If this difference is less than the cost of the promotion, then the ROI is negative and the promotion is not effective.

Practice Final Exam Questions

1. A survey of Americans planning long summer vacations revealed a mean planned expenditure of \$1,076. Assume that this mean is based on a random sample of 500 Americans who were planning long summer vacations and that the sample standard deviation was \$400. Find the standard error, margin of error, and a 95% confidence interval for the mean. *Answer: 17.89; 35.06; 1041 to 1111.*

2. A survey of your customers shows, to your surprise, that 42 out of 200 randomly selected customers were not satisfied with after-sale support and service. You estimate that 42/200 = 21% of all your customers are dissatisfied with these services.

   (a) Compute the standard error of the estimate.

   (b) Compute a 99% confidence interval the percentage of all your customers who are dissatisfied.

   Answer: (a) .02880; (b) $.21 \pm .07419$*.*

3. A manufacturer of detergent claims that the mean weight of a particular box of detergent is 3.25 pounds. A random sample of 64 boxes reveals a sample average of 3.238 pounds and a sample standard deviation of 0.117 pounds.

   (a) Using the .01 level of significance, is there evidence that the mean weight of the boxes is less than 3.25 pounds (one-sided test)? To receive full credit, state the null and alternative hypotheses, compute the P-value, and state your conclusion.

   (b) Now suppose that you wish to test a two-sided alternative. Compute the P-value for a two-sided alternative and state whether or not the result is significant at the .01 level.

   (c) Construct a 95% confidence interval for the mean.

   Answer: (a) $s_{\bar{x}} = .014625$*, P-value = .2061; (b) .4122; (c)* $3.238 \pm .02867$*.*

4. According to Census data, in 1950 the population of the U.S. amounted to 151.3 million persons, and 13.4% of them were living in the West. In 1980, the population was 226.5 million, and 19.1% of them were living in the West. We wish to study whether the percentage of people who live in the West has changed between 1950 and 1980. Is the difference in percentages practically significant? statistically significant? Or do these questions even make sense? Explain briefly. (Hint: try to identify the population, sample, parameter, and statistic) *Answer: Both of the percentages are from Censuses. Hypothesis tests make no sense when you have the entire population in your "batch" of data.*

5. A large shipment of air filters is received by Joe's Auto Supply Company. The air filters are to be sampled to estimate the proportion that are unusable. From past experience the proportion of unusable air filters has never been more than 10%. How large a random sample should be taken to estimate the true proportion of unusable air filters to within ± 7% with 99% confidence? For now, do not use the finite population correction. *Answer: (a) 122.*

6. T.C. Resort Properties has three hotels. They conducted a survey of dissatisfied recent guests and asked why they would not return. A cross tabulation from SPSS of responses by hotel is given below. The first number in each cell gives the observed count and the second number gives the adjusted standardized residual.

	Golden Palm	Palm Royale	Palm Princess	Total
Price	23 / −2.6	7 / −1.9	37 / 4.3	67
Location	39 / 3.4	13 / 1.0	8 / −4.3	60
Room	13 / −0.6	5 / −0.2	13 / 0.8	31
Other	13 / −0.3	8 / 1.5	8 / −0.9	29
Total	88	33	66	187

	Value	df	Sig
Pearson Chi-Square	27.410	6	.0000

(a) Conduct a chi-squared test of independence on the table. State the null and alternative hypotheses, the P-value, and your decision.

(b) If hotel and reason were independent of each other, how many people out of a sample of 187 would you expect to have stayed at the Golden Palm and checked “Room?”

(c) What reason appears to be the problem for the Golden Palm? Why?

(d) What reason appears to be the problem for the Palm Princess? Why?

Answer: (a) H_0 : rows and columns are independent versus H_1: rows and columns are dependent. P-value = .0000, reject; (b) 14.59; (c) Location, residual is 3.4; (d) Price, residual is 4.3.

7. In order to measure the effect of a storewide sales campaign on non-sale items, the research director of a national supermarket chain took a random sample of 13 pairs of stores that were matched according to average weekly sales volume. One store of each pair (the experimental group) was exposed to the sales campaign, and the other member of the pair was not. The results were measured over a weekly period. SPSS output is given below.

```
+---------------------+---------+------+-----------------+
|                     |   Mean  |   N  | Std. Deviation |
+---------------------+---------+------+-----------------+
|With sales campaign  | 62.838  |   13 |    20.031       |
|without campaign     | 59.192  |   13 |    19.490       |
+---------------------+---------+------+-----------------+
```

```
+-----------------------+-------+----------+-----------+--------+
|                       |  Mean | Standard | Standard  |  Sig   |
|                       |       | Deviation| Error Mean| 2-tail |
+-----------------------+-------+----------+-----------+--------+
| with sales campaign   | 3.646 |   3.175  |  .881     |  .001  |
| - without campaign    |       |          |           |        |
+-----------------------+-------+----------+-----------+--------+
```

(a) At the .05 level of significance, can the research director conclude that there is evidence that the sales campaign has increased the average sales of non-sale items? State the null and alternative hypotheses, the P-value, and your decision.

(b) (not responsible for this on an exam) Compute a 95% confidence interval for the difference between the means.

(c) What assumptions must you make to perform the test in (a) and construct the confidence interval in (b)?

Answer: (a) $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ versus, $P = .001$, reject H_0. (b) 1.727 to 5.565. $3.646 \pm 2.179 \times 0.881$ *= (1.73 to 5.56). You get full credit for using 1.96 or saying that you can't do it without a t table. (c) Population distribution is normal, since sample size so small, and random sample of stores.*

8. Returning to the defaulting customer case, there were 544 people who put exactly \$498 down and 0.1838% (0.001838) of these 544 people defaulted. Can we compute a 95% confidence interval as follows? **Explain.**

$$0.001838 \pm 1.96\sqrt{\frac{0.001838(1-0.001838)}{544}}$$

Answer: The normal approximately is not valid because $np < 5$, i.e., $0.001838 \times 544 = 1 < 5$. *Note that the lower end of the confidence interval is negative. You will have to use the binomial distribution to compute the confidence interval.*

9. As part of their training, air force pilots make two practice landings with instructors and are rated on performance. The instructors discuss the ratings with the pilots after each landing. Statistical analysis shows that pilots who make poor landings the first time tend to do better the second time. Conversely, pilots who make good landings the first time tend to do worse the second time. The conclusion: criticism helps the pilots while praise makes them do worse. As a result, instructors were ordered to criticize all landings, good or bad. Was this warranted by the facts? Answer yes or no, and explain briefly. *Answer: No, regression effect.*

10. Peterson and Wilson (1992) conducted a study relating to the manner in which a satisfaction question is posed. They asked a closed-end question about satisfaction with a household's primary vehicle in a telephone interview. An identical question was asked in each of two versions of the questionnaire. The only difference was that in one version participants were asked how "satisfied" they were, whereas in the other they were asked how "dissatisfied" they were; response categories were identical in the two question versions. The sample consisted of randomly selected middle-class automobile owners in two large metropolitan areas and respondents were assigned randomly to one of the two question-wording groups. The cross tabulation below gives the counts and adjusted residuals in parentheses.

X	Response	Question Wording **Satisfied**	 **Dissatisfied**	Total
1	Very dissatisfied	10 (−2.4)	23 (2.4)	33
2	Somewhat dissatisfied	12 (−1.5)	20 (1.5)	32
3	Somewhat satisfied	82 (1.2)	69 (−1.2)	151
4	Very satisfied	139 (0.9)	128 (−0.9)	267
	Total	243	240	483

The value of the chi-square test statistic was 8.675, giving a P-value of 0.034.

(a) Perform a chi-square test at the 5% level. State the null and alternative hypothesis, P-value, and your decision.

(b) Examine the adjusted standardized residuals and write *one sentence* discussing the nature of any dependence. Think before you write and lead with that which is most important. You will be graded, in part, on your ability to summarize the conclusions concisely.

(c) If response were independent of question wording, how many people would you expect in this sample of 483 people to have respondended with "Very Satisfied" if asked how "satisfied" he/she is with their car.

(d) If someone is asked how "dissatisfied" he/she is with the car, what is the chance that he/she answers "Very Satisfied?"

(e) If someone answers "Very satisfied," what is the chance he/she was asked the "dissatisfied" version of the question?

(f) Identify the independent and dependent variables in this experiment.

(g) Discuss threats to internal validity.

(h) Discuss threats to external validity.

(i) Alternatively we could assign the value 4 = "Very satisfied," . . . , and 1 = "Very dissatisfied" and compute the mean response for the two question wordings. To do so, what level of scaling must we assume that the response (X) variable has?

(j) We would like to perform an independent-sample t-test. State the null and alternative hypothesis. Be sure to define any parameters used in the statement of your hypotheses.

(k) Compute the mean of response for those asked the "Satisfied" question.

(l) Compute the sample standard deviation of response for those asked the "Satisfied" question.

(m) Briefly describe the shape of the distribution of X for those asked the "satisfied" question.

(n) The output below is extracted from SPSS. The difference is computed as the mean for those asked the "satisfied" question minus the mean of those asked the "dissatisfied question."

	t-test for Equality of Means				
	t	df	Sig (2-tailed)	Mean Difference	Std Error Difference
Equal variances assumed	2.287	481	.023	0.182	0.07959
Equal variances not assumed	2.284	455.7	.023	0.182	0.07970

Find a 95% confidence interval for the difference.

(o) Briefly discuss the differences between what you can conclude from this test compared with the chi-square analysis. Be concise. Lead with the headline.

(p) Marketers often use the "top-two-box score," which is the percentage of respondents who check one of the top two boxes. Find the top-two-box score for those asked the "satisfied" question.

(q) Find the top-two-box score for those asked the "dissatisfied" question.

(r) We wish to test whether the top-two-box scores are equal for the two groups, versus a two-sided alternative. State the null and alternative hypothsis, identifying all parameters used.

(s) Find the P-value to test the hypotheses from the previous part.

Answer: (a) H_0 *: Response independent of wording.* H_1 *: Response dependent on wording. P=0.034, reject. (b) Those asked the dissatisfied question were more likely to say very dissatisfied. (c)* $267 \times 243/483 = 134.3$*. (d) 128/240 = 0.53. (e) 128/267 = 0.479. (f) Independent variable wording, dependent variable response. (g) none. (h) Perhaps the fact that the sample consists of middle class auto owners is a problem, but this is a good study for the most part. (i) interval. (j)* μ_S *is the true mean for the satisfied question and* μ_D *is the mean of those asked the dissatisfied question.* $H_0 : \mu_S = \mu_D$ *versus* $H_1 : \mu_S \neq \mu_D$*. (k)* $(10 + 2(12) + 3(82) + 4(139))/243 = 3.44$*. (l)* $\sqrt{[10(1-3.44)^2 + 12(2-3.44)^2 + 82(3-3.44)^2 + 139(4-3.44)^2]/242} = 0.7711$*. (m) Left skewed. (n)* $0.182 \pm 1.96 \times 0.07970$ *(o) With the means test you can't see that those asked the dissatisfied question are more likely to check dissatisfied. (p)* $(82+139)/243) = 0.9095$*. (q)* $(69 + 128)/240 = 0.8208$*. (r)* $H_0 : \pi_S = \pi_D$ *versus* $H_1 : \pi_S \neq \pi_D$*. (s)* $\overline{p} = (151 + 267)/483 = 0.8654$*.* $S_{\hat{D}} = \sqrt{(1/243 + 1/240)0.8654(1-0.8654)} = 0.03106$*.* $P(\hat{D} > 0.08863) = P(Z > 0.08863/0.03106) = P(Z > 2.85) = 0.0022$*. So* $P = 0.0044 < 5\%$ *so reject* H_0*.*

11. You are considering a new landing page for your website and want to know if your new landing page increases the *conversion rate* over the old one. The conversion rate is the percentage of visitors who make a purchase. It is well established that the conversion rate of the current landing page is 5%. A test of the new page shows that, with 1000 randomly selected visitors from today's web traffic, the conversion rate of the new page is 6.6%.

(a) Identify the specific parameter being studied in this problem. State the null and research hypothesis for a two-sided test.

(b) Find the standard error of observed conversion rate.

(c) "Under the null hypothesis" (i.e., assuming H_0 true), find the standard deviation of the conversion rate observed in this test.

(d) Can you use the normal distribution to find the P-value for the hypothesis in part (a)? Why or why not? Be very specific to receive full credit.

(e) Find the P-value to test the hypothesis in part (a).

(f) State your conclusion about the new landing page using the 0.05 level.

(g) Find a 95% confidence interval for the conversion rate of the new system. Is the reference value from the hypothesis in part (a) contained in the interval?

(h) Briefly discuss the main threats to the validity of your conclusion.

Answer: *(a) Let* π *be the true conversion rate of the new landing page.* $H_0 : \pi = .05$ *versus* $H_1 : \pi \neq .05$. *(b)* $S_p = \sqrt{.066(.934)/1000} = .007851$. *(c)* $\sigma_p = \sqrt{.05(.95)/1000} = .006892$. *(d) Yes, because of the normal approximation to the binomial (which is a special case of the CLT). You can use it because* $1000(.05) = 50 > 5$. *(e)* $P(p > .06) \approx P(Z > (.066 - .05)/.006892) = P(Z > 2.32) = .0102$ *so* $P = .0204 < .05$ *so reject. (f) Reject. The new page is better. (g)* $0.066 \pm 1.96(.007851) = [0.05059, 0.08141]$. *The reference value* .05 *is not contained in the interval. (h) History, e.g., the 5% rate is based on just yesterday, which could be different from other days of the year (e.g., black Friday). This is related to timing. Note: take off points if they say things like mortality, instrument variation, main-testing effect, selection bias, mortality, surrogate situation, etc.*