Measuring Uncertainty and Confidence

Statistical Literacy and Critical Thinking
STAT 1600 – Course Pack

Statistics Computational Lab (SCL)

2018 - 2019 Edition

A general introduction to statistics with an emphasis on data analysis and graphical presentation. Extensive use will be made of the computer to prepare results. Topics may include: data collection, sampling and experimentation, measurement issues, descriptive statistics, statistical graphics, normal distribution, cross-classified data, correlation and association, formal statistical inferences, and resampling methods.

Contents

Contents

Contents

6

Chapter 1

Knowledge and Data

1.1 Objectives

After completing this chapter, students should be able to:

- Discern the difference between knowledge and data.

- Understand the process of acquiring knowledge from the data.

- Perceive the fallacies of design.

- Comprehend the meaning of the Type-I and Type-II errors.

1.2 Introduction

1.2.1 Why Data?

Understanding by the end user

First, data is not information. Researchers extract information from the data. The study of statistics will give us the tools (methods) to judge the validity of the data.

Designing studies

Most scientific journals, after 1966, started to report statistical results. For example, before 1966, FDA (Food and Drug Agency) approved many drugs with written statements from a physician saying that the drug was safe, after 1966 the FDA required statistical studies to

show that the proposed drug did not harm. Later, FDA expected that the drug would be therapeutic.

Conducting research

Unless a sample is biased (we will get to that later), it will reveal the truth about the population. For example, let's suppose that a drug development team after working on a compound for several months decided to try the experimental drug themselves to test its effects before the animal safety studies were complete. (By the way, this behavior is unethical.) The team concluded that it appeared to be safe; however, a mouse study later showed that mice in the high dose group started to canonicalize themselves.

Developing critical thinking skills

Critical thinking starts with asking these type of questions: How can I confirm the validity of these data? What can I infer from these data? Is the source of these data credible? And so forth.

1.3 Knowledge and Data

There is a difference between what we know, and what we think that we know. Are vaccines safe? Will the new tax rules bring more or less revenue? Do mandatory seat belt laws save lives? Will legalizing drugs lead to more addiction? Do alternative therapies work as effectively as traditional treatments? Does exercise help prevent illness and are people who generally exercise healthier? Do taller people make more money than shorter people? [Naranjo]

We can settle most of these issues by looking at data. What do the data tell us? Even then, some problems can only be partially answered by at best incomplete data. In this course, we will lay the groundwork for writing a protocol and will discuss some fundamentals of sound scientific studies and data analysis. In the process, we hope to become better consumers of information and better judges of what is real or not true, and what remains to be proven.

1.3.1 Building Knowledge Step-by-step

Scientific studies often go through the following steps. It might help to read the example in the next section, then read this again.

1. Conceptualize the problem

Suppose that you wish to study and the problem of interest. The wording is usually broad. The idea at first may be too complex to understand so we look for ways to explain and describe it.

For example, perhaps we are interested in a problem of interest within education. We will want to clarify and define education. Are we interested in investigating particular schools or grade levels? Perhaps we want to look at a specific component of education, like science or look at different methods of instruction. Clarifying specific concepts which we want to investigate and putting this into words helps bring focus to our study.

2. Operationalize the problem

As a researcher for your study you also need to come up with ways to narrow in on your problem of interest. What specifically do you want to measure? Here we are formulating the specific question we want to answer. We are also beginning to define our variables. We will need to know what we can measure to help answer our problem of interest.

Perhaps we are interested in looking at how well students are doing with several different methods of teaching science. We may look at measures of assessment like how well they did on an exam or overall in a class given the different teaching methodologies.

3. Design the study

Here we need to address all the components of how our study will be performed. We will likely need to identify the population of interest and to obtain a good representative sample of that population. We will need to figure out how we will select the sample and how many groups we may be comparing.

In our study of science education will we sample from several schools or several districts? If we take a sample of students from several schools, but from only one state, for example, our population would be students within that state. If we wanted to define our population as education within the continental U.S., we might need to pull randomly from schools across the U.S.

4. Collect the data

In the data collection stage, we must decide what instrument we are using (questionnaire, interview, observation). Once we have designed our study, we should be able to identify the main variables of interest. As we collect the data, we are gathering information to help us answer the question we have proposed.

In our example of science education, perhaps we are collecting scores on a particular assessment (exam) to see how well students are learning the material.

5. Analyze the data

Using the proper statistical methods and procedures for the data is an important step in the data analysis stage. Also, checking the assumptions behind these statistical methods is critical. Are we comparing means? Proportions? Are the differences in means/proportions statistically significant? How do we know if there is a significant difference (if the differences we find are not likely due to chance)?

In our exploration of teaching science perhaps we are comparing two different methodologies of teaching. We may then compare the overall means of the two groups.

6. Conclusions

Do your results generalize to a larger population? Check back to see how you defined the population so that you can generalize the results. Are you concluding cause-and-effect relationship, did you perform an experiment, or did you observe associations? In your conclusions, you will also want to identify limits to your study and perhaps suggest further research that may be done in your area.

As we conclude science teaching, perhaps we could only afford to do a smaller study within a particular district. We could then relate the results of our study, based on our sample of schools, to the population of the entire school district. We may not be able to make conclusions to the whole state or the U.S. since our sample was not representative of the whole state or the U.S. Depending on how the study was defined we may have only made observations of the teachers. If we were investigating two distinct types of instruction, then this falls more within the realm of an experiment. We, of course, would still need to identify and try to control for confounding variables, which might be difficult.

7. Disseminate results

How are we sharing our results? We need to find the best approach to sharing the results of our study.

Is this information going to be made available to the schools for future educational Improvement? Are we going to try to publish in a scientific or educational journal?

The Science Wheel Summarizes the above Steps

The Wheel of Science [Wallace, 1971] gives us a process to refine our thinking about the nature of social interactions. See Figure 1.1. We will define the four parts of the wheel below: Theory, Hypotheses, Observations, and Empirical Generalizations.
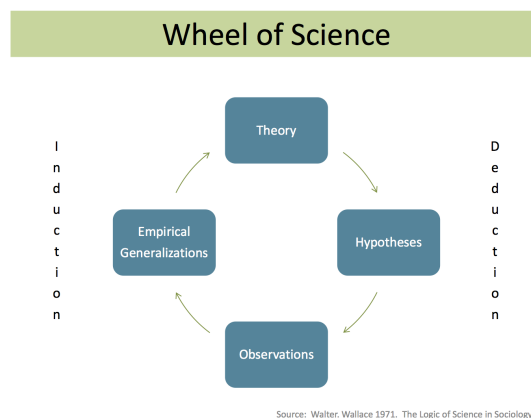
Figure 1.1: The Wheel of Science

1.3.2 Example

Let's look at another example. Suppose we are interested in comparing the following four popular weight loss diets: Zone (balance between carbohydrates, protein and fat), Atkins (low carbohydrate, high fat, unrestricted calories), LEARN (low fat, based on national guidelines), and Ornish (low fat, high carbohydrate, unrestricted calories). How would we design a comparison study?

1. Conceptualize the problem

Which weight loss program is most effective? Which one is the *most healthy*? Can unrestricted-calorie diets make you lose weight?

2. Operationalize the problem

How would we measure 'effective' and 'healthy'? Do we look at weight *loss* after two weeks? Two months? Two years? Do we compare the average weight *loss* or percentage of people who lost 15 pounds or more? How do we measure health? By cholesterol reduction? LDL cholesterol reduction? Blood pressure reduction? Glucose levels? Percentage who feel better?

3. Design the study

Where do we recruit our subjects for the study? How long will the study last? Will we include only overweight subjects? Do they choose which diet, or do we randomize the selection procedure? How do we ensure that they stay on the diet? Do we drop the cheaters from the study?

4. Collect the data

How many times will we measure their weights? How many times will we take blood samples? Urine samples? Do we send all samples to the same lab? Do we measure how strictly each subject adheres to their diet?

5. Analyze the data

Are there significant differences in average weight loss between diet groups? Are there significant differences in the percentage who lost weight? Are there differences in cholesterol changes, blood pressure changes, glucose level changes? Are there differences in long-term adherence rates to each diet?

6. Conclusions

Which diet would we recommend? Under what conditions? Are our subjects sufficiently representative to allow generalization? Are we sure the observed weight loss can be attributed to the diet?

7. Disseminate results

How would we present the results of the study? What tables and graphs would make the study easy to read and understand?

1.3.3 A final product

The following is a summary of a study reported in the Journal of the American Medical Association (JAMA), Vol. 297, No. 9 in March 2007. The paper is titled "Comparison of the Atkins, Zone, Ornish,and LEARN Diets for Change in Weight and Related Risk Factors Among Overweight Premenopausal Women The A TO Z Weight Loss Study: A Randomized Trial," by Christopher D. Gardner, PhD, Alexandre Kiazand, MD, Sofiya Alhassan, PhD, Soowon Kim, PhD, Randall S. Stafford, MD, PhD, Raymond R. Balise, PhD, Helena C. Kraemer, PhD, and Abby C. King, PhD. All authors were affiliated with the Stanford Prevention Research Center and the Department of Medicine, Stanford University Medical School.

Context

Popular diets, particularly those low in carbohydrates, have challenged current recommendations advising a low-fat, high-carbohydrate diet for weight loss. Potential benefits and risks have not been tested adequately.

Objective

To compare four weight-loss diets representing a spectrum of low to high carbohydrate intake for effects on weight loss and related metabolic variables. Design, Setting, and Participants Twelve-month randomized trial conducted in the United States from February 2003 to October 2005 among 311 free-living, overweight/obese (body mass index, 27-40) nondiabetic, premenopausal women. Intervention Participants were randomly assigned to follow the Atkins (n=77), Zone (n = 79), LEARN (n=79), or Ornish (n=76) diets and received weekly instruction for two months, then an additional 10-month follow-up.

Main Outcome Measures

Weight loss at 12 months was the primary outcome. Secondary outcomes included lipid profile (low-density lipoprotein, high-density lipoprotein, and non-high-density lipoprotein cholesterol, and triglyceride levels), the percentage of body fat, waist-hip ratio, fasting insulin and glucose

levels, and blood pressure. Outcomes were assessed at months 0, 2, 6, and 12. The Tukey studentized range test was used to adjust for multiple testing.

Results

Weight loss was greater for women in the Atkins diet group compared with the other diet groups at 12 months and mean 12-month weight loss was significantly different between the Atkins and Zone diets ($P < .05$). Mean 12-month weight loss was as follows: Atkins, -4.7 kg (95% confidence interval [CI], -6.3 to -3.1 kg), Zone, -1.6 kg (95% CI, -2.8 to -0.4 kg), LEARN, -2.6 kg (-3.8 to -1.3 kg), and Ornish, -2.2 kg (-3.6 to -0.8 kg). Weight loss was not statistically different among the Zone, LEARN, and Ornish groups. At 12 months, secondary outcomes for the Atkins group were comparable with or more favorable than the other diet groups.

Conclusions

In this study, premenopausal overweight and obese women assigned to follow the Atkins diet, which had the lowest carbohydrate intake, lost more weight and experienced more favorable overall metabolic effects at 12 months than women assigned to follow the Zone, Ornish, or LEARN diets. While questions remain about long-term effects and mechanisms, a low-carbohydrate, high-protein, high-fat diet may be considered a feasible alternative recommendation for weight loss.

The study tells us that all three groups lost weight. Over 12 months, the Atkins group lost the most weight without increasing cholesterol, blood pressure, the percentage of body fat, or fasting glucose levels.

The study does not tell us that if one stays on the Atkins diet long term (say, two years, or five years, or 10 years) their cholesterol, blood pressure, the percentage of body fat, or fasting glucose levels will remain favorable. If one 'cheats' on the Atkins diet periodically, what will be the effect on weight and metabolic profile? Will the same results be observed for men? For non-obese women?

Some studies are well designed; others are not. It is often up to us to make our judgments about what is true, instead of leaving this responsibility to journalists or bloggers. Whenever you hear a claim on television or in a magazine, ask yourself "What is the evidence?"

1.4 Common Fallacies

A fallacy is sometimes defined as a mistake in basic reasoning. The following are examples of fallacies:

1. **Lack of evidence fallacy**

   "There is no proof that the drug is unsafe." It allows claims to be made without providing any evidence, simply by shifting the burden of proof. The fallacy lies in the reasoning that lack of evidence means the contrary is true.

2. **Anecdotal evidence fallacy**

   "We give you testimonies of real people who . . ." improved their golf game, or improved their sex life, or lost weight, or got rid of acne or achieved financial success. They make claims without comparison studies. Take a golf infomercial with testimonies from five golfers that a new driver improved their distance and accuracy. Did it? These are five golfers out of how many that they approached? These are televised shots out of how many that each took? The fallacy lies in the reasoning that existence means prevalence.

3. **Correlation equals causation fallacy**

   For example, "Married people are happier than singles," may be wrongly interpreted as "Want to be happy? Get married." It may be that happy people are the ones who tend to get married, or that high earners tend to be both happy and married. The fallacy lies in the reasoning that "two things happening together" must mean one causes the other.

1.4.1 Some fallacies in interpreting evidence

In medical studies, we use clinical trials to see if a difference exists between treatments. We use statistical analysis to determine if there is a significant difference between the treatment types. Perhaps there is very little difference, and that difference can be explained as being due to chance. It could be there is no actual difference between treatment groups. It leads us to a Type I error which can commonly occur.

A **Type I error** occurs when the researcher falsely finds a difference between treatments where no actual difference exists. Often, we use 95% confidence intervals when analyzing for statistical significance. It means that 5% of the time we may conclude a significant difference, when in fact one doesn't exist! The researcher is thus willing to accept a 5% chance that the study conclusions could be wrong. Notice, however, 5% is very small compared to the 95% assurance we have that there is a significant difference.

Another type of error that may be made is a **Type II error**. Here the researcher fails to find a difference between the treatment groups when a true difference does exist! Perhaps we did a study with only a small sample, and we didn't see a difference between our treatment and control group. We would conclude there is no difference, which the treatment was not effective. Another lab might perform the same type of study but have a much large sample size, and once they analyze their data find that there is a true difference. If we fail to find a statistical difference when there is a difference, we have committed a Type II error. It also relates to the power of a study. The power of a study is related to the strength of the study. Can we detect an effect when there is an effect? How well was the study conducted? It also relates to our sample size. When we have a larger sample size, we will reduce the likelihood of committing a Type II error.

1.5 Key Words

• Correlation	• Knowledge
• Data	• Type I error
• Fallacies	• Type II error

1.6 Exercises

―――――――

Exercises

Exercise 1.6-1: What might be wrong about these headlines?

1. A study proclaims: "Slightly overweight people live longer than thin people."

2. A study states, "People who consider themselves depressed eat more chocolate than people who consider themselves otherwise."

3. A U.S. Census reported "More American women are living without a husband than with one." Additionally, women rated themselves happier when compared to the previous year's census. Can we then infer that living single is leading to greater happiness in women? What flaw is present here?

Exercise 1.6-2: A handful of people were recently polled on their preference of soda at a local shopping mall in the U.S. and the study indicated more people preferred Mountain Dew over Cola. It indicates Mountain Dew is exceeding Cola as the beverage of choice in mainstream America. What is wrong with this assertion?

Exercise 1.6-3: A new anti-wrinkle cream was given to several people who frequently purchase make-up products from a specific company. In a follow-up with these individuals, all of them assert the new product was highly effective. What fallacy is present here? Should the company mass produce this new product? If not, what else needs to be studied here?

Exercise 1.6-4: All the members (sample size, n=30) of a high school tennis team were given a new racket and were asked to report back on how the racket affected their performance during practices for the week. All the members of the tennis team then completed a survey, and everyone minus one individual reported that the new racket impacted their performance in a positive manner. Since we have a large enough sample size can we now conclude the new racket causes tennis players to perform better? Why or why not? Is there a fallacy here?

Chapter 2

Data Presentation

2.1 Objectives

After completing this part, students should be able to:

- Recognize the types of variables (categorical or numerical) and give examples of each type.

- Remember and explain the three levels of measurement and give examples of each level.

- Tell the difference between dependent and independent variables.

- Compute the relative frequency.

- Use graphical methods to display data.

2.2 Statistics and Data

Statistics refers to a collection of techniques and procedures for analyzing data. In this context, statistics may be considered a synonym of the data analysis. In a narrower context, 'statistics' is sometimes used as a synonym for the numbers themselves, as in 'demographic statistics' or 'baseball statistics.' 'Data' refers to any collection of measurements that a researcher makes on some number of subjects. Researchers often store these measurements in a row-and-column display called a spreadsheet. The following example shows a spreadsheet containing data taken on ten students in a class.

Student	Sex	Level	GPA	Class Hours Credit Hrs Taken	Transport	Hrs Slept
CLASS DATA						
1	M	Sophomore	3.10	32	Car	7
2	M	Junior	3.20	66	Car	8
3	F	Senior	3.49	94	Bus	8
4	M	Senior	2.68	89	Walk	10
5	F	Junior	3.73	69	Bicycle	8
6	F	Junior	3.39	59	Car	8
7	F	Senior	3.80	86	Walk	8
8	M	Junior	3.11	75	Car	8
9	F	Sophomore	3.10	27	Car	7
10	M	Senior	3.10	96	Walk	3

Each row represents a subject (or student, in this case), and each column represents a measurement. The measurement columns are also called variables because the values vary from one subject to another.

2.3 Classification of variables

2.3.1 Levels of Measurement

It is useful to distinguish between **four levels of measurements** for variables, from weakest to strongest: Nominal (no ordering), Ordinal (ordering exists, but not distance), Interval (distance exists, but not ratios), and Ratio (ratios exist).

1. **Nominal Variables**

   Nominal variables are categorical variables that have two or more categories without having any logical sequence or order. For example, Sex is a nominal variable, since 'Male' and 'Female' are just names of categories. Additional nominal variable examples include religious affiliation, language spoken, nationality, and blood type. There is no intrinsic ordering between these. We often use bar charts and pie charts to represent these data.

- You cannot perform arithmetic operations like addition, subtraction, multiplication, etc.
- No order or logical sequence is present
- Only Names
- Mean and median cannot be used for nominal variables.

2. **Ordinal Variables**

Ordinal variables are categorical variables with an explicit ordering or rank. A student's level of standing (freshman, sophomore, junior, or senior) is ordinal; they are also names of categories but, unlike sex, they are rank-ordered. However, we cannot subtract them, and distances do not make sense. Other examples of ordinal variables include Likert scale items in which we rank items. Completing surveys or even end of semester evaluations we can rank on a scale of 1-5 (often from strongly disagree to agree strongly).

- Be careful about the size of the difference between categories may not be equally spaced.
- Only order matters not the difference between categories
- We cannot use the mean for ordinal variables

3. **Interval Variables**

Interval variables are numerical variables where the difference between two values is meaningful. GPA is an interval measurement; we can subtract these measurements, and distances make sense. For example, the distance from 2.3-2.4 is the same distance as 3.7-3.8. However, ratios do not make sense; is 4.0 'twice as high' as 2.0? The answer is no. The grading system would work just as well on the scale (A, B, C)=(5.0, 4.0, 3.0) instead of (4.0, 3.0, 2.0).

4. **Ratio Variables**

Ratio variables are numerical variables with true zero. Many credit hours is a ratio measurement. A student who has completed 90 credit hours has twice as many as 45 credit hours and three times as many as 30 credit hours. Another example is temperature on the Kelvin scale, which has true zero. A Kelvin temperature of 400 is twice as high as 200 Kelvin.

It is useful to recognize a hierarchy of information in the sense that a measurement level contains an amount of information greater than or equal to the level below it. At lower levels of measurement, data analyses tend to be less sensitive and sophisticated. A statistical study should aim for the highest levels of measurement possible or affordable.

2.3.2 Numerical versus categorical variables

Interval and ratio variables together are often called **numerical** variables or **quantitative** because they provide a number which measures 'quantity' (how much, how many) of something.

Nominal and ordinal variables together are often called *categorical* variables because they classify rather than count or measure. It is tempting to think of categorical variables as 'non-numerical,' but sometimes they do consist of numbers. For example, 'Social Security number' consists of numbers, but are used more as labels rather than quantities. Hence, SS number is categorical.
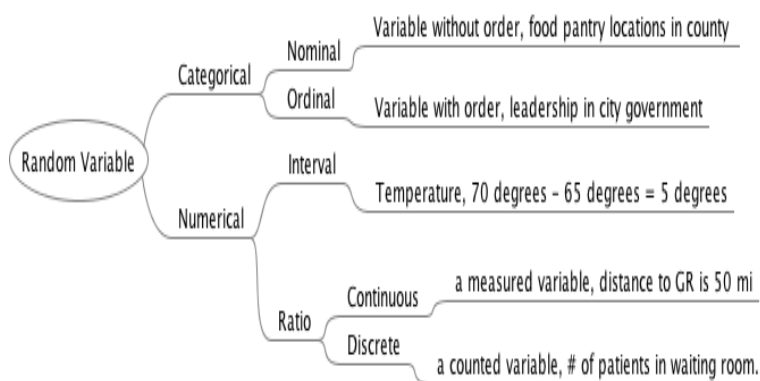
Figure 2.1: Mind map of a Random Variable

2.3.3 Dependent versus independent variables

Another classification for variables is dependent and independent. This distinction is relevant in studies that investigate cause and effect. The independent variable is the probable cause. We also refer to this as the predictor or explanatory variable. The dependent variable is the outcome variable that is affected by the independent variable.

In cause-and-effect studies, the **independent** variable is the probable cause. The **dependent** variable is the outcome being caused/affected.

The following is a description of a clinical study on the possible link between vaccines and autism. In particular, the study tries to find an association between the amount of a mercury-containing preservative in vaccines and measures of brain function in children. It has one independent variable (amount of mercury in vaccines) and 42 dependent variables (measures of brain function). Similar to two previously published studies, it found no evidence of neurologic problems in children exposed to mercury-containing vaccines. [Thompson, 2007]

Background

Others have hypothesized that early exposure to thimerosal, a mercury-containing a preservative used in vaccines and immune globulin preparations are associated with neuropsychological deficits in children.

Methods

William W. Thompson enrolled 1047 children between the ages of seven and ten years and administered standardized tests assessing 42 neuropsychological outcomes. We did not assess autism-spectrum disorders. We determined exposure to mercury from thimerosal from computerized immunization records, medical records, personal immunization records, and parent interviews. We obtained the information on potential confounding factors from the interviews and medical charts. We assessed the association between current neuropsychological performance and exposure to mercury during the prenatal period, the neonatal period (birth to 28 days), and the first seven months of life.

Conclusion

Our study does not support a causal association between early exposure to mercury from thimerosal-containing vaccines and immune globulins and deficits in neuropsychological functioning at the age of 7 to 10 years.

Among the 42 variables used to measure brain function tests on speech and language (comprehension of instructions, recalling sentences, stuttering), tests on verbal memory (no delay, short delay, long delay), tests on motor coordination (grooved pegboard, finger tapping), tests on behavior regulation (hyperactivity, inattentiveness), presence of tics (motor and phonics) and tests on intelligence (verbal IQ, full-scale IQ). Most of these are numerical test scores. Others are categorical, like the presence of stuttering (Yes-No) and tics (Yes-No).

Note that the researcher has a choice of whether to use numerical or categorical measures of brain function. Even for categorical outcomes, there is a choice between nominal (i.e., Yes-No), or ordinal (Frequently-Sometimes-Never). Many surveys like to use the ordinal five-point scale:

1. Strongly agree

2. Agree

3. Neither agree nor disagree

4. Disagree

5. Strongly disagree

It is called a *Likert item*. The scores from adding up several Likert items is said to be on a *Likert scale*. Western Michigan University's course evaluations use a Likert scale.

2.4 Summarizing Categorical Data

The Census Bureau conducts a separate nationwide survey on population and housing information. Called the American Community Survey (ACS), it is "designed to provide communities a fresh look at how they are changing." The ACS collects population and housing information every year instead of every ten years. The following data is a sample of the actual responses to the ACS from the state of Michigan and Indiana in 2008.

The variables 'State' and 'Type of Payment' are categorical. The variables 'No. of bedrooms,' 'Monthly Payment' and '12-month Household Income' are numerical.

| | Michigan and Indiana ACS Data | | | | |
| | A Partial List | | | | |
Household	State	Bedrooms	Monthly	Type	Income
1	Michigan	2	880	Rent	11200
2	Michigan	3	990	Mortgage	80800
3	Michigan	4	750	Mortgage	87600
4	Michigan	3	1400	Mortgage	94000
5	Michigan	4	1400	Mortgage	97000
6	Michigan	1	560	Rent	6000
7	Michigan	4	900	Mortgage	95000
8	Michigan	2	0	None	39000
9	Michigan	1	380	Rent	24370
10	Michigan	2	910	Rent	54500
⋮			⋮		⋮

Table 2.1: The full list is found in eLearning web-site.

2.4.1 Relative Frequency Table

A *relative frequency table* gives the count for each category and the relative frequency or percentage of time in which each category occurs. Here is a relative frequency table for Monthly Payment type.

	Frequency	Relative Frequency(%)
Mortgage	44	58
None	12	16
Rent	20	26

Table 2.2: Relative Frequency Table for Type of Monthly Payment

To compute Relative Frequency, divide Frequency by the total and multiply by 100%. For example,

$$RF = \frac{44}{76} \times 100 = 57.9$$

2.4.2 Bar Chart

A *bar chart* is a plot of the relative frequency table. We use bar charts for categorical data. Data scientists present a bar chart where the order of the categories in the bar chart is arbitrary. Sometimes they are presented in order of decreasing frequency for aesthetic purposes. Some computing packages present the categories alphabetically unless otherwise specified.

Bar Plot for Type of Payment
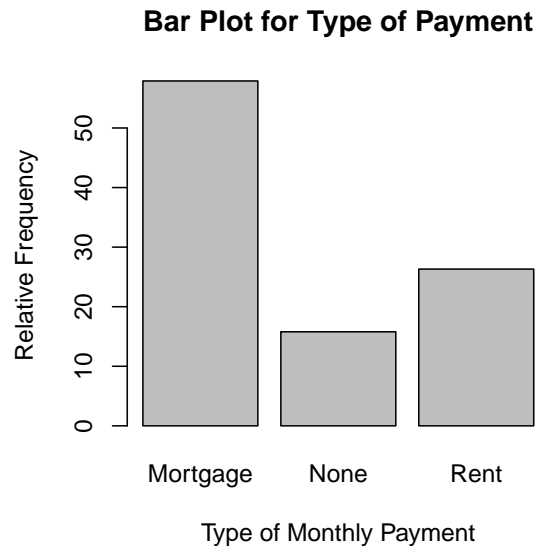
2.4.3 Pie Chart

A pie chart gives the same information as a bar chart, except in a circular shape. We also use pie charts for categorical data. Some people prefer the look of pie charts over bar charts. However, studies have shown that people have difficulty comparing the relative size of wedges. Look at the pie chart below. Is Rent less than half or more than half of Mortgage?

Bar Plot for Type of Payment
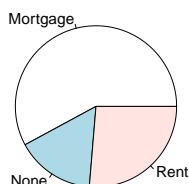
2.5 Summarizing numerical data

Consider the variable Monthly Payment in the ACS housing data on page 11. How do the numbers appear? How large are they? How variable is the data? Are there any extreme values? These questions are attempts at getting to know the data. In this section, we discuss tools for getting to know numerical data better. We have n = 64 observations available since 12 households (out of 76) do not have monthly payments. Here is a list of the monthly payments:

880 990 750 1400 1400 560 900 380 910 1200 1200 450 1200 1300 550 340 770 700 700 140 220 650 5200 970 740 500 700 1800 800 1200 1000 1200 200 710 1100 400 370 670 350 510 900 340 440 1500 420 490 2400 500 200 670 760 720 500 280 190 1000 250 700 380 740 850 530 290 230

Now, we sort the observations from smallest to largest:

140 190 200 200 220 230 250 280 290 340 340 350 370 380 380 400 420 440 450 490 500 500 500 510 530 550 560 650 670 670 700 700 700 700 710 720 740 740 750 760 770 800 850 880 900 900 910 970 990 1000 1000 1100 1200 1200 1200 1200 1200 1300 1400 1400 1500 1800 2400 5200

The smallest observation is $140. The largest is $5200, which is an **outlier**.

> In data analysis, an **outlier** is an observation that falls far from the rest of the data. It may need to be checked for correctness if there is any possibility of clerical error.

Can this be a typographical error that we need to correct? Looking closely at the data, we see it belongs to observation 28, which is a mortgage for a household with $358,000 annual income. We infer that it is probably correct, so we leave it alone. The simple act of sorting the data already gives us more information about monthly payment. For example, we can now see that the typical monthly payment falls around $700. We also see the extent to which the $5200 value is outlying. We now turn to better ways to visualize the **spread** of this data.

2.5.1 Stem-and-Leaf Plot

Graphs are helpful in summarizing information in the data, especially for large datasets. Below, we summarize monthly payment in a graph called a stem-and-leaf plot. The numbers on the left are stems, while the numbers on the right are called leaves.

There is one leaf for each observation, so there are 63 leaves in all (we left out the $5200 outliers, for now). Our smallest observation is $140. We write it on the stem '1' as a leaf of '4'. We can infer that the stems represent $100's, and the leaves represent $10's. We write the next largest value $190 on stem '1' as a '9'. $200 is written on stem '2' as '0', etc.

Stem and Leaf Display of Monthly
Payment (stem width $100)

Stem	Leaf
1	49
2	0023589
3	445788
4	02459
5	0001356
6	577
7	00001244567
8	058
9	00179
10	00
11	0
12	00000
13	0
14	00
15	0
16	

Stem-and-Leaf Display of Monthly
Payment (stem width=$200)

Stem	Leaf
0	49
2	0023589445788
4	024590001356
6	57700001244567
8	05800179
10	000
12	000000
14	000
16	
18	0
20	
22	
24	0
26	
28	
30	

You see why we left out $1800, $2400, and $5200. It would have made the graph extend beyond the page! We can make the graph more compact by merging stems. The plot below is about half the length because the intervals for each stem are twice as wide.

The stem '8' now contains both '8' and '9', so that the leaves are read as 800, 850, 880, 900, 900, 910, 970, 990. When comparing both plots, which stem-and-leaf plot is better? The

answer depends on the viewers' taste and is not a matter of right or wrong. In both cases, we see that monthly payments seem to spread out from around a center of $700.

Suppose we want to compare mortgage and rent. Do mortgage payments tend to be greater or less than rent? We can present stem-and-leaf for mortgage and rent side by side.

Stem and Leaf Display of Monthly Payment (stem width $100)

Mortgage			Rent	
1	9		1	4
2	0039		2	258
3	478		3	458
4	245		4	09
5	00135		5	06
6	7		6	57
7	0012457		7	0046
8	05		8	8
9	007		9	1
10	00		10	
11	0		11	
12	0000		12	0
13	0		13	
14	00		14	
15	0		15	
16			16	

In general, rent is smaller than mortgages and less variable. While mortgages seem centered around $700, rent is centered around $500. Furthermore, mortgages have a longer right tail, with a significant percentage over $1000. Plus, remember that we have removed a $1800 - $5200 mortgages from these plots.

2.5.2 Relative Frequency Table and Histogram

Earlier in this chapter, we presented a relative frequency table for categorical data. Relative frequency tables may also be used for numerical data. A relative frequency table for monthly payment is presented below. The class width is chosen to achieve a moderate number of class intervals.

There are three things to keep in mind when constructing a frequency table. First, decide how many classes (i.e., intervals) we want. This also determines the *class width*. Try to have 5 to 15 intervals, depending on how many observations we have. Second, decide where the first interval starts. Third, decide how to avoid boundary disputes.

The last item requires us to choose a **boundary convention**. For instance, the intervals in the frequency table above could have been written as follows:

	Frequency	Relative Frequency(%)
0-200	2	3
200-400	13	20
400-600	12	19
600-800	14	22
800-1000	8	12
1000-1200	3	5
1200-1400	6	9
1400-1600	3	5
1600-1800	1	2
2400-2600	1	2
5200-5400	1	2

Table 2.3: RF Table of Monthly Payment for ACS Housing Data

0-199
200-399
400-599
600-799
etc.

This way, it is easier to tell that $200 belongs to the second interval, not the first. However, the table looks more complicated, harder to read. To keep the intervals simple but avoid boundary disputes, include a footnote to the table that describes the boundary convention, i.e. "Intervals contain the left endpoint but not the right." Then, we know that $200 belongs to the second interval, not the first. Alternatively, we may use square braces and parenthesis.

$[0 - 200)$
$[200 - 400)$
$[400 - 600)$
$[600 - 800)$
etc.

It also means that the class intervals contain the left endpoint but not the right. Which method do we prefer?

The relative frequency table is a compact numerical way to present how the data is distributed. If we plot the frequencies as columns, the resulting plot is called a *histogram*.

The histogram and stem-and-leaf plots look alike, except that the stem-and-leaf plot has columns that go sideways instead of upwards. Stem-and-leafs are better if we want the data values themselves available from the plot. However, the histogram can handle large sample sizes easily and is more flexible in choosing class widths. For example, we may choose class

widths of $500, as follows: $[0, 500), [500, 1000), [1000, 1500), \ldots, [5000, 5500)$, which stem plots cannot do.

2.5.3 Dot Plot

For moderately large samples, a *dotplot* diagram is a quick way to see what the data appears. We present dotplots of the monthly payment for rent and mortgages below. As we observed using side-by-side stem plots, rent is smaller than mortgages and less variable.
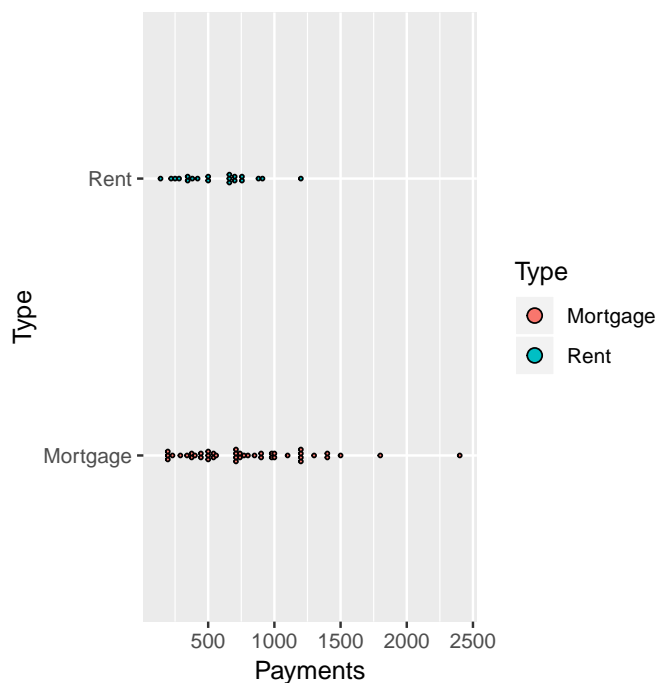
2.5.4 Box-and-Whisker Plot

A box-and-whisker or boxplot shows useful information about the data including the measures the central tendency and the variability of the distribution. First, order the data from smallest to largest. What is the range of the first (or smaller) half of the data? The smallest quarter of the data? The next quarter? The box-and-whisker plot or boxplot is a graphical picture of the distribution of quarters of the data. Consider once again the monthly payments from the ACS housing data. We list them here sorted from smallest to largest (minus the outlier ($5200)). 140 190 200 200 220 230 250 280 290 340 340 350 370 380 380 400 420 440 450 490 500

500 500 510 530 550 560 650 670 670 700 700 700 700 710 720 740 740 750 760 770 800 850 880 900 900 910 970 990 1000 1000 1100 1200 1200 1200 1200 1200 1300 1400 1400 1500 1800 2400

There are $n = 63$ observations. One-quarter of the data is 63/4=15.75, or approximately 16 observations. The boxplot below gives the range of each quarter of the data: the range of the first quarter (i.e., the lowest 16 monthly payments) is the left whisker. The range of the 2nd quarter is the left part of the box, the 3rd quarter the right part of the box, and the range of the last quarter is the right whisker.

Box of Monthly Payments
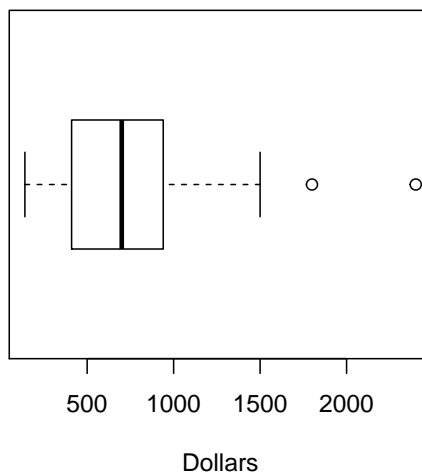
More precisely, the software draws vertical lines at (140, 400, 700, 970, 2400). Hence, the lowest 16 monthly payments lie within ($140, $400). The next set of 16 observations lies within ($400, $700). The third set lies within ($700, $970). Finally, the last 16 observations lie within ($970, $2400). The five values (140, 400, 700, 970, 2400) that divide the data into quarters and form the fences and whiskers of the boxplot are collectively called the five-number-summary of the data. They are often denoted as MIN, Q_1, MED, Q_3, and MAX respectively.

1. **MIN** is called the **minimum** and is the smallest of the ordered observations.

2. Q_1 is the upper boundary of the first quarter and is called the **first quartile**.

3. **MED** is the upper boundary of the second quarter and is called the **second quartile**. However, it also divides the data into lower and upper halves and is more often called the median.

4. Q_3 is the upper boundary of the third quarter and is called the **third quartile**.

5. **MAX** is the largest of the ordered observations and is called the **maximum**.

Boxplots are particularly useful for comparing two distributions side-by-side. Below are boxplots of mortgage data and rent drew parallel to each other on the same scale. When comparing data of mortgages and rental, we find that mortgages tend to be larger than rent since the median and quartiles are larger. There is also a considerable difference in the spread, with the mortgage data having a long right tail, evidence that mortgages have a higher ceiling than rent.

boxplot of Monthly Payments
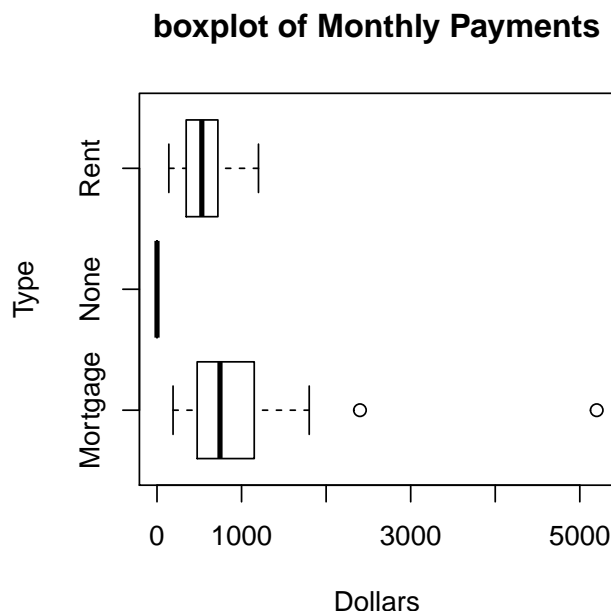
Different statistical computing packages have different ways of computing the quartiles, but the differences are minimal. In this class, we compute the quartiles as follows. First, arrange the observations from smallest (1st ordered observation) to largest (n^{th} ordered observation). Then

---

Q_1 is the .25(n+1)st ordered observation.
MED is the .50(n+1)st ordered observation.
Q_3 is the .75(n+1)st ordered observation.

If .25(n+1) is not an integer, take the average of the two adjacent ordered observations. Similarly, for MED and Q_3. Here again are the $n = 63$ ordered observations of monthly payments.

140 190 200 200 220 230 250 280 290 340 340 350 370 380 380 400 420 440 450 490 500 500 500 510
530 550 560 650 670 670 700 700 700 700 710 720 740 740 750 760 770 800 850 880 900 900 910 970
990 1000 1000 1100 1200 1200 1200 1200 1200 1300 1400 1400 1500 1800 2400

Since $.25(63+1)=16$, then Q_1 is the 16th ordered observation. Hence $Q_1 = 400$. Similarly, $0.50(63+1)=32$ so MED=700. Finally, $.75(63+1)=48$, so $Q_3 = 970$.

If $n = 64$, then $.25(64+1) = 16.25$, so that Q_1 would be the average of the 16th and 17th ordered observations. Therefore, if we return the \$5200 outlier we removed, then the five-number-summary would be (140, 410, 700, 980, 5200).

2.5.5 Symmetry and Skewness

The shape of the data is often described by its symmetry or non-symmetry (also called skewness). Here are stem-and-leaf plots for *symmetric*, *right skewed*, and *left-skewed* data.

Symmetric Data			Right-Skewed Data			Left-Skewed Data	
4	7		4	58		4	0
5	35		5	112345569		5	5
6	115		6	2445		6	2
7	11112		7	17		7	
8	467899		8			8	17
9	1199		9	2		9	2445
10	14		10	5		10	111245569
11	1		11	2		11	58

The term 'skew' refers to the direction of the most extended tail when you flip the stem-and-leaf sideways or draw a histogram. Here is the histogram of the left-skewed plot below.
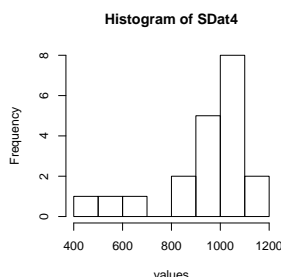
A long tail denotes extreme or outlying observations. The left-skewed histogram above contains observations that are small and outlying, while a right-skewed histogram would have observations that are large and outlying. A left skew also indicates that the smallest quarter

of observations will be more spread out. For some further examples of histograms depicting both symmetry and skewness, see below.
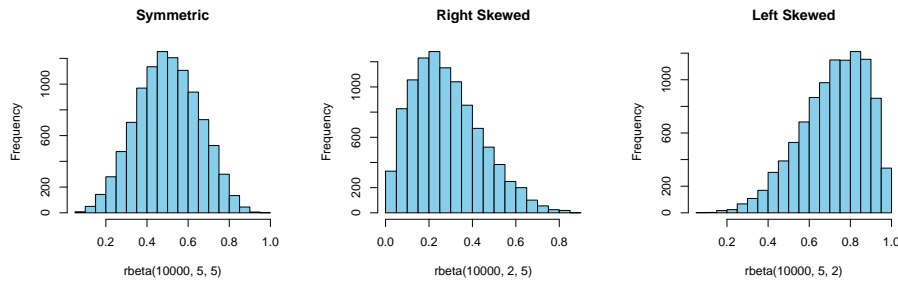
2.6 Key Words

- Data

- Dependent

- Graphical Methods
    - bar chart
    - box and whisker
    - dot plot
    - histogram
    - pie chart
    - stem-and-leaf

- Independent

- Statistics

- Tabular Methods
    - frequency distribution

- Variables
    - Categorical
    - Numerical

2.7 Exercise

————

Exercises

Exercise 2.7-1: Identify the following types of data as numerical or categorical. If numerical, further classify into interval or ratio.

1. The scores on exam one for Stat 1600.

2. Marital status

3. Annual income

4. Social Security Number

5. Cumulative GPA

6. Academic level (freshman, sophomore, junior, senior, other)

7. Quality (poor, fair, good, excellent)

8. Height (short, average, tall)

9. Age (years)

10. Grade $(A, B, C, \ldots)$

11. Color

12. Rating of eight local plays (poor, fair, good, excellent).

13. Times required for mechanics to do a tune-up.

```
2 |
3 | 4
4 | 56889
5 | 01112345567889
6 | 0124457
7 | 17
8 |
9 | 2
```

Exercise 2.7-2: The 6-year graduation rate for a random sample of 30 colleges and universities in the U.S. is displayed in the following stem-and-leaf plot. Note that the stem unit is 10%, and the leaf unit is 1%. For example, the maximum value is 92%.

1. Obtain the five-number summary.

2. Obtain a boxplot and histogram for 6-year graduation rate.

3. In our opinion, which of the three plots (stem-and-leaf, boxplot, histogram) best illustrates the data? Why?

Exercise 2.7-3: A manager of a car rental company took a random sample of 100 business days over the last fiscal year and recorded the number of cars rented per day. The frequency distribution of the data is given below.

Interval	Frequency	Relative Frequency
$(20, 25]$	4	
$(25, 30]$	11	
$(30, 35]$	23	
$(35, 40]$	31	
$(40, 45]$	15	
$(45, 50]$	10	
$(50, 55]$	6	

1. Fill in the Relative Frequencies above.

2. Draw a histogram by hand, with either the frequencies or relative frequencies as the vertical axis.

3. What interval does the median number of car rentals per day fall?

4. What percentage of business days had 30 or fewer car rentals?

5. How many business days had more than 45 car rentals?

Exercise 2.7-4: The following plots represent five different samples of data. For each, describe the shape.

Figure 2.2: Describe Shapes

a)

b)

c)

d)

4	58
5	011245569
6	2445
7	17
8	
9	2
10	5
11	0

e)

0	4
1	889
2	11467
3	457
4	7
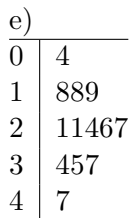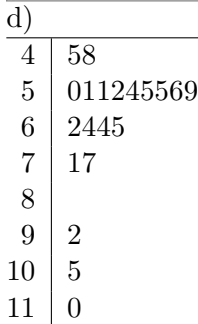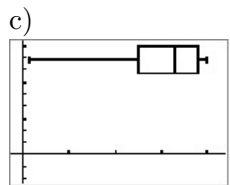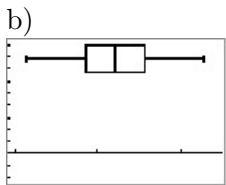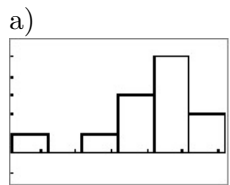
Answer:

Exercise 2.7-5: A survey contained a question regarding marital status. The respondent checked either single, married, divorced, separated or widowed. What is the level of measurement of marital status?

1. Ordinal

2. Interval-Ratio

3. Nominal

4. I don't know

Answer:

Exercise 2.7-6: Blood alcohol content (BAC) is the concentration of alcohol present in a person's blood. What is the level of measurement of BAC?

1. Ordinal

2. Interval-Ratio

3. Nominal

4. I don't know

Answer:

Exercise 2.7-7: What type of variable is the number of robberies reported in June 2014 in Kalamazoo County?

1. Attribute

2. Discrete

3. Continuous

4. Qualitative

Answer:

Exercise 2.7-8: The team of researchers took the following items from a 2007 survey of drug use among young UK children Fuller [2008]. In how many occasions have you used or taken Cannabis? Determine whether the data are categorical or numerical and the level measurement.

_____	Never
_____	Once
_____	2 - 8 occasions
_____	9 - 15 occasions
_____	More than 15 occasions

Answer:

Exercise 2.7-9: The researchers asked students the following questions from a 2007 survey of drug use among young UK children Fuller [2008]. Write the number of glasses of liquors (e.g., Baileys, gin, tequila, vodka) drunk in the last seven days. _____. Determine whether they are categorical or numerical and the level of measurement.

Answer:

Exercise 2.7-10: The General Social Survey (GSS), conducted by the National Opinion Research Center at the University of Chicago, is a primary source of data on social attitudes in the U.S. Once each year, researchers from UC interview 1500 adults in their homes all across the country. They ask subjects their opinions about sex and marriage, attitudes toward women, welfare, foreign policy, and many other issues Trustees [2016].

The population for the GSS is

1. the 1500 persons interviewed.

2. the University of Chicago.

3. the list of questions asked.

4. all adult residents of the U.S.

Answer:

Exercise 2.7-11: A recent issue of the *New England Journal of Medicine* reported a study of all 122,754 infants born over an 8.5 year period at Parkland Hospital in Dallas, Texas, leaving out multiple births and babies with birth defects. The researchers wanted to know if there is a specific birth weight below which infant death and illness increases sharply.

The independent variable in the study is

1. death and illness.

2. infants (leaving out multiple births, etc.)

3. birth weight.

4. Parkland Hospital.

Answer:

Exercise 2.7-12: For the following research project, classify all variables in terms of levels of measurement and reveal whether they're continuous and discrete. As researchers, we should select the correct statistical method: *single variable descriptive statistic* or *a multivariable descriptive statistic* or

inferential statistics. Recall that it is common for some projects to require more than one type of method.

Ten years ago, a state re-instituted the execution for first-degree murder. Researchers asked, did this change of policy reduce capital crime? A team of researchers gathered data on the number of homicides in the state for two-year periods before and after the policy change.

Answer:

Exercise 2.7-13: For the following research project, classify all variables regarding levels of measurement and reveal whether they're continuous and discrete. We should also select which statistical method should be used: *single variable descriptive statistic* or *a multi-variable descriptive statistic* or *inferential statistics*. Recall that it is common for some projects to require more than one type of method.

Our relative is running for a seat on the city commission and has hired us to poll a random sample of voters about their concerns. Specifically, she wants a profile of the electorate that will tell her:

- what percentage belong to each political party,

- what percentage are male or female, and

- what percentage favor or oppose the widening of the main street in the city.

Answer:

Exercise 2.7-14: Descriptive statistics allow social researchers to

1. quantify the strength and direction of the relationship between two variables,

2. reduce thousands of individual numbers into a few easily understood numbers.

3. graphically display the number of respondents according to their gender.

4. all of the choices

Answer:

Chapter 3

Location and Spread

3.1 Objective

After completing this part, students should be able to:

- Explain the concept of measures of central tendency and interpret the information they provide.

- Calculate, describe and compare the three most commonly used measures of central tendency: means, medians, and modes.

- Understand other measures of central tendency.

- Select appropriate measure of central tendency based on the level of measurement and characteristics of the data (distribution).

- Explain the concept of measures of dispersion and the information they convey.

- Compute and explain the following:
  - The range ($\mathbf{R}$).
  - The interquartile range ($\mathbf{IQR}$)
  - The sample variance (denoted as s^2).
  - The population variance (denoted as σ^2).
  - The sample standard deviation (denoted as s).
  - The population standard deviation (denoted as σ).

- Select the measure of dispersion appropriate for our current problem.

3.2 Estimates of Center

Suppose that a random sample of two-bedroom apartments in the Kalamazoo area yields the following data on monthly rent:

$635, $525, $500, $800, $650, $750, $555, $500, $670, $675

How much would we say is the average rent for two-bedroom apartments in Kalamazoo? In this chapter, we will discuss the sample average and several alternatives to the average when estimating 'average value' in a population.

3.2.1 The Sample Mean

The sample mean is the statistical term for 'average of the sample.' For the example above, the sample mean (denoted $\bar{X}$) is: The sample mean is the statistical term for 'average of the sample.' For the example above, the sample mean (denoted $\bar{X}$) is:

$$\bar{X} = \frac{635 + 525 + 500 + 800 + 650 + 750 + 555 + 500 + 670 + 675}{10} = \$626$$

so that the average rent in Kalamazoo may be estimated as $626. Note that this is an *estimate* based on a *sample*. Therefore, it is subject to *sampling error*. Sampling error means that due to the luck of the draw, the sample average likely missed the true population average. More precisely, the two-word term 'sampling error' refers to $|\bar{X} = \mu|$, the distance by which the *sample* average $\bar{X}$ misses the population average μ.

The advantages of the mean:

- It is easy to understand and simple calculate.

- It is based on all the values.

- It is not based on the position in the series.

The disadvantage of the mean:

- It is always affected by small and large values.

3.2.2 The Sample Median

There are alternative ways to estimate average rent. We can compute the sample median, instead of the sample mean. We have discussed the sample median (denoted MED) earlier in Chapter 1. It is computed as follows.

1. Let n represent the total number of observations.

2. Order the n observations from smallest to largest.

3. Then calculate $0.50(n + 1)$.

4. If $0.50(n + 1)$ is an integer, then MED is the $0.50(n + 1)st$ ordered observation.

5. If $0.50(n + 1)$ is not an integer, then MED is the average of the two adjacent ordered observations.

6. There are $n = 10$ observations in our rental data. We first order them from smallest to largest.

$$500, 500, 525, 555, 635, 650, 670, 675, 750, 800$$

Now, $0.50(n + 1) = 0.50(10 + 1) = 5.5$. Since this is not an integer, we average the 5th and 6th largest observations:

$$MED = \frac{(635 + 650)}{2} = \$642.50$$

The advantage of the median is that it is more robust than the mean, i.e., the median is not as affected by extreme values (outliers).

3.2.3 The Trimmed Mean

Since the mean uses all observations in the calculation, it can be strongly affected by outlying small and large values. What happens to the mean when the smallest value \$500 is replaced by \$400? It will become smaller. What happens to the median? It remains unchanged.

The trimmed mean is a compromise estimator that looks a lot like a mean but is less sensitive to extreme values. The 10%-trimmed mean removes the lowest 10% and highest 10% of the data, then take the sample mean of the remaining data. In the rental example, 10% of the data is one observation. We remove the lowest observation (\$500) and the largest observation (\$800), and take the mean of the remaining eight observations:

$$\bar{X} = \frac{635 + 525 + 650 + 750 + 555 + 500 + 670 + 675}{10} = \$620$$

What if 10% of the data is not an integer? For example, if $n = 23$, then 10% of n is 2.3. Since we cannot remove 2.3 observations, we will remove three observations from each end (to ensure at least 10% protection) and take the average of the middle 17 observations.

3.2.4 The Median of Pairwise Averages

The median of pairwise averages is another compromise between the mean and median. As the name implies, we replace the observations by pairwise averages, then take the median of those. Here again are the ten ordered observations:

$$500, 500, 525, 555, 635, 650, 670, 675, 750, 800$$

The first observation (\$500) is paired with all ten observations (including itself). The second observation (also \$500) has already been paired with the first observation, so it paired with only nine observations (including itself). The third observation (\$525) has already been paired with the first two observations, so it is paired with only eight observations (including itself). Proceeding in this pattern, we get the following pairwise averages:

$\frac{500+500}{2}$	$\frac{500+500}{2}$	$\frac{500+525}{2}$	$\frac{500+555}{2}$	$\frac{500+635}{2}$	$\frac{500+650}{2}$	$\frac{500+670}{2}$	$\frac{500+675}{2}$	$\frac{500+750}{2}$	$\frac{500+800}{2}$
	$\frac{500+500}{2}$	$\frac{500+525}{2}$	$\frac{500+555}{2}$	$\frac{500+635}{2}$	$\frac{500+650}{2}$	$\frac{500+670}{2}$	$\frac{500+675}{2}$	$\frac{500+750}{2}$	$\frac{500+800}{2}$
		$\frac{525+525}{2}$	$\frac{525+555}{2}$	$\frac{525+635}{2}$	$\frac{525+650}{2}$	$\frac{525+670}{2}$	$\frac{525+675}{2}$	$\frac{525+750}{2}$	$\frac{525+800}{2}$
			$\frac{555+555}{2}$	$\frac{555+635}{2}$	$\frac{555+650}{2}$	$\frac{555+670}{2}$	$\frac{555+675}{2}$	$\frac{555+750}{2}$	$\frac{555+800}{2}$
				$\frac{635+635}{2}$	$\frac{635+650}{2}$	$\frac{635+670}{2}$	$\frac{635+675}{2}$	$\frac{635+750}{2}$	$\frac{635+800}{2}$
					$\frac{650+650}{2}$	$\frac{650+670}{2}$	$\frac{650+675}{2}$	$\frac{650+750}{2}$	$\frac{650+800}{2}$
						$\frac{670+670}{2}$	$\frac{670+675}{2}$	$\frac{670+750}{2}$	$\frac{670+800}{2}$
							$\frac{675+675}{2}$	$\frac{675+750}{2}$	$\frac{675+800}{2}$
								$\frac{750+750}{2}$	$\frac{750+800}{2}$
									$\frac{800+800}{2}$

There will be 55 pairwise averages in all. Notice that the sample size, n, went from ten observations to 55 observations after we do this. We now need to figure the median and remember, the MED is the $0.50(n + 1)st$ ordered observation. So, we need to use 55 now as our n. It gives us a value of 28. We need order the pairwise averages we have calculated above, and the median will be the 28th observation. The median of these averages is thus 625 dollars.

3.3 Estimate of Spread (or Uncertainty or Variation)

Recall the data on the monthly rent of two bedroom apartments in the Kalamazoo area:

$$500, 500, 525, 555, 635, 650, 670, 675, 750, 800$$

If a future student asked us "What should a two-bedroom apartment cost us in rent?" how should we reply? Knowing that the data averages $626, we might say something like "Around $626, give or take ... (?)." This second number, the give or take, is important because it says how much uncertainty there is in our guess. In other words, the student's rent will probably miss $626, but by how much? For a second example, if we were going to San Francisco for a couple of days in August, and we want to know what clothes to bring, it is not enough to know that the average temperature is 68 degrees. If it was 68 degrees, give or take 15 degrees, then we will need a sweater. If it is 68 degrees, give or take 30 degrees, we might need a winter coat.

Variation presents itself everywhere. Consider weight loss. The 77 subjects in the Atkins group lost an average of 10.5 pounds 'give or take' 15 pounds. Compare this to an average of 3.5 pounds for the Zone diet 'give or take' 14 pounds. How about in bowling? In his first seven games in a tournament in Indiana, Walter Ray Williams Jr. averaged 228, 'give or take' 34. Notice how the 'give or take' number seems to complete the description.

In this chapter, we will discuss the sample standard deviation, which is typically used as the 'give or take' number to describe spread in a group of numbers.

3.3.1 The Sample Standard Deviation

The standard deviation (or SD) for the monthly rent data is computed using a series of steps. Notice in the last row we have the average, the sum of the deviations from the mean, which will always be 0, and the Sum of Squares (SS) which is the sum of the squared deviations.

$$SD = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{9.764 \times 10^4}{9}} \approx 104.2$$

It is helpful to get a feel for the numbers in the second column. The first value 9 says that '635 is nine above average'. The next number -101 says that '525 is 101 below average.' Ignoring signs, the numbers in the second column represent the 'give or take' from the average. To get the actual SD, we take the Sum of Squares and divide by $(n-1)$, and last, of all take the square root of that answer. It gives us a final SD of 104.158. We say that monthly rent averages $626, 'give or take' $104.

	Rent	Deviations	SS
1	635	9	81
2	525	-101	10201
3	500	-126	15876
4	800	174	30276
5	650	24	576
6	750	124	15376
7	555	-71	5041
8	500	-126	15876
9	670	44	1936
10	675	49	2401
Sum	6260	0	97640

Look at the second column again. Ignoring signs, these are the distances from the average. Which rental is closest to the average? The answer is $635, which missed by 9. Which rental is farthest from the average? $800, which missed by 174. What is the average 'size of the miss?' It should be somewhere between the smallest and largest, right? It is how we interpret the SD: "On the average, the monthly rentals miss the mean by $104."

Taken together, the mean and SD allow comparison of both relative size and spread of two groups of numbers. In the professional bowling tournament in Indiana that he won in November of 2008, here are Walter Ray Williams Jr.'s first seven games and last seven games (including the final):

Games		Mean	SD
First seven	163,231,224,238,279,239,226	228.6	34.3
Last seven	246,244,247,248,237,258,246	246.6	6.2

What do the mean and SD tell us? He averaged higher in the end but was also more consistent with a give-or-take of only 6 points! His earlier games had more massive swings: from a low of 163 to a high of 279, resulting in the SD of 34.

3.3.2 Effect of Multiplication and Addition by a Constant

Recall that monthly rent for apartments average $626 with an SD of $104. If the student plans to get a roommate and pay only half the rent, how much does he expect to pay? If we are thinking of dividing both numbers by two, i.e., $313 give or take $52, this is correct.

$$\text{Original rent: } 626 \pm 104 \qquad\qquad \text{Half the rent: } 313 \pm 52$$

Now suppose that the student does not plan to get a roommate, but his parents have agreed to contribute $100 to rent each month. How much does the student expect to pay after a subsidy of $100? If we are thinking of subtracting $100 from both numbers, i.e., $426 'give or take' $4, there is an error in our thinking. Here the SD remains the same, i.e.,

Original rent: 626 ± 104 Subsidized rent: 526 ± 104

If we are not convinced, consider the data itself on the following table.

	Rent	Deviations	SS
1	317.5	4.5	20.2
2	262.5	-50.5	2550.2
3	250.0	-63.0	3969.0
4	400.0	87.0	7569.0
5	325.0	12.0	144.0
6	375.0	62.0	3844.0
7	277.5	-35.5	1260.2
8	250.0	-63.0	3969.0
9	335.0	22.0	484.0
10	337.5	24.5	600.2
Sum	3130.0	0.0	24410.0

Table 3.1: Calculating the SD when rent is halved

$$SD = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{2.441 \times 10^4}{9}} \approx 52.08$$

In general, when the data is **multiplied or divided by a positive constant**, the **same thing happens** to both the average and the SD

$$SD = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{9.764 \times 10^4}{9}} \approx 104.16$$

In general, when we add or subtract a **constant** to the data, the same thing happens to both the average, but the **SD remains unchanged.**

	Rent	Deviations	SS
1	535.0	9.0	81.0
2	425.0	-101.0	10201.0
3	400.0	-126.0	15876.0
4	700.0	174.0	30276.0
5	550.0	24.0	576.0
6	650.0	124.0	15376.0
7	455.0	-71.0	5041.0
8	400.0	-126.0	15876.0
9	570.0	44.0	1936.0
10	575.0	49.0	2401.0
Sum	5260.0	0.0	97640.0

Table 3.2: Calculating the SD when rent is reduced by 100 dollars

3.4 Exercises

—————

Exercises

Exercise 3.4-1: The carbon monoxide measures (in mgs) are given on 25 brands of cigarettes.

13.6 16.6 23.5 10.2 5.4 15.0 9.0 12.3 16.3 15.4 13.0 14.4 10.0 10.2 9.5 1.5 18.5 12.6 17.5 4.9 15.9 8.5 10.6 13.9 14.9

Calculate the

1. Mean

2. median

3. 10% trimmed-mean

4. standard deviation

5. mean and standard deviation (in grams)

Exercise 3.4-2: Calculate the median of pairwise average for the first 6 observations which are listed below:

13.6 16.6 23.5 10.2 5.4 15.0

Exercise 3.4-3: Daily high temperature for a given day are provided for the past 10 years

2007	2008	2009	2010	2011
59	50	49	13	41
2012	2013	2014	2015	2016
46	51	53	58	47

Find the following statistics for temperature:

1. Range

2. Mean

3. Median

4. Mode

5. Remove the temperature of the year 2010 from the data set and re-calculate the Mean and compare it with part (3)? Is the mean closer to the median and why?

Exercise 3.4-4: Joshua has been working on programming and updating a Website for his company for the past 12 months. The following numbers represent the number of hours Joshua has worked on this Website for each of the past 12 months:

24, 25, 31, 40, 48, 40, 36, 50, 38, 35 ,42, 112

1. Calculate the mean

2. Calculate the median

3. Decide if its symmetric, skewed to the right or to the left

4. Decide which measure of center provides the most relevant information about the distribution? Why?

Exercise 3.4-5: In a study of warp breakage during the process of weaving fabric (*Technometrics*, 1982: 63). Ten specimens of yarn were tested. The number of cycles of strain to breakage was determined for each yarn specimen, resulting in the following data:

$131, 195, 137, 42, 337, 829, 149, 55, 180, 203$

1. Calculate the mean

2. Calculate the median

3. Calculate the standard deviation

4. Decide if its distribution (symmetric, skewed to the right or to the left)

5. Decide which measure of center provides the most relevant information about the distribution? Why?

Exercise 3.4-6:   The weekly budgets for groceries of six students are as follows:
$30, $35, $40, $28, $35, $25.
Compute the mean, median, and mode, respectively.

1. 188, 30, 40

2. 32.167, 35, 35

3. 32.167, 32.5, 35

4. 188, 30, 35

Answer:

Exercise 3.4-7:   John got 16 and 22 on his two Stat 2830 quizzes. What score must he have on his next quiz to have the mean of exactly 20 for his three quizzes?

1. 20

2. 22

3. 18

4. 24

Answer:

Exercise 3.4-8:   The following table lists the average number of cars per 1000 population for eight nations. Compute the three central tendency variables (mean and median) for these data.

Nation	cars per 1000 population
United States	820
Canada	607
China	83
Russia	293
Japan	591
Mexico	275
Spain	593
United Kingdom	519

1. Which is greater in value?

2. Is there a positive skew in the data?

3. How do you know?

Answer:

Exercise 3.4-9:   You are a researcher for a mid-sized city. Using Table 3.8, you have collected six types of variables from a random sample of students from a large university. These variables include:

Select the appropriate measures of central tendency for each variable:

1. their region of birth of country

2. the extent they support marijuana legalization (7=strong, 4=neutral, 1=weak)

3. the weekly amount of money spent on cafeteria food

4. number of movies they watched per week

5. quality of cafeteria food at their university (10=excellent, 0=bad)

6. religious affiliation.

Answer:

Table 3.3: Find appropriate measures of central tendency

Student	Birth	Expense	Movies	Food	Religion
a	West	43	4	6	Cath.
b	MW	51	3	5	Other
c	South	65	14	0	Other
d	South	52	0	10	Prot.
e	North	48	1	6	Jew
f	North	62	5	8	Prot.
g	MW	47	7	1	None
h	South	45	10	2	Cath.
i	North	39	14	7	Prot.
j	North	33	0	10	Prot.

Exercise 3.4-10: Using the dataset from Table 3.3, determine the value of the central tendency measures for each variable.

Answer:

Exercise 3.4-11: For the next two questions, use the following scenario. As a leader of a Kalamazoo social services agency that employs 20 staff members, you are concerned that your staff has an increased caseload of clients compared to 10 years ago. The case load of each worker is reported in the following table (Table 3.9) for years 2005 and 2015.

2005	50	64	55	64	60	53	56	50	51	45
	46	46	50	56	63	65	53	54	58	69
2015	57	47	46	59	59	50	57	52	41	66
	65	52	57	75	65	76	43	56	52	67

Table 3.4: Central Tendency

Has the average caseload increased, decreased, or stayed the same?

Answer:

Exercise 3.4-12: The admissions department at WMU gave 25 randomly selected freshmen a national prejudice survey. The racial prejudice index scores will be used in three years to see if higher education affects prejudice.

45	30	35	30	42
50	43	40	32	48
9	13	10	11	11
40	26	39	38	44
32	37	41	27	22

Table 3.5: Freshmen Racial Prejudice Index

Calculate the mean and median scores of these data.

Answer:

Exercise 3.4-13: The same 25 students completed the same survey during their senior year. Compute the mean and median for this second set of scores, and compare them to the scores from four years earlier. What happened?

50	27	31	35	41
11	45	50	37	43
11	9	10	20	10
35	10	30	41	40
15	30	40	26	21

Table 3.6: Senior Racial Prejudice Index

Calculate the mean and median scores of these data.

Answer:

Exercise 3.4-14: The next two questions refer to the following scenario. A local social service agency has started a sex education course for teen girls. The girls took a 20-question exam for general information about sex, anatomy and physiology upon entry and again after completing the course. Table 3.7 has the listing of scores of a random sample of 15 girls.

Calculate the mean and median of the pre and post test scores and comment on the results.

Answer:

Table 3.7: Pre and Post Test Results

case	posttest	pretest	difference
1	12	8	4
2	13	7	6
3	12	10	2
4	19	15	4
5	8	10	-2
6	17	10	7
7	12	3	9
8	11	10	1
9	7	5	2
10	12	15	-3
11	21	13	8
12	5	4	1
13	15	10	5
14	11	8	3
15	20	12	8

Chapter 4

Threats to Valid Comparisons

4.1 Objective

After completing this part, students should be able to:

1. Understand the difference between association and cause.

2. How confounders can affect the results of a survey.

4.2 Hidden Confounder

In 1992, a research study "Lower extremity fractures in motor vehicle collisions: Influence of direction of impact and seatbelt use" by Dischinger, P., Cushing, B. and Kerns, T. was published in the 36th Proceedings of the Association for the Advancement of Automotive Medicine. It involved data analysis of the trauma-center population in Maryland. Some of the conclusions were:

1. there was a higher incidence of lower extremity injury in frontal collisions,

2. seatbelt use was not effective in preventing lower extremity fractures, and

3. there was a higher incidence of lower extremity fracture among women.

The conclusion (3) is exciting. It begs the follow-up question: "Why do women have higher rates of leg fractures?" Is it because they drive faster, or apply brakes more slowly, or have weaker bones? It turns out that these are false questions – they presume that gender is the variable that causes higher leg fractures.

Researchers proposed an explanation in a follow-up study with the same lead author: "Lower extremity fractures in motor vehicle collisions: The role of driver gender and height" by

Abstract: In a previous study it was noted that there was a higher incidence of lower-extremity fractures among women drivers. Analyses were based on a linkage between trauma registry and police crash report data. The present study addresses the issue of whether the differences noted are attributed to driver gender or are merely a reflection of differences in driver height. An inverse association was noted between driver height and the incidence of lower-extremity fractures. Those shorter than average (5'7") for this population had a 64% increase in a lower-extremity fracture, which can be mainly attributed to ankle/tarsal injuries. Thus, the incidence of these injuries appears to be a function of driver height, with an increase among shorter drivers, most of whom are women.

It turns out that height was the culprit, but it initially looked like gender because height and gender have a strong link. The following *pathway graph* describes the true relationship:

$$\text{Height} \implies \text{Leg Fracture}$$
$$\updownarrow$$
$$\text{Gender}$$
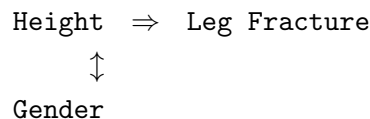
But the association between height and gender led us to believe the wrong relationship:

$$\text{Gender} \implies \text{Leg Fracture}$$

The symbol $\Rightarrow$ represents "cause-and-effect" while $\updownarrow$ represents "association."

In this situation, leg fracture rate is the outcome variable gender is the probable cause. The hidden variable height is called a **confounder**, or a confounding variable.

Confounding Variable:
A confounder or confounding variable is a third variable that is associated with both the probable cause and the outcome. It can lead us to a wrong conclusion about the cause-and-effect relationship.

4.2.1 Apples and Oranges

Now, why do unknown confounders belong in a chapter on threats to valid comparisons? They are one of the most significant sources of (often unknowingly) invalid comparisons. It seemed fair to compare leg fracture rates of men and women, didn't it? What's wrong with that? Unfortunately, concerning leg fracture rates, it was a case of comparing apples to oranges – women as a group is shorter than men! Of course, in the 1992 study, the investigators did not know that height would make a difference. Women also tend to weigh less, smoke less, drink less, have longer hair, and wear higher heels. Which of these would make a difference, i.e., be potential confounders?

Confounders are a big problem in comparison studies. Some confounders may remain hidden, but it is critical that the researcher identify and control for potential confounders as much as possible. Does smoking cause lung cancer? When comparing smokers to nonsmokers, it is easy to show that smokers have higher lung cancer rates. But as a group, they exercise less than nonsmokers. They drink more coffee than nonsmokers. Can the exercise or the coffee or the combination be the culprit? Furthermore, smokers tend to be male, older, and drive in the winter with their car windows open. There are plenty of confounding variables even in this comparison.

4.2.2 In the news

The STATS website (http://www.stats.org/) is dedicated to correcting "scientific misinformation in the media and public policy resulting from bad science, politics, or a simple lack of information or knowledge." The following is an excerpt from an article written by Rebecca Goldin and Jing Peng in August 2010. What is the confounding variable? The probable cause? The outcome variable?

> **If you take Viagra, will you get an STD?**
> Rebecca Goldin Ph.D. and Jing Peng, August 2, 2010
>
> Judging from recent headlines, it seems clear: "Sex Diseases Tripled in Men 40 or Older Taking Viagra, Cialis, Study Says" reports Bloomberg; "Older Viagra Users More Likely to Get STDs" says the Chicago Sun-Times, presumably comparing older Viagra users with older non-Viagra users. And Health Day was even more explicit, saying "Drugs Like Viagra Linked to Higher Rates of STDs." The next logic behind the claim seems persuasively apparent: if men who take Viagra are having more sex, then they have inevitably increased their risk of catching a sexually transmitted disease (STD). But is this the case?
>
> The study compared men who took erectile dysfunction (ED) drugs with those who did not. Researchers found that ED drugs such as Viagra are linked to higher rates of STDs among older men, but not older women, especially after the

introduction of Viagra in 1998. A study published in the July 6 issue of the Annals of Internal Medicine was the first to examine the relationship between ED drugs and STDs. But its findings turn out to be far different than media accounts would have you believe.

The main problem in the coverage is the direct suggestion that taking Viagra is associated with STDs, as opposed to being the sort of person who takes Viagra.

There is a critical distinction. To see what it means, let's go backward in time and compare STD rates among people who plan to take Viagra but haven't and people who are not planning on taking Viagra. This quick comparison is what the authors of the research article did; they combed medical records for people who filled prescriptions for ED drugs and compared STD rates for these people before filling their prescription to the STD rates of people who did not subsequently fill a prescription for Viagra. It turns out that the rate of STDs is higher among people who intend to take it. In other words, the drug is absent, but those people who will take Viagra within a year are already at higher risk of STDs.

In fact, compared to those who don't take ED drugs, those who plan to take Viagra had a slightly higher rate for STDs, an odds ratio (OR) of 2.80; 95% confidence interval CI, 2.10 to 3.75, than those who actually take it (OR) 2.65, CI, 1.84 to 3.81, though the difference was not significant. It means that the drug had no discernable effect on STD rates for this group of men.

The paper goes on with detailed analysis, but the main point has been stated. To summarize the conclusions of the study:

1. Taking ED drugs do not increase the rate of STD's

2. The type of people who take ED drugs are different from people who do not.

$$\text{Person type} \quad \Rightarrow \quad \text{STD}$$
$$\updownarrow$$
$$\text{ED drugs}$$
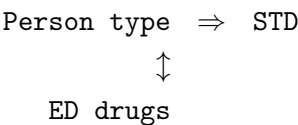
instead of the (admittedly more sensational) relationship implied by the headlines.

4.3 Key Words

• Association	• Confounders
• Cause	• Effect

4.4 Exercises

Exercises

Exercise 4.4-1: A study says, "Slightly overweight people live longer than thin people."

1. The headline implies cause-and-effect, not just association, i.e. "if you are thin, you should try to gain weight." When comparing lifespan of slightly overweight and thin people, can you think of possible confounders?

2. Using one confounder from your answer(s) in (1), draw a pathway graph depicting the possible relationship between confounder, possible cause, and outcome.

Exercise 4.4-2: Generally, a study that appeared in the American journal of cardiology (March 15, 2003) found out that "heart attack survivors who owned a dog had better heart function post-heart attack than those who did not own a dog."

1. Does a dog help a heart heal faster? Can you think of possible confounders?

2. Using one confounder from your answer(s) in (1), draw a pathway graph depicting the possible relationship between confounder, possible cause, and outcome.

Exercise 4.4-3: A study says, "People who consider themselves depressed eat more chocolate than people who consider themselves otherwise."

1. The headline implies cause-and-effect, not just association, i.e., "if you are depressed, you will tend to eat more chocolate." When comparing chocolate consumption of depressed versus not-depressed, can you think of possible confounders?

2. Using one confounder from your answer(s) in (1), draw a pathway graph depicting the possible relationship between confounder, possible cause, and outcome.

Chapter 5

Study Designs

5.1 Objective

After completing this part, students should be able to:

1. Grasp the basic of good experimental design, especially of confounders

2. Understand the importance of randomized trials (RT).

3. Comprehend the meaning of double-blind, randomized controlled trials (RCT).

4. Discern the observational studies.

5. Perceive the need for case-controlled studies.

6. Grasp the need for case-crossover studies.

5.2 Randomized trials

In the previous chapter, we covered unknown confounders, one of the most critical threats to valid comparisons in statistics. In this chapter, we will explore elements of study design that help statisticians deal with the danger of bias posed by hidden confounders. One of the most important aspects of many study designs is appropriate randomization in the selection of subjects for the study, and to different groups within the study. Without such randomization, results of data analyses can be biased. Let's see by way of example how randomization can help us address the threat of results biased by hidden confounders.

Recall the diet study discussed previously: "Comparison of the Atkins, Zone, Ornish, and LEARN Diets for Change in Weight and Related Risk Factors Among Overweight Premenopausal Women The A TO Z Weight Loss Study: A Randomized Trial" by C.D. Gardner, et al. (JAMA, Vol. 297, pp. 969-77, March 2007)

> **Objective:**
> To compare four weight-loss diets representing a spectrum of low to high carbohydrate intake for effects on weight loss and related metabolic variables. Design, Setting, and Participants Twelve-month randomized trial conducted in the United States from February 2003 to October 2005 among 311 free-living, overweight/obese (body mass index, 27-40) nondiabetic, premenopausal women. Intervention participants were randomly assigned to follow the Atkins (n=77), Zone ($n = 79$), LEARN ($n = 79$), or Ornish ($n = 76$) diets and received weekly instruction for 2 months, then an additional 10-month follow-up.

Here is an example of a randomized trial or randomized experiment. Subjects entering the test are randomized, using a virtual roll of a die, into one of several treatment groups.

The randomization into groups eliminates the main threat to valid comparisons: confounders. How? By making the comparison groups similar to each other in all aspects, except for the treatment. Think about it. If you ask for volunteers to each treatment group, then most of the meat-eaters would go to the Atkins group, and most of the vegetarians would go to the Ornish group. Would it then surprise you if one group had, say, more Asians than another group? Or more smokers? Or less physical activity? Randomized assignments, instead of volunteering, is the best way to get balanced groups in the factors that we think might matter and achieves balance in all other factors we have not even considered.

Table 5.1 on page 64 presents the characteristics of the subjects after randomization, but before treatment has started. Observe how randomization has achieved a right balance between groups in demographic and anthropometric variables, and known health risk factors. All other variables, essential or not, like IQ, shoe size, and preference for country music would also tend to be balanced, thus providing a level of protection against unknown confounders.

> To safeguard against potential **confounders**, comparison groups should be **similar** in all factors except for treatment itself. **Randomization** is the best way to achieve the goal.

5.2.1 Double-blind randomized controlled trials (RCT)

Scientific studies should be randomized as described above whenever possible. But it might surprise you to learn that randomization is just one way to address hidden confounders and

bias in clinical trials. The present section will cover *controls and blinding*, two critical elements in study design that address confounders and bias.

Thus far, we've discussed how vital randomly assigning subjects in a study to groups within that study is to overcoming bias, but study designers should be aware that there is some luck involved in randomization. It is not always likely, but it is possible, that even the best randomization techniques leave unknown confounders present across all *subjects* within the study. We may be selecting subjects for the study from a population that all have a confounding characteristic in common. In such a case, it would not help us much to randomly assign subjects to different groups of the study since all the groups would end up exhibiting the same bias regardless as to how we assigned the subjects. Imagine attempting to conclude whether a medication is effective in treating the common cold by studying only subjects who have some particular resistance to the cold. No patients would get severe colds, or if they did, they would not manifest many cold symptoms like coughing, sneezing, or a runny nose. Sounds hard to study such subjects, right? Yes, the immune response confounder makes these subjects hard to explore, but this problem is solvable. Scientists need more *control* in assessing such matters and their hidden confounders, so control *groups* are added to scientific studies to correct for this problem.

Control groups are groups of subjects within the study that are not expected to exhibit the only associations being studied. For instance, control groups in medical studies are often groups of patients given inert placebos or "sugar pills" instead of active medications. Since the study directors do not provide these groups with the active medication, such groups are not expected to exhibit effects as strong as the treatment effects of those that do receive medications. We wouldn't expect patients who receive sugar pills to miraculously feel relief from their cold symptoms in the way patients who receive a new alternative to pseudoephedrine might. Even when such placebo groups do exhibit effects like the expected treatment effects, which can be due to the placebos themselves (placebo effect) or other hidden confounders (different resistance to the cold), the presence of these control groups within the study gives researchers a valid point of comparison. It is true that both the placebo and medicated groups selected from our cold-resistant population of patients would be asymptomatic. However, even though nobody is sneezing too much, comparing the medicated and unmedicated groups allows scientists to determine if it's all just the patients' robust immune responses, if the psychological impact of taking a pill is responsible, or if the medication is exceptionally effective at treating the cold.

When one of the comparison groups in a study is a control group or placebo, randomized trials are sometimes called *randomized trials*, or randomized controlled experiments. As much as possible, researchers try to conduct *double-blind* randomized controlled trials, where neither doctor nor patient knows what treatment the patient is receiving. In the common cold example, some patients get a sugar pill or placebo while others get a new alternative medication thought to treat cold symptoms. The critical point to remember here is that in a double-blind

This is a line graph comparing the average effects of a medicine and placebo on two groups.
Is there enough of a difference between medicine and placebo to market the drug as effective?
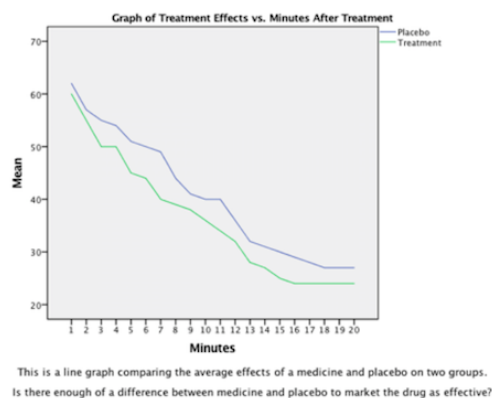
controlled experiment both the *patients* and *doctors* have no idea if a patient is receiving the sugar pill/placebo or the cold medicine. This method gives each patient a pill that looks identical regardless as to whether it's the inert placebo or the active drug.

Figure 5.1: **Placebo** or **Medicine**. Can you spot the difference? – Single Blinding. Can your doctor? – Double Blinding.

This double-blinding is essential because it offers additional protection against bias. In such studies, all groups have the same frame of mind (as opposed to knowing you are not getting the new drug). Similarly, the experimenter has the same frame of mind evaluating patients from each group. You might wonder, isn't it enough to single-blind the study, to make sure that just the patients are unaware whether they received the placebo or medicinal pill? No, it is not. Imagine an experimenter who administers both the placebos and medications but does so know which is which. That researcher might unintentionally treat these two groups differently, perhaps just by spending an extra minute with patients from the medicated treatment group or by approaching them with a little more enthusiasm. It's this sort of unwitting behavior that can invite bias into the study.

5.3 Observational studies

In many cases, when researchers, who want to compare Rx effects, cannot assign treatments to subjects, they cannot be randomized. The study involving women and leg fractures is an example. We were comparing leg fracture rates of two comparison groups, men, and

women. Whenever a new subject enters the study (by having a car accident), we observe what gender they belong to, instead of randomly assigning it. Studies like these are called observational studies, as opposed to randomized experiments. In the hierarchy of scientific evidence, observational studies are not as reliable as randomized trials, which is the gold standard. Since the subjects of observational studies assign themselves to their group, there may be a selection bias that leads to confounders (like the women in the leg-fracture-study being shorter). We can control the effects of known confounders in the analysis. For example, we can compare leg fracture rates of men and women with the same heights. Investigators will need to anticipate potential confounders and control for them. We enumerate common reasons for nonrandomized studies:

1. Assigning treatment is impossible (e.g., to compare fracture rates between men and women, we cannot randomize subjects into the comparison groups).

2. Assigning treatment is unethical (e.g., to compare cancer rates of smokers and nonsmokers, we do not want to randomize subjects into smoker-nonsmoker comparison groups)

3. Assigning treatment is impractical (e.g., the outcome is a rare event like cancer or stroke, and a randomized trial would need too many subjects and too much time). In cases like these, a case-control study is generally the way to go.

> **Observational** studies are conducted when **randomization** to treatment groups is **impossible**, **unethical**, or **impractical**.

Returning to the leg-fracture example, we stick to observational studies in such a case for two reasons. As described above, it's impossible to perform the researcher any other way with the data collected. But we also *should not* elect to change the study, to collect the data in a different and randomized manner. In other words, we should not even try to get the bias-reducing benefits of randomization because of ethical constraints. To perform randomized clinical trials in this study of leg fractures, we would need to randomly select subjects whose legs *would* be broken for the sake of the study. Causing suffering like this is prohibited by several codes of ethics like the Nuremberg Code, a system of research ethics drafted after the Nuremberg trials of Nazi war criminals. These criminals included scientists who designed and performed inhumane experiments. Beyond the official ban, though, such research is wrong in its causing undue suffering and its infringement upon fundamental human rights.

Another reason we may conduct an observational study is that they can be more practical compared to randomized controlled trials. It can be due to the constraints of a rare disease, or the cost of a randomized study would entail. A researcher might want to study whether regular exercise can prevent a rare form of cancer. Randomization and control here would require that we randomly select patients first, group them along exercise programs (say regular, irregular, and no exercise), and wait to see how the different groups respond (do the exercise groups have significantly fewer cases of cancer?). This study might not work out at all because, since

the cancer is rare, it's possible that no or very few of our subjects will get cancer. It also may take many years before cancer would appear.

This study would be quite time-consuming and labor-intensive. Imagine all the financial resources that would be required to do all that exercise-monitoring over all the patients and years! This experiment brings us to our second point that observational studies are practical because they tend to be less costly than randomized controlled trials. Companies often use statistics to generate revenue, but they tend to do this not by randomized experiments, but by analyses of marketing data they collect through sales or surveys. Nonrandom observational studies like analyses of sales data are a relaxed approach in that all they require are a data collection apparatus and a statistician. For instance, clothing designers want to keep up with trends in consumer spending but might sell hundreds of different garments to thousands of outlets across the world. When sales of one item begin to slip, it might be difficult to notice amidst all the chaos of such a business enterprise. A statistician can help here by analyzing the proportions of revenue generated by each garment from month to month and making recommendations about which products to push when. The statistician might even find a style or color-based patterns that repeat themselves seasonally, allowing the company to adjust before their sales ever begin slipping.

5.3.1 Case-control studies

Instead of randomizing subjects into diet groups and then comparing the weight loss outcome, a case-control study would look for people in the population who lost weight, and then ask them what diet they used. Thus, you start with the *outcome* and then work back to the type of treatment. These are also classified as *retrospective* studies because they look back and compare weight loss or disease rates of various procedures. Randomized experiments are necessarily *prospective*, in the sense that you randomize treatment and then later see which groups lost more weight or had more disease.

We frequently use case-control studies because they are cheaper and easier to conduct since it generally requires a survey of a database, instead of expensive and time-consuming recruitment and handling of subjects. There are plenty of successful case-control success stories in the scientific literature.

The first study formally linking lung cancer to smoking was a 1950 case-control study "Smoking and Carcinoma of the Lung" by Richard Doll and A. Bradford Hill(British Medical Journal, 1950 September 30; 2(4682): page 739–748). Using patients in 20 hospitals in London, they found that lung-cancer patients had higher rates of smokers than a comparable control group of patients. For example, among district hospitals, 48 out of 98 lung cancer patients smoked 15 or more cigarettes daily. In contrast, only 30 out of 98 non-cancer patients smoked 15 or more cigarettes daily.

In general, case-control studies can conclude a link or 'association' but are not able to prove 'causation.' However, case-control studies provide initial evidence that can generate resources for more rigorous studies like double-blind, randomized controlled trials. In the case of smoking and lung cancer, randomized trials are unethical, but given the eventual results of multiple studies, the scientific community now accepts that smoking causes lung cancer.

5.3.2 Case-crossover studies

Sometimes, a subject can be its control. This method is called a case-crossover study because of subjects in the treatment group "crossover" to the control group. An example is a 1997 study linking cell phone use to car accidents: "Association between cellular-telephone calls and motor vehicle collisions" by D.A. Redelmeier and R.J. Tibshirani (The New England Journal of Medicine, 1997 Feb 13; Vol 336, pp. 453-8). The subjects were people who reported a collision to the North York Collision Reporting Centre between July 1, 1994, and August 31, 1995, Among these, 742 had cell phones and consented to participate in the study. Instead of asking each person whether they were using their cell phone during the time of the collision (an unreliable method), the investigators examined their detailed phone billing records. The results of the experiment follow below:

> Overall, 170 subjects (24 percent) had used a cellular telephone during the 10-minute period immediately before the collision, 37 (5 percent) had used the telephone during the same period on the day before the collision, and 13 (2 percent) had used the telephone during both periods. The crude analysis indicated that cellular-telephone activity was associated with a relative risk of a motor vehicle collision of 6.5 (95 percent confidence interval, 4.5 to 9.9). The primary analysis, adjusted for intermittent driving, indicated that cellular-telephone activity was associated with a quadrupling of the risk of a motor vehicle collision (relative risk, 4.3; 95 percent confidence interval, 3.0 to 6.5).

5.4 Key Words

- randomized trial
- double-blind randomized trial
- observational studies
- case-control studies
- case-crossover studies

Table 5.1: Baseline Participant Characteristics

	Atkins	Zone	LEARN	Ornish
n	77	79	79	76
Demographics				
Age, y	42(5)	40(6)	40(7)	42(6)
Education, y	16(2)	16(2)	16(2)	16(2)
Race/ethnicity, No.(%)				
Asian/Pacific Islander	7(9)	9(11)	6(8)	8(10)
Black	2(3)	7(9)	6(7)	4(5)
Hispanic	7(9)	8(10)	7(9)	11(14)
White	59(76)	52(66)	59(75)	52(69)
Other	2(3)	3(4)	1(1)	1(1)
Smokers, No.(%)	2(3)	4(5)	4(5)	3(4)
Physical activity	34(6)	34(6)	34(5)	35(7)
(kcal/kg per day)				
Anthropometrics				
Weight, kg	86(13)	84(12)	85(14)	86(10)
Body fat, %	41(6)	40(6)	38(6)	40(6)
Body mass index	32(4)	31(3)	31(4)	32(3)
Waist-hip ratio	.843(.067)	.841(.068)	.839(.066)	.840(.060)
Cardio. disease risk factors				
LDL-C, mg/dL	109(29)	114(32)	104(29)	111(27)
HDL-C, mg/dL	53(14)	52(11)	51(11)	50(11)
Triglycerides, mg/dL	125(78)	123(98)	119(73)	118(62)
Non-HDL-C, mg/dL	134(33)	139(39)	127(34)	135(33)
Ratio of total cholestrol to HDL-C	3.7(1.0)	3.8(1.1)	3.6(1.0)	3.8(1.0)
Fasting insulin, U/mL	10(7)	10(7)	10(8)	10(5)
Fasting glucose, mg/dL	92(9)	94(20)	96(17)	93(13)
Blood pressure, mmHg				
Systolic	118(11)	115(13)	116(12)	116(10)
Diastolic	75(8)	74(9)	75(9)	75(8)
Metabolic syndrome, No. (%)	22(29)	20(25)	29(37)	27(36)

5.5 Exercises

——————

Exercises

Exercise 5.5-1: Give some examples of studies where:

1. a randomized trial is impossible

2. a randomized trial is impractical because of length of time the study would require

3. a randomized trial is impractical because of the number of subjects the study would require

Exercise 5.5-2: Suppose a doctor is interested in investigating the causes of a sporadic disorder that occurs in only 0.00001% of men. What type of study would be necessary to establish the origins of the disease? Is this type of research an appropriate choice for such a disease firmly? Why or why not?

Exercise 5.5-3: What is a double-blind study? Give an example where double-blinding is not possible.

Exercise 5.5-4: Some recent research has shown that patients who know they have taken nothing more than an inert sugar pill still experience the placebo effect. Briefly describe a study design that might replicate these results. What type of study would you choose? Should we consider different groups?

Exercise 5.5-5: What is the advantage of case-control studies over clinical trials?

Exercise 5.5-6: A football organization is concerned about the number of injuries to its athletes during games. The organization designs an observational study to help decide on rule changes that will reduce the risk of injury to players. They collect data and observe that a very high proportion of injuries incurred during the first moments of the game during kickoff. The organization thus entertains the idea of changing the rules governing kickoff. Are there any hidden confounders that the organization should address before making changes to kickoff rules? What are they, and how might they be treated?

Exercise 5.5-7: What is the advantage of clinical trials over case-control studies?

Chapter 6

The Normal Distribution

6.1 Objective

After completing this part, students should be able to:

> - Define and explain the concept of the **Normal Curve**.
>
> - Convert empirical scores to z-scores and use z-scores and the Normal Distribution curve table to find areas above, below and between points of the curve and express them regarding their probabilities.
>
> - Utilize the standard normal distribution to solve probability (chance) problems.

6.2 Using the normal curve

Best Buy wants to know how many smart fitness and GPS watches (like Fitbit, Garmin, and Apple) to order for their Kalamazoo location. Past data show that this time of year, they sell an average of 36 fitness watches per month, with a standard deviation of 8.

Problem 1.

If they order 40 watches, what is the probability that they will run out of stock?

Problem 2.

Given the cost of running out of stock and the storage cost of keeping too many, the manager decides to order enough smartwatches to cover customer demands 90% of the time, i.e., there should be 10% or less likelihood that they run out of stock. How many should they order?

Knowing the average and SD of a process (36 ± 8) gives us some understanding of what to expect. However, the sales situation above requires the computation of probabilities, or chances, as in the "chance of demand exceeding 40." This desired probability is the shaded area to the right of 40 under the histogram of demand (see the first graph in Figure 4.1). Does this look like 0.20 of the total area? 0.30? 0.40?

Figure 6.1: Probability of Demand exceeding 40

Probability of Demand > 40
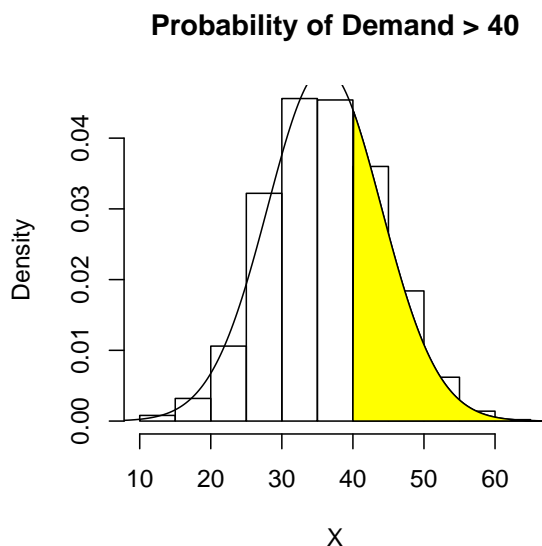
The normal curve (or bell curve) is useful in helping us calculate probabilities like these. If we smooth out the top of the histogram in Figure 4.1, it will look like a normal curve. More precisely, it will look like a normal curve with a mean of 36 and an SD of 8, denoted as $N(36, 8)$. So the shaded area on the left will be approximated by the shaded area under the $N(36, 8)$ curve on the right. Do the two shaded regions look the same?

Of course, the areas look similar because, in our example, the histogram on the left looks like a normal curve. It does not have to. If it does not, then the approximation of areas using the normal curve will be wrong. We should use the normal curve approximation with caution.

The Standard Normal or Z Curve

The standard normal curve (or Z-curve) looks like Figure 4.2, and has the following properties:

1. center is at zero

2. The area under the curve satisfies the following:
   The area between -1 and +1 is 0.68
   The area between -2 and +2 is 0.95
   The area between -3 and +3 is 0.997
   The area between $-\infty$ and $+\infty$ is 1.00

Figure 6.2: Normal Z-curve

The Z–Curve
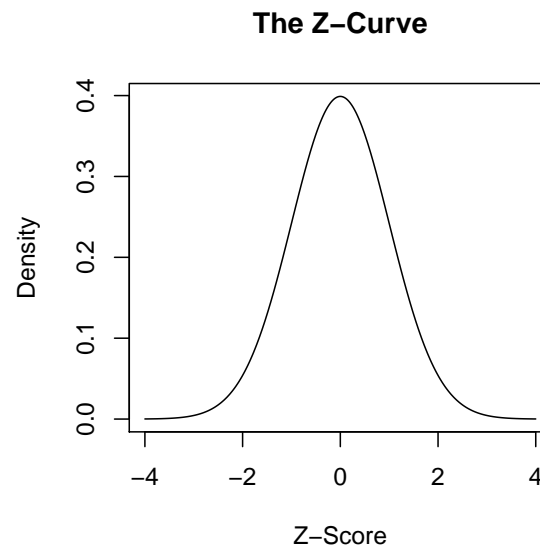
In general, any area under the curve can be found using the Z-table on page 46 at the end of this chapter.

Using the Z-table, find ...

For practice

1. the area to the left of 1.2

2. the area to the left of 1.25

3. the area to the right of 1.25

4. the area to the left of -1.25

5. the area between 1.25 and 2.50

Are we ready to find the area under the normal curve in Figure 4.1? Not yet. The horizontal axis in Figure 4.1 is wrong for the Z-table, it does not have a mean of 0 and an SD of 1. Instead, it has a mean of 36 and an SD of 8. The trick is to replace the $N(36, 8)$ horizontal axis with a $N(0, 1)$ axis, labeled Z. See Figure 4.3. Note that $Z = 1$ whenever X is one SD above the mean. Similarly, $Z = -1$ whenever X is one SD below the mean. In general, Z is related to X as follows:

$$Z = \frac{X - mean}{SD} \qquad (6.1)$$

Figure 6.3: Areas within one and two SD's of the Mean *N(36, 8)*

What is the probability of demand exceeding 40? This is the area under the curve to the right of $X = 40$ or $Z = 0.5$ (see Figure 4.4). Using the Z-table, this area is $1 - 0.6915 = 0.3085$.

6.3 Calculating percentiles

We restate Problem 2 given at the start of the chapter.

Figure 6.4: $P[X \geq 40]$

Problem 2. Given the cost of running out of stock and the value of having too many, the manager decides to order enough smartwatches to cover customer demands 90% of the time, i.e., there should be 10% or less likelihood that they run out of stock. How many should they order?

We want to estimate the number of smartwatches so that there is only a 10% chance of running out. Note that this is the 90th percentile of demand.

Looking at Figure 4.5, we see that we need to shade the upper 10% of the area under the histogram. What X value corresponds to the boundary? When we use the normal table, the Z-value 90th percentile is 1.28. What X value corresponds to Z = 1.28? We solve equation 4.1 backward.

$$
\begin{aligned}
1.28 &= \frac{X - 36}{8} \\
36 + 1.28(8) &= X - 36 + 36 \\
36 + 1.28(8) &= X \\
X &= 46.24
\end{aligned}
$$

The manager should have at least 46 smartwatches in the store.

6.4 Calculating symmetric tail areas

In later sections, we will need to calculate tail areas (or *P*-values). For example, how frequently does a normal variable fall outside of 2.25 SD from the center? We know a random variable following a normal distribution falls within 1 SD from the center 68% of the time, and hence outside of 1 SD from the center only 32% of the time. What about 2.25 SD?

Figure 6.5: The 90th Percentile: $P[X \geq a] = 0.10$

Probability of Demand > 46
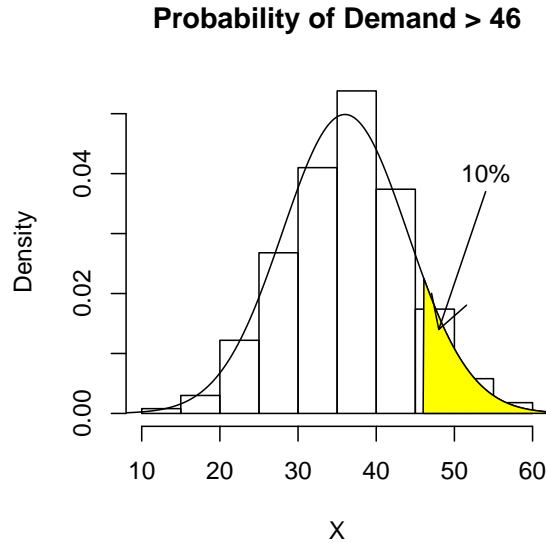
The area outside of 2.25 SD from the center is the combined area left of -2.25 and right of 2.25. From the Z-table, the area to the right of 2.25 is 1 - 0.9878 = 0.0122. Therefore, the combined tail areas is 0.0122 + 0.0122 = 0.0244. A random variable following a normal distribution falls outside of 2.25 SD from the center only 2% of the time.

What percentage of time does
a normal variable fall outside:

1. 0.4 of the mean

For practice

2. 1.4 SD of the mean

3. 2.4 SD of the mean

4. 3.4 SD of the mean

6.5 The Empirical Rule

Figure 4.3 provides a useful interpretation of the SD. We state it as follows:

> **Empirical Rule:**
> If the data histogram is approximately bell-shaped, expect around
> 68% of the observations will fall within **one** SD of the mean
> 95% of the observations will fall within **two** SD of the mean
> 99.7% of the observations will fall within **three** SD of the mean

6.6 Key Words

• area under the curve	• normal curve
• empirical rule	• normal Z-curve

6.7 Exercises

—————

Exercises

Exercise 6.7-1: Researchers conducted a study to determine if any link existed between cellular phone usage and the development of brain cancer (don't worry, no connection was found). Data from this study indicate that the daily cell phone usage for all users is approximately normally distributed with mean 2.4 hours and standard deviation of 1.1 hours.

1. What proportion of cell phone users are on their phones between 1 hour and 3 hours per day?

2. To be safe, suppose you decide to be in the 5th percentile of cell phone users concerning monthly usage. How much time can you spend on your phone per day?

Exercise 6.7-2: On the average, a watch battery is known to last for two years (24-months), with a standard deviation of 9 months. Assume a normally distributed population.

1. What percentage of watch batteries last more than six months?

2. What is the lifespan of a watch battery which lasts longer than 60% of all batteries?

3. What proportion of watch batteries that last shorter than two years or longer than 3 1/2 years (42 months)?

Exercise 6.7-3: Suppose 95% of data coming from a normally distributed population falls between 4 and 35. Based on the empirical rule, what is the standard deviation of this sample of data?

Exercise 6.7-4: Let's assume a normally distributed random variable with a mean of 10 and standard deviation of 2.

1. What is the probability the value is greater than 6?

2. What is the probability the value is less than 12?

3. What is the probability the value is between 6 and 12?

4. 33% is above what value?

5. 33% is below what value?

Exercise 6.7-5: The stock price for Coca-Cola (KO) is normally distributed with a mean of 42.14 and standard deviation of 1.43.

1. What is the probability that the stock price is between 39.88 and 46.01?

2. What is the probability that the stock price is above 40?

3. What is the probability that the stock price is below 40?

Exercise 6.7-6: The average adult female height is 63.8 inches with a standard deviation of 2.40. Assume the distribution is approximately normal.

1. What proportion of adult female heights is below 72?

2. 25% of adult females are greater than what height?

Exercise 6.7-7: Values in the table give the area under the curve to the left of the z-values on the margins. The upper margin gives the hundredth digit of z. For example:

1. The area to the left of 0.0 is?

2. The area to the left of 0.2 is?

3. The area to the left of 0.25 is?

4. The area to the left of 2.25 is?

Exercise 6.7-8: Test scores for 120 students were found to have a mean of 32. Suppose the upper bound of the middle 99.7% is 41. What is the standard deviation of this data?

1. 2

2. 2.5

3. 3

4. 4

Answer:

Exercise 6.7-9: Given a distribution of male weights has a mean of 153 lbs and standard deviation of 11 pounds, how many standard deviations away from the average weight is 173 lbs?

1. −1.82

2. 1.82

3. 1.82 lbs.

4. You cannot calculate the z-score with the information provided.

Answer:

Exercise 6.7-10: Given the distribution of male weights has a mean of 155 and standard deviation of 12, what percent of men weight less than 143 pounds?

1. 0.1587

2. 0.3413

3. 15.87

4. 34.13

Answer:

Exercise 6.7-11: Given the distribution of male poundage with a mean of 155 and standard deviation of 12, what value represents the 90th percentile for weights of men (in pounds)?

1. 155.00

2. 170.36

3. 179.00

4. 189.13

Answer:

Exercise 6.7-12: Given the distribution of male poundage with a mean of 155 and standard deviation of 12, what value represents the 25th percentile for weights of men (in pounds)?

1. 155.00

2. 163.10

3. 146.91

4. 131.00

Answer:

Exercise 6.7-13: The mean and standard deviation are 8.30 and 3.224, respectively. What countries are in the bottom 15 percent? Hint: find the value of the 15th percentile and compare it to the percentage of GDP.

Exercise 6.7-14: The mean and standard deviation are 8.30 and 3.224, respectively. What percentage of countries 2013 Expenditures on Health as a share of GDP had values below 7.3? Hint: find the probability of the mean being less than 7.3 and compare it to the Percentage of GDP.

Exercise 6.7-15: The mean and standard deviation are 8.30 and 3.224, respectively. What percentage of countries 2013 Expenditures on Health as a share of GDP had values above 9.3? Hint: find the probability of the mean being greater than 9.3.

Exercise 6.7-16: The mean and standard deviation are 8.30 and 3.224, respectively. What percentage of countries 2013 Expenditures on Health as a share of GDP had values between 7.3 and 9.3? Hint: find the probability of the mean being between 7.3 and 9.3.

Chapter 7

The Binomial Distribution

7.1 Objective

After completing this part, students should be able to:

- Use the binomial distribution to compute probabilities.

- Make use of the expected value and SD of a Binomial Random Variable.

- Apply Binomial Probabilities Using the Normal Curve.

- Understand that some Approximations Are Better Than Others.

7.2 Binomial Probabilities

A sequence of n observations is called a binomial process if

1. each observation results in one of two possible outcomes (which we call success and failure)

2. the probability of success is p, and the probability of failure is q = 1 - p for all observations

3. the observations are independent of each other.

Obs 1	Obs 2	Obs 3	$\cdots$	Obs n
$p \swarrow \searrow q$	$p \swarrow \searrow q$	$p \swarrow \searrow q$		$p \swarrow \searrow q$
S F	S F	S F		S F

Let X denote the total number of successes among the n observations. Then X is called a **binomial random variable** with parameters n and p. The following are all binomial random variables.

Example 1

A Stat 1600 multiple choice quiz has five questions, with each question having five choices. Let X be the number of correct answers (C) by someone who is guessing on all questions. Then X is a binomial random variable with parameter values n = 5 and p = 0.2.

Ques 1	Ques 2	Ques 3	Ques 4	Ques 5
.2 ╱╲ .8	.2 ╱╲ .8	.2 ╱╲ .8	.2 ╱╲ .8	.2 ╱╲ .8
C W	C W	C W	C W	C W

Example 2

Available data shows that 40% of telephone respondents agree to be interviewed for market research surveys. Suppose the polling organization Reliable Research randomly selects and dials telephone numbers until it reaches 50 respondents. Let X be the number of respondents (out of the 50) who agree to be interviewed. Then X is a binomial random variable with parameter values n = 50 and p = 0.40.

Example 3

Historically, 20% of buyers at Best Buy who purchase smart fitness and GPS watches (like Fitbit, Garmin, and Apple) also purchase the Geek Squad's protection plan. Suppose Best Buy sold 300 smart fitness watches during the previous quarter. Let X be the number of extended protection plans that the retailer sold along with the 300 smartwatches. Then X is a Binomial random variable with parameter values n = 300 and p = 0.20.

7.3 Computing Binomial Probabilities

In Example 1, the number of correct guesses may be 0, 1, 2, 3, 4, or 5. How likely can a guesser get all five questions right? The answer is 0.0003, or about three times in 10000 attempts. How about the likelihood of getting 2 out of 5 items right? The answer is .2048, about a fifth of the time. The following **probability distribution table** gives the likelihood or probability of each possible value of X.

$P[X = 0]$	$P[X = 1]$	$P[X = 2]$	$P[X = 3]$	$P[X = 4]$	$P[X = 5]$
0.32768	0.4096	0.2048	0.0512	0.0064	3.2×10^{-4}

You can compute these probabilities yourself by successively substituting $j = 0, 1, 2, 3, 4,$ and 5 in the formula

$$P[X = j] = \frac{5!}{j!(5 - j)!} .2^j .8^{5-j}, \text{ where } j = 0,1,2,3,4,5$$

Remember that $0! = 1$ and $(0.2)0 = 1$. This formula is called the **binomial probability distribution function (pdf)** for $n = 5$ and p= 0.2. To compute the probabilities for Examples 2 and 3, you will need the binomial pdf for general n and p:

$$P[X = j] = \frac{n!}{j!(n - j)!} p^j q^{n-j}, \text{ where } j = 0,1,2, \cdots , n$$

Example 3

(Cont.): The retailer sold ten smart fitness and GPS watches in one day. What is the probability that customers purchased three extended Geek Squad protection plans? Using the equation above with $n = 10, p = 0.20,$ and $j = 3$, we get

$$P[X = 3] = \frac{10!}{3!(10 - 3)!} .2^j .8^{10-3} = 0.201$$

Five percent of customers who rent video games from Gamers Retro Rental who return the rental late. What is the probability that customers will return their rental late if 30 customers borrow a video game during the last hour?

For practice
1. 2 will be returned late

2. none will be returned late

3. 2 or fewer will be returned late

4. 5 or more will be returned late

7.4 Expected Value and SD of a Binomial Random Variable

Suppose last quarter; Best Buy sold 300 Smart fitness watches. If there is 0.20 likelihood of selling an extended protection plan with each smartwatch, the number of extended protection plans sold last quarter should be around 60 give or take 7 or so (we will compute this later).

We call the first number the expected value of the number of protection plans sold; the second number is the standard deviation. Recall that X, the number of protection plans sold, is a **Binomial random variable**. The **expected value**, denoted **E(X)**, of a binomial random variable X with parameters n and p is computed as:

$$E[X] = np$$

The expected value $E(X)$ is also called the average or mean of X, and denoted μ. The standard deviation, denoted $SD(X)$ or σ_x, of a binomial random variable is computed as:

$$SD(X) = \sigma_x = \sqrt{npq}$$

Returning to the example, since $n = 300$ and $p = 0.20$, we have $E(X) = 300(0.20) = 60$, and $SD(X) = \sqrt{300(0.20)(0.80)} = 6.93$, or approximately 60 ± 7.

The SD for random variables is interpreted similarly to the SD for a sample. If the store sells 300 smartwatches sets every quarter, they won't sell precisely 60 Geek Squad protection plans every time; sometimes they will sell more, sometimes they will sell less. By how much more, and how much less? The answer is, "By 7, on average." Similarly, a baseball player with 0.200 batting average won't necessarily get 60 hits in 300 at-bats. We expect him to get 60 hits, give or take seven hits.

For practice   Suppose that 5% of video games rented at Gamers Retro Rental incur a late rental fee. If 700 videos were rented last week, the number that will incur a late rental fee should be around _____ give or take _____.

7.5 Computing Binomial Probabilities Using the Normal Curve

Beyond the empirical rule, we may apply the normal curve to approximating binomial probabilities. The key image is a plot of binomial probabilities as a histogram. For example, Figure 5.1 is a histogram of binomial probabilities for $n = 30$ and $p = 0.4$.

The height of the rectangle over, say 10, is its probability $P[X = 10$. However, since the width of the square is 1, then

$$P[X = 10] = \texttt{height of rectangle over 10 = area of rectangle over 10}$$

What about $P[X \le 10]$? This probability corresponds to the total area of the rectangles over and the left of 10 (shaded rectangles in Figure 5.2)

The total area equals 0.2915 using the binomial probability function to compute the area (probability) of each shaded rectangle. However, this requires repeated applications of the binomial formula (11 times, in fact). We may calculate a quick approximation of the desired probability by replacing the rectangles with a curve! See shaded area under the curve in Figure 5.2.

There are infinitely many normal curves, which one do we use to replace the rectangles? Answer: the one with the same mean and SD as the rectangles! The (binomial) rectangles have a **mean**:

$$\mu = np = (30)(.4) = 12$$

The (binomial) rectangles have a SD:

$$\sigma = \sqrt{npq} = \sqrt{(30)(.4)(.6)} = 2.68$$

Using the normal curve with the same mean and SD, the area to the left of 10.5 is 0.2877, which is a close estimate of the actual area under the curve is 0.2915.

Similarly, $P[X = 14] = 0.1101$ is approximated by 0.1115, the area under the normal curve between 13.5 and 14.5.

7.6 Some Approximations Are Better Than Others

Examine the shape of the binomial histogram for n = 20 and p = 0.10 in Figure 5.3.

Since the curve is right-skewed, we will get a poor approximation of areas if we replace the rectangles by a normal curve. If the value of p were 0.90 instead of 0.10, the binomial histogram would be left-skewed. It is typical behavior of binomial histograms whenever p is either too close to 0 or too close to 1. When is it 'safe' to use the normal curve to approximate binomial probabilities? A convenient rule of thumb is as follows:

> The **Normal Curve gives reasonable approximations** of binomial probabilites whenever both $np > 5$ and $nq > 5$.

Figure 7.1: Histogram of Probabilities for Binomial n = 30 and p = 0.4

Histogram of Probabilities
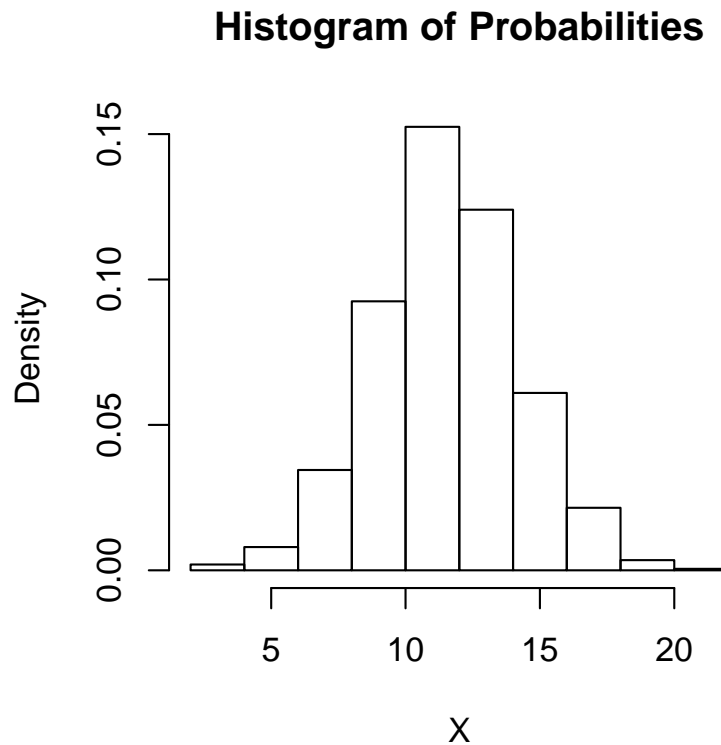
We remind the reader that the normal curve approximations are still approximations. The binomial formula should be used to calculate exact probabilities whenever possible.

7.7 Key Words

- binomial distribution
- binomial experiment
- binomial random variable
- binomial tables
- discrete random variables

Figure 7.2: Approximating $P[X \leq 10]$ using Curve instead of Rectangles

Probability of Demand No More than 10
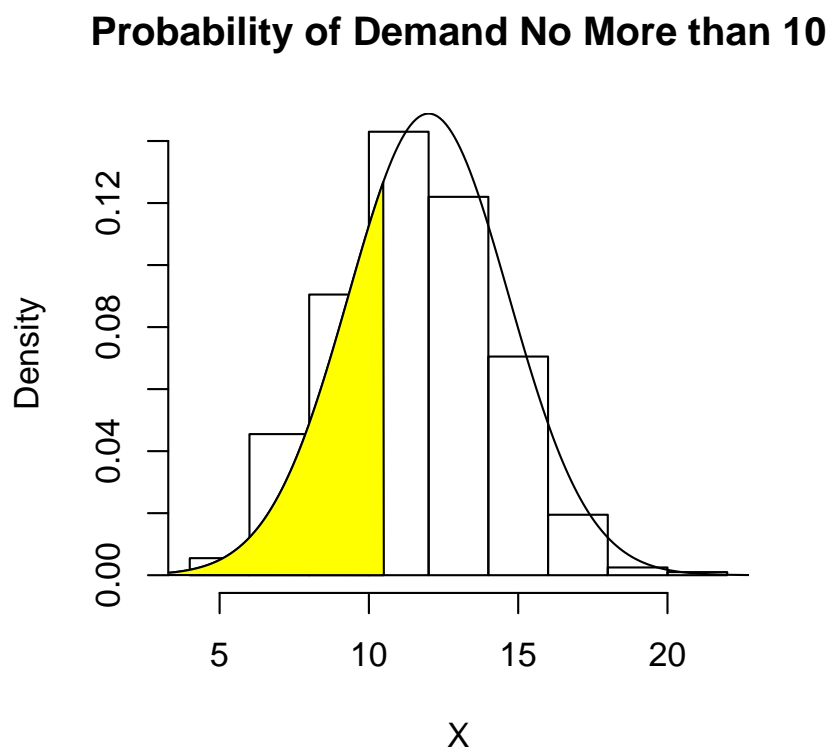
Figure 7.3: Histogram of Probabilities for Binomial n = 20 and p = 0.10

Histogram of Probabilities
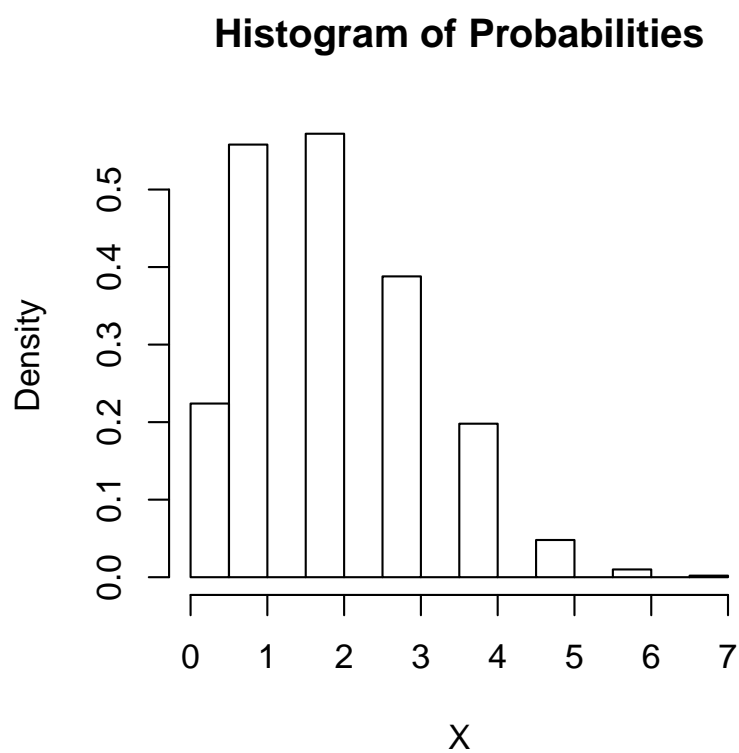
7.8 Exercises

————————

Exercises

Exercise 7.8-1: A binomial random variable with a probability of success of 0.5 and ten observations. Assume each observation is independent.

1. What is the mean and standard deviation?

2. What is the probability that there are more than five successes?

3. What is the probability that there are fewer than five successes?

4. What is the probability that there are between 1 and 3, inclusive?

Exercise 7.8-2: A statistics exam contains ten questions with five multiple choice options per question. By guessing on all questions,

1. What is the probability that at least two questions correct?

2. What is the probability that at most two questions correct?

3. What is the probability that between 1 and 3?

Exercise 7.8-3: Apple's smartphone market share is 0.146. If Apple conducts a nationwide survey of 651 smartphone users,

1. What is the probability that at least 100 of the people are Apple users?

2. What is the probability that at most 100 of the people are Apple users?

3. What is the probability that between 80 and 120?

Exercise 7.8-4: When rolling a die ten times,

1. What is the probability of rolling a six no more than three times?

2. What is the probability that no less than 3 times?

Exercise 7.8-5: The career batting average of Ty Cobb is 0.3664. If Ty Cobb had eight at-bats during a doubleheader,

1. What is the probability that he gets at least seven hits?

2. What is the probability that at most one hits?

3. What is the probability that between 4 and 6?

Exercise 7.8-6: Fill in the blanks. The probability of picking the Powerball number is 0.0254. If 50 tickets are purchased, around _____ of tickets will be winners, give or take _____. Assume each pick is independent.

Exercise 7.8-7: Fill in the blanks. The probability of a defective light bulb is 0.04. If purchase order of 200 bulbs is submitted, the number of defective light bulbs in the shipment is around _____, give or take _____. Assume each light bulb is independent.

Chapter 8

Sampling Distribution of the Proportion

8.1 Objective

After completing this part, students should be able to:

- Describe and use the sampling distribution of the proportion.

8.2 The Sample Proportion

Suppose a student guesses at the answer to every question in a 300-question exam. If he gets 60 questions correct, then his proportion of correct guesses is $60/300 = 0.20$. If he gets 75 questions correct, then his proportion of correct guesses in $75/300 = 0.25$. The proportion of correct guesses is simply the number of correct guesses divided by the total number of questions.

Similarly, if Best Buy's Geek Squad replacement plan sells 60 extended warranties with 300 smartwatches sold, then its protection plan sales rate is $60/300 = 0.20$. If it sold 75 protection plans, this is a sales rate of $75/300 = 0.25$. The warranty or protection plan sales rate is merely the number of warranties sold divided by the total number of smartwatches sold.

Now, let X denote the number of successes out of a sample of n observations. If each response is a success with probability p independently of the other observations, then X is a binomial random variable with parameters n and p. Furthermore, the proportion of successes in the sample is also a random variable and is computed as

$$\hat{p} = \frac{X}{n} = \frac{Number\ of\ successes}{Number\ observations\ in\ the\ sample}$$

Since X is expected to be around np give or take $\sqrt{npq}$, then X/n is expected to be around np/p give or take $\sqrt{npq}/n$, or p give or take $\sqrt{pq}/\sqrt{n}$. Make sure that you agree with the last statement before moving on. It may help to think of this analogy: Suppose annual rainfall in Kalamazoo is expected to be around 24 inches give or take 6 inches. How do we change the measurement from inches to feet? We divide both numbers by 12! In feet, annual rainfall in Kalamazoo is expected to be around 24/12 give or take 6/12, or 2 feet give or take 0.5 feet. Now read the first sentence of this paragraph one more time.

Going back to the Best Buy example, the number of protection plans sold is expected to be around 60 ± 7. Thus, the proportion of plans sold is supposed to be about $60/300 \pm 7/300$, or 0.20 ± 0.02.

We summarize the formulas for the mean and SD of X and $\hat{p}$ in the following table.

Random Variable	Mean	SD
X	np	$\sqrt{npq}$
$\hat{p}$	p	$\sqrt{\frac{pq}{n}}$

Excercise 1   If the local Best Buy store sold 1200 smartwatches last year,

1. the proportion of sets sold with extended protection plans should be around 0.20, give or take _____.

2. the percentage of watches sold with extended protection plans should be around 20%, give or take _____.

Data analysis sometimes involves percentages instead of proportions. Proportions and percentages are two ways of saying the same thing (e.g., we refer to 1/5 as either 0.20 or 20%). How do we convert a proportion to a percentage? We multiply by 100%. To avoid repetition, we present all statistical formulas in proportions. As Exercise 1 shows, we will convert all the answers to percentages in the end.

Excercise 2   Historically, 5% of customers return their video game rentals from Gamers Retro Rental late.

1. Gamers Retro Rental rented out 100 video games yesterday. The percentage that will be returned late should be around 5%, give or take _____.

2. Gamers Retro Rental rented out 700 video games last week. The percentage that will be returned late should be around 5%, give or take _____.

Exercise 2 is an example of the law of large numbers. A more straightforward illustration involves a coin toss. If you toss a coin repeatedly, which tends to get closer to 50% heads: 100 tosses or 700 tosses? The correct answer is 700. As the number of tosses increases, the closer we expect to get to 50%. The reason for this, as the Exercise shows, is the smaller give or take value. The sample size, n, lies in the denominator of the SD of $\hat{p}$. Therefore, the larger the sample size, the smaller the SD, which happens to be the give-or-take value.

The Law of Large Numbers for Sample Percentages:
The sample percentage tends to get closer to the true percentage as the size increases.

8.2.1 The Sampling Distribution of p-hat is Approximately Normal

Since $\hat{p} = \frac{X}{n}$, the sampling distribution of $\hat{p}$ looks the same as that of X except for different numbers on the horizontal axis. For $n = 30$ and $p = 0.4$, the probability histogram of X and $\hat{p}$ is shown in Figure 6.1.

Figure 8.1: Probability of Histogram and p-hat

Histogram of n=30 & p=.4
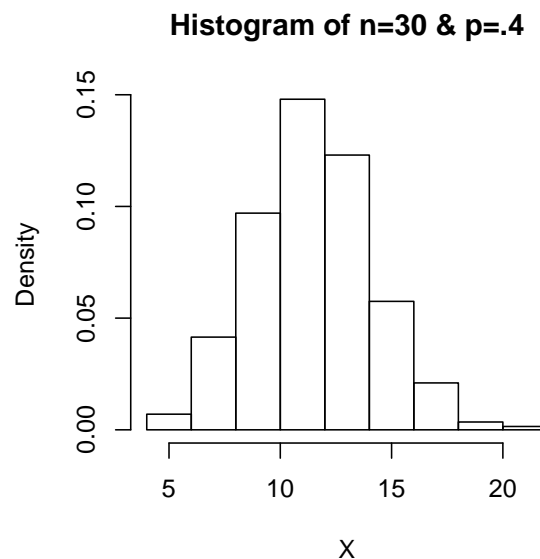
Therefore, like the binomial, the sampling distribution of $\hat{p}$ may be approximated by a normal curve with the correct mean and SD.

Example: Toss a fair coin 50-times. What is the chance of getting 60% or more heads?

Solution: The question is equivalent to 'What is the probability that $\hat{p}$ exceed 0.60?' Using the mean and SD given in the formula above with $n = 50$ and $p = 0.50$, the sample proportion is expected to be around 0.50 give or take $\sqrt{\frac{0.5(0.5)}{50}}$ or 0.50 give or take 0.07. With a mean of 0.50 and a SD 0.07, the area to the right of 0.60 (under the normal curve with mean 0.50 and SD 0.07) is 0.0766. Thus, the proportion of heads will exceed 0.60 fewer than 8 percent of the time.

Example 3   If Best Buy sold 1200 smartwatches last year, the percentage of smartwatches sold with extended protection plans is expected to be around 20%, give or take _____. Estimate the likelihood that it sold protection plans with more than 25% of those watches.

8.3 Estimating the Population Proportion p

The Best Buy computations in the previous section assume that we know the protection plan sales rate is $p = 0.20$. In data analysis, population parameters like p are typically unknown and estimated from the data. Consider estimating the proportion p of the current WMU graduating class who plan to go to graduate school. Suppose we take a sample of 40 graduating students and suppose that 6 out of the 40 are planning to go to graduate school. Then our estimate is $\hat{p} = \frac{6}{40} = 0.15$ of the graduating class plan to go to graduate school. Now $\hat{p}$ is based on a sample, and unless we got fortunate, chances are the 0.15 estimate missed. By how much? On the average, a random variable misses the mean by one SD. From the previous section, the SD of $\hat{p}$ equals $\sqrt{\frac{pq}{n}}$. It follows that the expected size of the miss is $\sqrt{\frac{pq}{n}}$. This last term is the *standard error of estimation of the sample proportion*, or simply **standard error (SE)** of the proportion.

However, since we don't know p, we cannot calculate this SE. In a situation like this, statisticians replace p with $\hat{p}$ when calculating the SE. The resulting quantity is called the *estimated standard error of the sample proportion*. In practice, however, the word "estimated" is dropped and estimated SE is called merely the SE.

The population p is **estimated using the sample proportion** $\hat{p}$. This estimate tends to miss by an amount called the **standard error (SE)** of $\hat{p}$.

$$SE_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

Fill in the following blanks.

Exercise 4

1. If 6 out of 40 students plan to go to graduate school, the proportion of all students who plan to go to graduate school is estimated as _____. The standard error of this estimate is _____.

2. If 54 out of 360 students plan to go to graduate school, the proportion of all students who plan to go to graduate school is estimated as _____. The standard error of this estimate is _____.

Exercise 4 shows the effect of increasing the sample size on the SE of the sample proportion. Multiplying the sample size by a factor of 9 (from 40 to 360) makes the SE decrease by a factor of 3. In the formula for the SE of $\hat{p}$, the sample size appears (i) in the denominator, and (ii) inside a square root. Therefore, multiplying the sample size by a specific factor divides the SE of $\hat{p}$ by the square root of that factor.

> As the sample size n **increases**, the $SE_{\hat{p}}$ **decreases** like the square root of the sample size.

8.4 Estimating Population Proportion Using Intervals

Variables tend to miss their expected value but should be within *one* SD 68% of the time, and within 1.96 SD 95% of the time. Since the SE of $\hat{p}$ is simply an estimate of the SD, then we can write $|\hat{p} - p| \leq 1.96(SE)$ or that p is inside the interval $\hat{p} \pm 1.96(SE)$ 95% of the time. In other words, the interval $\hat{p} \pm 1.96(SE)$ contains the true value of p with 95% certainty. This method gives us an interval estimate of p.

> **95% Confidence Interval for p:**
> A 95% confidence interval estimate for the population proportion p is given by
>
> $$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

The term $1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$ is the 95% margin of error.

Try the following problems.

Exercise 5

1. If 6 out of 40 students plan to go to graduate school, the proportion of all students who plan to go to graduate school is estimated as _____. The margin of error is _____.

2. Calculate a 95% confidence interval estimate for the true proportion p of WMU students who plan to go to graduate school.

3. If 54 out of 360 students plan to go to graduate school, calculate a 95% confidence interval estimate for the true proportion p of WMU students who plan to go to graduate school.

8.5 Sample Size for Estimating the Population Proportion

If 9 out of 25 randomly selected WMU students live in Southwest Michigan, the 95% confidence interval for the true proportion is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.36 \pm 0.19$. This result says that the true proportion can be as low as 0.17 or as high as 0.55. If we wanted to reduce the margin of error from 0.19 to some value M, then we set the formula for margin of error equal to M, i.e. $M = 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$. Solving for n gives the result we need.

To be 95% confident that the sample proportion is within a distance of M of the true proportion p, choose a sample size equal to

$$n = (1.96)^2 \frac{\hat{p}\hat{q}}{M^2}$$

where $\hat{p}$ is an estimate based on historical data or a pilot study. The quantity M is called the 95% margin of error for p.

Example:

Suppose we want to reduce the margin of error for estimating the population proportion from 0.19 to 0.10. Using the estimate $\hat{p} = 0.36$ based on the initial sample, the sample size we need is: $n = (1.96)^2 \frac{(0.36)(0.64)}{0.10^2} = 89$. To verify that this is the correct sample size, the 95% confidence interval would be computed (if the sample proportion remained at 0.36) as $0.36 \pm (1.96)^2 \frac{(0.36)(0.64)}{(0.10)^2} = 89$. To verify that this is the correct sample size, the 95%

confidence interval would be computed (if the sample proportion remained at 0.36) as $0.36 \pm (1.96)\sqrt{\frac{0.36(0.64)}{89}} = 0.36 \pm 0.10$.

8.6 Key Words

- binomial distribution
- binomial experiment
- binomial random variable
- binomial tables
- discrete random variables

8.7 Exercises

Exercises

Exercise 8.7-1: Suppose that 20% of students in a large university are graduate students. If a random sample of 125 students are randomly selected, what is the probability that 25% or more of the sample are graduate students?

Exercise 8.7-2: Let's say that we take a sample of 100 observation has 20 successes.

1. What is the estimate of the population proportion?

2. What is the standard error of this estimate?

3. What is the 95% margin of error?

4. What is the 95% confidence interval?

Exercise 8.7-3: A sample of 100 individuals showed that 20% experienced gastrointestinal problems after consuming 10 grams of sorbitol, a common artificial sweetener. Attach a standard error to this estimate.

Exercise 8.7-4: When a flight experiences fewer no-shows than expected, some passengers are
'bumped' from their flights (are denied boarding). These incidents can reflect poorly on customer satisfaction. Suppose United Airlines (for example) would like to estimate the actual proportion of involuntarily bumped passengers across all domestic flights in the industry. In a pilot study of 500 domestic passengers, 33 were involuntarily bumped.

1. What is the estimate of the population proportion?

2. What is the standard error of this estimate?

3. What is the 95% margin of error?

4. What is the 95% confidence interval?

Exercise 8.7-5: An appliance manufacturer offers maintenance contracts on its major appliances. A manager wants to know what fraction of buyers of the company's convection ovens are also buying the maintenance contract with the oven. From a random sample of 120 sales slips, 31 of the oven buyers opted for the contract.

1. The proportion of customers who buy the contract along with their oven is estimated as _____.

2. Calculate a standard error for the estimate in (a).

3. Calculate a 95% confidence interval estimate for the true proportion of customers who buy the contract along with their oven.

Exercise 8.7-6: Wiley Publications has determined that out of a sample of 5,511 of its publications for 2012, 1,754 of them are pirated in some form.

1. What is the estimate of the population proportion?

2. What is the standard error of this estimate?

3. What is the 95% margin of error?

4. What is the 95% confidence interval?

Exercise 8.7-7: Researchers who were concerned if doctors were consistently adjusting dosages for the weight of elderly patients studied 2000 prescriptions. They found that for 600 of the prescriptions, the doctors failed to change the dosages.

1. Doctors fail to adjust dosage for an estimated _____ percent of prescriptions.

2. Calculate a standard error for the percentage in (a).

3. Calculate a 95% margin of error for the percentage in (a).

4. Calculate a 95% interval estimate for the true proportion

5. Calculate a 95% confidence interval for the true percentage of prescriptions where doctors fail to adjust dosages.

Chapter 9

Comparing Two Proportions

9.1 Objective

After completing this part, students should be able to:

- Explain the logic of estimation and the role of the sample, sampling distribution and the population.

- Make and interpret intervals for sample proportions.

- Describe the relationships between trust levels, sample size, and the width of the confidence interval.

9.2 Estimating the difference between independent proportions

Has retention rate at WMU changed over time? Suppose that a random sample of 200 entering students in 1989 showed 74% were still enrolled three years later. Another random sample of 200 entering students in 1999 showed that 66% were still enrolled three years later. This difference constitutes an 8% change in 3-year retention rate. However, the 8% difference is based on random sampling and is only an estimate of the actual difference. What is the likely size of the error of estimation?

Changing notation from the percentage to proportions and taking the difference of 0.74 - 0.66, we get 0.08 to compare retention rates. The proportions of 0.74 and 0.66 are *independent* proportions, in the sense that we base them on separate and independent groups of students. The SE of the difference is

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2}$$

Whenever the two proportions are independent. Applying equation SE of $\hat{p} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ twice, we have $SE_{\hat{p}_1} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}}$ and $SE_{\hat{p}_2} = \sqrt{\frac{\hat{p}_2\hat{q}_2}{n_2}}$. Substituting into the formula above, $SE_{(\hat{p}_1-\hat{p}_2)} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2}$, we get:

> **Standard Error of the Difference between two independent Proportions**
>
> $$SE_{(\hat{p}_1-\hat{p}_2)} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Continuing with the retention rate example, we let $\hat{p}_1 = 0.74$, $\hat{p}_2 = 0.66$, $n_1 = 200$, $n_2 = 200$ so that

$$SE_{(\hat{p}_1-\hat{p}_2)} = \sqrt{\frac{0.74(0.26)}{200} + \frac{0.66(0.34)}{200}} = 0.045$$

Thus, the difference in retention rate is estimated by $0.74 - 0.66 = 0.08$ with a standard error of 0.045. Changing notation back to a percentage and with less technical language, the drop-in retention rate is estimated to be 8%, give or take 4.5% or so. We also could have computed $SE_{(\hat{p}_1-\hat{p}_2)}$ in three steps. First by using $SE_{\hat{p}}$ twice,

$$SE_{\hat{p}_1} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}} = \sqrt{\frac{0.74(0.26)}{200}} = 0.031$$

$$SE_{\hat{p}_2} = \sqrt{\frac{\hat{p}_2\hat{q}_2}{n_2}} = \sqrt{\frac{0.66(0.34)}{200}} = 0.033$$

Then using $SE_{(\hat{p}_1-\hat{p}_2)} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_1}^2} = \sqrt{(0.031)^2 + (0.033)^2} = 0.045$

9.2.1 Using a confidence interval

The difference of two proportions is a random variable with an expected value and spread. The 68% and 95% rules apply, i.e. the estimated difference $\hat{p}_1 - \hat{p}_2$ should be close to the true value – within *one* SE 68% of the time, and within 1.96 SE's 95% of the time. Following the same reasoning as before,

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96(SE_{(\hat{p}_1-\hat{p}_2)})$$

should contain the true difference $p_1 - p_2$ with 95% level of confidence. Substituting

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

, we get the following formula:

95% Confidence Interval for $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96 \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

For retention rate, the difference between 1989 and 1999 the department estimated as 0.08 with SE = 0.045. Therefore, a 95% confidence interval for the change is

$$0.08 \pm 1.96(0.045)$$

or $0.08 \pm 0.088 = (-0.008, 0.168)$. Rounding off to $(-0.01, 0.17)$, we say that the drop-in retention rate from 1989 to 1999 is between -0.01 `and` 0.17 with 95% confidence. Note that **zero** is contained or is not been excluded from the interval, making it still a possibility that there is no real change in retention rate, just chance variability.

9.3 Statistical significance

Let $\hat{p}_1$ be the proportion of heads in 50 tosses of a coin. Let $\hat{p}_2$ be the proportion of heads in the next 50 tosses of the same coin. Will $\hat{p}_1$ and $\hat{p}_1$ be equal? Not likely. They will tend to differ, due to "the luck of the draw" or chance variability.

The table below shows partial data from an occupation survey by the Census Bureau. In this 2009 survey, regular 'cooks' were a separate classification from 'chefs or head cooks.' Note that even though 37% of cooks were women, only 16% of chefs or head cooks were women. Is the difference just luck of the draw, or due to something else besides chance?

	Women	Men	Total
Cooks	441	762	1203
Chefs or Head Cooks	45	245	290

Statistics help decision-making in cases like these by assessing how much chance variability to expect between two proportions. Following the calculations of the section above, we have

$$\hat{p}_1 - \hat{p}_2 = \frac{441}{1203} - \frac{45}{290} = 0.37 - 0.16 = 0.21$$

with a standard error

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{0.37(0.63)}{1203} + \frac{0.16(0.84)}{290}} = 0.026$$

The 95% confidence interval for $p_1 - p_2$ is

$$0.21 \pm 1.96(0.026) = (0.16, 0.26)$$

Thus, even allowing for 1.96 SE's of chance variability, the actual difference between proportions is at least 0.16 (and could be as large as 0.26). This difference means that the interval does not contain or it excluded 0 from the range of possible values. When this happens, statisticians say that the differences are *statistically significant.*

If the confidence interval for $\hat{p}_1 - \hat{p}_2$ excludes **zero**, then the difference is **statistically significant.**

9.3.1 The P-value

For convenience, let us continue the example of the previous section.

Question   If chance alone was at work, how likely will we get a difference of 0.21 between two proportions?

Answer   Very small, less than 0.0001 (or 1 in 10,000).

The 'likelihood of getting 0.21 by chance' is called a P-value. The fact that it is minimal means we should exclude the option that 'chance alone is at work.'

The actual probability calculation is beyond the scope of this class. Let's say that random variables very rarely go past 4 SE's from their expected values (less than 1 in 10,000 times). Since the SE for the difference is 0.026, the observed difference $\hat{p}_1 - \hat{p}_2 = 0.21$ is not just 1, nor 2, but 8 SE's from 0. This difference cannot be chance alone. Something else is at work.

The Rule for p-value:
If the p-value ≤ 0.05, the difference is *statistically significant.*
If the p-value ≤ 0.01, the difference is called *highly significant.*

In the occupation example, we can say that the percentage of women head chefs is lower than that of regular cooks. Furthermore, the difference is highly significant.

9.3.2 Risk ratio and odds ratio

In clinical studies, statisticians frequently take ratios of proportions or probabilities (instead of differences). There are several reasons for this idea. Sometimes, the disease or medical event of interest is quite rare, i.e., $p_1 = 0.0008$. If a new treatment reduces the probability of getting the disease to 0.0006, the difference in probabilities is quite small and hard to assess $((p_1 - p_2) = 0.0002)$. In the meantime, the ratio $\frac{p_2}{p_1} = \frac{0.0006}{0.0008} = 0.75$ means that the risk of getting the disease under the new treatment has been reduced by 25%.

A technical reason for using ratios is that a ratio can control for other variables such as age and race.

9.3.3 Risk ratio

Following is the abstract of the study "Safety and Efficacy of a Recombinant Hepatitis E Vaccine."

Background

Hepatitis E virus (HEV) is a significant cause of viral hepatitis. We evaluated the safety and efficacy of an HEV recombinant protein (rHEV) vaccine in a phase 2, randomized, double-blind, placebo-controlled trial.

Methods

In Nepal, we studied 2000 healthy adults susceptible to HEV infection who were randomly assigned to receive three doses of either the rHEV vaccine or placebo at months 0, 1, and 6. Active (including hospital) surveillance was used to identify acute hepatitis and adverse events. The primary endpoint was the development of hepatitis E after three vaccine doses.

Results

A total of 1794 subjects (898 in the vaccine group and 896 in the placebo group) received three vaccine doses; researchers followed the total vaccinated cohort for a median of 804 days. After three vaccine doses, hepatitis E developed in 69 subjects, of whom 66 were in the placebo group. The vaccine efficacy was 95.5% (95% confidence interval [CI], 85.6 to 98.6). In an intention-to-treat analysis that included all 87 subjects in whom hepatitis E developed after the first vaccine dose, nine subjects were in the vaccine group, with a vaccine efficacy of 88.5% (95% CI, 77.1 to 94.2). Among subjects in a sub-group randomly selected for analysis

of injection-site findings and general symptoms (reactogenicity sub-group) during the 8-day period after the administration of any dose, the proportion of subjects with adverse events was similar in the two study groups, except that injection-site pain was increased in the vaccine group ($p = 0.03$).

Conclusion

In a high-risk population, the rHEV vaccine was effective in the prevention of hepatitis E.

The data given in the 'Results' part of the abstract may be summarized as follows.

	hepatitis E		
	Yes	No	Total
Vaccine	3	895	898
Placbo	66	830	896

We may present data for comparing proportions in a 2 by 2 table.

	Disease		
	Yes	No	Total
Exposure	a	b	$a + b$
No Exposure	c	d	$c + d$

The risk ratio also called the relative risk is the ratio of probabilities

$$RR = \frac{\frac{P(Disease)}{Exposure}}{\frac{P(Disease)}{NoExposure}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{p_1}{p_2}$$

For our example, we can define and calculate the risk ratio as

$$RR = \frac{\frac{P(Disease)}{Exposure}}{\frac{P(Disease)}{NoExposure}} = \frac{\frac{3}{898}}{\frac{66}{896}} = \frac{0.00334}{0.07366} = 0.045$$

It means that getting the vaccine reduces our risk to only 4.5% of the original, or has 95.5% efficacy.

9.3.4 A 95% confidence interval for risk ratio

We write the confidence interval formulas for the natural logarithm of RR. Consider a 2 by 2 table as before.

	Disease		
	Yes	No	Total
Exposure	a	b	$a + b$
No Exposure	c	d	$c + d$

We will calculate the confidence interval for RR in 4 steps.

1. Calculate the **natural log (ln)** of the risk ratio:

$$ln(RR) = ln\left(\frac{\frac{a}{a+b}}{\frac{c}{c+d}}\right)$$

2. Calculate the standard error of ln(RR) as follows:

$$SE_{ln(RR)} = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}}$$

3. A 95% confidence interval for ln(RR) is given by

$$[ln(RR) - 1.96(SE), ln(RR) + 1.96(SE)]$$

4. Finally, a 95% confidence interval for RR is given by

$$\left[e^{ln(RR)-1.96(SE)}, e^{ln(RR)+1.96(SE)}\right]$$

Returning to our example,

1.
$$ln(RR) = ln(0.045) = -3.101$$

2.
$$SE = \sqrt{\frac{1}{3} + \frac{1}{66} - \frac{1}{898} - \frac{1}{896}} = \sqrt{0.3462} = 0.5884$$

3.
$$[-3.101 - 1.96(0.5884), -3.101 + 1.96(0.5884)] = [-4.254, 1.948]$$

4.

$$\left[e^{-4.254}, e^{-1.948}\right] = [0.014, 0.143]$$

With 95% confidence, the risk of getting hepatitis with the vaccine is only 1.4% to 14.3% of placebo. It means that the vaccine reduces our risk by as low as 85.7% or as high as 98.6%. Now read the Results section of the abstract again. They say "The vaccine efficacy was 95.5% (95% confidence interval [CI], 85.6 to 98.6)." The two sets of the numbers match, (slight discrepancy due to rounding error.)

9.3.5 Odds ratio

The odds of an event occurring is

$$Odds = \frac{\texttt{Probability that event occurs}}{\texttt{Probability that event doesn't occur}} = \frac{p}{q}$$

For example, if you win a game a 20% of the time ($p = 0.20$), then our odds of winning is $\left(\frac{0.20}{0.80}\right) = \frac{1}{4}$. We say that we have a *1-in-4* odds of winning, or we win once for every four times we lose. If we win 80% of the time, the odds are $\frac{0.80}{0.20} = 4$. It means we have *4-in-1* odds of winning, or we win four times for every one time we lose. Here is a table of odds corresponding to various probabilities.

Probability	Odds
0.10	$1/9 = 0.11$
0.20	$1/4 = 0.25$
0.50	$1/1 = 1.00$
0.80	$4/1 = 4.00$
0.90	$9/1 = 9.00$

Unlike probabilities, odds can be greater than 1. The odds ratio is just the ratio of two odds (usually for comparing two groups).

$$Odds\,Ratio = \frac{\texttt{Odds of Group 1}}{\texttt{Odds of Group 2}} = \frac{\frac{p_1}{q_1}}{\frac{p_2}{q_2}}$$

Returning to the hepatitis E study, recall the disease occurrence data:

The disease rate for each group is

	hepatitis E		
	Yes	No	Total
Vaccine	3	895	898
Placbo	66	830	896

$$\texttt{Odds(Hepatitis|Placebo)} = \frac{0.07366}{(1 - 0.07366)} = 0.07952$$

$$\texttt{Odds(Hepatitis|Vaccine)} = \frac{0.00334}{(1 - 0.00334)} = 0.00335$$

$$\texttt{Odds Ratio} = \frac{\texttt{Odds(Hepatitis|Placebo)}}{\texttt{Odds(Hepatitis|Vaccine)}} = \frac{0.07952}{0.00335} = 23.7$$

We say that "the odds of getting hepatitis is 24 times greater if we remain unvaccinated."

Odds ratios are generally easier to interpret if they are more significant than one. We can always ensure this by choosing which group to put in the numerator, i.e., the one with more substantial odds.

It is essential to understand that the odds ratio is not a ratio of likelihood or probabilities. If the disease rates for men and women are 0.80 and 0.40, respectively, then the odds ratio is

$$\frac{\frac{0.80}{0.20}}{\frac{0.40}{0.60}} = 6.00$$

In this example, men are twice as likely to get the disease but have six times the odds.

9.3.6 A 95% confidence interval for odds ratio

We write the confidence interval formulas for odds ratios (OR) for the natural logarithm, similarly to risk ratios. Consider a 2 by 2 table as before.

	Disease		
	Yes	No	Total
Exposure	a	b	$a + b$
No Exposure	c	d	$c + d$

The odds of disease occurrence in the exposed group is $\left[\frac{\frac{a}{a+b}}{\frac{b}{a+b}}\right] = \frac{a}{b}$. Similarly, the odds in the unexposed group is $\frac{c}{d}$. Hence, the odds ratio of disease occurrence is

$$\text{OR} = \frac{\texttt{Odds(Disease|Exposed)}}{\texttt{Odds(Disease|Not Exposed)}} = \frac{a/b}{c/d} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 \times (1-p_2)}{p_2 \times (1-p_1)}$$

If $OR < 1$, we can put the 'not exposed' group in the numerator, so that $OR > 1$ (the resulting odds ratio should, of course, be interpreted accordingly). In our hepatitis example, we can use the following table:

	hepatitis E		
	Yes	No	Total
Placbo	66	830	896
Vaccine	3	895	898

and get $OR = ((66)(895))/((830)(3)) = 23.7$ ("Placebo group has 24 times the odds of getting hepatitis"). If we use

	hepatitis E		
	Yes	No	Total
Vaccine	3	895	898
Placbo	66	830	896

then $OR = (3)(830)/(66)(895) = 0.04$ ("Vaccine group has 0.04 times the odds of getting hepatitis").

In either case, we will calculate the confidence interval for odds ratio in 4 steps.

1. Calculate the natural log of the odds ratio:

$$ln(OR) = ln\left(\frac{a \times d}{b \times c}\right)$$

2. Calculate the standard error of ln(OR) as follows:

$$SE_{ln(OR)} = \sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}$$

3. A 95% confidence interval for ln(OR) is given by

$$[ln(OR) - 1.96(SE), ln(OR) + 1.96(SE)]$$

4. Finally, a 95% confidence interval for OR is given by

$$e^{ln(OR)-1.96(SE)}, e^{ln(OR)+1.96(SE)}$$

Returning to the hepatitis E study, recall the disease occurrence data:

	hepatitis E		
	Yes	No	Total
Placbo	66	830	896
Vaccine	3	895	898

The odds ratio of getting hepatitis is

$$\texttt{Odds Ratio} = \frac{\texttt{Odds(Hepatitis|Placebo)}}{\texttt{Odds(Hepatitis|Vaccine)}} = \frac{66 \times 895}{3 \times 830} = 23.7$$

so that

1.
$$ln(OR) = ln(23.7) = 3.165$$

2.
$$SE_{ln(OR)} = \sqrt{\frac{1}{66} + \frac{1}{830} + \frac{1}{3} + \frac{1}{895}} = \sqrt{0.3508} = 0.5923$$

3.
$$[3.165 - 1.96(0.5923), 3.165 + 1.96(.5923)] = [2.004, 4.326]$$

4.
$$[e^{2.004}, e^{4.326}]$$

With 95% confidence, the odds of unvaccinated subjects getting hepatitis is between 7 and 76 times greater than vaccinated subjects.

9.4 Key Words

• α (Alpha)	• Confidence level
• Bias	• Efficiency
• Confidence interval	• margin of error

9.5 Exercises

Exercises

Exercise 9.5-1: In a study of drug usage by students at a large university, researchers obtained the data regarding hard liquor experience of
smokers and nonsmokers.

	Drug Use		
	Once or more	Never	Total
Smokers	23	18	41
Nonsmokers	15	56	71

1. Estimate the difference in the percentage
   of drug use between smokers and non-smokers.

2. Calculate a standard error for our estimate in (1).

3. Calculate a 95% confidence interval for the difference in the percentage of drug use between smokers and nonsmokers.

4. Estimate the risk ratio of drug use between smokers and nonsmokers.

5. Calculate a standard error for the natural log of our estimate in 4.

6. Calculate a 95% confidence interval for the risk ratio of drug use between smokers

7. Estimate the odds ratio of drug use between smokers and nonsmokers.

8. Calculate a standard error for the natural log of our estimate in 7.

9. Calculate a 95% confidence interval for the odds ratio of drug use between smokers and nonsmokers.

10. Interpret the above confidence intervals in parts 3, 6, and 9. Which are significant, and which are not? Why or why not?

Exercise 9.5-2: Time magazine reported the result of a telephone poll of 800 adult Americans. The reporter asked Americans the following question: "Should the federal tax on cigarettes be raised to pay for health care reform?"

	Federal Tax	
	on Cigarettes	
Status	Yes	No
Smoker	41	154
Nonsmoker	351	254

1. Estimate the difference in the percentage of Americans who supported the federal tax on cigarettes between smokers and non-smokers.

2. Calculate a standard error for our estimate in (1).

3. Calculate a 95% confidence interval for the difference in the percentage of Americans who supported the federal tax on cigarettes between smokers and non-smokers.

4. Estimate the risk ratio of Americans who supported the federal tax on cigarettes between smokers and non-smokers.

5. Estimate the odds ratio of Americans who supported the federal tax on cigarettes between smokers and non-smokers.

Exercise 9.5-3: The age at which a woman gives birth to her first child may be an essential factor in the risk of later developing breast cancer. An international study conducted by WHO selected women with at least one birth and recorded if they had breast cancer or not and whether they had their first child before their 30th birthday or after. In a sample of 3220 women who had their first child after their 30th birthday, 683 developed breast cancer. Whereas, in a sample of 10245 women who had their first child before their 30th birthday, 1483 developed breast cancer.

1. Estimate the difference in the percentage
   of developing breast cancer between women who had their first child after their 30th birthday and before their 30th birthday.

2. Calculate a standard error for our estimate in (1).

3. Calculate a 95% confidence interval for the difference in the percentage developing breast cancer between women who had their first child after their 30th birthday and before their 30th birthday.

4. Estimate the risk ratio of developing breast cancer between women who had their first child after their 30th birthday and before their 30th birthday.

5. Estimate the odds ratio of developing breast cancer between women who had their first child after their 30th birthday and before their 30th birthday.

Exercise 9.5-4: In a sample of 200 surgeons, 15% thought the government should control health care. Whereas, in a sample of 200 general practitioners, 21% thought the same.

1. Estimate the difference in the percentage of those who think the government should control health care between surgeons and general practitioners.

2. Calculate a standard error for our estimate in (1).

3. Calculate a 95% confidence interval for the difference in the percentage of those who think the government should control health care between surgeons and general practitioners.

4. Estimate the risk ratio of those who think the government should control health care between surgeons and general practitioners.

5. Estimate the odds ratio of those who think the government should control health care between surgeons and general practitioners.

Exercise 9.5-5: A Pew Research Center poll asked randomly selected subjects if they agreed with the statement that "It is morally wrong for married people to have an affair." Among the 386 women surveyed, 347 agreed and among the 359 men, 305 agreed. Use a 5

percent significance to test if the difference in significant?

Given the computer output, what should we conclude with 95 percent confidence interval?

CI for Two Proportion			
Sample	X	N	Sample p
1 (women)	347	386	0.898964
2 (men)	305	359	0.849582

Difference = p(1) - p(2)
Estimate for difference: -0.0493816
95% CI for difference: (0.00172287, 0.0970402)
Test for difference = 0 (vs not = 0): Z = 2.04,
p-val = .042

1. Determine the difference in the percentage of those who thought that it is wrong for married people to have an affair.

2. Calculate a standard error for our estimate in (1).

3. Determine a 95% confidence interval for the difference in the percentage of those who thought that it is wrong for married people to have an affair.

4. Estimate the risk ratio of those who thought that it is wrong for married people to have an affair.

5. Estimate the odds ratio of those who thought that it is wrong for married people to have an affair.

Answer:

Chapter 10

Sampling Distribution of the Mean

10.1 Objectives

After completing this part, students should be able to:

- Understand the "wall" between sample and population

- Define and explain the behavioral properties of the sample average

- Estimate the population mean

- Estimate, the population mean, using intervals

- Determine the sample size for estimating the population mean

10.2 Behavioral Properties of the Sample Average

What percentage of adult men are between 5'6" and 6' tall? Population surveys have shown that men's heights are approximately normally distributed with mean 5'9" and SD 3". Thus, the percentage of men between 5'6" and 6' is estimated as 68%, the percentage within 1 SD of the mean. See Figure 10.1 below.

If the population of men is randomly assigned into groups of 9, and the average heights are computed for each group, what percentage of groups average between 5'6" and 6' in height? Is the answer approximately 68%? No. In fact, more than 99% of the groups will average between 5'6" and 6', even though only 68% of individuals do. Why? Because averages tend to include tall, short and medium heights – therefore averages tend to fall closer to the middle than individuals. (Think about this: We put in a hat the names of all the men in the class.

Figure 10.1: The Percentage of Men's Heights between 66 and 72 inches
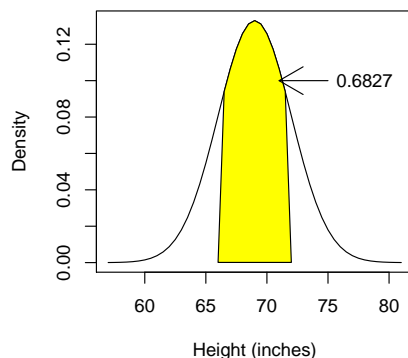
We will win \$20 if the names we draw average over 6 feet tall. Would we instead draw 1 or 2 names? What are our chances of winning if we draw nine names?)

The following is a small simulation study of the behavior of the sample mean. Fifty samples are drawn (each containing $n = 9$ individuals) from a population with mean 69 inches and SD 3 inches. For each sample, we calculate the average. Observe that none of the samples average over 71 inches, even though many individuals do.

Table 10.1: Heights of 50 samples of Nine Men

Sample	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Ave
1:	65.5	66.8	68.9	67.8	71.9	66.8	71.0	73.1	62.6	68.27
2:	68.6	71.2	72.6	64.3	70.9	70.0	69.0	69.8	62.4	68.75
3:	67.4	67.9	67.1	68.2	70.7	68.3	67.2	68.7	67.0	68.04
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
SD:	3.17	2.56	2.66	3.25	2.94	2.56	3.10	3.33	3.62	0.92
Mean:	68.53	68.18	68.92	68.51	69.26	68.72	68.66	68.86	68.11	68.64

The first lesson of this chapter says:

Averages are less variable than individuals

Do we see this in the simulation study? To make it easier to see, look at the SD of each column. The SD of individuals tend to be around 3.0 (the actual value), but the SD of averages is much smaller. How much smaller? We now state the central lesson of the chapter.

$$\boxed{\text{SD of } \bar{X} = \frac{\text{SD of individuals}}{\sqrt{n}}}$$

Since the individuals in the simulation study have SD of 3 inches, the SD of 9-member averages is:

$$\boxed{\text{SD of } \bar{X}_{n=9} = \frac{\text{SD of individuals}}{\sqrt{n}} = \frac{3}{\sqrt{9}} = 1.0}$$

Therefore, the real value for the SD of the last column is 1.0. The simulated SD is 0.92, which is close.

Example 1

Suppose that men's heights are normally distributed with a mean 5'9" and a SD 3".

1. What percentage of men are over 5'11" tall?

2. Select a man at random. What is the probability that he is over 5'11" tall?

3. If we calculate the average height for all possible samples of size nine that we can take, what percentage of averages will exceed 5'11"?

4. Given one randomly selected sample of size 9, what is the probability that the average height will exceed 5'11"?

5. Given a sample of size 25, what is the probability that the average height will exceed 5'11"?

6. Given a sample of size nine, the average height of the sample will exceed _____ with probability 0.90

7. 90% of samples of size nine will have an average height exceeding _____.

10.3 Estimating the Population Mean

Consider estimating the average GPA (call this μ) of the approximately 23,000 WMU undergraduates. In the absence of the complete database, we may wish to estimate μ by taking a random sample of, say, $n = 25$ students and computing the sample average (call this $\bar{X}$). Suppose $\bar{X} = 3.05$. Now, unless we got fortunate with the random sample, chances are the 3.05 estimate missed the true value μ. By how much?

On the average, a random variable misses its expected value by one standard deviation. We expect $\bar{X}$ to miss by how much? one SD of $\bar{X}$. Using the equation above, we can estimate this by $\frac{S}{\sqrt{n}}$, since the sample standard deviation S estimates the SD of individuals.

The population mean (μ) is estimated using the sample mean ($\bar{X}$). The estimate tends to miss the μ by an amount called the *standard error (SE)* of the mean which is calculated as $\frac{S}{\sqrt{n}}$:

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

Returning to the GPA example, suppose that the sample of $n = 25$ students yielded an average GPA of 3.05 and a standard deviation of 0.40. Then the WMU population average GPA is estimated as 3.05 with a standard error of $\frac{S}{\sqrt{n}} = \frac{0.40}{\sqrt{25}} = 0.08$.

Example 2

What is the average length of stay for undergraduate students at WMU?

1. Suppose 25 graduating students were randomly selected and asked about their length of stay. Suppose that the sample averaged 5.3 years, with an SD of 1.5 years. Then the WMU average stay is estimated as _____ years give or take _____ years or so.

2. A second sample of 100 students were interviewed. The mean and SD for the second sample were also 5.3 years and 1.5 years, respectively. Calculate an estimate for the WMU average stay and provide a standard error for your estimate.

Similar to the SE for proportions, the formula for the SE of the mean has the sample size (i) in the denominator, and (ii) inside the square root sign. Therefore, increasing the sample size by a factor of 4 makes the standard error decrease by a factor of $\sqrt{4}$.

The standard error of the mean decreases like the square root of the sample size (n).

10.4 Estimating the Population Mean Using Intervals

Variables tend to miss their expected value, but should be within 1 SD 68% of the time, and within 1.96 SD's 95% of the time. Changing notation for SD to SE, we get

$$|\bar{X} - \mu| \leq 1.96(SE)$$

95% of the time, where the SE is given in equation ??? 11.1. As a consequence,

$$\mu \text{ is inside the interval } \bar{X} \pm 1.96(SE)$$

95% of the time. This method gives an interval estimate of μ.

95% Confidence Interval for μ
A 95% Confidence Interval estimate for the population mean μ is given by

$$\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$$

The term $1.96 \frac{S}{\sqrt{n}}$ is called the 95% margin of error.

10.5 Sample Size for Estimating the Population Mean

If we wanted to reduce the margin of error to some value M, then we set the formula for margin of error equal to M, i.e., $M = 1.96 \frac{S}{\sqrt{n}}$. Solving for n gives the result we need.

In order to be 95% confident that the sample mean is within a distance M of the true population μ, choose a sample size equal to

$$n = \frac{1.96^2 S^2}{M^2}$$

where S is the standard deviation based on historical data or a pilot sample. The quantity M is called the 95% margin of error for μ.

10.6 Key Words

- α (Alpha)   • Confidence level

- Bias   • Efficiency

- Confidence interval   • margin of error

10.7 Exercises

———————

Exercises

Exercise 10.7-1: The national average on a science test for tenth graders has a mean of 210 and a standard deviation of 28.

1. What percentage scored over 220? (You may assume that the histogram of scores looks approximately like the normal curve.)

2. One tenth grader is randomly selected. What is the chance that he/she scored over 220?

3. A random sample of 40 tenth graders are selected. What is the chance that this group will average over 220? (Should this be smaller, larger, or approximately equal to (2)?)

4. A larger sample of 100 tenth graders is selected. What is the probability that this group will average over 220? (Should this be smaller, larger, or approximately equal to (3)?

Exercise 10.7-2: A sample of 16 observations is taken from a distribution with mean of 10 and standard deviation 2.

1. Suppose the sample a mean is 10.5. What is the standard error of this estimate?

2. What is the 95% margin of error?

3. What is the 95% confidence interval?

4. What is the probability the sample mean is greater than 9?

5. What is the probability the sample mean is less than 9?

6. What is the probability the sample mean is between 8 and 10?

7. 33% of sample means are above what value?

8. 33% of sample means are below what value?

Exercise 10.7-3: Safe Skies Airline took a random sample of 25 flights to estimate the average time that arriving passengers wait for luggage at the carousel. The sample average was found to be 16.2 minutes with a standard deviation of 4 minutes. The population average waiting time is estimated as _____ minutes give or take _____ minutes or so.

Exercise 10.7-4: The normal human body temperature is on average $98.6^o F$ with a standard deviation of $1.0^o F$. Researchers take a random sample of 35.

1. What is the distribution of the sample mean?

2. What is the mean of this estimate?

3. What is the standard error of this estimate?

Exercise 10.7-5: A manufacturing company's profits depend on the cost of materials. One material of interest is carbon fiber, which is used to make golf shafts and fishing rods. The cost per pound (in dollars) was recorded for ten randomly selected days from the first six months of 2002. The data follow:

$$7.6, 7.8, 8.8, 7.3, 6.6, 7.5, 6.7, 8.6, 7.4, 7.7$$

1. Calculate an estimate for the average cost per pound during the first six months of 2002.

2. Calculate a standard error for our estimate in (1).

Exercise 10.7-6: The average annual rainfall in Mawsynram, India (the wettest place on Earth) is 467.35 inches with a standard deviation of 5.12 inches. A random sample of 100 is taken.

1. What happens to the distribution of sample means as the sample size increases?

2. What is the mean of this distribution?

3. What is the standard error of this distribution?

Chapter 11

Comparing Two Means

11.1 Objective

After completing this part, students should be able to:

- Compare two population averages when the samples are independent using confidence intervals.

- Compare two population averages when the data are dependent (paired) using confidence intervals.

11.2 Estimating the Difference between Independent Means

Suppose we sample two means $\bar{x}_1$ and $\bar{x}_2$ from two populations with unknown means μ_1 and μ_2. Think of it as a picture:
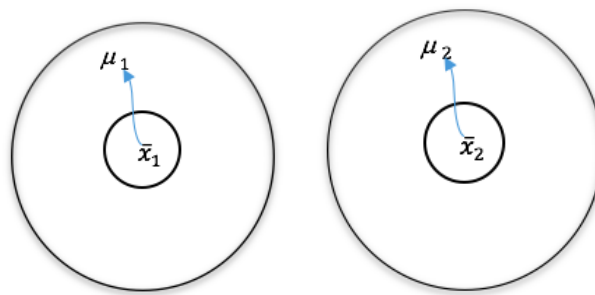
As you can see, the sample circles including $\bar{x}_1$ and $\bar{x}_2$ are smaller than the entire population circles including μ_1 and μ_2. In other words, we do not have all the information we need to be sure about the exact values of μ_1 and μ_2. However, as we have learned, statistics allows for

sample means to estimate population means as depicted by the arrows. What's new in this chapter is that here we learn statistics can do more than allow us to estimate values for μ_1 and μ_2 separately. We can also use statistics to investigate how the population means differ from one another in tandem. We do this by performing *hypothesis tests* on and using confidence intervals centered around the difference between the sample means $\bar{x}_1$ and $\bar{x}_2$. Let's now turn to some formulaic examples of such tests and intervals.

Is there "grade inflation" in WMU? How does the average GPA of WMU students today compare with, say ten years ago? Suppose a random sample of 100 student records from 10 years ago yields an average sample GPA of 2.90 with a standard deviation of 0.40. A random sample of 100 current students today yields a sample average of 2.98 with a standard deviation of 0.45. The difference between the two-sample means is 2.98- 2.90 = .08. Is this proof that GPA's are higher today than ten years ago? Well ... first, we need to account for the fact that 2.98 and 2.90 are not the correct averages but that we compute the average from random samples. Therefore, 0.08 is not the exact difference, but merely an estimate of the actual difference. By how much will it miss?

Note that we took differences $2.98 - 2.90$ to compare average GPA. The two averages 2.98 and 2.90 are independent, in the sense that we base them on separate and independent groups of students. The SE of the difference is

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} \tag{11.1}$$

whenever the two means are independent. Equation (11.1) is similar to the equation from chapter 7, the $SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{(SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2)}$, for proportions.

Now, applying the equation from chapter 10, $SE_{\bar{X}}$ which is calculated as $\frac{s}{\sqrt{n}}$ twice, we get

> The standard error of the difference between two independent means.
>
> $$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{11.2}$$

Continuing with the example, let $\bar{X}_1 = 2.98$ and $\bar{X}_2 = 2.90$. Then the sample standard deviations are $S_1 = 0.45$ and $S_2 = 0.40$. The sample sizes are $n_1 = 100$ and $n_2 = 100$.

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.45^2}{100} + \frac{0.40^2}{100}} = 0.06$$

Therefore, we can state the conclusion of the study as follows: "The average GPA of WMU students today is .08 higher than 10 years ago, give or take .06 or so." We also could have used equation (11.1) instead of (11.2) in calculating the standard error:

$$
\begin{aligned}
SE_{\bar{X}_1} &= \frac{S_1}{\sqrt{n_1}} = \frac{0.45}{\sqrt{100}} = 0.045 \\
SE_{\bar{X}_2} &= \frac{S_2}{\sqrt{n_2}} = \frac{0.40}{\sqrt{100}} = 0.040 \\
SE_{\bar{X}_1 - \bar{X}_2} &= \sqrt{0.045^2 + 0.040^2} = 0.06
\end{aligned}
$$

11.2.1 Using a confidence interval

The following two sections discuss the formulas and concepts necessary for calculation and interpretation (respectively) of confidence intervals on the difference between two independent means. Let's start with the concepts and then proceed to some formulas and examples.

Confidence Interval - Concepts:
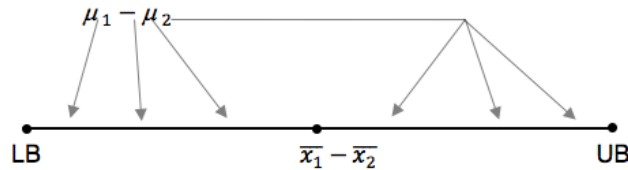
The interval between lower and upper bounds (LB, UB) includes some possible values of $\mu_1 - \mu_2$ (depicted by the many arrows). We need an interval around the central estimate $\bar{x}_1 - \bar{x}_2$ because of variation between samples and populations (which was depicted using concentric circles at the start of this chapter). This interval can be thought of as the "wiggle room" needed to estimate $\mu_1 - \mu_2$ using only $\bar{x}_1 - \bar{x}_2$.

Here we show how to use this interval, and it is necessary to talk about some basic properties of a difference first. For any two numbers A and B, there are three possibilities when evaluating their difference:

1. If $A - B$ is a positive number, then A is greater than B. Consider the numbers 4 and 3. If we take $4 - 3 = 1$, the answer is greater than zero.

2. If $A - B$ is a negative number, then B is higher than A. For instance if we take $3 - 4 = -1$, the answer is negative.

3. If $A - B$ is 0, then A = B. Consider the numbers 4 and 4: $4 - 4 = 0$.

The same kind of reasoning holds for all the possible values of $\mu_1 - \mu_2$ between LB and UB depicted above:

1. If all the values from LB to UB are positive, then μ_1 is significantly greater than μ_2.

2. If all the values from LB to UB are negative, then μ_1 is significantly less than μ_2.

3. If zero is between LB and UB (inclusive), then μ_1 is not significantly different from μ_2.

Note that this is to say that the same reasoning and terminology outlined in the section on statistical significance in chapter 7 apply in the new case of a difference. If the confidence interval for $\mu_1 - \mu_2$ does not contain 0, then 0 has been effectively excluded from the range of possible values.

> When the confidence interval for $\mu_1 - \mu_2$ does not contain **zero**, we say that the difference is statistically **significant**.

Confidence Interval - Calculations and Examples:

The difference of two means is a random variable with expected value and spread. The 68% and 95% rules apply, i.e. the estimated difference of $\bar{x}_1 - \bar{x}_2$ should be within 1 SE of the true value 68% of the time, and within 1.96 SE's 95% of the time. Following the usual reasoning,

$$(\bar{X}_1 - \bar{X}_2) \pm 1.96 SE_{(\bar{X}_1 - \bar{X}_2)}$$

should contain the true difference $(\mu_1 - \mu_2)$ with 95% confidence. Substituting (11.2), we get the following formula.

> 95% confidence interval for $(\mu_1 - \mu_2)$:
>
> $$(\bar{X}_1 - \bar{X}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (11.3)$$

For grade inflation, we have

$$(2.98 - 2.90) \pm 1.96\sqrt{\frac{0.45^2}{100} + \frac{0.40^2}{100}}$$

$$0.08 \pm 1.96(0.06)$$

$$(-0.04, 0.20)$$

We say that the difference in GPA averages is between $-.04$ and $.20$ with 95% confidence. Note that 0 has not been excluded, making simple chance variability a viable explanation for the observed difference.

11.2.2 Statistical Significance

Let us revisit the diet study mentioned earlier. The following table contains the mean changes in body mass index (weight in kilograms divided by height in meters squared) for the Atkins, Zone and Ornish diets. Now, compare the Atkins and Zone diets at 12 months:

Table 11.1: Mean Changes (SD) in Body Mass Index by Diet Group & Time

Time (months)	Atkins ($n = 77$)	Zone ($n = 79$)	Ornish ($n = 76$)
2	-1.60(0.98)	-0.76(0.99)	-0.95(0.90)
6	-2.16(2.14)	-0.73(0.90)	-0.85(1.60)
12	-1.65(2.54)	-0.53(2.00)	-0.77(2.14)

On the average, Atkins lost how much more body mass index points than Zone? The estimate of $(\mu_1 - \mu_2)$ is

$$(\bar{X}_1 - \bar{X}_2) = (-0.53) - (-1.65) = 1.12$$

The standard error is

$$SE_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{(2.00)^2}{79} + \frac{(2.54)^2}{77}} = 0.367$$

The 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$1.12 \pm 1.96(0.367)$$

$$(0.84, 1.84)$$

Thus, even allowing for 1.96 SE's of chance variability, Atkins lost at least .40 more body mass index points than Zone (and could be as large as 1.84). When the confidence interval for $(\mu_1 - \mu_2)$ does not contain zero, we say that the difference is statistically significant.

11.2.3 The P-value

Continuing the example of the previous section, we might ask "Can't the difference in averages $\bar{x}_1 - \bar{x}_2 = 1.12$ be explained by chance variability, rather than diet effect?"

The answer is "Yes, 1.12 can occur by chance, but with very tiny probability." How small? Well, if the actual difference were 0, and the SE is 0.367, then the value 1.12 is

$$\frac{1.12}{0.367} = 3.05$$

SE's from the expected value using the normal curve, random variables fall as far as 3.05 or more SE's from the expected value with approximately 0.0022 probability. Since this number (also called the P-value) is quite small, it makes it hard to believe that the actual difference is zero. Hence, we conclude that statistically, the two means are different. Alternatively, we can say that the means are *significantly* different.

Table 11.2: Weight in pounds before and after after 12 months on diet

Subject	Before	After
1	180	155
2	192	187
3	205	194
4	166	176
5	220	205
6	177	172
7	189	173
Ave:	189.9	180.3
SD:	18.1	16.4

11.3 Paired data (before-and-after)

In this section, we will discuss a common problem in data analysis: comparing before and after measurements. Consider the possible weight loss data in Table 11.2.

Using the notation of the previous section estimating the difference between independent means, we have

$$\bar{X}_1 = 189.9 \qquad \bar{X}_2 = 180.3$$
$$S_1 = 18.1 \qquad S_2 = 16.4$$
$$n_1 = 7 \qquad n_2 = 7$$

What is the estimate of mean weight change after 12 months? $\bar{x}_1 - \bar{x}_2 = 9.6$ pounds, right? What is the standard error of this estimate? Using equation (11.2)

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(18.1)^2}{7} + \frac{(16.4)^2}{7}} = 9.2 \text{ pounds}$$

right? **wrong!**

We should not use equations (11.1) or (11.2) because the two means are **not** independent, i.e., they are not calculated on independent samples. The use of the plural 'samples' is in itself wrong because we do not have two samples, we only have one! We need to watch out for this. How many samples are there? Before-and-after data generally consist of only one sample of subjects, each measured twice.

So how do we calculate an estimate and standard error of average weight loss? By calculating the amount of change from Before to After. The computed value amounts to taking differences, as shown in Table 11.3.

Table 11.3: Weight in pounds before and after after 12 months on diet

Subject	Before	After	Difference
1	180	155	25
2	192	187	5
3	205	194	11
4	166	176	-10
5	220	205	15
6	177	172	5
7	189	173	16
Ave:			9.6
SD:			11.1

Compare Table 11.3 to Table 11.2. We have reduced the summary statistics to a single sample, appropriately. The relevant statistics are now:

$$\bar{X} = 9.6, S = 11.1, n = 7$$

What does the sample mean $\bar{X} = 9.6$ estimate? It estimates the average change, right? To be specific, it estimates the average weight loss from Month 0 to Month 12. What is the standard error of the estimate? Since it is just another average, the appropriate procedure is given by

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{11.1}{\sqrt{7}} = 4.2 \texttt{ pounds}$$

Completing the analysis, the 95% confidence interval for average change is

$$9.6 \pm 1.96(4.2)$$

$$(1.4, 17.8)$$

Since the interval does not contain zero, the average weight change is statistically significant.

11.3.1 Paired Data

Paired data are data in which natural matchings occur. For example, when researchers collect two measurements (one before treatment and one after treatment) from a subject, then we have paired data, and the analysis should follow as described above.

For example, we might want to compare the head injury of drivers versus passengers in car crashes. In a study, automobiles were crashed into a wall at a speed of 35 MPH with dummies in the driver and front passenger seat. The head injury criterion (HIC) was measured. Following is a selection of cars and their HIC values.

Company	Driver	Passenger
Acura Integra 87	599	597
Audi 80 89	600	515
Chevrolet Camaro 91	585	583
Ford Escort 87	551	418
Honda Accord LX 91	562	539
Toyota Corolla Fx 88	593	397
Volvo 740 GLE 88	519	445

Since we uniquely match each pair of observations (e.g., 599 and 597) to each other (driver and passenger HIC in the same car and crash), this is paired data and deserves paired data analysis.

11.4 Key Words

- independent sample experiment

- paired difference experiment

- confidence intervals

- sampling distribution

11.5 Exercises

_____

Exercises

Exercise 11.5-1:   Do credit cards with no annual fee charge higher interest rates (APR) than cards that have annual fees? Among 29 cards surveyed, 17 had no annual fees while 12 charged an annual fee. Among the cards with no annual fee, the average APR was 19% (SD=8%). Among cards with an annual fee, the average APR was 17% (SD=3%).

1. Estimate the difference in APR.

2. Calculate a standard error for your estimate in (1).

3. Calculate a 95% confidence interval for the difference in APR between the two groups.

4. Are interest rates significantly higher for cards with no annual fees?

5. What is the P-value for comparing the two averages? Is the difference significant?

Exercise 11.5-2:   100 students graduating with Bachelor degrees in engineering make an average of $70,000 with a standard deviation of $5,000 when entering the workforce. 68 students graduating with Bachelor degrees in statistics make an average of $65,000 with a standard deviation of $3,000 when entering the workforce. Assuming the two samples of students are independent, answer the following questions:

1. What is the difference between the sample averages (engineers – statisticians)?

2. What is the standard error of the estimate for the true difference in average entrance salaries between all engineering and statistics BA graduates?

3. What is a 95% confidence interval for the difference in sample averages?

4. Do these statistics suggest that average entrance salaries for stats vs. engineering students are significantly different?

Exercise 11.5-3:   A Junior at Southwest Michigan college is debating whether to pursue an MBA after her Bachelor degree in management. She interviews some people she knows in the workforce and was able to obtain their salaries. The annual salaries (in dollars) are summarized in the following table:

Degree	Aver	SD	Sample Size
Bachelor	$48,286	$416	12
Master	$59,496	$675	7

1. Estimate the difference in average salary between the two groups.

2. Calculate a standard error for our estimate in (1).

3. Calculate a 95% confidence interval for the average difference in salary.

4. Are MBA salaries significantly higher?

5. What is the P-value for comparing the two averages? Is the difference significant?

Exercise 11.5-4: A group of charity workers employed by a large foundation track the donations made to the foundation every month. They note that the average donations for August were somewhat higher than the average donations for September, and they want to know if this fact should worry them moving forward. Performing the correct statistical test could determine whether the average difference in donations between August and September was significant. However, they must first decide whether their August and September donations are independent or not. If we were advising the charity workers, what would we tell them?

Exercise 11.5-5: A new gasoline additive is supposed to make gas burn more cleanly and increase gas mileage in the process. Consumer Protection Anonymous conducted a mileage test to confirm this. They took seven of their cars, filled it with regular gas, and drove it on I-94 until it was empty. They repeated the process using the same cars, but using the gas additive. The recorded gas mileage follows:

Additive	1	2	3	4	5	6	7
Without	22	15	18	28	12	25	18
With	26	19	17	34	17	25	22

1. Calculate the mean difference in mileage between the two fuel types.

2. Calculate a standard error for our estimate in (1).

3. Calculate a 95% confidence interval for the mean mileage difference.

4. Does the data support the claim of higher gas mileage?

Exercise 11.5-6: The group of charity workers from the above question decides on a statistical test based on the sage wisdom we previously offered. They perform a test for significance of the average difference, and it yields a p-value of 0.50. As their stats advisor, how would we interpret this p-value concerning whether the average difference in donations for August versus September was significant?

129

Chapter 12

Categorical Variables: Association or Independence

12.1 Objective

After completing this part, students should be able to:

- Compute expected frequencies.

- Understand the difference between association and independence.

- Test for a statistical association.

12.2 Association versus independence in an r x c table

Is there an association between gender and height? Yes, males tend to be taller than females. A more formal way of saying this is 'height distribution for males tends to be different from females.' Is there an association between shoe size and height? Yes, 'height distribution for men who wear size 12 is different from those who wear size 8.' Is there an association between GPA and height? No, 'height distribution tends to be the same for 3.0 students as well as 3.5 students.'

Two variables A and B are said to be **associated** if the distribution of B tends to change with the level of the A variable. Otherwise, they are said to be **independent** variables.

Therefore, height is associated with gender and shoe size, but independent of GPA.

If we are thinking, "association and independence are the same," we are almost correct. The difference is about design. In the test of independence, we collect observational units

at random from a population, and the two categorical variables are observed for each unit. In the test of association, we collect the data by randomly sampling from each sub-group **separately**. (Say, 100 Democrat, 100 Republican, 100 Independent, and so on.) The null hypothesis is that each sub-group shares the same distribution of another categorical variable. (Say, "chain smoker", "occasional smoker", "non-smoker".) The difference between these two tests is subtle yet important.

Now consider the following 3 by 4 table. Researchers followed 189 students entering a business school program as part of attrition (i.e., drop out, transfer) study. The students were cross-classified according to 4 categories of high school GPA $(2.0 - 2.5, 2.5 - 3.0, 3.0 - 3.5, 3.5 - 4.0)$ and three categories of attrition outcomes ('did not return for the 2nd year,' 'returned for second but not for a 3rd year,' 'returned for 3rd year'). Is there an association between HS GPA and college attrition?

Table 12.1: Retention versus HS GPA

| Returned | GPA | | | |
	$2.0 - 2.5$	$2.5 - 3.0$	$3.0 - 3.5$	$3.5 - 4.0$
No –2nd yr	25	3	4	6
No – 3rd yr	14	7	4	6
Yes – 3rd yr	41	7	28	44

To analyze whether attrition and GPA are independent, we will analyze whether attrition distribution remains the same regardless of GPA level. Let us start by looking at the 1st column (worst HS grades) and 4th column (best HS grades). Do the distributions look the same? The answer seems to be 'no' - a bigger proportion of the 1st column never returned for their second year. In other words, the value '25' in the very first cell is too large, implying that 'poor grades seems to be associated with first-year attrition.' If grades and attrition were independent, Table 12.1 should have looked more like Table 12.2.

Table 12.2: Expected counts (if independent)

| Returned | GPA | | | |
	$2.0 - 2.5$	$2.5 - 3.0$	$3.0 - 3.5$	$3.5 - 4.0$
No –2nd yr	16	3	7	11
No – 3rd yr	13	3	6	9
Yes – 3rd yr	51	11	23	36

The table shows expected counts under independence. Observe that the row and column totals of the two tables are the same

Furthermore, note that 20.1% of the data is in the first row, 16.4% in the second row, and 65.5% in the third row. If we apply the same percentage breakdown to each column, we get

Table 12.3: Expected counts (if independent)

Returned	GPA				Total	Percent
	$2.0 - 2.5$	$2.5 - 3.0$	$3.0 - 3.5$	$3.5 - 4.0$		
No – 2nd yr					38	(20.1%)
No – 3rd yr					31	(16.4%)
Yes – 3rd yr					120	(63.5%)
Total	80	17	36	56	189	(100%)

$80 \times .201 = 16.08$	$17 \times .201 = 3.42$	$56 \times .201 = 7.24$	$56 \times .201 = 11.26$
$80 \times .164 = 13.12$	$17 \times .164 = 2.79$	$56 \times .164 = 5.90$	$56 \times .164 = 9.18$
$80 \times .635 = 50.80$	$17 \times .635 = 10.80$	$56 \times .635 = 22.86$	$56 \times .635 = 35.56$

Rounding off gives us the expected frequencies in Table 12.2.

12.2.1 Testing for statistical association

Statisticians will conclude 'independence' if Tables 12.1 and 12.2 are close and conclude 'association' if they are far from each other. We measure closeness and fairness by subtraction and squaring, as follows:

$$\chi^2 = \frac{(25 - 16.08)^2}{16.08} + \frac{(3 - 3.42)^2}{3.42} + \cdots + \frac{(44 - 35.56)^2}{35.56} = 23.42$$

Note that if Tables 12.1 and 12.2 are the same, then the χ^2 statistic (pronounced 'chi-square') in (12.1) will be zero (0). If the two tables are far apart, the χ^2 statistic will be large. Statisticians use the following rule.

If the $\chi^2 > b$, then conclude statistic association

Otherwise, conclude independence. The number b is called a critical value and depends on the dimensions of the table. Let r be the number of rows, and c be the number of columns. Let

$$df = (r - 1) \times (c - 1)$$

be a parameter called the degrees of freedom. Then b is given by the following table.

In our example on grades and attrition, we have r=3 rows and c=4 columns, so that

df	1	2	3	4	5	6	7	8	9	10
b	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31

$$df = (3-1) \times (4-1) = 6$$

so, the line between statistical association and independence is drawn at b=12.59. Since $\chi^2 = 23.42$ from (12.1), then $\chi^2 > 12.59$. We conclude that there is significant association between high school GPA and college attrition rate.

12.3 Key Words

- chi-square test for indep
- contingency table
- homogeneity
- multinomial experiment

12.4 Exercises

_____

Exercises

Exercise 12.4-1: In a study of drug usage by students at a large university, data was obtained regarding hard liquor experience of smokers and nonsmokers.

Hard-Liquor Use	Nonsmokers	Smokers
Once or more	15	23
Never	56	18

Is liquor use associated with smoking? Conduct a chi-square test to assess significance of association.

Exercise 12.4-2: 2. During the filming of an original comedy special, Netflix monitored whether audience members who received free tickets laughed during the show (LDS). The results follow:

	LDS		
Free Tickets	Yes	No	Total
Yes	17	1	18
No	28	43	71
Total	45	44	89

1. Calculate the expected count for audience members who did not receive a free ticket and did not laugh during the show.

2. Calculate the expected count for audience members who did not receive a free ticket and did laugh during the show.

3. Calculate the expected count for audience members who did receive a free ticket and did laugh during the show.

4. Calculate the expected count for audience members who did receive a free ticket and did not laugh during the show.

5. Calculate the chi-square test statistic.

6. What do you conclude?

Exercise 12.4-3: A study investigating the association between size of cars and country found the following frequency counts:

	US	Japan	UK	France
Economy	21	24	33	55
Compact	27	35	37	40
Full size	36	11	12	4
Luxury	15	3	7	8

Is there evidence of a significant relationship between size of car and country, or are the two variables independent?

Exercise 12.4-4: Suppose Netflix held another special, collected data, and had a statistician calculate and interpret the chi square test statistic. However, this time, the statistician found insignificant differences between observed and expected counts for all those who did and did not laugh with and without free tickets. What is the appropriate conclusion in this case?

Exercise 12.4-5: Computer-controlled cameras are being used to ticket automobile drivers for speeding and running red lights. These devices are operated by private firms

and have an incentive to pull in as many drivers as they can. Although approximately 70% of the motorists stoically accept and pay these tickets, others resent this procedure and fight the ticket. A frequency table with marginal totals is given below.

	Volation		
Ticket	Run Red Light	Speeding	Total
Pay			140
Fight			60
Total	60	140	200

1. Compute the table of expected frequencies.

2. Suppose we know that $1/3$ of those who were ticketed for running a red light fought the ticket. Is this enough information to conduct a test of association or independence between the two variables?

3. Using the information in (b), compute the chi-square statistic for testing independence or association between the two variables.

4. What is the correct degrees of freedom to use?

5. What is the conclusion of your test?

Chapter 13

Correlation

13.1 Objective

After completing this part, students should be able to:

- Interpret a scatter plot.

- Interpret the correlation coefficient r and the coefficient of determination r^2.

The following data appeared in the Wall Street Journal in 1984. Advertisements were selected by an annual survey conducted by Video Board Tests, Inc., a New York ad-testing company, based on interviews with 20,000 adults who were asked to name the most outstanding TV commercial they had seen, noticed, and liked. We based the retained impressions on a survey of 4,000 adults, in which regular product users were asked to cite a commercial they had seen for that product category in the past week. 'TV Ad Budget' was the 1983 advertising budget in \$ millions. 'Impressions' is the estimated number of million impressions per week.

Figure 13.1 shows a scatterplot of Impressions Score (Y) versus Ad Spending (X). Note that the points seem to fall around a line sloped upwards loosely. We say that there is a positive linear association or a linear relationship between spending and the number of impressions made.

If the points fall around a straight line sloped downwards, we say that there is a negative association.

Researchers often express the direction and the strength of association in a single number called the (Pearson) *correlation coefficient*. Typically denoted by r, the correlation coefficient r is a number between -1 and $+1$, inclusive. A value of $r = 0$ means that no linear association exists; the points either look like a random scatter or fall around a horizontal line. A value of $r = +1$ indicates a perfectly linear relationship; all the points fall on a straight line sloped

Company	TV Ad Budget	Impressions
Miller Lite	50.1	32.1
Pepsi	74.1	99.6
Stroh's	19.3	11.7
Fed'd Express	22.9	21.9
Burger King	82.4	60.8
Cola-Cola	40.1	78.6
McDonald's	185.9	92.4
MCI	26.9	50.7
Diet Cola	20.4	21.4
Ford	166.2	40.1
Levi's	27.0	40.8
Bud Lite	45.6	10.4
ATT/Bell	154.9	88.9
Calvin Klein	5.0	12.0
Wendy's	49.7	29.2
Polandoid	26.9	38.0
Shasta	5.7	10.0
Meow Mix	7.6	12.3
Oscar Meyer	9.2	23.4
Crest	32.4	71.1
Kibbles 'n Bits	6.1	4.4

upwards. If $r = -1$, all the points fall on a straight line sloped downwards. The correlation between TV Ad Budget and Impressions in the data above is $+0.65$.

Figure 13.2 on the following page shows various scatterplots with different correlations. Although 0.5 is halfway between 0 and 1, note that the plot corresponding to $r = 0.5$ barely shows a pattern of association. In practice, plots can show correlations up to 0.3 purely by accident (e.g., the correlation between GPA and, say, shoe size). When correlation reaches $+1$ or -1, all points fall on a straight line.

13.2 Computing the Pearson Correlation Coefficient

The following numerical example shows how to calculate r. We will break the calculations down into four steps.

1. Calculate the mean and SD of X and Y.

Plot of Impressions vs. Ad Budget
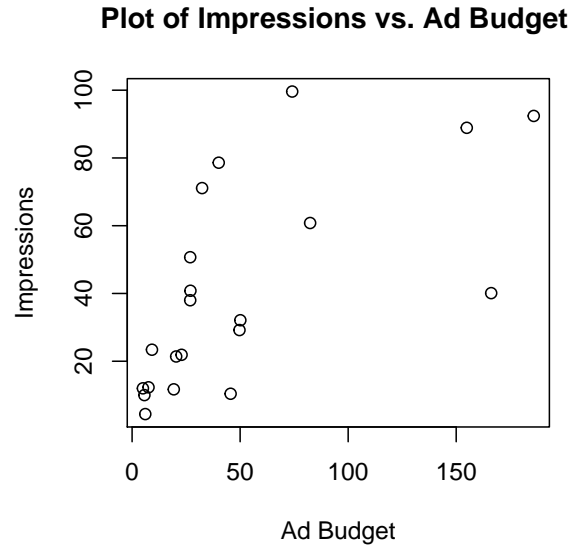
Sample	X	Y
1	1	3
2	3	9
3	4	7
4	4	9
5	5	15
6	7	11
Average	4	9
SD	2	4

2. Calculate the Z-scores from equation (4.1) on page 42.

$$Z_x = \frac{X - \bar{X}}{S_x} = \frac{X - 4}{2} \ \text{ and } \ Z_y = \frac{Y - \bar{Y}}{S_y} = \frac{Y - 9}{4}$$

3. Multiply the Z-scores and add up.

4. Finally, the correlation is the sum divided by $(n - 1)$.

$$r = \frac{3.75}{6 - 1} = 0.75$$

We summarize the whole process in the following formula.

Z_x	Z_y	$Z_x Z_y$
-1.5	-1.5	2.25
-0.5	0	0
0	-0.5	0
0	0	0
0.5	1.5	0.75
1.5	0.5	0.75
Average		3.75

The Correlation coefficient r:

$$r = \frac{\sum Z_x Z_y}{n-1} \qquad (13.1)$$

Consider the Ad Spending example at the start of this chapter. Many of the (X, Y) points are simultaneously above average since companies that have higher than average Advertising Spending also have higher than average Impressions.

Both $X - \bar{X}$ and $Y - \bar{Y}$ are positive for these companies; therefore, Z_x and Z_y are both positive, and the product $(Z_x)(Z_y)$ is positive for these companies. Most of the remaining companies have lower than average Spending and lower than average Impressions. Both Z_x and Z_y are negative for these companies, but the product $(Z_x)(Z_y)$ is still positive! Hence the numerator in (13.1) tends to be a large positive number for the Ad Spending data.

If the points were sloped downwards, then high X-values tend to go with low Y-values, and many points have a negative product $(Z_x)(Z_y)$. This procedure shows how the correlation formula (13.1) works.

If X and Y tend to be simultaneously above average or simultaneously below average, then the correlation coefficient will be **positive**.
If the data is paired above-average X values with below-average Y values, then the correlation coefficient will be **negative**.

13.3 Key Words

- correlation coefficient
- correlation plots
-
- Pearson Correlation
- scatter plot

Figure 13.1: Correlation plots (from top to bottom, r = -0.5391, 0.6254, -0.9467, 0.9960)
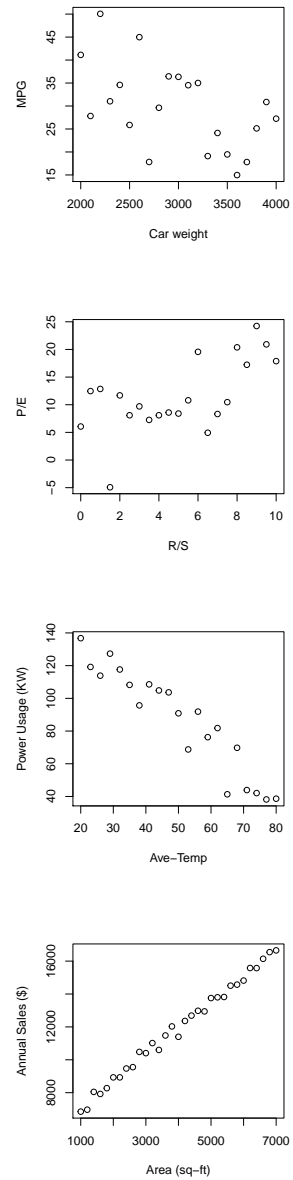
13.4 Exercises

————————

Exercises

Exercise 13.4-1: Consider the following data.

Table 13.1:

X	Y
-2	0
2	3
5	10
-1	1
6	15

1. Compute the correlation between X and Y.

2. Compute the correlation between Y and X.

3. Add 5 to Y, so the new values are 5, 8, 15, 6, 20. Now compute the correlation between X and Y. Is the correlation smaller, larger, or the same as before?

4. Multiply Y by 5, so the new values are 0, 15, 50, 5, 75. Now compute the correlation between X and Y. Is the correlation smaller, larger, or the same as before?

5. Multiply Y by -1, so that the new values are 0, -3, -10, -1, -15. Now compute the correlation between X and Y. Is the correlation smaller, larger, or the same as before?

Exercise 13.4-2: Suppose we have two variables, X and Y. If an increase in X results in the same increase in Y, then what is the correlation coefficient r?

Exercise 13.4-3: Suppose we have two variables, X and Y. If a decrease in X results in the same increase in Y, then what is the correlation coefficient r?

Exercise 13.4-4: Suppose that we plot two variables, X, and Y in a scatterplot, and we observe that the points appear to be clustered around a line of best fit that is tilted upward from left to right. What is the possible range of the correlation coefficient r?

Exercise 13.4-5: Suppose we have two variables, X and Y. The points (x, y) are plotted in a scatterplot, and it is observed that the points appear to be clustered around a line of best fit that is tilted downward from left to right. What is the possible range of the correlation coefficient r?

Exercise 13.4-6: Suppose we have two variables, X, and Y. The products of all their respective Z scores are calculated, and none of them are negative. What can we conclude about the correlation coefficient r?

Exercise 13.4-7: Suppose we have two variables, X and Y. The products of all their respective Z scores are calculated, and all the products are negative. What can we conclude about the correlation coefficient r?

Chapter 14

Linear Regression

14.1 Objective

After completing this part, students should be able to:

- Elucidate the simple linear regression model.

- Detect the least square point estimates of the slope and y-intercept.

- Report the assumptions behind simple linear regression and calculate the standard error.

- Evaluate the significance of the slope and y-intercept.

- Compute and elucidate the confidence interval for the mean value and the prediction interval for individual value.

- Put into words the concepts of the total, unexplained, and explained variation.

- Use correlation and regression techniques to evaluate two-variable relationships.

14.2 Simple Linear Regression

The following table contains data on winning bid price for 12 Saturn cars on eBay in July 2002. The car mileage is also given, and the cars have been arranged in increasing order of Miles.

Here is a scatterplot of the data. Since Price depends on Miles (not the other way around), we let Price be the Y-variable, or the *response variable*. Miles is the X-variable, or the *explanatory variable*.

Scatterplot of Price ($) vs. Miles

Car	Miles	Price ($)
1	9300	7100
2	10565	15500
3	15000	4400
4	15000	4400
5	17764	5900
6	57000	4600
7	65940	8800
8	73676	2000
9	77006	2750
10	93739	2550
11	146088	960
12	153260	1025
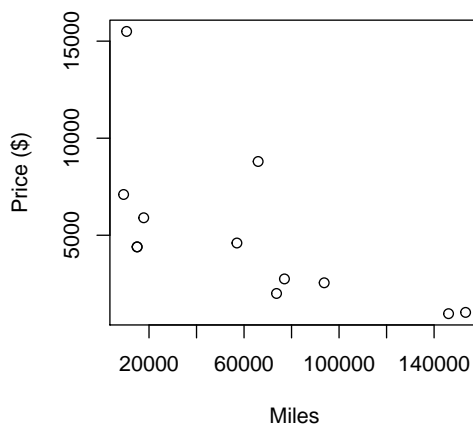
Problem:

Based on the data, how much do we expect to get for a Saturn car that has been driven 60,000 miles?

Simple linear regression is a data analysis technique that tries to find a *linear* pattern in the data. We then use this line for prediction.

Notice that the points seem to fall around a *straight line* sloping downwards. Can we draw this line? We will discuss one way to do this, called the *least squares* (LS) method. For now, suppose that the LS line has already been computed (we will do this later). The LS line is overlayed on the scatterplot looks like Figure 14.1.

The formula for this line, in the form $Y = a + bX$, is

$$\texttt{Predicted Price} = 8136 + (-0.05127)(\texttt{Miles})$$

The *slope* of the line is -0.05127, which means that predicted Price tends to drop 5 cents for every additional mile driven or about $512.70 for every 10,000 miles. The *Y-intercept* of the line is $ 8136; this should not be interpreted as the predicted price of a car with 0 mileage because the range of the data does not include cars with 0 miles.

Figure 14.1: Least Squares (LS) Regression Line is overlayed on Scatterplot
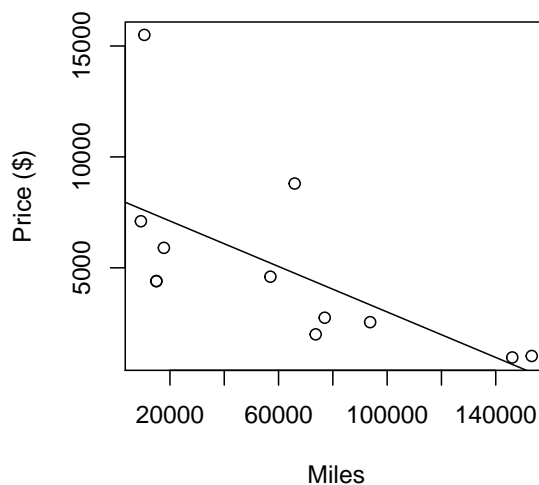
We can now use the line to predict the selling price of a car with 60000 miles. What is the height or Y value of the line at $X = 60000$? The answer is

$$\texttt{Predicted Price} = 8136 + (-0.05127)(60000) = \$5059.80$$

alternately, about \$5000 or \$5100 or so.

14.3 Calculating the Least Squares Regression Line

One way to calculate the regression line is to use these five statistics:

$$\bar{X}, S_{\bar{X}}, \bar{Y}, S_{\bar{Y}}, \texttt{ and } r$$

(i.e., the mean and SD of X, the mean and SD of Y, and the correlation between X and Y.)

The least square regression line is given by the equation

$$\texttt{Predicted} = a + bX$$

where the slope b and the intercept a are calculated as

$$b = r\frac{S_y}{S_x} \tag{14.1}$$

$$a = \bar{Y} - b\bar{X} \tag{14.2}$$

Next, we will perform the calculations for the Saturn Price data.

Car	Miles	Price ($)
1	9300	7100
2	10565	15500
3	15000	4400
4	15000	4400
5	17764	5900
6	57000	4600
7	65940	8800
8	73676	2000
9	77006	2750
10	93739	2550
11	146088	960
12	153260	1025
Average	61195	4999
SD	50989	4079
r	-0.641	

Using the formulas for slope and intercept in equation (14.1 and 14.2)

$$b = r\frac{S_y}{S_x} = -0.641\frac{4079}{50989} = -0.05127$$

$$a = \bar{Y} - b\bar{X} = 4999 - (-0.05127)(61195) = 8136$$

so that the regression line is

$$PREDICTED = a + bX = 8136 + (-0.05127)X$$

Regarding the original variable names, the regression line is

$$\texttt{Predicted Price} = 8136 + (-0.05127)\texttt{Miles}$$

14.4 More on Simple Regression

Why is this called the least squares line? Example best shows the answer.

Car	Miles	Price ($)	PRED	RES = Y - PRED
1	9300	7100	7659.35	-559.35
2	10565	15500	7594.49	7905.51
3	15000	4400	7367.11	-2967.11
4	15000	4400	7367.11	-2967.11
5	17764	5900	7225.41	-1325.41
6	57000	4600	5213.82	-613.82
7	65940	8800	4755.47	4044.53
8	73676	2000	4358.85	-2358.85
9	77006	2750	4188.13	-1438.13
10	93739	2550	3330.24	-780.24
11	146088	960	646.36	313.64
12	153260	1025	278.66	746.34

The first car has $Miles = 9300$. What is its predicted price? The predicted value is

$$\texttt{Predicted Price} = 8136 + (-0.05127)(9300) = 7659.35$$

This predicted value missed the actual selling price $Y = 7100$. By how much? By

$$\texttt{Residual} = 7100 + 7659.35 = -559.35$$

The negative value means actual value is too low. This difference is called the residual.

Small residuals (ignoring the sign) are good because this means the prediction was close (Car 1 above was predicted well, but Car 2 was not – the selling price is almost double what was predicted). Therefore, a prediction line is okay if it gives residuals that are as small as possible.

> The sum of squared residuals is
>
> $$SSE = (-559.35)^2 + \cdots + (746.34)^2 = 107805718.50$$

and is a measure of 'overall size' of the residuals. In the Saturn Price data, $SSE = 107,805,718$.

> The least square line given by (14.1) will have a smaller SSE than any other straight line.

This means that if you use any other intercept and slope combination besides $(a, b) = (8136, .05127)$, the new set of predicted values and residuals will give an SSE that is larger than, or at best equal to 107,805,718.

14.5 A 95% Confidence Interval for Slope

Is there a linear relationship between X and Y? It seems evident that selling price (Y) responds to a car's mileage (X), but in science, relationships are often not too noticeable and need confirmation by data. For example, does an individual's systolic blood pressure (Y) tend to increase with their cholesterol level (X)? Is there a relationship between one's total number of years of education (X) and income (Y)? In this section, we will investigate the strength of linear relationships by looking at the slope estimate. Since the slope represents how much Y responds to changes in the X-value, we will calculate a 95% confidence interval for the slope, and examine whether it excludes 0. If it does, then we can rule out the likelihood that the slope is 0. Thus, we conclude that there is a significant linear relationship between X and Y.

We start by stating the formula for standard error:

> The slope estimate b tends to miss the true value β by an amount called the *standard error* of the slope, denoted SE of b and calculated as:
>
> $$SE_b = \sqrt{\frac{(1 - r^2)S_y^2}{(n - 2)S_x^2}} \tag{14.3}$$

The interval estimate is the familiar $b \pm 1.96(SE)$. It is formally calculated as follows.

A 95% confidence interval estimate for the slope of the regression line is given by:
The slope estimate b tends to miss the true value β by an amount called the *standard error* of the slope, denoted SE of b, and calculated as:

$$b \pm 1.96\sqrt{\frac{(1-r^2)S_y^2}{(n-2)S_x^2}} \qquad (14.4)$$

If this interval excludes 0, then the likelihood of zero slopes is ruled out, and we conclude that there is a significant linear relationship between X and Y.

Returning to our Saturn car price example, recall that $b = -0.05127$. The standard error of this estimate is

$$SE_b = \sqrt{\frac{(1-(-0.641)^2)(4079)^2}{(12-2)(50989)^2}} = 0.01942$$

The 95% confidence interval is

$$-.05127 \pm 1.96(0.01942)$$

$$(-.09, -.01)$$

Since this interval excludes 0, we conclude a significant relationship between car mileage and selling price.

14.6 Key Words

- correlation coefficient
- correlation plots
-
- Pearson Correlation
- scatter plot

14.7 exercises

_____

Exercises

Exercise 14.7-1: Consider the following data:

X	Y
-2	0
2	3
5	10
-1	1
9	5

1. Calculate the regression line for predicting Y from X.

2. Draw the scatterplot with an overlaid regression line.

3. Add 5 to Y, so the new values are 5, 8, 15, 6, 20. Calculate the new regression line.

4. Multiply Y by 5, so the new values are 0, 15, 50, 5, 75. Calculate the new regression line.

Exercise 14.7-2: Several children are observed, and their ages (in years) and vocabularies (the estimated number of words that each child knows) are recorded. A child psychologist wants to create a model that relates these two variables.

1. Which variable should be explanatory, and which response?

2. A regression equation is calculated as Y = 836X + 451. What is the slope of this regression equation?

3. A regression equation is calculated as Y = 836X + 451. What is the intercept of this regression equation?

4. If a 12-year-old child knows 8,000 words, then what is the residual for this child based on the above regression equation?

5. If a child has a negative residual, then do they fall above or below the predicted number of words known for average children their age (above or below the regression line)?

6. If a 14-year-old child knows 14,000 words, then what is the residual for this child based on the above regression equation?

7. If a child has a positive residual, then do they fall above or below the predicted number of words known for average children their age (above or below the regression line)?

Exercise 14.7-3: Moviegoers are monitored for their level of anxiety while watching a new horror movie billed to be the "scariest movie of all time." The intensity of a scene in the movie and anxiety are both measured on numerical scales of 0 to 100. A producer for the movie finds that the correlation coefficient between intensity and anxiety is 0.63, the standard deviation of intensity is 30, and the standard deviation of anxiety is 35.

1. What is the slope of the regression equation that predicts anxiety based on intensity?

2. How do you properly interpret the slope calculated in part a?

3. If the intercept is 15, then what is the regression equation?

4. What is the predicted value of anxiety for a scene measured at 90 intensity units?

5. If a moviegoer experiences anxiety measured at 82 units during a scene measured at 90 intensity units, then what is the residual for that moviegoer?

6. How do you properly interpret the residual in part 5?

Exercise 14.7-4: How well does the number of beers a student drinks predict his or her blood alcohol content (BAC)? Sociology researchers, at Ohio State University, wanted to know if there is a relationship between the amount of beer consumed and BAC. The researchers assigned the number of cans of beer to each student. After each student had consumed the assigned number of beers, thirty minutes later, an officer of the law measured the students BAC. Anonymous [2016] One student drank nine beers. You see from the scatter plot that his BAC was about

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.012	.013		-.892	.387
	No.beers	.017	.003	.875	6.746	.000

a. Dependent Variable: BAC

Figure 14.2: Number of Beers vs. BAC

1. 0.014

2. 0.14

3. 1.4

4. 14

5. 140

Answer:

Exercise 14.7-5: How well does the number of beers a student drinks predict his or her blood alcohol content (BAC)? Sociology researchers, at Ohio State University, wanted to know if there is a relationship between the amount of beer consumed and BAC. The researchers assigned the number of cans of beer to each student. After each student had consumed the assigned number of beers, thirty minutes later, an officer of the law measured the students BAC. Anonymous [2016] A scatterplot of the data appears below. The scatterplot shows
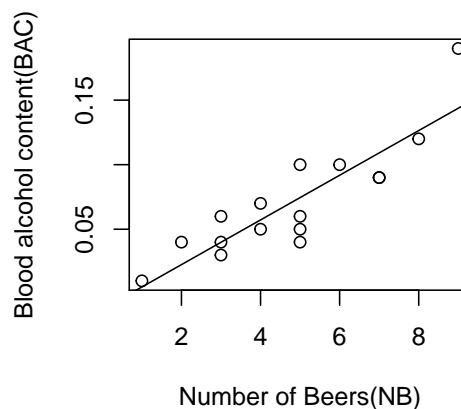
Figure 14.3: No. Beers vs. BAC

1. a weak negative relationship.

2. a moderately high negative correlation.

3. almost no connection.

4. a small positive correlation.

5. a moderately high positive straight-line relationship between some beers and BAC.

Answer:

Exercise 14.7-6: How well does the number of beers a student drinks predict his or her blood alcohol content (BAC)? Sociology researchers, at Ohio State University, wanted to know if there is a relationship between the amount of beer consumed and BAC. The researchers assigned the number of cans of beer to each student. After each student had consumed the assigned number of beers, thirty minutes later, an officer of the law measured the students BAC. Anonymous [2016] A scatterplot of the data appears below. A plausible value of the correlation between number of beers and blood alcohol content, based on the scatterplot, is

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.875ᵃ	.765	.748	.02184

a. Predictors: (Constant), No.beers

Figure 14.4: Number of Beers vs. BAC

1. $r = -0.875$.

2. $r = -0.765$.

3. r close to 0.

4. $r = 0.765$.

5. $r = 0.875$.

Answer:

Exercise 14.7-7: STEP 1: In the next six tasks, we will use the data from GSS2014 in this exercise. We have been asked to examine the relationship between a person's height and a person's income. Here we have *income* which is an interval-ratio type of variable while *height* is also an interval-ratio type of variable. These types of variables require that we use a Simple Linear Regression (SLR) analysis. So step 1 in the process of analysis is to choose the an independent and dependent variables. Note: In the GSS2014 dataset, *height* is known as HEIGHT and *income* is known as rincom06.

Scatterplot of Height vs. Income
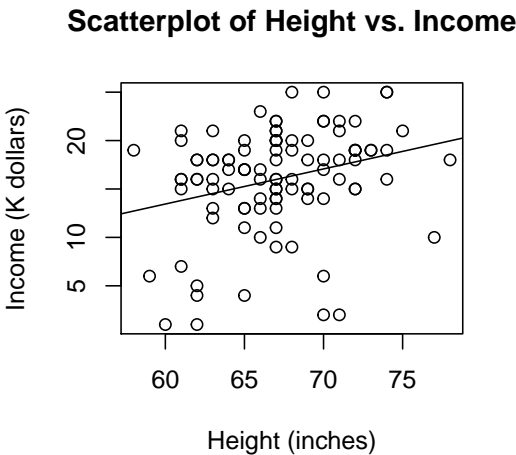
Figure 14.5: Respondent's Height vs. Income

```
Call:
lm(formula = Income ~ Height)

Residuals:
    Min    1Q  Median    3Q    Max
-15.431 -2.257   1.062  3.591  8.656

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.2948     8.4679  -0.980  0.32972
Height       0.3623     0.1261   2.874  0.00496 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.168 on 98 degrees of freedom
Multiple R-squared:  0.07775,Adjusted R-squared:  0.06834
F-statistic: 8.262 on 1 and 98 DF,  p-value: 0.004965
```

Answer:

Exercise 14.7-8: STEP 2: Using the information from exercise four, we must state the hypotheses about its relationship between *income* and *height*. Based on our experience, how should we state our hypotheses?

Answer:

Exercise 14.7-9: STEP 3: Using the information from exercise four, determine the coefficient of determination. Recall that this value is also referred to as r-squared.

Answer:

Exercise 14.7-10: STEP 4: Using the information from exercise four, record the independent, dependent variables, and the correlation coefficient.

Answer:

Dep. Var.	Independent 1. _____
1. _____	_____

Exercise 14.7-11: STEP 5: Using the information from exercise four, describe the results of the independent variable. Identify variables that we tested, their strength and direction of the relationship. We should distinguish the relationship in general terms and refer the statistical value in parentheses. Also not whether the hypotheses were supported.

Answer:

Exercise 14.7-12: STEP 1: In the next six tasks, we will use a sample of 100 subjects from the GSS2014 data in this exercise. We have been asked to examine the relationship between a person's income, age, and "not married" and a person's happiness. Here we have variables which are interval-ratio type of variable. These types of variables require that we use a Simple Linear Regression (SLR) analysis. So step 1 in the process of analysis is choosing independent and dependent variables. Note: In the GSS2014 dataset, *happy* is known as GENERAL HAPPINESS, *income06* is known as TOTAL FAMILY INCOME, *age* is known as AGE OF RESPONDENT, and *absingle* is known as NOT MARRIED.

	1	2	3	4
1	1.00	0.13	-0.14	0.05
2	0.13	1.00	-0.04	0.13
3	-0.14	-0.04	1.00	-0.16
4	0.05	0.13	-0.16	1.00

153

```
Call:
lm(formula = Happiness ~ Age + Income + NotMarried)

Residuals:
    Min     1Q Median     3Q    Max
-1.1178 -0.6868 0.1506 0.3450 1.3119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.812601  0.358840   5.051 2.09e-06 ***
Age         0.004425  0.003526   1.255   0.212
Income     -0.016172  0.011968  -1.351   0.180
NotMarried  0.008675  0.135283   0.064   0.949
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6622 on 96 degrees of freedom
Multiple R-squared:  0.03691,Adjusted R-squared:  0.006818
F-statistic: 1.227 on 3 and 96 DF,  p-value: 0.3044
```

What are the dependent and independent variables?

Answer:

Exercise 14.7-13: STEP 2: Using the information from exercise nine, we must state the hypotheses about its relationship between *Happiness*, *income*, and *age*. Based on our experience, how should we state our hypotheses?

Answer:

Exercise 14.7-14: STEP 3: Using the information from exercise 12, determine the coefficient of determination.

Answer:

Exercise 14.7-15: STEP 4: Using the information from exercise 12, record the independent, dependent variables, and the correlation coefficients.

	Independent			
Dep. Var.	1. ____	2. ____	3. ____	4. ____
1. ____	____			
2. ____	____	____		
3. ____	____	____	____	
4. ____	____	____	____	____

Exercise 14.7-16: STEP 5:

Using the information from exercise 12, record the independent, dependent variables, and the correlation coefficients.

Answer:

Exercise 14.7-17: STEP 1: In the next six tasks, we will use all of the data from GSS2014 in this exercise. We have been asked to examine the relationship between a person's income, age, and "not married" and a person's happiness. Here we have variables which are interval-ratio type of variable. These types of variables require that we use a Simple Linear Regression (SLR) analysis. So step 1 in the process of analysis is choosing independent and dependent variables. Note: In the GSS2014 dataset, *happy* is known as GENERAL HAPPINESS, *income06* is known as TOTAL FAMILY INCOME, *age* is known as AGE OF RESPONDENT, and *absingle* is known as NOT MARRIED.

```
Call:
lm(formula = Happiness ~ Age + Income + NotMarried)

Residuals:
    Min     1Q Median     3Q    Max
-1.2435 -0.6599 0.1324 0.2881 1.3956
```

	1	2	3	4
1	1.00	0.02	-0.00	-0.01
2	0.02	1.00	0.00	0.02
3	-0.00	0.00	1.00	0.06
4	-0.01	0.02	0.06	1.00

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3173217  0.0923286  25.099   <2e-16 ***
Age          0.0001357  0.0009617   0.141    0.888
Income      -0.0264172  0.0028697  -9.205   <2e-16 ***
NotMarried  -0.0290415  0.0336223  -0.864    0.388
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6378 on 1516 degrees of freedom
Multiple R-squared:  0.05344,Adjusted R-squared:  0.05156
F-statistic: 28.53 on 3 and 1516 DF,  p-value: < 2.2e-16
```

What are the dependent and independent variables?

Answer:

Exercise 14.7-18: STEP 2: Using the information from exercise 12, we must state the hypotheses about its relationship between *Happiness*, *income*, and *age*. Based on our experience, how should we state our hypotheses?

Answer:

Exercise 14.7-19: STEP 3: Using the information from exercise nine, determine the coefficient of determination.

Answer:

Exercise 14.7-20: STEP 4: Using the information from exercise 12, record the independent, dependent variables, and the correlation coefficients.

Dep. Var.	Independent			
	1. ____	2. ____	3. ____	4. ____
1. ____				
2. ____	____			
3. ____	____	____		
4. ____	____	____	____	

Exercise 14.7-21: STEP 5:

Using the information from exercise 12, record the independent, dependent variables, and the correlation coefficients.

Chapter 15

Workshops

Exercises

Exercise 15.-1:

<div style="text-align: right">

Workshop 1A, Submitted by:

</div>

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

Table 15.1: A partial list of students registered for Stat 1600 during Fall 2016 of their Major course of study.

Major	A	B	C	D	Web	Total	RF
Anthropology	2			1		3	
Art		5	3	1		9	
Aviation			1	1		2	
Biology		3		1	1	5	
Business	1					1	
Communication	5	1	4	3	4	17	
Criminal Justice		3	2	1		6	
Data Science		1				1	
Education		2			2	4	
English	2	3			1	6	
Foreign Lang.		1	1			2	
Geography		1			1	2	
Graphic Design	1		1			2	
History	1		1			2	
Journalism		1	1		1	3	
Mathematics		2				2	
Music	2		7	3	3	15	
Nursing		1				1	
Physics				1		1	
Psychology	12	9	9	8	6	44	
Social Work	6	6	6	5	3	26	
Sociology		1		1		2	
Total	32	40	36	26	22	156	

(Class spans columns A, B, C, D, Web)

Using the data from Table 15.1,

1. Construct a relative frequency table for total student majors (column above).

2. The highest percentage of students fall under what major?

3. What percentage of students are Art majors?

4. What percentage of students' majors fall in both Sociology and Psychology?

5. How does the percentage of students who are communications majors in Class A compare to the percentage of communication majors overall for total students?

Exercise 15.-2:

Workshop 1B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Table 15.2: A partial list of students registered for Stat 1600 during Fall 2016 of their Major course of study.

			Class				
Major	A	B	C	D	Web	Total	RF
Anthropology	2			1		3	
Art		5	3	1		9	
Aviation			1	1		2	
Biology		3		1	1	5	
Business	1					1	
Communication	5	1	4	3	4	17	
Criminal Justice		3	2	1		6	
Data Science		1				1	
Education		2			2	4	
English	2	3			1	6	
Foreign Lang.		1	1			2	
Geography		1			1	2	
Graphic Design	1		1			2	
History	1		1			2	
Journalism		1	1		1	3	
Mathematics		2				2	
Music	2		7	3	3	15	
Nursing		1				1	
Physics				1		1	
Psychology	12	9	9	8	6	44	
Social Work	6	6	6	5	3	26	
Sociology		1		1		2	
Total	32	40	36	26	22	156	

Using the data from Table 15.2,

1. Using the column 'RF' above, construct a relative frequency table for student majors in the Web Class only.

2. The highest percentage of students fall under what major?

3. What percentage of students are Music majors?

4. What percentage of students are found in the majors of Psychology, Social Work and Sociology (combined)?

5. How does the percentage of students who are education majors in the Web class compare to the percentage of education majors overall for total students? Which is higher?

Exercise 15.-3:

Workshop 1C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Hospital-acquired infections are a serious concern for patients in hospitals and can be caused by bacterial, viral, and fungal pathogens. Suppose hospitals in the state of Michigan reported the following number of infections and deaths from infection below for the past year.

Table 15.1: Hospital-acquired Infections

Type of infection	Number of Infections	Number of Deaths
Pneumonia	184	79
Bloodstream	258	65
Surgical site	361	12
Urinary tract	833	103
Other	287	54

Using the data from Table 15.3,

1. What type of variable is 'type of infection' numerical or categorical?

2. We can further describe this variable as _____.

3. Construct a relative frequency table for the number of infections (show percentages to two places past the decimal).

4. Complete a bar chart for the relative frequency table above complete with title and axis Labels.

5. What other type of graph can we use for this variable type?

Exercise 15.-4:

Workshop 2A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The average test grades of 19 students are as follows (on a scale from 0 to 100, with 100 being the highest score):

91 (A), 95 (A), 96 (A), 82 (B), 94 (A), 75 (C), 91 (A), 79 (C), 92 (A), 100 (A), 89 (B), 93(A), 91 (A), 86 (B), 93 (A), 72 (C), 74 (C), 93 (A), 90 (A)

Table 15.2: Distribution of Grades

grades	frequency	relative frequency
A		
B		
C		

1. Fill in the frequency and relative frequency table above

2. Create a bar chart and pie chart for the above data

3. What percentage of students got a grade of 'A'?

4. Looking at the bar chart in part a, identify the shape of the data (symmetric, right-skewed, left-skewed)?

Exercise 15.-5:

Workshop 2B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The following 50 test scores for Stat 1600 students are found below. Based on the intervals provided, complete the relative frequency table and answer the following questions.

84 88 76 44 80 83 51 93 69 78
49 55 78 93 64 84 54 92 96 72
97 37 97 67 83 93 95 67 72 67
86 76 80 58 62 69 64 82 48 54
80 69 62 67 66 73 77 82 84 86

Table 15.3: Distribution of Test Scores

Grades	frequency	relative frequency
30-39		
40-49		
50-59		
60-69		
70-79		
80-89		
90-99		

1. Fill in the frequency and relative frequency above.

2. Complete a stem and leaf plot with the stem representing the 10's place and using 3-9.

3. Draw a histogram for the test scores using the intervals in the relative frequency table above.

4. What percentage of students had scores of 59 or less?

Exercise 15.-6:

Workshop 2C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The numbers of absences for 20 Stat 1600 students are shown below (on a scale from 0 to 30).

Sample: 10, 0, 1, 3, 15, 6, 2, 1, 0, 21, 25, 11, 9, 7, 4, 5, 12, 28, 17, 8

Sorted: 0, 0, 1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 17, 21, 25, 28

Table 15.4: Distribution of Absences

Number of absences	Frequency	Relative frequency
0-9		
10-19		
20-29		

1. Fill in the frequency and relative frequency above.

2. Complete a stem and leaf plot with the stem representing the 10's place and using 0-2.

3. Draw a histogram for the absences using the intervals in the relative frequency table above.

4. What percentage of students had number of absences less than 20?

163

Exercise 15.-7:

<div align="right">

Workshop 3A, Submitted by:

</div>

Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____

The number of ice cream cones sold during a single day in the middle of August was tracked for surrounding stores. Consider the following stem and leaf and answer the questions below. The stem represents '10's' and the leaves represent '1's'. For instance, the first number represented is 14 (14 ice cream cones were sold at one store).

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

1 | 46
2 | 2358
3 | 0446779
4 | 1359
5 | 46
```

Give the five-number summary of the data. :

1. Make sure to show your work!

   a) What is the MIN?

   b) What is Q1?

   c) What is the MED?

   d) What is Q3?

   e) What is the MAX?

2. Draw a boxplot for the data.

3. Describe the shape of the distribution (symmetric, right skewed, or left skewed).

4. Ice Cream sales for a single day in January were tracked among several shops. Would we use a histogram or bar chart to represent this data? Why? Draw the correct chart (histogram or bar chart).

Cold Stone Creamery	18
Culver's	16
Ritter's	9
Treat Street	3
Y'OPA	20
Yo Go Delites	24

Exercise 15.-8:

Workshop 3B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The age at inauguration for each U.S. President is provided below:

52, 69, 64, 46, 54, 47, 70, 61, 57, 57, 58, 57, 61, 54, 68, 51, 49, 64, 57, 50, 48, 65, 52, 56, 46, 54, 56, 55, 51, 54, 51, 60, 62, 43, 55, 56, 61, 49, 51, 47, 55, 55, 54, 42, 51

https://www.loc.gov/rr/program/bib/inaugurations/

1. Give the five-number summary of the data:

    a) What is the MIN?

    b) What is Q1?

    c) What is the MED?

    d) What is Q3?

    e) What is the MAX?

2. Draw a boxplot for the data.

3. Finish the relative frequency table below.

Age	Frequency	Relative frequency
40 - 44		
45 - 49		
50 - 54		
55 - 59		
60 - 64		
65 - 69		
70 - 74		

4. Construct a histogram based on the given intervals for age at inauguration. Describe the shape of the distribution (symmetric, right skewed, or left skewed).

5. Do some extra research: Who was the youngest president at inauguration? Who was the oldest president at inauguration?

Exercise 15.-9:

Workshop 3C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The independence years of some Caribbean and African countries are shown below:

1804, 1847, 1910, 1922, 1941, 1951, 1956, 1956, 1956, 1956, 1956, 1957, 1958, 1958,1960, 1960, 1960, 1960, 1960,1960, 1960, 1960, 1960, 1961, 1961, 1962, 1962, 1966, 1966, 1973, 1974, 1975, 1977, 1978, 1979, 1979, 1980, 1981, 1981, 1983, 1990, 1993

$n = 42$

1. Give the five-number summary of the data:

   a) What is the MIN?

   b) What is Q1?

   c) What is the MED?

   d) What is Q3?

   e) What is the MAX?

2. Draw a boxplot for the data.

3. Finish the relative frequency table below.

Age	Frequency	Relative frequency
$1800 - 1839$		
$1840 - 1879$		
$1880 - 1919$		
$1920 - 1959$		
$1960 - 1999$		

4. Construct a histogram based on the given intervals for independence year. Describe the shape of the distribution (symmetric, right skewed, or left skewed).

5. Do some extra (but fun) research: Which country has the minimum (first) independence year on this list?

Exercise 15.-10:

<div align="right">

Workshop 4A, Submitted by:

</div>

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

A local bank has been monitoring daily average wait-times for its customers for four weeks. The manager believed that she could improve the wait-times by having one(1) permanent clerk and three(3) floating clerks, instead of the current three (3) permanent clerks. The floating clerks would do other tasks except when the lines became excessive. Last week the manager implemented the new system.

Old Policy: 5, 7, 10, 11, 18, 4, 4, 4, 14, 21, 0, 7, 8, 9, 18, 6, 6, 12, 22, 25
New Policy: 4, 3, 9, 13, 17

Were waiting times longer using the "Old Policy" or the "New Policy?" To answer this question, calculate the following for both the old and new policies.

1. Draw comparison dotplots for the samples above.

2. For each sample, calculate the mean AND median. Locate (place) these on your dotplots, clearly distinguishing between the two.

3. What is the mode for the old policy?

4. Calculate the 10% trimmed mean for the old policy. How does this change from the regular mean?

Exercise 15.-11:

Workshop 4B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Annual chocolate sales in billions of dollars from 6 countries are provided below.

$2.5, $1.9, $4.2, $1.2, $3.3, $19.1

For the data above calculate:

1. Mean

2. Median

3. Median of Pairwise Averages

Exercise 15.-12:

Workshop 4C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Weekly costs of lattes for 5 WMU graduate students are provided below.

$3.47, $12, $7.52, $10.5, $4.5

1. Mean

2. Median

3. Median of Pairwise Averages

Exercise 15.-13:

Workshop 5A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The average MCAT scores for Michigan medical schools are provided below.

504, 507, 507, 516, 508, 514

1. Calculate the overall mean

2. Calculate the SD

3. How does Michigan's mean and SD compare to the national average of matriculants (those enrolled in medical schools) 508.7 and SD of 6.9? Which has greater variability?

http://www.mcattestscores.com/usmedicalschoolsmcatscoresGPA.html

https://www.aamc.org/download/321494/data/factstablea16.pdf

Exercise 15.-14:

Workshop 5B, Submitted by:

Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____

Hourly wage of personnel within a small company are as follows:

Staff	Salary (dollars)
Owner	30
Manager	21
Worker	12
Worker	11
Worker	9

1. Calculate the average hourly wage of the staff.

2. Calculate the SD

3. A similar company has a mean hourly wage of $17 per hour with a SD of 9.75. Which has greater variability? How does that relate to the spread of the data?

Exercise 15.-15:

Workshop 5C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The data below are the weights of six (6) randomly selected kids in a daycare measured in pounds

13.6, 16.6, 23.5, 10.2, 5.4, 15.0

1. What is the average weight of the six kids in pounds?

2. Suppose the weights of the six kids are measured in kilograms instead. Calculate the average weight of the kids in kilograms (Hint: 1 pound = 0.453592 kilograms).

3. What is the standard deviation of the weight of the kids in pounds?

Exercise 15.-16:

Workshop 6A, Submitted by:

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

The following is an excerpt from a New York Times article by Nicholas Bakalar (Feb. 10, 2009).

> Lengthy television viewing in adolescence may raise the risk for depression in young adulthood, according to a new report.
>
> The study, in the February issue of The Archives of General Psychiatry, published by the American Medical Association, found a rising risk of depressive symptoms with increasing hours spent watching television.
>
> Researchers used data from a larger analysis of 4,142 adolescents who were not depressed at the start of the study. After seven years of follow-up, more than 7 percent had symptoms of depression.
>
> But while about 6 percent of those who watched less than three hours a day were depressed, more than 17 percent of those who watched more than nine hours a day had depressive symptoms.

1. The word 'raise' in the title implies cause-and-effect, not just association. When comparing lengthy viewers with those who watch less, what possible confounders can you think of?

2. Using one confounder from your answer(s) in (1), draw a pathway graph depicting the possible relationship between confounder, possible cause and outcome.

Exercise 15.-17:

Workshop 6B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The following is an excerpt from a Reuters article by Amy Norton (July 2011).

> Are kids safer with grandparents driving?
>
> When children are involved in a car accident, they are less likely to be injured if grandma or grandpa are driving rather than mom or dad, a new study suggests.
>
> The study, published in the journal Pediatrics, looked at five years' worth of data on U.S. car crashes involving children younger than 16.
>
> It found that when grandparents were at the wheel, 0.7 percent of children were injured in the crash. That compared with 1 percent when parents were driving.
>
> When the researchers considered other factors – such as the circumstances and severity of the crash – kids riding with grandparents were half as likely to be injured, versus those riding with their parents.

1. The title implies cause-and-effect, not just association. When comparing injury rates between parents driving and grandparents driving, what possible confounders can you think of?

2. Using one confounder from your answer(s) in (1), draw a pathway graph depicting the possible relationship between confounder, possible cause and outcome.

Exercise 15.-18:

Workshop 6C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Frequency of playing violent video games has at times been linked with and seen as a possible cause for those committing violent acts. Much debate still looms over whether playing violent video games can lead to a desensitization of violence and a lack of compassion for others.

One article that explores these ideas is titled, The Impact of Degree of Exposure to Violent Video Games, Family Background, and Other Factors on Youth Violence, by the authors Whitney Decamp and Christopher Ferguson. The authors respond to the concern that violence in video games might contribute to youth violence. There is still no consensus on whether violent video games may be contributing to violence in youth. The authors sought to study this issue in an ethnically diverse sample of youth in eighth ($n = 5133$) and eleventh grade ($n = 3886$). Questionnaire surveys were given to the youth to assess a link. Independent or predictor variables included the playing of violent video games, and an inclination towards violent video games. Dependent variables included questions pertaining to hitting with the intention to hurt another and taking part in a fight.

Mixed results were reported. The authors summarized that violent video game play was not significant in 5 of the models used. There was, however, a positive correlation between games and violent behavior for the models without controls. When the authors controlled for the propensity and context, however, this suggested the association to be more spurious (not being what it claims to be). What other factors may be impacting this outcome?

1. What possible confounders can you think of for a study of this nature?

2. Using one confounder from your answer(s) in (1), draw a pathway graph depicting the possible relationships between confounder, possible cause and outcome.

Exercise 15.-19:

Workshop 6A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Jet design researchers have come up with two new designs for the wings of a rocket plane. They think both designs are an improvement over the old design but want to know which of their new designs performs better at high altitudes. Specifically, they want to know which new design results in more maneuverability at or above 100,000 feet above sea level. They decide to design a randomized trial to study this.

1. The researchers include a control group in their design plan. Describe a possible control group. Why did they add it to the study?

2. Several pilots and planes are lined up for this study. How could the researchers randomize their experiment to avoid confounders?

3. One way to measure maneuverability is by surveying pilots about how they thought the jets handled. Could a design that asks pilots this question be blinded? If so, how?

4. Could the design from (2) be double-blinded? If so, how?

5. What makes single and double blinding important?

Exercise 15.-20:

Workshop 7B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Some students at WMU were curious about whether partying influences GPA. To find out, they surveyed 100 of their WMU peers, asking each peer only two questions: 1) How many hours a week do you party? 2) What is your GPA? The students then compiled their data and performed an appropriate statistical analysis. They concluded that partying does influence GPA, i.e., the more you party, the lower your GPA.

1. What kind of study design is this?

2. What are advantages of this study design?

3. What are disadvantages of this study design?

4. Did the students overlook anything that might have biased or confounded their results?

5. Suppose the students wanted to correct for something they overlooked. How could they improve their study design?

Exercise 15.-21:

Workshop 7C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Yogurt is known to contain 'friendly bacteria' known generally as probiotics which can benefit the gastrointestinal system. Many strains of bacteria can be found in yogurt, with some of the most common being, *Lactobacillus bulgaricus*, *Streptococcus thermophilus*, and *Lactobacillus acidophilus*. Further research has been done suggesting probiotics boost the immune system and may even help in weight management and reduction of disease such as cancer. A group of researchers decided to investigate this using a questionnaire assessing yogurt consumption (yes/no) and asking questions pertaining to the health of the individuals.

1. First of all, what kind of study design is this?

2. What are advantages of this type of study design?

3. What are disadvantages of this study design?

4. The researchers then separated the results of the surveys for those who said 'yes' they consume yogurt to those who said 'no' they do not consume yogurt. They compared the two groups using these as a treatment and control group. What is wrong with this assumption?

5. Suppose the researchers found that those consuming yogurt had better health and concluded that *Lactobacillus bulgaricus* caused a reduction of weight, improved GI and Immune systems. What are two major problems in this?

6. What may be a bias or lead to confounding in this study?

7. How could the researchers improve their study design in order to make a conclusion about the effectiveness of *Lactobacillus bulgaricus* and immune system function?

Exercise 15.-22:

Workshop 8A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Debt Poet's Society

According to a Sallie Mae survey and credit bureau data, in 2008, college students carried an average of $3,173 debt on their credit cards (USA TODAY, April 13, 2009). Suppose that current credit card debts for all college students have a normal distribution with a mean of $3,173 and a standard deviation of $800.

1. Find the probability that credit card debt for a randomly selected college student below $2,109.

2. Find the probability that credit card debt for a randomly selected college student is above $3,605.

3. Find the probability that credit card debt for a randomly selected college student is between $2,109 and $3,605.

Exercise 15.-23:

The Social Network

According to an article published on the Web site www.PCMag.com, Facebook users spend an average of 190 minutes per month checking and updating their Facebook pages (Source: http://www.pcmag.com/article2/0,2817,2342757,00.asp). Suppose that the current distribution of time spent per month checking and updating a member's Facebook page is normally distributed with a mean of 190 minutes and a standard deviation of 53.4 minutes. For a randomly selected Facebook member, determine the probability that the amount of time that he or she spends per month checking and updating his or her Facebook page is. . .

1. More than 180 minutes

2. Less than 120 minutes

3. Between 120 and 180 minutes

Exercise 15.-24:

Workshop 8C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The amount paid out by Ajax Insurance Company in a year is normally distributed with mean 3 billion dollars and standard deviation 0.5 billion dollars. The company's executives consider it a bad year if they must pay out more than 4.25 billion dollars.

1. Determine the probability of a bad year.

2. What is the probability that the company will have a good year?

3. What is the probability that the company paid out between 2 billion dollars and 3.5 billion dollars?

Exercise 15.-25:

Workshop 9A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Debt Poet's Society

According to a Sallie Mae survey and credit bureau data, in 2008, college students carried an average of $3,173 debt on their credit cards (USA TODAY, April 13, 2009). Suppose that current credit card debts for all college students have a normal distribution with a mean of $3,173 and a standard deviation of $800.

1. 10% of the credit card debt for college students is above what value?

2. 10% of the credit card debt for college students is below what value?

3. The middle 80% of the credit card debt for college students is between what two values?

4. How likely would it be for a college students to have credit card debt of $4,500 or more?

Exercise 15.-26:

Workshop 9B, Submitted by:

Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____

Tommy Wait, a minor-league baseball pitcher, is notorious for taking an excessive amount of time between pitches. In fact, his times between pitches are normally distributed with a mean of 36 seconds and a standard deviation of 2.5 seconds.

1. 10% of the time between pitches is above what value?

2. 10% of the time between pitches is below what value?

3. The middle 80% of the time between pitches is between what two values?

4. How likely would it be for the time between pitches of 40 seconds or more?

Exercise 15.-27:

Workshop 9C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The amount paid out by Ajax Insurance Company in a year is normally distributed with mean 3 billion dollars and standard deviation 0.5 billion dollars. The company's executives consider it a bad year if they must pay out more than 4.25 billion dollars.

1. 18% of the amount paid out by Ajax Insurance Company is below what amount?

2. There is a 5% chance that the company will have to pay more than an amount x (in billions of dollars). What is the value of x?

3. The middle 60% of the amount paid out by the company is between what two values (in billions of dollars)?

Exercise 15.-28:

Workshop 10A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Workshop 8: Drugstore Cowboy

According to an October 27, 2006 article in *Newsweek*, 60% of Americans said that they take expired medicines. Suppose that this result is true of the current population of Americans.

1. Let x be a binomial random variable that denotes the number of American in a random sample of 4 who have taken expired medicines. What are the possible values that x can assume?

2. Find the probability that exactly 2 Americans in a sample of 4 have taken expired medicines.

3. Find the probability that at least 3 Americans in a sample of 4 have taken expired medicines.

4. Find the probability that at most 1 American in a sample of 4 have taken expired medicines.

Exercise 15.-29:

Workshop 10B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The Social Network

According to a Harris Interactive poll, 50% of American college graduates have Facebook accounts. Suppose that this result is true for the current population of American college graduates.

`http://www.harrisinteractive.com/harris_poll/pubs/Harris_Poll_2009_04_16.pdf`

1. Let x be a binomial random variable that denotes the number of American college graduates in a random sample of 5 who have Facebook accounts. What are the possible values that x can assume?

2. Find the probability that exactly 3 American college graduates in a sample of 5 have Facebook accounts.

3. Find the probability that at least 3. American college graduates in a sample of 5 have Facebook accounts.

4. Find the probability that at most 1 American college graduate in a sample of 5 have Facebook accounts.

Exercise 15.-30:

Workshop 10C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The Bureau of Labor Statistics reports that in May 2016 that 10.5% of people who work in the Kalamazoo metropolitan area are employed in production occupations (i.e., making things). Suppose that a random sample of 5 Kalamazoo workers is taken. Let X be a random variable which denotes the number of workers in the sample who are employed in production occupations.

1. What are the possible values which X can take?

2. Find the probability that exactly 2 of the 5 workers in the sample will be employed in production occupations.

3. Find the probability that at least 3 of the 5 workers in the sample will be employed in production occupations.

4. Find the probability that at most 1 of the 5 workers in the sample will be employed in production occupations.

Exercise 15.-31:

Workshop 11A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Credit where Credit is Due

According to a May 27, 2009 Minneapolis Star-Tribune article, 78% of U.S. households have at least one credit card.

Source: http://www.startribune.com/politics/45797562.html

1. For a random sample of 500 U.S. households, what is the expected value and standard deviation of the number of U.S. households with at least one credit card.

2. Using the normal approximation, find the probability that in a random sample of 500 U.S. households, more than 375 households have at least one credit card.

3. Using the normal approximation, find the probability that in a random sample of 500 U.S. households, less than 385 households have at least one credit card.

4. Using the normal approximation, find the probability that in a random sample of 500 U.S. households, 376 to 384 households have at least one credit card.

Exercise 15.-32:

Workshop 11B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Tongue Twister

A 2007 article states that 4.8% of U.S. households are "linguistically isolated," which means that all members of the household aged 14 years and older have difficulty speaking English. Assume that this percentage is true for the current population of U.S. households.

(Source: http://www.antara.co.id/en/arc/2007/9/12/
five-percent-of-us-families-dont-speak-english-report/)

1. For a random sample of 750 U.S. households, what is the expected value and standard deviation of the number of U.S. households classified as "linguistically isolated."

2. Using the normal approximation, find the probability that in a random sample of 750 U.S. households, more than 45 would be classified as "linguistically isolated."

3. Using the normal approximation, find the probability that in a random sample of 750 U.S. households, less than 45 would be classified as "linguistically isolated."

4. Using the normal approximation, find the probability that in a random sample of 750 U.S. households, exactly 45 would be classified as "linguistically isolated."

Exercise 15.-33:

Workshop 11C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The Bureau of Labor Statistics reports that in May 2016 that 10.5% of people who work in the Kalamazoo metropolitan area are employed in production occupations (i.e., making things).

1. For a random sample of 200 Kalamazoo workers, what is the expected number who will work in production occupations? What is the standard deviation of this?

2. Using the normal approximation, find the probability that in a random sample of 200 Kalamazoo workers, more than 25 will work in production occupations.

3. Using the normal approximation, find the probability that in a random sample of 200 Kalamazoo workers, fewer than 15 will work in production occupations.

4. Using the normal approximation, find the probability that in a random sample of 200 Kalamazoo workers, between 15 and 25 (inclusive) will work in production occupations.

Chapter 16

Workshops

Exercises

Exercise 16.-1:

Workshop 12A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

According to a survey conducted by Pew Research Center in June 2009, 110 people ages 18 to 29 years said that religion is very important to them. Suppose this result is based on a sample of 250 people ages 18 to 29 years.

1. What is the sample estimate of the proportion of all people ages 18 to 29 years who will say that religion is very important to them?

2. What is the standard error of this estimate?

3. Using the normal curve and assuming your estimate from a. is representative of the population proportion, what is the probability that more than 45% of people ages 18 to 29 years say that religion is very important to them?

Exercise 16.-2:

Workshop 12B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Don't Drink and Drive

According to a Harris Interactive survey of 600 adults conducted in April 2009, 150 adults do not drink alcohol.

1. What is the sample estimate of the proportion of all adults who do not drink alcohol?

2. What is the standard error of this estimate?

3. Using the normal curve and assuming your estimate in a. is representative of the population proportion, what is the probability that more than 24% of adults do not drink alcohol?

Exercise 16.-3:

Workshop 12C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

J. S. Carlton et al. conducted a survey of 698 biological scientists from Big 10 universities, asking respondents "When compared with pre-1800's levels, do you think that mean global temperatures have generally risen, fallen, or remained relatively constant?" 652 respondents said global temperatures had risen.

1. What is the point estimate for the proportion of Big 10 biological scientists who believe in global warming?

2. What is the standard error of this estimate?

3. Using the normal curve and assuming your estimate in a. is representative of the population proportion, what is the probability that less than 93% of Big 10 biological scientists believe in global warming?

Exercise 16.-4:

Workshop 13A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

According to a survey conducted by Pew Research Center in June 2009, 110 of people ages 18 to 29 years said that religion is very important to them. Suppose this result is based on a sample of 250 people aged 18 to 29 years.

1. What is the *standard error* of the sample proportion?

2. What is the *margin of error* of your sample proportion at a 95% confidence level?

3. What is the *confidence interval* for the sample proportion at a 95% confidence level?

4. If a larger random sample of 1000 such people is taken - and the sample proportion stays the same - what is the new standard error of the sample proportion?

5. Compare the two standard errors. Which is bigger, by how much?

197

Exercise 16.-4:

Workshop 13A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

According to a survey conducted by Pew Research Center in June 2009, 110 of people ages 18 to 29 years said that religion is very important to them. Suppose this result is based on a sample of 250 people aged 18 to 29 years.

1. What is the *standard error* of the sample proportion?

2. What is the *margin of error* of your sample proportion at a 95% confidence level?

3. What is the *confidence interval* for the sample proportion at a 95% confidence level?

4. If a larger random sample of 1000 such people is taken - and the sample proportion stays the same - what is the new standard error of the sample proportion?

5. Compare the two standard errors. Which is bigger, by how much?

Exercise 16.-5:

Workshop 13B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Don't Drink and Drive

According to a Harris Interactive survey of 600 adults conducted in April 2009, 150 adults do not drink alcohol.

1. What is the *standard error* of the sample proportion?

2. What is the *margin of error* of your sample proportion at a 95

3. What is the *confidence interval* for the sample proportion at a 95

4. If a larger random sample of 2400 adults is taken - and the sample proportion stays the same - what is the new standard error of the sample proportion?

5. Compare the two standard errors. Which is bigger, by how much?

Exercise 16.-6:

Workshop 13C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

J. S. Carlton et al. conducted a survey of 698 biological scientists from Big 10 universities, asking respondents "When compared with pre-1800's levels, do you think that mean global temperatures have generally risen, fallen, or remained relatively constant?" 652 respondents said global temperatures had risen.

1. What is the *standard error* of the sample proportion?

2. What is the *margin of error* of your sample proportion at a 95

3. What is the *confidence interval* for the sample proportion at a 95

4. If a larger sample of 1500 is taken - and the sample proportion stays the same - what would be the new standard error?

5. Compare the two standard errors. Which is bigger, by how much?

Exercise 16.-7:

Workshop 14A, Submitted by:

Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____

Live and Let Dye

In an experimental study, 44 rats were exposed to a high dose of red dye, 44 rats were exposed to a low dose of red dye, and it was then observed how many in each group developed cancer.

	Cancer		
Dose Level of Red Dye	Yes	No	Total
High	14	30	44
Low	4	40	44
Total	18	70	88

1. Estimate the difference in percentage of developing cancer between high dose level of red dye and low dose level of red dye.

2. Calculate a standard error for your estimate in (1).

3. Calculate a 95% confidence interval for the difference in percentage of developing cancer between high dose level of red dye and low dose level of red dye.

Exercise 16.-8:

Gone Fishin'

Consider the following data from the Chicago Western Electric Study (Shekelle et al., 1981). In the following table 466 men were classified based on death from coronary heart disease (CHD) and fish consumption.

	Death from CHD		
Daily Fish Consumption	Yes	No	Total
High	34	227	261
Low	42	163	205
Total	76	390	466

1. Estimate the difference in percentage of death from CHD between high daily fish consumption and low daily fish consumption.

2. Calculate a standard error for your estimate in (1).

3. Calculate a 95% confidence interval for the difference in percentage of death from CHD between high daily fish consumption and low daily fish consumption.

Exercise 16.-9:

<div align="right">

Workshop 14C, Submitted by:

</div>

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

Making the Cut

The following table shows how many prospective students of a sample of 200 students were admitted to WMU last year, and if the student's GPA was below 2.8 or not. Out of the sample, 60 applicants were rejected and 140 were accepted.

	Rejected from WMU		
GPA Below 2.8	Yes	No	Total
Yes	45	20	65
No	15	120	135
Total	60	140	200

1. Estimate the difference in proportion of being rejected from WMU between having below a 2.8 GPA and not having below a 2.8 GPA.

2. Calculate a standard error for your estimate in (a).

3. Calculate a 95% confidence interval for the difference in proportion of being rejected from WMU between having below a 2.8 GPA and not having below a 2.8 GPA.

Exercise 16.-10:

Workshop 15A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Live and Let Dye

In an experimental study, 44 rats were exposed to a high dose of red dye, 44 rats were exposed to a low dose of red dye, and it was then observed how many in each group developed cancer.

	Cancer		
Dose Level of Red Dye	Yes	No	Total
High	14	30	44
Low	4	40	44
Total	18	70	88

1. Estimate the risk ratio of developing cancer between high dose level of red dye and low dose level of red dye.

2. Calculate a 95% confidence interval for the risk ratio of developing cancer between high dose level of red dye and low dose level of red dye.

Exercise 16.-11:

Workshop 15B, Submitted by:

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

Name: _____   Signature: _____

Gone Fishin'

Consider the following data from the Chicago Western Electric Study (Shekelle et al., 1981). In the following table 466 men were classified based on death from coronary heart disease (CHD) and fish consumption.

	Death from CHD		
Daily Fish Consumption	Yes	No	Total
High	34	227	261
Low	42	163	205
Total	76	390	466

1. Estimate the risk ratio of death from CHD between high daily fish consumption and low daily fish consumption.

2. Calculate a 95% confidence interval for the risk ratio of death from CHD between high daily fish consumption and low daily fish consumption.

Exercise 16.-12:

Workshop 15C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Making the Cut

The following table shows how many prospective students of a sample of 200 students were admitted to WMU last year, and if the student's GPA was below 2.8 or not. Out of the sample, 60 applicants were rejected and 140 were accepted.

GPA Below 2.8	Rejected from WMU		
	Yes	No	Total
Yes	45	20	65
No	15	120	135
Total	60	140	200

1. Estimate the risk ratio of being rejected from WMU between having below a 2.8 GPA and not having below a 2.8 GPA.

2. Calculate a 95% confidence interval for the risk ratio of being rejected from WMU between having below a 2.8 GPA and not having below a 2.8 GPA.

Exercise 16.-13:

Workshop 16A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Live and Let Dye

In an experimental study, 44 rats were exposed to high dose of red dye and 44 rats were exposed to low dose of red dye and then it was observed how many in each group developed cancer.

	Cancer		
Dose Level of Red Dye	Yes	No	Total
High	14	30	44
Low	4	40	44
Total	18	70	88

1. Estimate the odds ratio of developing cancer between high dose level of red dye and low dose level of red dye.

2. Calculate a 95% confidence interval for the odds ratio of developing cancer between high dose level of red dye and low dose level of red dye.

Exercise 16.-14:

Workshop 16B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Gone Fishin'

Consider the following data from the Chicago Western Electric Study (Shekelle et al., 1981). In the following table 466 men were classified based on death from coronary heart disease (CHD) and fish consumption.

| Daily Fish Consumption | Death from CHD | | |
	Yes	No	Total
Low	42	163	205
High	34	227	261
Total	76	390	466

1. Estimate the odds ratio of death from CHD between low daily fish consumption and high daily fish consumption.

2. Calculate a 95% confidence interval for the odds ratio of death from CHD between low daily fish consumption and high daily fish consumption.

Exercise 16.-15:

<div style="text-align: right">

Workshop 16C, Submitted by:

</div>

Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____

Making the Cut

The following table shows how many prospective students of a sample of 200 students were admitted to WMU last year, and if the student's GPA was below 2.8 or not. Out of the sample, 60 applicants were rejected and 140 were accepted.

	Rejected from WMU		
GPA Below 2.8	Yes	No	Total
Yes	45	20	65
No	15	120	135
Total	60	140	200

1. Estimate the odds ratio of being rejected from WMU between having below a 2.8 GPA and not having below a 2.8 GPA.

2. Calculate a 95% confidence interval for the odds ratio of being rejected from WMU between having below a 2.8 GPA and not having below a 2.8 GPA.

Exercise 16.-16:

Workshop 17A, Submitted by:

Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____

Historically throughout the United States, families of four spend about $239 a week on groceries and food with a standard deviation of 50. Assume a random sample of 250 households is taken in the United States.

1. What is the expected value of the sample mean.

2. What is the Standard error of this estimate?

3. What is the probability that the mean weekly household spending is more than $242?

4. What is the probability that the mean weekly household spending less than $235?

5. What is the probability that the mean weekly household spending is between $235 and $242?

6. If an even larger sample of 2500 households was taken and the SD remains the same, what is the new Standard Error of the estimate? Why did this occur?

Exercise 16.-17:

Workshop 17B, Submitted by:

Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____

Too Fast, Not Enough Furious
The above chart shows the average cost of annual auto insurance premiums in 2008 for drivers with no accidents, one accident, two accidents, and so on. According to this chart, the average annual auto insurance premium for drivers with no accidents was $1,387 in 2008.
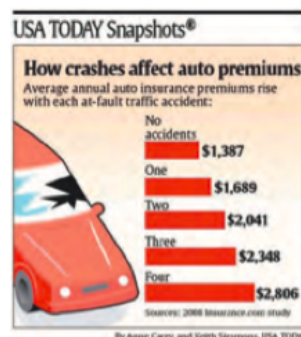
1. What is the estimated mean annual auto insurance premium for drivers with no accidents?

2. What is the standard error of the estimate assuming this is based on a random sample of 400 drivers? Assume that the standard deviation of auto insurance premiums for all current drivers with no accidents is $180.

3. Assuming the estimate in part a. is the mean for all drivers with no accidents, what is the probability that the mean annual auto insurance premium for drivers with no accidents is greater than $1,395?

4. If a larger random sample of 1600 such drivers was taken and the SD remains the same, what is the new standard error of the estimate?

5. Assuming the estimate in part a. is the mean for all drivers with no accidents, what is the probability that the mean annual auto insurance premium for drivers with no accidents is greater than $1,395 using the new standard error?

Exercise 16.-18:

Workshop 17C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Historically throughout the United States college students watch about 35 hours a month of television with a standard deviation of 5 hours. Assume a random sample of 25 college students is taken in West Michigan.

1. What is the expected value of the sample mean.

2. What is the Standard error of this estimate?

3. What is the probability that the mean monthly hours spent watching tv for the college students is more than 36.5?

4. What is the probability that the mean monthly hours spent watching tv is less than 34?

5. What is the probability that the mean monthly hours watching tv is between 34.5 and 36?

6. If an even larger sample of 250 students was taken and the SD remains the same, what is the new Standard Error of the estimate? Why did this occur?

Exercise 16.-19:

Workshop 18A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Historically throughout the United States families of four spend on the average of $239 a week on groceries and food with a standard deviation of $50. Families of 3, however, spend on average $220 with a standard deviation of $45 Assume these statistics are based on a random sample of 250 households with four members, and on a random sample of 285 households with three members.
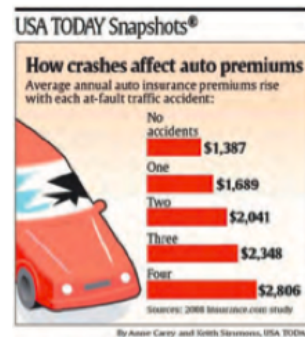
1. Estimate the difference in the average spending of a family of four and a family of three. (Use the family of four first in your equation).

2. What is the Standard error of this estimate?

3. What is the Margin of Error for your estimate?

4. What is the 95% Confidence Interval for your estimate? Is it significant?

Exercise 16.-20:

Workshop 18B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The adjacent chart shows the average cost of annual auto insurance premiums in 2008 for drivers with no accidents, one accident, two accidents, and so on. According to this chart, the 2008 average annual auto insurance premium for drivers with no accidents was $1,387, and the premium for drivers with one accident is $1,689. Assume that this dataset is partly based on a random sample of 60 drivers with no accidents, and a random sample of 56 drivers with one accident. Assume also that the standard deviations of the samples with no accidents and one accident are $150 and $178, respectively. Finally, assume the samples are independent.

USA TODAY Snapshots®

How crashes affect auto premiums
Average annual auto insurance premiums rise with each at-fault traffic accident:

No accidents $1,387
One $1,689
Two $2,041
Three $2,348
Four $2,806

Sources: 2008 insurance.com study

By Anne Carey and Keith Simmons, USA TODAY

1. What is the difference in means between average premiums for drivers with no accidents and drivers with one accident?

2. What is the standard error of the above estimate?

3. What is the Z critical value for a 95% confidence interval?

4. The upper bound of this confidence interval is how far from the central point estimate?

5. What are the lower and upper bounds of this confidence interval?

6. What can be concluded about the significance of the difference?

Exercise 16.-21:

<div align="center">

Workshop 18C, Submitted by:

</div>

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

The summary statistics for Verbal SAT are provided below:

Graduate	n	Mean	SD
No	13	466.9	89.94
Yes	28	480	67.33

1. Estimate the difference in average Verbal SAT scores between the two groups.

2. Calculate a standard error for your estimate in (1).

3. Calculate a 95% confidence interval for the difference between the two means. Does your interval exclude 0? Is your difference statistically significant?

Exercise 16.-22:

Workshop 19A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

A special vitamin concoction which claims to increase muscle strength is tested on 5 weight lifters. Athletes are selected and the weight they can bench press is recorded before and then after two weeks of taking the vitamin regimen.

Athlete	Before	After
1	220	206
2	181	179
3	210	218
4	231	235
5	201	203

1. Estimate the mean difference in weight lifted before and after the vitamin regimen.

2. Calculate a standard error for your estimate in (1).

3. Calculate a 95% confidence interval for the difference. Does your interval exclude 0? Is your difference statistically significant?

Exercise 16.-23:

Workshop 19B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

In 1968, the Labor Force Participation rate of women was measured in 5 different cities. The rates in those same cities were measured again in 1972.

City	1972	1968
N.Y.	0.45	0.42
L.A.	0.50	0.50
Chicago	0.52	0.52
Philadelphia	0.45	0.45
Detroit	0.46	0.43

1. Estimate the mean difference in participation of women between 1968 and 1972.

2. Calculate a standard error for your estimate in (1).

3. Calculate a 95% confidence interval for the difference. Does your interval exclude 0? Is your difference statistically significant?

Exercise 16.-24:

Workshop 19C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

A certain local grocery chain known for great prices is testing the effectiveness of a new product placement locations in volume of sales for their flagship household cleaning product UltraClean Spray. The grocery chain uses 5 different stores and records a week worth of sales with the old lay out and a week with the new layout. The results are recorded below.

Store	Old Layout # sold	New Layout # sold
1	10	14
2	23	30
3	9	5
4	13	24
5	18	19

1. Estimate the mean difference in items sold before the new layout and after the new layout.

2. Calculate the Standard Error for your estimate in (1), use a standard deviation of 5.718.

3. Calculate a 95% confidence interval for the difference. Is your difference statistically significant?

4. Interpret your results, did the new layout significantly decrease, significantly increase sales or neither?

Exercise 16.-25:

Workshop 20A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

In an experimental study, 44 rats were exposed to high dose of red dye and 44 rats were exposed to low dose of red dye and then it was observed how many in each group developed cancer.

Cancer	Dose Level of Red Dye High	Low	Total
Yes	14	4	18
No	30	40	70
Total	44	44	88

1. Fill in the table of expected counts:

Cancer	Dose Level of Red Dye High	Low	Total
Yes			18
No			70
Total	44	44	88

2. Calculate the chi-square test statistic.

3. What is the correct degrees of freedom?

4. Based on the degrees of freedom, what is the critical value?

5. What can we conclude from this test? Is the dose level of red dye related to cancer occurrence?

Exercise 16.-26:

Workshop 20B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Consider the following data from the Chicago Western Electric Study (Shekelle et al., 1981). In the following table 466 men were classified based on death from coronary heart disease (CHD) and fish consumption.

Death from CHD	35 gm of Daily Fish Consumption		
	No	Yes	Total
Yes	42	34	76
No	163	227	390
Total	205	261	466

1. Fill in the table of expected counts:

Death from CHD	35 gm of Daily Fish Consumption		
	No	Yes	Total
Yes			76
No			390
Total	205	261	466

2. Calculate the chi-square test statistic.

3. What are the degrees of freedom for the critical value?

4. What is the critical value?

5. What is the correct conclusion? Are daily fish consumption and death from CHD related?

Exercise 16.-27:

Workshop 20C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Officer Smiley collects data on her department's ticketing patterns. She includes variables like gender of ticketing officers (Male or Female) and the type of offense for which the ticket is written (civil infraction, misdemeanor, felony, etc.). She wonders if there is any association between the gender of a ticketing officer and the type of offense. She generates the following data summary to conduct a test at the 5% significance level:

Actual	Civil Infraction	Felony	Misdemeanor	Total
Female	6	9	11	26
Male	12	8	8	28
Total	18	17	19	54

Expected	Civil Infraction	Felony	Misdemeanor	Total
Female	8.666	8.185	9.148	26
Male	9.333	8.815	9.852	28
Total	18	17	19	54

1. What is the value of the chi-square test statistic?

2. What are the degrees of freedom for the critical value?

3. What is the critical value for her test?

4. What is the correct conclusion of this test? Are the gender of a ticketing officer and the type of offense on a ticket related?

Exercise 16.-28:

Workshop 21A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

An intern at a courthouse is curious about whether there is any association between two numerical variables: the age of defendants in years, and the sentence that they receive in years. The courthouse collects data on both variables for five cases:

	Age	Sentence
1	18	2
2	21	4
3	35	8
4	26	8
5	40	9

1. Calculate the means of Age and Sentence.

2. Given that the standard deviations of Age and Sentence are respectively 9.301 and 3.033, finish the following table of the Z-scores for Age and Sentence:

	Z_{Age}	$Z_{Sentence}$
1	-1.075	-1.385
2		
3		
4		
5		

3. Finish the following table of products of Z scores for Age and Sentence:

	$Z_{Age} \times Z_{Sentence}$
1	1.489
2	
3	
4	
5	

4. Calculate Pearson's Correlation Coefficient r.

Exercise 16.-29:

Workshop 21B, Submitted by:

Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____
Name: _____	Signature: _____

Session musicians in a studio are worried about their tuning. To get just the right mix, they decide to measure the frequency of the notes they play while recording. They begin by measuring frequency in Hertz at the note "middle C" (0 in the table), and work their way up by regular intervals (whole steps) from there. From "middle C" through the first four intervals, they produce the following data:

Intervals	Frequency (Hz)
0	261.626
1	293.665
2	329.628
3	369.994
4	415.305

1. Calculate the means for Intervals and Frequency.

2. Given that the standard deviations for Intervals and Frequency are respectively 1.581 and 60.807, finish the following table of products of Z-scores:

	$Z_{Interval} \times Z_{frequency}$
0	1.507
1	0.420
2	
3	
4	

3. Calculate Pearson's Correlation Coefficient r.

4. Interpret the correlation coefficient r within the context of this dataset. What is the relationship between intervals and frequencies?

5. If the musicians' frequency measurements are all off by 50 hertz, such that every frequency measurement is a uniform 50 hertz greater than its actual value, what would happen to the correlation coefficient r?

Exercise 16.-30:

Workshop 21C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Here we present numerical data on the first five households from the Rent/Mortgage dataset in chapter 2 of this coursepack:

	No. of Bedrooms	Monthly Payment	12-month HH Income
1	2	880	11200
2	3	990	80800
3	4	750	87600
4	3	1400	94000
5	4	1400	97000
$\bar{x}$	3.2	1084	74120
s	0.84	300.72	35,721.03

Using the above information. . .

1. Calculate the correlation coefficient r between No. of Bedrooms and Monthly Payment.

2. Calculate the correlation coefficient r between Monthly Payment and 12-Month HH Income.

3. Which set of two variables exhibits the *strongest* relationship: Monthly Payment & No. of Bedrooms, or Monthly Payment & 12-month HH Income? Why?

Exercise 16.-31:

Workshop 22A, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Data on the numbers of absences and final grades are shown below for 10 randomly selected students in a statistics class.

Student	Number of absences	Final Grade (%)
1	7	82
2	2	87
3	13	43
4	9	62
5	4	77
6	3	93
7	1	95
8	15	33
9	8	55
10	6	60
$\bar{x}$	6.8	68.7
s	4.613988	21.35962

1. Calculate the regression line for predicting final grade based on number of absences.

2. Draw the scatterplot with an overlaid regression line.

3. Using your regression line for predicting final grade from number of absences, what is the estimated final grade of a student who has been absent 5 times?

4. Calculate a 95% confidence interval for the slope of the line. Is the slope statistically significant?

Exercise 16.-32:

Workshop 22B, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Data obtained are found below for the age of 8 randomly selected patients and their systolic blood pressure.

Patient	Age (x)	Blood Pressure (y)
A	22	120
B	31	121
C	43	127
D	48	125
E	57	130
F	62	144
G	67	152
H	73	149
$\bar{x}$	50.375	133.5
s	17.75981	12.8619

1. Calculate the regression line for predicting systolic blood pressure based on age.

2. Draw the scatterplot with an overlaid regression line.

3. Using your regression line for predicting blood pressure from age, what is the estimated blood pressure of a person of 53 years of age?

4. Calculate a 95% confidence interval for the slope of the line. Is the slope statistically significant?

Exercise 16.-33:

Workshop 22C, Submitted by:

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Name: _____ Signature: _____

Session musicians in a studio are worried about their tuning. To get just the right mix, they decide to measure the frequency of the notes they play while recording. They begin by measuring frequency in Hertz at the note "middle C" (0 in the table), and work their way up by regular intervals (whole steps) from there. From "middle C" through the first four whole step intervals, they produce the following data:

Intervals	Frequency (Hz)
0	261.626
1	293.665
2	329.628
3	369.994
4	415.305

1. Write the full regression equation, using Intervals as explanatory and Frequency as response. The standard deviations of Intervals and Frequency are, respectively, 1.581 and 60.807. The correlation coefficient is 0.998

2. Interpret the slope of this regression line in context.

3. The session musicians take another measurement and find that the fifth whole step interval up from middle C, the A sharp, has an actual frequency of 466.164 Hz. Using your line from part a, calculate the residual for the predicted fifth interval from middle C.

4. Calculate a 95% confidence interval for the slope of the line. Is the slope statistically significant?

Bibliography

Anonymous. Regression exercise. web-page, 04 2016.

Elizabeth Fuller, editor. *Drug use, smoking and drinking among young people in England in 2007*. The Health and Social Care Information Centre, 35 Northampton Square, London, EC1V 0AX, 2008.

Joshua Naranjo. Knowledge building/research 1 knowledge and data.

William W. Thompson. Early thimerosal exposure and neuropsychological outcomes at 7 to 10 years. *New England Journal of Medicine*, 357:1281–1292, September 2007.

Board Trustees. General social survey (gss). www.norc.org/Research/Projects/Pages/general-social-survey.aspx, 02 2016. URL www.norc.org/Research/Projects/Pages/.

Walter Wallace. *The Logic of Science in Sociology*. Aldine-Atherton, Chicago, 2nd edition, 1971.

Index