A Handbook of Statistics

An Overview of Statistical Methods

Darius Singpurwalla

Darius Singpurwalla

A Handbook of Statistics

An Overview of Statistical Methods

A Handbook of Statistics: An Overview of Statistical Methods
1st edition

ISBN 978-87-403-0542-5

Contents

Preface

This book was written for individuals interested in learning about the practice of statistics without needing to understand the theoretical foundations of the field. The book's serves as a handbook of sorts on techniques often used in fields such as business, demography, and health.

The book is divided into five main sections:

1. Introduction
2. Descriptive statistics
3. Probability
4. Statistical inference
5. Regression and correlation

The first section consists of one chapter that covers the motivation for studying statistics and some basic definitions used throughout the course. Sections two through four are further divided into chapters covering a specific concept or technique in support of the main section. For instance, the descriptive statistics section is broken down into two sections, one focusing on graphical techniques used to summarize data and the other focusing on numerical summaries of data sets. The techniques introduced in these sections will be illustrated using real world examples. Readers can practice the techniques demonstrated in the chapters with practice problems that are included within the chapters. The primary software package used to carry out the techniques introduced in this book is Microsoft Excel.

1 Statistics and Statistical Thinking

Statistics is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical information. Statistics is used in several different disciplines (both scientific and non-scientific) to make decisions and draw conclusions based on data.

For instance:

- In the pharmaceutical industry, it is impossible to test a potential drug on every single person that may require it. But, what drug companies can do, is consult a statistician who can help them draw a sample of individuals and administer the drug to them. The statistician can then analyze the effects of the drug on the sample and generalize their findings to the population. If the drug fails to alleviate any negative symptoms from their sample, then perhaps the drug isn't ready to be rolled out to the population.
- In business, managers must often decide whom to offer their company's products to. For instance, a credit card company must assess how risky a potential customer is. In the world of credit, risk is often measured by the measuring the chance that a person will be negligent in paying their credit card bill. This clearly is a tricky task since we have limited information about this individual's propensity to not pay their bills. To measure risk, managers often recruit statisticians to build statistical models that predict the chances a person will default on paying their bill. The manager can then apply the model to potential customers to determine their risk and that information can be used to decide whether or not to offer them a financial product.

As a more concrete example, suppose that an individual, Dave, needs to lose weight for his upcoming high school reunion. Dave decided that the best way for him to do this was through dieting or adopting an exercise routine. A health counselor that Dave has hired to assist him, gave him four options:

1. The Atkins Diet
2. The South Beach Diet
3. A diet where you severely reduce your caloric intake
4. Boot Camp, which is an extremely vigorous, daily, exercise regimen.

Dave, who understands that using data in making decisions can provide additional insight, decides to analyze some historical trends that his counselor gave him on individuals (similar in build to Dave) that have used the different diets. The counselor provided the weights for different individuals who went on these diets over an eight week period.

	Weight							
Diet	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Atkins	310	310	304	300	290	285	280	284
South Beach	310	312	308	304	300	295	290	289
Reduced Calorie	310	307	306	303	301	299	297	295
Boot Camp	310	308	305	303	297	294	290	287

Average Weight Loss on Various Diets across Eight Weeks (initial weight listed in week 1)

Based on these numbers, which diet should Dave adopt?

What if Dave's goal is to lose the most weight? The Atkins[1] diet would seem like a reasonable choice as he would have lost the most weight at the end of the eight week period. However, the high protein nature of the diet might not appeal to Dave. Also, Atkins seems to have some ebb and flow in the weight loss (see week 7 to week 8). What if Dave wanted to lose the weight in a steadier fashion? Then perhaps boot camp might be the diet of choice since the data shows an even, steady decline in weight loss. This example demonstrates that Dave can make an educated decision that is congruent with his own personal weight loss goals.

There are two types of statistics that are often referred to when making a statistical decision or working on a statistical problem.

Definitions

Descriptive Statistics: Descriptive statistics utilize numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present the information in a convenient form that individuals can use to make decisions. The main goal of descriptive statistics is to describe a data set. Thus, the class of descriptive statistics includes both numerical measures (e.g. the mean or the median) and graphical displays of data (e.g. pie charts or bar graphs).

Inferential Statistics: Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data. Some examples of inferential statistics might be a z statistics or a t-statistics, both of which we will encounter in later sections of this book.

1.1 Descriptive Statistics

Dave can use some basic descriptive statistics to better understand his diet and where there might be opportunity to change some of his dietary habits. The table below displays the caloric contents of the foods that Dave has eaten this past week.

	Sunday	Monday	Tuesday	Wednesday
Breakfast	4 Coffees	Protein Plate	Skip	Fruit
Calories	200	175	0	200
Lunch	Pizza and Wings	Salad	Super Burrito	Salad
Calories	1200	375	750	375
Dinner	Leftovers	Frozen Dinner	Frozen Dinner	Frozen Dinner
Calories	1000	275	300	300
Snack	3 Beers	3 Beers	3 Beers	3 Beers
Calories	300	300	300	300

Caloric Intake Log for Sunday–Wednesday

What information can be gleaned from the above caloric intake log? The simplest thing to do is to look at the total calories consumed for each day. These totals are shown below.

	Sunday	***Monday***	***Tuesday***	***Wednesday***
Breakfast	4 Coffees	Protein Plate	Skip	Fruit
Calories	200	175	0	200
Lunch	Pizza and Wings	Salad	Super Burrito	Salad
Calories	1200	375	750	375
Dinner	Leftovers	Frozen Dinner	Frozen Dinner	Frozen Dinner
Calories	1000	275	300	300
Snack	3 Beers	3 Beers	3 Beers	3 Beers
Calories	300	300	300	300
Daily Total	**2700**	**1125**	**1350**	**1175**

Caloric Intake Log (with totals) for Sunday–Wednesday

What new information does providing the totals add?

- The heaviest calorie day was by far on Sunday.
- Enjoying the three beers per night is consuming approximately 20% of Dave's calories on each day (except Sunday).
- Dave tends to have higher calorie lunches on the days that he skipped breakfast (or just had coffee).

The third point is an interesting take away from the above analysis. When Dave skips breakfast, his lunches are significantly higher in caloric content than on the days that he eats a breakfast. What Dave can take away from this is quite clear – he should start each day by eating a healthy breakfast as this will most likely lead to eating less total calories during the day.

1.2 Inferential Statistics

The main goal of inferential statistics is to make a conclusion about a population based off of a sample of data from that population. One of the most commonly used inferential techniques is *hypothesis testing*. The statistical details of hypothesis testing will be covered in a later chapter but hypothesis testing can be discussed here to illustrate inferential statistical decision making.

Key Definitions

Experimental Unit: An object upon which data is collected.

Population: A set of units that is of interest to study.

Variable: A characteristic or property of an individual experimental unit.

Sample: A subset of the units of a population.

What is a statistical hypothesis? It is an educated guess about the relationship between two (or more) variables. As an example, consider a question that would be important to the CEO of a running shoe company: Does a person have a better chance of finishing a marathon if they are wearing the shoe brand of the CEO's company than if they are wearing a competitor's brand? The CEO's hypothesis would be that runners wearing his company's shoes would have a better chance at completing the race since his shoes are superior. Once the hypothesis is formed, the mechanics of running the test are executed. The first step is defining the variables in your problem. When forming a statistical hypothesis, the variables of interest are either dependent or independent variables.

Definition

Dependent Variable: The variable which represents the effect that is being tested

Independent Variable: The variable that represents the inputs to the dependent variable, or the variable that can be manipulated to see if they are the cause.

In this example, the dependent variable is whether an individual was able to complete a marathon. The independent variable is which brand of shoes they were wearing, the CEO's brand or a different brand. These variables would operationalized by adopting a measure for the dependent variable (did the person complete the marathon) and adopting a measure for the kind of sneaker they were wearing (if they were wearing the CEO's brand when they ran the race).

After the variables are operationalized and the data collected, select a statistical test to evaluate their data. In this example, the CEO might compare the proportion of people who finished the marathon wearing the CEO's shoes against the proportion who finished the marathon wearing a different brand. If the CEO's hypothesis is correct (that wearing his shoes helps to complete marathons) then one would expect that the completion rate would be higher for those wearing the CEO's brand and statistical testing would support this.

Since it is not realistic to collect data on every runner in the race, a more efficient option is to take a sample of runners and collect your dependent and independent measurements on them. Then conduct your inferential test on the sample and (assuming the test has been conducted properly) generalize your conclusions to the entire population of marathon runners.

While descriptive and inferential problems have several commonalities, there are some differences worth highlighting. The key steps for each type of problem are summarized below:

Elements of a Descriptive Statistical Problem

1) Define the population (or sample) of interest
2) Select the variables that are going to be investigated
3) Select the tables, graphs, or numerical summary tools.
4) Identify patterns in the data.

Elements of an Inferential Statistical Problem

1) define the population of interest
2) Select the variables that are going to be investigated
3) Select a sample of the population units
4) Run the statistical test on the sample.
5) Generalize the result to your population and draw conclusions.

1.3 Types of Data

There are two main types of data used in statistical research: qualitative data and quantitative data. Qualitative data are measurements that cannot be measured on a natural numerical scale. They can only be classified into one or more groups of categories. For instance, brands of shoes cannot be classified on a numerical scale, we can only group them into aggregate categories such as Nike, Adidas, or K-Swiss. Another example of a qualitative variable is an individual's gender. They are either male or female and there is no ordering or measuring on a numerical scale. You are one or the other. Graphs are very useful in studying qualitative data and the next chapter will introduce graphical techniques that are useful in studying such data.

Quantitative data are measurements that can be recorded on a naturally occurring scale. Thus, things like the time it takes to run a mile or the amount in dollars that a salesman has earned this year are both examples of quantitative variables.

1.4 Chapter 1 Problems

1) Discuss the difference between descriptive and inferential statistics.
2) Give an example of a research question that would use an inferential statistical solution.
3) Identify the independent and dependent variable in the following research question: A production manager is interested in knowing if employees are effective if they work a shorter work week. To answer this question he proposes the following research question: ***Do more widgets get made if employees work 4 days a week or 5 days a week?***
4) What is the difference between a population and a sample?
5) Write about a time where you used descriptive statistics to help make a decision in your daily life.

2 Collecting Data & Survey Design

The United States has gravitated from an industrial society to an information society. What this means is that our economy relies less on the manufacturing of goods and more on information about people, their behaviors, and their needs. Where once, the Unites States focused on producing cars, we now mine individual's buying habits to understand what kind of car they are most likely to buy. Instead of mass producing cereal for people to eat, we have focus groups that help us determine the right cereal to sell in different regions of the country. Information about people's spending habits and personal tendencies is valued in today's economy and because of this reliance on information, the work that statisticians do is highly valued. But statisticians cannot do their work without data and that is what this chapter is about – the different ways that data are collected.

Definition

Survey: A method of gathering data from a group of individuals often conducted through telephone, mail or the web.

While surveys remain the primary way to gather information, there are several other methods that can be as effective as surveying people. For instance, the government provides several data sets made available from the surveys that they conduct for the U.S. The website www.data.gov compiles several government sponsored data sets for individuals to use for their own research purposes. In addition to the government, researchers often make their own data sets available for others to use. Outside of data sets, focus groups and in-depth personal interviews are also excellent sources for gathering data. Each of these different sources, have benefits and drawbacks.

As stated above, existing data sets can come from government sponsored surveys[2] or researchers who have made their data available to others. These are known as secondary sources of data. The benefits of using secondary sources of data are that they are inexpensive and the majority of the work of collecting and processing the data has already been done by someone else. The drawbacks are that the data might not be current and the data that has been collected might not directly answer the research question of interest.

Personal interviews are another method of collecting data. Personal interviews involve one person directly interviewing another person. They are primarily used when subjects are not likely to respond to other survey methods such as an online or paper and pencil survey. The advantages of personal interviews are that they are in depth and very comprehensive. That is, the interviewer has the ability to probe the respondent to get the detail required to fully answer the questions being asked. In addition, response rates to personal interviews are very high. The disadvantages of personal interviews are that they are very expensive to conduct in comparison to other survey methods. In addition, they can be time consuming and because of the time requirements to conduct a personal interview, the sample sizes yielded from this type of data gathering are small.

Another data collection method, the focus group, is a research technique that is often used to explore people's ideas and attitudes towards an item of interest. They are often used to test new approaches or products and observer in real time, potential customer's thoughts on the product. A focus group is usually conducted by a moderator, who leads a discussion with the group. The discussion is often observed by an outside party who records the reactions of the group to the product. The advantage of a focus group is that the researcher can collect data on an event that has yet to happen such as a new product or a potential new advertising campaign. The disadvantage of the focus group is that the group may not be representative of the population and the issue of small sample sizes comes into play again. Therefore generalizations of the conclusions of the focus group to the larger population are invalid. But since the goal of conducting a focus group is rarely to perform a valid statistical analysis, they are useful for getting a feel for what people's initial reactions to a product are.

The primary way to collect data is through the questionnaire-based survey, administered either through mail, phone or more frequently these days, the internet. In a questionnaire based survey, the researcher distributes the questionnaire to a group of people and then waits for the responses to come back so he/she can analyze the data. There are several advantages to conducting a survey. The first is that they are a cost effective way of gathering a great deal of information. The questionnaire can also cover a wide geographic area, thus increasing the chances of having a representative sample of the population. The main disadvantage of the mail/internet survey is that there is a high rate of non-response. Also, there is little to no opportunity to ask follow up questions.

There are several key principles to keep in mind when developing a questionnaire. The first one is to define your goals upfront for why you are doing the survey and write questions that contribute to achieving those goals. Second, response rates are often maximized by keeping the questionnaire as short as possible. Therefore, only ask questions that are pertinent to your study and avoid questions that are only "interesting to know". It also helps to formulate your analysis plan (i.e. how you plan to analyze the data) before writing the questionnaire. Doing this will help keep your goals in mind as you move through writing the questionnaire.

Just as there are keys to constructing a good questionnaire, there are several elements to writing a good question. First and foremost, a good question is clearly written and evokes a truthful, accurate response. One dimensional questions, questions that only cover one topic, are well suited to do this. Take for example the following.

What kind of computer do you own?

A) Macintosh

B) PC

At first glance this may seem like a good question. It only covers one topic and the odds are good that the responder will not be confused when reading it. The major error to the question is that it does not cover all possible responses. What if the respondent does not own a computer? There is no option to report that response in the question. Therefore, a better question might be:

What kind of computer do you own?

A) Macintosh

B) PC

C) Neither

2.1 Sampling Methods

Sampling is the use of a subset of a population to represent the whole. If one decides to use sampling in their study, there are several methods available to choose from. The two types of sampling methods are probability sampling, where each person has a known non-zero probability of being sampled, and non-probability sampling, where members are selected in a non-random methodology.

There are several kinds of probability sampling methods. Three are mentioned here: random sampling, stratified sampling and systematic sampling. In random sampling, each member of the population has an equal and known chance of being selected. Selection is done by essentially picking out of a hat. In stratified sampling, the population is subdivided by one or more characteristics and then random sampling is done within each divide. Stratified sampling is done when a researcher wants to ensure that specific groups of the population are selected in the sample. For instance, if one is selecting a sample from an undergraduate student body and wants to make sure that all classes (freshman, sophomore, junior, senior) then they might stratify by class and then randomly sample from each class. Systematic sampling, also known as 1 in K sampling is where the researcher selects every "k"th member from an ordered list.

Most non-probability sampling methods are utilized when random sampling is not feasible. Random sampling might not be feasible because the population is hard to contact or identify. For instance, one non probability sampling method is convenience sampling. Just as the technique's name sounds, the sample is selected because it was convenient to the researcher. Perhaps they run their study on the students in their class, or their co-workers. Another method is judgment sampling. This is where the researcher selects the sample based on their judgment. The researcher is assumed to be an expert in the population being studied. Snowball sampling relies on referrals from initial subjects to generate additional subjects. This is done in medical research, where a doctor might refer a patient to a researcher. The patient then refers another patient that they know and so on and so forth until the researcher has enough subjects to run their study.

2.2 Chapter 2 Exercises

Answer the following:

1) Why is it important to state the purpose of a survey before you conduct it?
2) You want to determine the extent to which employees browse the web during work hours at a certain company. What is your population?
3) What kind of problems do you for see if you conduct a survey on government regulations on the stock exchange right after a financial crisis?
4) Explain what is wrong with the following question: "Don't you agree that we have wasted too much money on searching for alien life?"
5) Bob wants to survey cellular telephone shoppers, so he goes the local mall, walks up to people at random and asks them to participate in his survey. Is this a random sample?
6) Explain why a "call in" poll to an evening news program constitutes an unscientific survey.

Rewrite the following questions:

7) Are you satisfied with your current auto insurance (yes or no)?
8) Which governmental American policy decision was most responsible for recent terrorist attacks?

3 Describing Data Sets

Populations can be very large and the data that we can collect on populations is even larger than the population itself because of the number of variables we collect on each population member. For instance, credit bureaus collect credit data on every credit eligible person in the United States. An online shopping hub such as Amazon.com, collects data on not just every purchase you make, but every product that you click on. The ability to amass such large amounts of data is a great thing for individuals who use data to run their business or study policy. They have more information with which to base their decisions on. However, it is impossible for most humans to be able to digest it in its micro form. It is more important to look at the big picture than it is to look at every element of data that you collect.

Statisticians can help out decision makers by summarizing the data. Data is summarized using either graphical or numeric techniques. The goal of this chapter is to introduce these techniques and how they are effective in summarizing large data sets. The techniques shown in these chapters are not exhaustive, but are a good introduction to the types of things that you can do to synthesize large data sets. Different techniques apply to different types of data, so the first thing we must do is define the two types of data that we encounter in our work.

Quantitative and qualitative data/variables (introduced in chapter 1) are the two main types of data that we often deal with in our work. Qualitative variables are variables that cannot be measured on a natural numerical scale. They can only be classified into one of a group of categories. For instance, your gender is a qualitative variable since you can only be male or females (i.e. one of two groups). Another example might be your degree of satisfaction at a restaurant (e.g. "Excellent", "Very Good", "Fair", "Poor). A quantitative variable is a variable that is recorded on a naturally occurring scale. Examples of these might be height, weight, or GPA. The technique that we choose to summarize the data depends on the type of data that we have.

3.1 Summarizing Qualitative Data

The first technique that we will look that summarizes qualitative data is known as a frequency table. The frequency table provides counts and percentages for qualitative data by class. A class is one of the categories into which qualitative data can be classified. The class frequency is the number of observations in the data set falling into a particular class. Finally, the class relative frequency is the class frequency divided by the total number of observations in the data set.

We can illustrate the frequency table using a simple example of a data set that consists of three qualitative variables: gender, state, and highest degree earned.

- Gender, needs no explanation, one can be either male or female.
- The variable state is the state (e.g. Virginia, Maryland) that the person lives in.
- Finally, the highest degree earned is the highest degree that a person in our data set has earned.

Let's say that this data set is 20 small business owners.

ID	Gender	State	Highest Degree Earned
1	M	VA	Law
2	M	MD	Law
3	M	IA	PhD
4	F	ID	HS Degree
5	F	VA	MBA
6	M	MD	BA
7	M	DC	MS
8	F	CA	MBA
9	F	CA	MBA
10	F	MD	MS
11	F	VA	MS
12	M	IA	MS
13	M	AK	MBA
14	M	AK	MBA
15	M	AR	MBA
16	F	AK	MBA
17	F	DC	PhD
18	F	MD	PhD
19	F	MD	Law
20	M	VA	Law

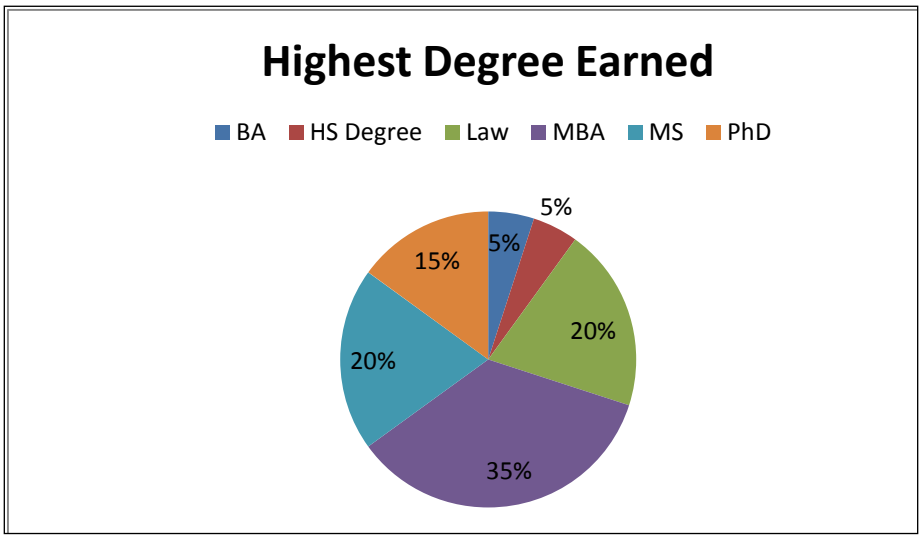
Frequency Table for Highest Degree Earned

Class	Class Frequency	Relative Frequency
BA	1	5.00%
HS Degree	1	5.00%
Law	4	20.00%
MBA	7	35.00%
MS	4	20.00%
PhD	3	15.00%
Grand Total	**20**	**100.00%**

The table below the raw data summarizes the data by the variable "Highest Degree Earned". The first column shows the different levels that the variable can take. The second and third columns show the frequency and relative frequency, respectively for each class. For this data set, the MBA degree is the most represented degree with 7 individuals holding MBA's. Finally, the third column shows the relative frequency, or the percentage of people with that degree. MBA's have 35%. The frequency table summarizes the data so that we can better understand the data set without having to look at every record. Most of our small business owners have MBA's with some other advanced degree a close second. In fact, 90% of the small business owners have some sort of advanced degree.

3.2 Graphical Techniques for Describing Data

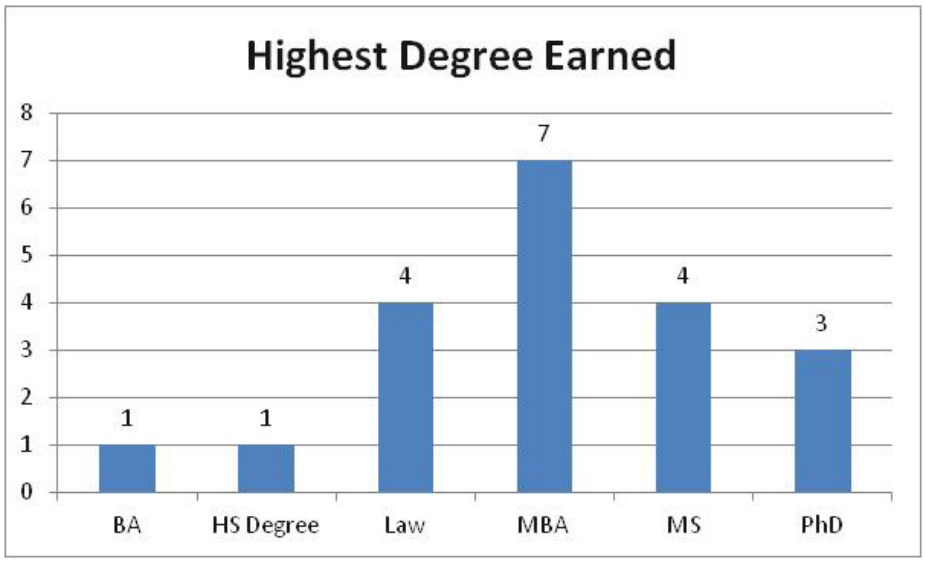
The data in the example can also be used to illustrate a pie chart. A pie chart is useful for showing the part to the whole in a data set. The pie chart is a circular chart that divided into different pieces that represents the proportion of the class level to the whole.



The pie chart provides the same information that the frequency chart provides. The benefit is for the individuals who prefer seeing this kind of data in a graphical format as opposed to a table. Clearly the MBA is most highly represented. Even without the labels, it would be easy to see that the MBA degree is the most represented in our data set as the MBA is the largest slice of the pie.

Frequency Histogram

The final type of graph that we will look at in this chapter is the frequency histogram. The histogram pictorially represents the actual frequencies of our different classes. The frequency histogram for the "Highest Degree Earned" variable is shown below.



In the frequency histogram, each bar represents the frequency for each class level. The graph is used to compare the raw frequencies of each level. Again, it is easy to see that MBA dominates, but the other higher level degrees (Law, MS, and PhD) are almost equal to each other.

3.3 Measures of Central Tendency

The central tendency of the set of measurements is the tendency of the data to cluster or center about certain numerical values. Thus, measures of central tendency are applicable to quantitative data. The three measures we will look at here are the mean, the median and the mode.

The mean, or average, is calculated by summing the measurements and then dividing by the number of measurements contained in the data set. The calculation can be automated in excel.

=average()

For example, if we have the data points 5,3,8,6 in cells A1:A4

=average(A1:A4) = 5.5

The median of a quantitative data set is the middle number when the measurements are arranged in ascending or descending order. The formula for the median in Excel is

=median()

Continuing in the above example

=median(A1:A4) = 5.5

The mode is the measurement that occurs most frequently in the data. The mode is the only measure of central tendency that has to be a value in the data set.

=mode()

Since the mean and median are both measures of central tendency, when is one preferable to the other? The answer lies with the data that you are analyzing. If the dataset is skewed, the median is less sensitive to extreme values. Take a look at the following salaries of individuals eating at a diner.

Name	Salary
Bob	$12,00
Tim	$13,000
Judy	$13,500
Jane	$16,000
Ralph	$20,000

The mean salary for the above is ~$15K and the median is ~$14K. The two estimates are quite close. However, what happens if another customer enters the diner with a salary of $33 MM.

Name	Salary
Bob	$12,000
Tim	$13,000
Judy	$13,500
Jane	$16,000
Ralph	$20,000
Jay Sugarman	$33,000,000

The median stays approximately the same (~$14K) while the mean shoots up to ~$5MM. Thus, when you have an outlier (an extreme data point) the median is not affected by it as much as mean will be. The median is a better representation of what the data actually looks like.

3.4 Measures of Variability

Variance is a key concept in statistics. The job of most statisticians is to try to understand what causes variance and how it can be controlled. Variance can be considered the raw material in which we do statistical analysis for. Some types of variance that can be considered as beneficial to understanding certain phenomena. For instance, you like to have variance in an incoming freshman class as it promotes diversity. However, variance can also negatively impact your process and it is something you want to control for and eliminate. For instance, you want to minimize variability in a manufacturing process. All products should look exactly the same.

The job of a statistician is to explain variance through quantitative analysis. Variance will exist in manufacturing processes due to any number of reasons. Statistical models and tests will help us to identify what the cause of the variance is. Variance can exist in the physical and emotional health of individuals. Statistical analysis is geared towards helping us understand exactly what causes someone to be more depressed than another person, or why some people get the flu more than others do. Drug makers will want to understand exactly where these differences lie so that they can target drugs and treatments that will reduce these differences and help people feel better.

There are several methods available to measure variance. The two most common measures are the variance and the standard deviation. The variance is simply the average squared distance of each point from the mean. So it is really just a measure of how far the data is spread out. The standard deviation is a measure of how much, on average each of the values in the distribution deviates from the center of the distribution.

Say we wish to calculate the variance and the standard deviation for the salaries of the five people in our diner (the diners without Jay Sugarman). The excel formulas for the variance and the standard deviation are:

=var.p(12000,13000,13500,16000,20000)

=stdev.p(12000,13000,13500,16000,20000)

The variance is 8,240,000 and the standard deviation is $2870.54. A couple of things worth noting:

- The variance is not given in the units of the data set. That is because the variance is just a measure of the spread of the data. The variance is unit free.
- The standard deviation is reported in the units of the dataset. Because of this, people often prefer to talk about the data in terms of the standard deviation.
- The standard deviation is simply the square root of the variance.
- The formulas for the variance and the standard deviation presented above are for the population variances and standard deviations. If your data comes from a sample use:

=var.s()

=stdev.s()

The standard deviation has several important properties:

1. When the standard deviation is 0, the data set has no spread – all observations are equal. Otherwise the standard deviation is a positive number.
2. Like the mean, the standard deviation is not resistant to outliers. Therefore, if you have an outlier in your data set, it will inflate the standard deviation.
3. And, as said before, the standard deviation has the same units of measurements as the original observations.

3.5 Combining the Mean and the Standard Deviation

The mean and the standard deviation can be used to gain quite a bit of insight into a data set. Say for instance that Dave and Jim, his co-worker, are in charge of filling up 2 liter growlers of beer. Since this is a manual process, it is difficult to get exactly two liters into a growler. To assess how far each of you is off, you decide to precisely measure the next 100 growlers that you fill up. The mean and the standard deviation are reported below.

Employee Name	Mean	Standard Deviation
Dave	2.01 liters	.05
Jim	1.95 liters	.01

The data tells us that, on average, Dave's pours are a bit closer to the mark of two liters when he fills his beer growler. However, his pours have a higher standard deviation than does Jim which indicates that his pours fluctuate more than Jim's does. Jim might pour less, but he consistently pours about the same amount each time. How could you use this information? One thought you might have is that from a planning purpose, you know that Jim is performing more consistently than Dave. You know you are going to get slightly below the 2 liter goal. You can confidently predict what his next pour will be. On the other hand, on average Dave gets closer to the mark of 2 liters, but his pours vacillate at a much higher rate than does Jim.

The Empirical Rule is a powerful rule which combines the mean and standard deviation to get information about the data from a mound shaped distribution.

Definition

The Empirical Rule: For a mound shaped distribution

- 68% of all data points fall within 1 standard deviation of the mean.
- 95% of all data points fall within 2 standard deviations of the mean.
- 99.7% of all data points fall within 3 standard deviations of the mean.

Put another way, the Empirical Rule tells us that 99.7% of all data points lie within three standard deviations of the mean. The empirical rule is important for some of the work that we will do in the chapters on inferential statistics later on. For now though, one way that the empirical rule is used is to detect outliers. For instance suppose you know that the average height of all professional basketball players is 6 feet 4 inches with a standard deviation of 3 inches. A player of interest to you stands 5 foot 3 inches. Is this person an outlier in professional basketball? The answer is yes. If we know that the distribution of the height of pro basketball players is mound shaped, then the empirical rule tells us that 99.7% of all player's heights will be within three standard deviations of the mean, or 9 inches. Our 5 foot 3 inches player is beyond 9 inches below the mean, which indicates he is an outlier in professional basketball.

The last concept introduced in this chapter is the z-score. If you have a mound shaped distribution the z-score makes use of the mean and the standard deviation of the data set in order to specify the relative location of a measurement. It represents the distance between a given data point and the mean, expressed in standard deviations. The score is also known as "standardizing" the data point. The excel formula to calculate a z-score is

=standardize(data, mean, standard deviation)

Large z-scores tell us that the measurement is larger than almost all other measurements. Similarly, a small z-score tells us that the measurement is small than all other measurements. If a score is 0, then the observation lies on the mean. And if we pair the z-score with the empirical rule from above we know that:

- 68% of the data have a z-score between -1 and 1.
- 95% of the data have a z-score between -2 and 2.
- 99.7% of the data have a z-score between -3 and 3.

Example

Suppose 200 steelworkers are selected and the annual income of each is determined. This is our sample data set. The mean is $34,000 and the standard deviation is $2,000. Joe's annual income is $32,000. What is his sample z-score?

=standardize(32000,34000,2000)

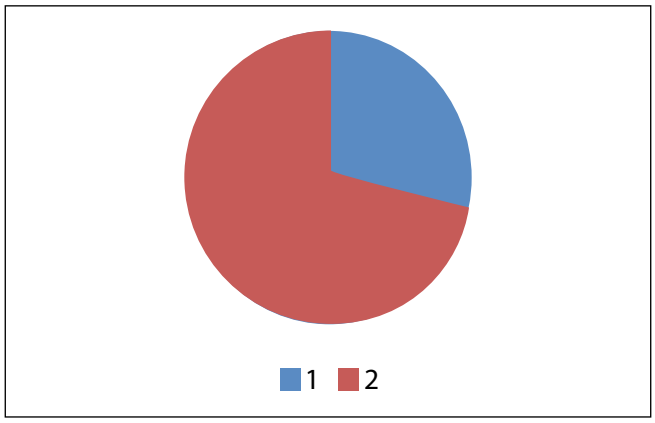
= -1.0

Joe's income falls one standard deviation below the average of all the steelworkers.

3.6 Chapter 3 Exercises

3.6.1 Summarizing Qualitative Data

1) You survey 20 students to see which new cafeteria menu they prefer, Menu A, Menu B, or Menu C. The results are A, A, B, C, C, C,A,A,B,B,B,C,A,B,A,A,C,B,A,B,B. Which menu did they prefer? Make a frequency table carefully defining the class, class frequency and relative frequency. Explain your answer.
2) A PC maker asks 1,000 people whether they've updated their software to the latest version. The survey results are 592 say yes, 198 say no and 210 do not respond. Generate the relative frequency table based off of the data. Why would you need to include the non-responders in your table, even though they don't answer your question.
3) What is the benefit of showing the relative frequency (the percentages) instead of the frequency in a table?
4) The pie chart below represent a random sample of 76 people, 22 females and the remaining are males. How could the pie chart below be improved?



5) Suppose 375 individuals are asked what type of vehicle they own: Ford, BMW, or Porsche. The results are shown below.
   a) Make a relative frequency table of the results.
   b) Create a pie chart of the results.
   c) Create a histogram of the results.

3.6.2 Measures of Central Tendency Questions

6) Does the mean have to be one of the numbers in a data set? Explain.
7) Does the median have to be one of the numbers in a data set? Explain.
8) Find the mean and the median for the following data set: 1, 6, 5, 7, 3, 2.5, 2, -1, 1, 0.
9) How does an outlier affect the mean and the median of a data set?

3.6.3 Measures of Variability Questions

10) What is smallest standard deviation that you can calculate and what would it mean if you found it?
11) What is the variance and the standard deviation of the following data set: 1, 2, 3, 4, 5.
12) Suppose you have a data set of 1,2,2,3,3,3,4,4,5 and you assume this sample represents a population.
    a) Explain why you can apply the Empirical Rule to this data set.
    b) Where would "most of the values" in the population fall based on this data set?
13) Suppose a mound shaped data set has a mean of 10 and a standard deviation of 2.
    a) What percent of the data falls between 8 and 12?
    b) About what percentage of the data should lie above 10? above 12?
14) Exam scores have a mound shaped distribution with a mean of 70 and standard deviation of 10. Your score is 80. Find and interpret your standard score.
15) Jenn's score on a test was 2 standard deviations below the mean. The mean class score was 70 with a standard deviation of 5. What was Jenn's original exam score?

4 Probability

The foundation for inferential statistics is probability. Without grasping some of the concepts of probability, what we do in inferential statistics will not have any meaning. Thus, this chapter (and the next one) introduces several concepts related to probability but from more of an intuitive perspective than a mathematical one. While there will be some mathematical formulas introduced in this section, the chapter's main purpose is to be able to help the reader understand the intuition behind the probabilistic concepts that relate to inference.

Before defining probability it is important to review the concept of uncertainty. Uncertainty is the study of something that you do not know. Uncertainty can pertain to an individual or a group of people. For example, the residents of Washington D.C. (a group) have a degree of uncertainty about whether or not it will rain tomorrow based on what the weatherman tells us (i.e. the weatherman reports a 65% chance of rain tomorrow). Additionally, uncertainty may not be universal. People can have different assessments of the chances of an event happening or not. That means, while you and I are uncertain about the chances of rain tomorrow, a weatherman may be quite certain about rain since he or she studies the weather for a living.

If uncertainty can vary across different people and groups, then there must exist a way to quantify uncertainty like we do with height, weight, and time. By quantifying uncertainty, we can discuss it in precise terms. For instance, if I say that I feel strongly about the chances of the Washington Nationals winning the World Series this year and my friend says he feels *very* strongly about the Nats chances, this seems to say my friend feels stronger about the Nats World Series chances than I do. But what precisely does this mean? It is hard to say without putting numbers behind the statements. For instance, my definition of strongly might be greater than my friends definition of very strongly.

The most commonly used measurement of uncertainty is probability. Thus, a definition for probability is that it is a measure of uncertainty. Put another way, probability is the likelihood that something will happen. Probability is usually expressed in percentage format. Your weatherman might tell you that there is a 65% chance of rain, which means you should probably pack your umbrella. If we go back to our Nats example, my strong belief that the Nats will win the World Series means a lot more to someone if I tell them that I think the Nats have a 60% chance. And if you tell them that my friend's very strong belief means the Nats have a 90% chance of winning the series makes the difference in our beliefs quantifiable.

The idea behind a statistical experiment ties in nicely with our discussion of uncertainty.

Definition

A statistical experiment is an act or process of observation that leads to a single outcome that cannot be predicted with certainty.

A statistical experiment is similar to any other kind of experiment that exists in biology, business, or engineering. You have a hypothesis you would like to test and you do not know what the result of the experiment will be. Think of some of the most common games of chance that you might play: flipping a coin, rolling a six sided dice, or pulling cards from a deck. Flipping a coin will yield heads or tails, yet you don't know which side of the coin will land up. Rolling a dice will yield one of six outcomes. Finally, you might pull several cards out of a deck, and wonder how many face cards (jack, queen, king, or ace) you will pull. Those are all examples of simple statistical experiments.

A statistical experiment consists of a process of observation that leads to one outcome. This outcome is known as a simple event. The simple event is the most basic outcome of any experiment and it cannot be decomposed into more basic outcomes.

Exercise: List the simple events associated with each of the following experiments:

1) Flipping one coin and observing the up-face.
2) Rolling one die and observing the up-face.
3) Flipping two coins and counting the number of heads on the up-face of the coins.
4) Will an individual respond to my direct marketing ad or not?

Answer 1) The simple events associated with flipping a coin and observing the up-face are: heads or tails.

Answer 2) The simple events associated with rolling a die and observing the up-face are: 1, 2, 3, 4, 5, 6.

Answer 3) The simple events associated with flipping two coins and counting the number of heads is: 0,1,or 2.

Answer 4) The simple events associated with an individual's response to a directing marketing ad is: yes, they will respond, or no, they will not.

If the outcome of a statistical experiment is one of several potential simple events, then the collection of all simple events is known as the sample space.

Definition

Sample Space: The collection of all the simple events for a statistical experiment. The sample space is denoted in set notation as a set containing the simple events. S:{E_1, E_2,…E_n}.

Exercise: List the sample space associated with each of the following experiments:

1) Flipping one coin and observing the up-face.
   Answer 1) ***S: {H,T}***

2) Rolling one die and observing the up-face.
   Answer 2) ***S:{1,2,3,4,5,6}***

3) Flipping two coins and counting the number of heads on the up-face of the coins.
   Answer 3) ***S: {0,1,2}***

4) Will an individual respond to my direct marketing ad or not?
   Answer 4) ***S: {Y,N}***

The final step in our statistical experiment set up is to assign probabilities to each simple event in our sample space. This can be done one of three ways:

1. Previous knowledge of the experiment.
2. Subjectively assigning probabilities.
3. Experimentation.

Assigning probabilities based on your previous knowledge of the experiment means that the experiment is familiar to us and we have enough knowledge about the experiment to predict the probability of each outcome. The best example of this kind of probability assignment is rolling a dice. We know that for a fair die, each outcome has equal probability of 1/6 or 17%. Thus, each simple event is assigned 17% probability. The assignment to the die rolling experiment (written in sample space terminology) is below:

S:{P(1)=17%; P(2)=17%; P(3)=17%; P(4)=17%; P(5)=17%; P(6)=17%;}

Similarly, if we have been working in direct marketing for a long time we might be able to assign probabilities of responding to a direct marketing package based on our knowledge of the business. If we know that the response rate for a certain offer is about 6%, we can use this to assign a probability of response.

S:{P(Y)=6%; P(N)=94%}

But what if we do not have any knowledge of the potential simple event probabilities? The first option we have is to assign the probabilities subjectively. That is, we use the knowledge of the event that is at our disposal to make an assessment about the probability. Say for instance that we have an opportunity to invest in a start up business. We want to better understand the risks involved in investing our money so we decide to assign probabilities to the simple events of the experiment of the success of the business. Since we do not have any expertise in this type of probabilistic assignment, we can make a subjective assignment by looking at different pieces of information that we have about the business. For instance, how strong is senior management, what is the state of the economy, and have similar business succeeded or failed in the past. Using these pieces of information, we can make an assignment. If the components look good, we might make assign their success at 40%.

S:{P(S)=40%; P(F)=60%}

The last option to assign probabilities is through experimentation. That is, we perform the experiment multiple times and assign the probabilities based on the relative frequency of the simple event happening. Imagine if we did not know that each face of a die had an equal opportunity of coming up. If we needed to assign probabilities to each outcome, we could roll the die several times and see the frequency of the values. Theoretically, we should see each face come up with the same frequency and then assign the probabilities based on them.

S:{P(1)=17%; P(2)=17%; P(3)=17%; P(4)=17%; P(5)=17%; P(6)=17%;}

There are two important rules to be mindful of when assigning probabilities to simple events. These rules are stated below.

The Rules of Probability

Rule #1: All simple event probabilities must be between 0 and 1.

Rule #2: The probabilities of all sample points must sum to 1.

Recall the sample space from the die tossing experiment.

$$S:\{P(1)=17\%;\ P(2)=17\%;\ P(3)=17\%;\ P(4)=17\%;\ P(5)=17\%;\ P(6)=17\%;\}$$

Each probability is between 0 and 1 (equal at 17%). The sum of the probabilities for the simple events is one.

An event, denoted by A, is the collection of simple events within a sample space. To calculate the probability of an event, sum the probabilities of the simple events that are in the sample space of A. The steps for calculating the probabilities of events are summarized below.

The Steps for Calculating the Probabilities of Simple Events

1) Define the experiment.
2) List the simple events.
3) Assign probabilities to the simple events.
4) Determine the collection of sample points contained in the event of interest.
5) Sum the sample point probabilities to get the event probability.

Example 1:

A fair die is tossed and the up face is observed. If the face is even you win $1, otherwise (if the face is odd) you lose a $1. What is the probability that you will win?

Per the above steps, the first thing to do is to define the experiment. In this case, the experiment is that we roll the dice and observe the up face. The simple events and sample space are S:{1,2,3,4,5,6}. The assignment of probabilities is then:

$$S:\{P(1)=17\%;\ P(2)=17\%;\ P(3)=17\%;\ P(4)=17\%;\ P(5)=17\%;\ P(6)=17\%;\}$$

The last step is to determine which of the simple events make up the event of interest. In this case, the event that we win. We win the game if the observed up face is even. Thus, we win if we roll a 2, 4, or 6. Written in notation A :{2,4,6}. And the probabilities associated with these simple events are

$$S:\{P(2)=17\%;\ P(4)=17\%;\ P(6)=17\%\}$$

Summing the probabilities of the simple events gives: .17+.17+.17 = 50%.

Probability related experiments exist outside of tossing die and flipping coins. Take for example the frequency table below which details the number of failures for a certain process for different failure categories

Management System Cause Category	Number of Incidents
Engineering and Design	27
Procedures and Practices	24
Management and Oversight	22
Training and Communication	10
Total	**83**

System Failures

What if we were interested in calculating the probability of an incident being a management related failure? The simple events are each type of incident that could occur and the sample space is the collection of these four simple events (engineering and design, procedures and practices, management and oversight, and training and communication. The probabilities will be assigned using the relative frequency of each of these events. Written in set notation the sample space is

S:{P(Engineering and Design)=32%; P(Procedures and Practices)=29%; P(Management and Oversight)=26%; P(Training and Communication)=12%}

Thus, the probability of an error due to management and oversight is 26%. Similarly, the probability of an incident not related to a management is the sum of the probabilities of other incidents 32%+29%+12% = 74%.

4.1 Union & Intersection of Events

Events can also be combined in two ways: unions and intersections. A union of two events A and B is the event that occurs if either A or B or both occur on a single performance of the experiment. It is denoted by 'U'. A U B consists of all the sample points that belong to A or B or both. When thinking about the union of two events, think about the word "or" and think about adding the probabilities together.

The intersection of two events A and B is the event that occurs if both A and B occur on a single performance of the experiment. The intersection of A and B is denoted by A∩B. The intersection consists of all sample points that belong to both A and B.

The union and the intersection can be demonstrated using a simple example with a deck of cards. There are 52 cards with four suits in a standard deck of cards. Face cards consist of jacks, queens, kings, and aces. Knowing this, answer the following questions:

Exercise

1) What is the probability of selecting a 9 from a deck of cards?
A: There are 9 heart cards in a standard deck, one for each suit. Therefore, the probability of selecting a nine is 4/52.

2) What is the probability of selecting a face card?
A: There are four face cards for every suit in a standard deck. Thus the probability of selecting a face card is 16/52.

3) What is the probability of selecting a 9 **and** a heart card?
A: In this case, there is only one card that is both a 9 and a heart card – the nine of hearts. Thus the probability is 1/52.

4) What is the probability of selecting a 9 or a face card?
A: There are four "9"s and 16 face cards. Thus, the probability is 20/52.

4.2 The Additive Rule of Probability

The additive rule of probability tells us how to calculate the probability of the union of two events. The additive rule of probability states that the probability of the union of events A and B is the sum of the probabilities of events A and B minus the probability of the intersection of events A and B – that is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

For example, what is the probability that a card selected from a deck will be either an ace or a spade?

The probability of an ace is 4/52, the probability of a spade is 13/52, and the probability of getting and ace and a spade is 1/52. Therefore, using the addition formula

$$(4/52)+(13/52) - (1/52) = (16/52)$$

Mutually exclusive events are two events whose intersection has no sample points in common. That is $(A \cap B) = 0$. Some intuitive examples of mutually exclusive events are skipping class and receiving a perfect score in a class attendance grade. An example of non-mutually exclusive events might be that there are clouds in the sky and it is raining, since clouds can be indicative of rain. So when you have mutually exclusive events, the addition formula resolves to just the sum of the probabilities, the intersection terms drops off.

A contingency table is one particular way to view a sample space. It is similar to a frequency table (from chapter 2) but the counts in the table are based off of two variables. Say for instance, 100 households were contacted at one point in time and each household was asked if in the coming year they planned to purchase a big-screen TV. At that point they could answer yes or no. Then a year later each household was called back and asked if they made a purchase and they could answer yes or no. So, our analysis consists of two variables: plan to purchase and did you purchase. In a contingency table one variable makes up the rows and the other makes up the columns.

	Did you purchase?		
Plan to Purchase	No	Yes	Total
No	65	10	75
Yes	6	19	25
Total	71	29	100

The data presented in the contingency table shows that 65 people say they did not plan to purchase a TV but did so anyway. Similarly, there were 10 people who said they had no plans to purchase a TV but then did so. Usually we use some notation for each response. For example, we might use A to mean yes, there was a plan to purchase and the A' means there is no plan to purchase. (You could use any letter for this.). Similarly, B could mean they did purchase and B' could mean they did not purchase. Using the data from the table, we could calculate the simple probabilities of purchasing a TV and plans to purchase a TV.

$$P(A) = 25/100 = 25\%$$

$$P(A') = 75/100 = 75\%$$

$$P(B) = 29/100 = 29\%$$

$$P(B') = 71/100 = 71\%$$

Some of the more complex events can be calculated using the contingency table. For instance, or the probability of planning to buy a TV and then purchasing it:

$$P(A \cap B) = 19/100 = 19\%.$$

Practical Example

A study done by Eharmony.com shows that 80% of all subscribers will lie on more than 3 questions on their forms, 30% of all subscribers have a negative experience and 25% have lied and had a negative experience. What are the probabilities of a new person coming into E-harmony lying (event A) about their profile, having a negative experience (event B) or both?

P(A) = 80%

P(B) = 30%

P(A U B)= 80%+30%-(25%) = 85%

4.3 Conditional Probability

Conditional probability is the probability of some event A, given that event B occurs. The keyword in conditional probabilities is the word "given" and the symbol to designate the conditional probability is the "|". For instance:

- What is the probability that it will rain today given there are clouds in the sky?
- What is the probability that the Dow Jones will rise given the Fed has just cut rates?

Those are both forms of conditional probability questions. The mathematical formula for calculating the conditional probability for events A and B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

To illustrate this, consider event A, the event that we observe an even up face (2, 4, or 6) on the roll of a die and B is the event that a value of 3 or less has already been rolled. What is the conditional probability of A given B?

$$P(A|B) = \frac{1/6}{1/2} = 1/3$$

Going back to our E-harmony and lying example, to prove to its customers that the E-harmony methodology actually does work, the owners commission a study. They want to show that if you lie on E-harmony you are more likely to have a negative experience, so it is in your best interest to be honest.

Recall that P(Lie) = 80% and P(Negative Exp) = 25% and, P(Lie and Negative Exp.) = 60%. Using the conditional probability formula,

$$P(A|B) = \frac{60\%}{80\%} = 75\%$$

Clearly, the chances of you having a negative experience on E-harmony is enhanced if you lie on your application.

4.4 Multiplicative Rule of Probability

Our final rule helps us to calculate the intersection of two events using the multiplicative rule. Recall the conditional probability formula for probability. Multiply both sides by the denominator, P(B). Doing this would leave us with the following.

$$P(B) * P(A|B) = P(A \cap B)$$

Then, simply reorder the terms to put the intersection on the left hand side of the equation.

$$P(A \cap B) = P(B) * P(A|B)$$

4.4.1 Independence of Events

Lastly, an important definition is independence. Events A and B are independent if the occurrence of B does not alter the probability that A has occurred. Written using formulas:

$$P(A|B)=P(A)$$

$$P(B|A)=P(B)$$

The benefit to having independent events is that it makes the calculation of the intersection rule much easier – All you have to do is multiply the two events together. To derive this, recall the formula for the intersection of two events:

$$P(A \cap B) = P(B) * P(A|B)$$

But we know that if the events are independent then

$$P(A|B) = P(A)$$

Therefore

$$P(A \cap B) = P(B) * P(A)$$

4.5 Chapter 4 Problems

1) An experiment results in one of the following sample points: E1, E2, E3, E4, or E5. Find P(E3) if P(E1)=.1, P(E2)=.2, P(E4)=.1, and P(E5)=.1

2) The sample space for an experiment contains five sample points with probabilities as shown in the table below.

Sample Points	Probabilities
1	.05
2	.20
3	.30
4	.30
5	.15

Find the probabilities of

A: {Either 1, 2, or 3 occurs}.

B: {Either 1,3,or 5 occurs}.

C: {4 does not occur}

3) The contingency table below looks at age by income class

Age	<$25,000	$25,000-$50,000	>$50,000
<30 years old	5%	12%	10%
30–50 years old	14%	22%	16%
>50 years old	8%	10%	3%

A is the event that {A respondent's income is >$50,000}

B is the event that {A respondents age is 30 or more}

a) Find the P(A) and P(B)
b) Find P (AUB)
c) Find P(A∩B)

4) For two events A and B, P(A) = .4, P(B) = .2 and P(A∩B)=.1
   a) Find P(A|B)
   b) Find P(B|A)
   c) Are A and B independent events?

5 Random Variables and Probability Distributions

In the previous chapter, we studied the basic rules of probability and how to calculate the probabilities of simple and complex events. We learned about statistical experiments and how to answer questions related to these experiments by writing out the sample space and identifying the simple events that made up the larger event of interest. We also learned how to calculate the probabilities for the union and intersection of the events. Finally, we learned about the concepts of independence and conditional probability.

But what do we do in situations where it is not possible to write out the sample space and assign probabilities to simple events? What if we were interested in calculating the probability of selecting a person at random that has a specific weight from a large population[3]? Or what if we were interested in picking a single, specific grain of sand from a beach? In these situations it is impossible to write out a sample space. Statistical random variables, defined below, are used to solve such problems as those listed above.

Definitions

Random Variable: A variable that assumes numerical values associated with the random outcomes of an experiment where one and only one numerical value is assigned to each sampled point. There are two types of random variables: discrete or continuous.

Discrete Random Variables: Random variables that can assume a countable number of values.

Continuous Random Variables: Random variables that can assume any value corresponding to any of the points contained in one or more intervals (i.e. values that are uncountable).

A discrete random variable is one that can only take on a countable number of distinct values such as 0,1,2,3,….They are often counts (but this is not necessary). If a random variable can only take a finite number of distinct values, it must be discrete. Some examples of discrete random variables include the number of defective light bulbs in a box, the number of children in a family, or the number of at bats it will take for Alex Rodriguez to hit his next home run.

Continuous random variables can assume any value corresponding to any of the points contained in one or more intervals. They are usually measurements. Things like heights, weights, and time are continuous random variables. Specific examples of continuous random variables might be the time it takes to complete a race or the length of time between arrivals at a hospital clinic.

Examples

Are the following examples of discrete or continuous random variables?

1) The number of newspapers sold by the New York Times.
   Discrete. The newspapers that are sold is a countable value.

2) The amount of ink used in printing a Sunday edition of the New York Times.
   Continuous. Ink is something that is measured, not counted.

3) The actual number of ounces in a one-gallon bottle of laundry detergent.
   Discrete. The question asks for the number of ounces in a one-gallon bottle

4) The number of defective parts in a shipment of nuts and bolts.
   Discrete. Again, this question is asking for the number of defective parts.

5) The number of people collecting unemployment insurance each month.
   Discrete. Again, it is the number of people.

Now that we have a working definition for what a random variable is, we can tie it back into our study of probability by introducing the concept of the probability distribution for a discrete random variable.

Definition

Probability Distribution: The probability distribution of a discrete random variable is a graph, table or formula that specifies the probability associated with each possible value the random variable can assume.

The basic rules of probability apply to a probability distribution the same way they did in previous chapters. That is:

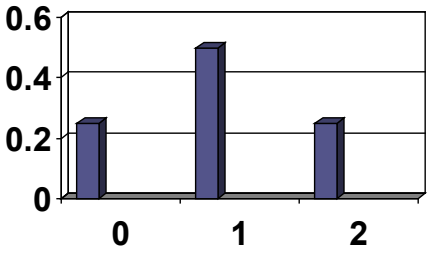
- The sum of all probabilities in a distribution sum to 1.
- Each value has a probability between 0 and 1.

The probability distribution for a discrete random variable can be represented in one of two ways: either as a table or as a graph. The figure below shows a tabular representation for a simple probability distribution that has three outcomes.

x	p(x)
0	¼
1	½
2	1/4

Tabular Representation for a Probability Distribution

Below, shows the same distribution, yet represented graphically, in a histogram



Graphical Representation for a Probability Distribution

And just as the distributions that we studied in chapter two had means and variances, probability distributions have them as well. For a probability distribution, the mean of the distribution is known as the expected value. The expected value intuitively refers to what one would find if they repeated the experiment an infinite number of times and took the average of all of the outcomes. Mathematically, it is calculated as the weighted average of each possible value. The weights are the probability of the event occurring. Like the means we looked at in chapter 2, it is a measurement for where the distribution is centered. The expected value is not necessarily the event with the highest probability and since we are taking a weighted average, the expected value is not necessarily one of the outcomes of the experiment.

The formula for calculating the expected value for a discrete random variable x, denoted by μ, is

$$\mu=\sum xp(x)$$

The variance of a discrete random variable x, denoted by σ^2 is

$$\sigma^2=E[(x-\mu)^2]=\Sigma(x-\mu)^2p(x)$$

The standard deviation is the square root of the variance.

Example

Find the mean and variance for the following probability distribution for the random variable, X.

X	0	1	2	3	4	5
P(x)	.002	.029	.132	.309	.360	.168

To calculate the expected value of this probability distribution, take the weighted average of the distribution:

(0*0.002) + (1*0.029) + (2*0.132) + (3*0.309) + (4*0.360) + (5*0.168) = 3.5.

To calculate the variance of the distribution:

$((0-3.5)^2*0.002)+ ((1-3.5)^2*0.029)+ ((2-3.5)^2*0.132)+ ((2-3.5)^2*0.132)+ ((3-3.5)^2*0.309)+ ((4-3.5)^2*0.360) +((5-3.5)^2*0.168) = 1.345$

Many experiments can be modeled by a few common probability distributions. The trick is to recognize which distribution your experiment follows. Once recognized, you can use well established properties about the distribution to calculate probabilities. This can save a considerable amount of time and effort as you no longer have to determine sample spaces and simple events. We will look at two well known distributions in the remainder of the chapter.

5.1 The Binomial Distribution

The binomial distribution is a well known discrete probability distributions that many experiments have been modeled after. It is characterized by five attributes.

1) The experiment consists of n identical trials.
2) There are only two possible outcomes on each trial which are denoted as S and F, for success and failure respectively.
3) The probability of S remains the same from trial to trial. This probability is denoted as "p" and the probability of failure is (1-p), or "q".
4) The trails are independent. This means that the results of one trail have no bearing on the next.
5) The binomial random variable x is the number of successes in n trails.

Thus, if an experiment meets the above criteria, then you can safely assume that your experiment follows a binomial distribution. As an example, suppose you bet your friend that you can hit the bulls eye on the dart board 6 out of 10 times. You want to calculate the probability that you can win the bet. The experiment looks like it is a binomial experiment, but to be certain you evaluate it against the criteria listed in the above box. The experiment consists of 10 identical trials (throwing a dart at the board) so the first condition is satisfied. Since you can only either hit the bulls eye (a success) or miss it (a failure) condition two is satisfied. The probability of a success theoretically should remain the same from one trail to the next. That is, over the course of 10 trails, your success rate should not change dramatically, thus condition three is met. The trails are independent, meeting condition four. Finally, the random variable in question is the number of times you hit a bulls eye in 10 trails. In this case, you are interested in the probability of hitting the bulls-eye six times.

Once we recognize that an experiment can be modeled with a binomial random variable, we can calculate the probability of interest directly using the following formula.

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

where p=probability of success; (1-p) is the probability of failure; n is the number of trials; and x is the number of successes in n trials. The capitol X in the P(X=x) statement denotes the random variable X and the little x is the value that the random variable can take. The term $\binom{n}{x}$ is known as a combinatric term.

The term is pronounced as "n choose x". The term is calculated using the factorial formula shown below:

$$\binom{n}{x} = \frac{n!}{k!(n-k)!}$$

where n! is the product of all possible integers less than or equal to n. For example, 5! = 5x4x3x2x1.

Continuing with our dart throwing example above say that you know that the probability of you throwing a bulls eye is 30%. This is your probability of success. Thus, (1-p) or the probability of failure is 70%. The number of trials, n = 10 and the number of successes you are interested in is 6, which is x. Therefore, plugging these numbers in:

$$\binom{10}{6} .3^6 (.6)^4$$

The resulting probability is 0.0367, or about 4%. So your odds of winning the bet are not very good.

Instead of using the above formula, binomial probabilities can easily be calculated in Excel. The function binomdist gives the direct probability for a binomial experiment such as the one above. binomdist takes four arguments: the number of successes, the number of trials, the probability of success and whether or not you want the function to return the exact probability or the cumulative probability (always set this to "false" to calculate the exact form of the binomial probability). The syntax is as follows:

=bindomdist(6,10,.3,false)

Example

A machine that produces stampings for automobile engines is malfunctioning and producing 10% defectives. The defective and non-defective stampings proceed from the machine in a random manner. If the next five stampings are tested, find the probability that three of them are defective. So our number of trials is five, our probability of success is .1, and we are interested in finding if three of them are defective. Plugging this into our excel function

=binomdist(3,5,.1,false) = 0.0081.

So the chances are quite low that we get 3 defects out of the next 5 tested.

To calculate the mean and variance of a binomial probability distribution use the following formulas:

$$\text{mean} = n^{*}p$$

$$\text{variance} = n^{*}p^{*}q$$

$$\text{standard deviation} = \sqrt{(n * p * q)}$$

where n is the number of trials and p is the probability of success.

Example

In the machine example from above, what is the mean and variance of the probability distribution?

mean = n*p = 5*.1 = 0.5

variance = n*p*q = 5*.1*.9 = 0.45.

Thus, recognizing a probability distribution can be helpful in solving problems related to statistical experiments. By recognizing the properties of the binomial experiment we could calculate specific probabilities as well as calculate the mean and variance of the distribution.

5.2 The Normal Distribution

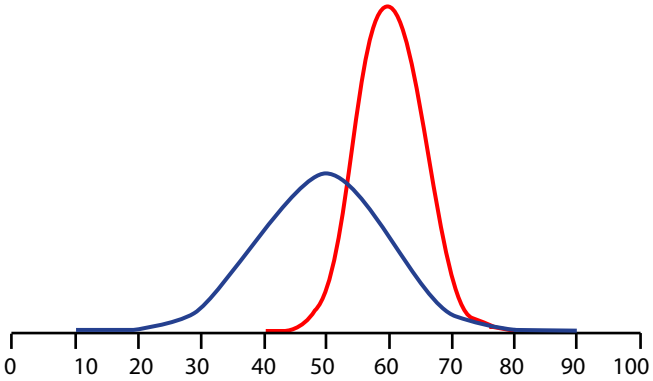
The next distribution that we will investigate is the normal distribution. It is one of the most famous statistical distributions in use and we will use it extensively throughout the rest of this book. The normal distribution is a continuous probability distribution. Thus, it applies to random variables that are measured not counted. Several phenomenon are modeled with the normal distribution. For instance, heights of people are normally distributed as well as possible blood pressure levels for people.

The best way to understand the normal distribution is to study its properties.

The Properties of the Normal Distribution

- A continuous random variable X has a normal distribution if its values fall into a smooth curve that is bell shaped.
- Every normal distribution has its own mean (denoted μ) and its own standard deviation (denoted σ). The normal distribution is defined by its mean and standard deviation.
- Its shape of the normal distribution is symmetric around the mean.
- The mean, median, and the mode of a normal distribution are equal.
- The area under the curve is 1.
- Normal distributions are denser in the center and less so in the tails.
- Since the normal distribution is mound shaped, it follows the empirical rule from chapter three. Thus, 68% of the area of a normal distribution is within one standard deviation of the mean; 95% of the data are within 2 standard deviations of the mean; 99.7% of the data are within 3 standard deviations of the mean.

What do we mean by the second property that each normal distribution is defined by its mean and standard deviation? The figure below shows two normal distributions.



Both the red distribution and the blue distribution are bell shaped and symmetric around the mean. What sets them apart however is their location (mean) and spread (standard deviation). The red normal has a larger mean than the blue distribution. Conversely, the blue distribution has a larger standard deviation than the red distribution.

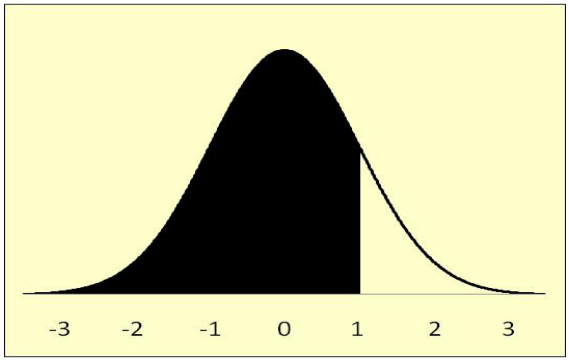
How do we calculate normal probabilities using the normal distribution? Calculating probabilities for the normal distribution is not as simple as it was for calculating them for a binomial random variable. The formula for calculating the exact probability for an event is complicated and requires using mathematics beyond the scope of this book.

Instead, we take advantage of what's known as the standard normal distribution to do so. The standard normal distribution is a special kind of normal distribution with mean 0 and standard deviation of 1. It is known as the z-distribution because z represents the number of standard deviations an observation is from the mean. and it is one of the most commonly referenced distributions in statistics.

We can use Excel to find the probabilities of interest for a standard normal distribution. Recall from earlier chapters that the z-score is the number of standard deviations away from the mean and that for a mound shaped distribution 99% of the data falls between a z score of -3 and 3. Excel will provide the area under the normal curve (the probability) to the left of a given value of z with the following formula

=normdist(z)

Excel will accept both positive and negative values for z.



Pictorial Representation of the Standard Normal Distribution

Here, the picture above represents the area to the left of the z-score of 1 – the value that Excel returns. In probabilistic terms, this is

$$P(Z<1) = \text{normdsist}(1).$$

Similarly, if we were interested in finding the probability that z is greater than 1 the formula is

$$P(Z>1)= 1 - \text{normdist}(1).$$

Or if we were interested in finding the probability that Z is between 1 and -1

P(-1< Z <1)= P(1> z > 1) = normdist (b) – normdist (a)

Exercises

Find the following probabilities

1) P(Z<1.28)
   =normdist(1.28)

2) P(Z>-1.64)
   =1-normdist(-1.64)

3) P(-1.64<Z<1.28)
   =normdist(1.28) – normdist(-1.64)

4) P(Z<0)
   =normdist(0)

5) P(1.36<Z<.54)
   =normdist(1.36) – normdist(.54)

However, most normal distributions don't have a mean of 0 nor a standard deviation of 1. This issue is easily taken care of by standardizing the variable. This means, we convert the random variable to a z random variable by subtracting the mean of the distribution and dividing by its standard deviation. As a specific example, say we want to find the probability that a normally distributed random variable has a mean of 100 and a standard deviation of 10 and we want to find the probability that X is less than 110. In probability terms P(X<110). First standardize the random variable

$$= P\left(Z < \frac{110-100}{10} \right)$$

=P(Z<1)

=normdist(1)

=.8413.

You can do this process in Excel without having to do the extra step of standardizing. To do this plug the value of interest, mean and standard deviation into the Excel formula

=normdist(X, mean, standard deviation, 1)

where the "1" tells Excel that you are looking for the area under the curve. So in the example above:

=normdist(110,100,10,1) = 0.8413

Example

Assume the length of time, x, between charges of a cellular phone is normally distributed with a mean of 10 hours and a standard deviation of 1.5 hours. Find the probability that the cell phone will last between 8 and 12 hours.

=normdist(12,10,1.5,1) – normdist(8,10,1.5,1) = .8175

Exercises to try on your own

- X~N(m=85, s=3)
  - P(X < 80) =normdist(80,85,3,1) = .0478
  - P(X > 89) =1-normdist(89,85,3,1) = .0912
  - P(80 < X < 90) =normdist(90,85,3,1)-normdist(80,85,3,1) =.9044
  - P(X > 100) =1-normdist(100,83,3,1) = 7.28×10^{-9}

- X~N(m=25, s=.5)
  - P(X < 23.5) =normdist(23.5,25,.5,1) = .0013
  - P(X > 24.25) =1-normdist(24.25,25,.5,1) = .9332
  - P(24.3 < X < 25.2) =normdist(25.2,25,.5,1)-normdist(24.3,25,.5,1) = .5747

5.3 Chapter 5 Problems

1) For a math test, you have a 10 question multiple choice exam. If each question has four choices and you guess on each question (with equal probability) what is the probability of getting exactly 7 questions?
   a) Does this fit the binomial paradigm? Why?
   b) Calculate the probability of getting exactly 7 questions correctly.
   c) Calculate the mean and standard deviation of this distribution.
2) What is the probability of obtaining exactly eight heads in ten tosses of a fair coin?
3) On a given day, the probability that your family eats breakfast together is 2/5. What is the probability that during any 7 day period, your family eats breakfast together at least six times?
4) What is the expected value for a probability distribution that has the following parameters: n=25 and p=0.7. What is the standard deviation of said distribution?
5) Which of the following are examples of binomial experiments?
   a) Tossing a coin 20 times to see how many tails occur.
   b) Asking 200 people if they watch the daily news.
   c) Rolling a die to see if a five appears.

6) Most graduate schools require that applicants take a standardized test for admission. Scores on the standardized test are normally distributed with a mean of 527 and a standard deviation of 112. What is the probability of an individual scoring above 500 on the exam?
7) Calculate the following probabilities based on the parameters given below:
   $X \sim N(\mu=75, \sigma=1.5)$
   a) $P(X < 70)$
   b) $P(X > 85)$
8) The distance that a new car can go on a tank of gas averages 266 miles with a standard deviation of 16 miles. What proportion of all cars will last between 240 miles and 270 miles?
9) The annual salaries of employees in a large company are normally distributed with a mean of $40,000 and a standard deviation of $10,000.
   a) What percent of people earn less than $40,000?
   b) What percent of people earn between $45,000 and $65,000?
   c) What percent of people earn more than $70,000?
10) For a normally distributed random variable, what percentage of data points lie within two standard deviations from the mean?

6 The Sampling Distribution

In the previous chapters we "knew" what a probability distribution looked like. That is, we were given the mean and standard deviation for a normal distribution and we were the given the probability of success for the binomial distribution. We could then use this information to answer questions we had about probabilities. However, usually we don't know what the mean or probability of success is and, in fact, most problems are trying to solve for this measure.

Definition

Parameter: A numerical descriptive measure of a population. Because it is based on the observations in the population, its value is almost always unknown and we want to estimate its value.

Sample Statistic: Numerical descriptive measure of a sample that are calculated from the observations in the sample (e.g. mean, standard deviation)

But how do we estimate these unknown population parameters? We often use the information that we gather from samples to calculate sample statistics which we use to estimate the unknown population parameters. The table below shows some common sample statistics that we use to estimate population parameters

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Variance	S^2	σ^2
Standard Deviation	S	σ
Binomial Proportion	$\hat{p}$	p

Sample Statistics Used to Estimate Population Parameters

The sampling distribution of a sample statistic, calculated from a sample of n measurements, is the probability distribution of the statistic. This is often a difficult concept to understand so it helps to work through it with an example. Say we are interested in knowing the average amount of money spent on groceries for a family of four in a month. It is not reasonable to go out to every family of four in the country and collect this data. But what I can do is select a sample of households, say 30, collect their monthly grocery bill and take the average. So I take my sample of 30 and get an average of $350. My friend also decides to perform the experiment and he goes out, collects a sample of 30 and gets a different monthly average, say $310.

The average grocery amount spent clearly depends on who is selected in the sample. We might happen to select a house that spends a significant amount of money on their monthly grocery bill, say $1075. Needless to say, if we ran this experiment a very large number of times and covered every possible 30 household sample in our population and then plotted the results, the resulting histogram would constitute the sampling distribution of all possible means.

Let's look at the sampling distribution in action using a smaller data set. The probability distribution below shows a population of measurements that can assume values of 1, 6, 11, and 16 each of which occurs with the same probability (25%)

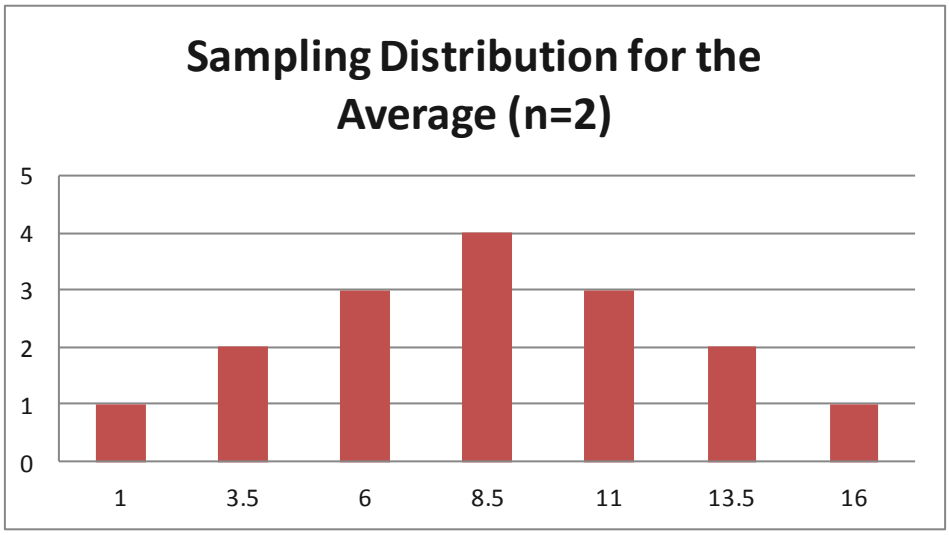
X	1	6	11	16
P(x)	.25	.25	.25	.25

Say we are interested in the sampling distribution for the mean for samples of n=2. As a first step, we identify every possible sample of two and take the mean. The table below shows the results of the first step.

First Draw	Second Draw	Mean
1	1	1
1	6	3.5
1	11	6
1	16	8.5
6	1	3.5
6	6	6
6	11	8.5
6	16	11
11	1	6
11	6	8.5
11	11	11
11	16	13.5
16	1	8.5
16	6	11
16	11	13.5
16	16	16

Then in step 2, create the frequency table (or histogram) for each of the resulting means. In this case, the means can take values of 1, 3.5, 6, 8.5, and 13.5.

Mean	Count	Frequency
1	1	6%
3.5	2	13%
6	3	19%
8.5	4	25%
11	3	19%
13.5	2	13%
16	1	6%



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

The sampling distribution for the mean of this data set looks like a normal distribution. This leads to some interesting properties about the sampling distribution of the mean. Assuming that a random sample of n observations has been selected from a population. First, is that the mean of the sampling distribution equals the mean of the sampled population. Second is the standard deviation of the sampling distribution equals the standard deviation of the sampled population divided by the square root of the sample size. This is known as the standard error of the mean. The formula for the standard error is presented below

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

There are two key theorems related to sampling distributions worth mentioning. We will only state the theorems here, not provide mathematical proof of them. The first theorem is that if a random sample of n observations is selected from a population with a normal distribution, the sampling distribution of x-bar will be a normal distribution.

The second theorem is known as the central limit theorem**.** The central limit theorem states that for a sufficiently large (n>30) random sample the sampling distribution of x-bar will be approximately normal with mean mu and standard error of sigma/sqrt(n). The larger the sample size, the better will be the normal approximation to the sampling distribution of x-bar. The central limit theorem is one of the most important theorems in statistics. Since we often sample from a population where you do not know the distribution of the underlying population, it says that if our sample size is more than 30, we have sampling distribution that is approximately normal. We then know several properties about this distribution, namely its mean and standard error.

7 Confidence Intervals

Let's start this chapter with a question. How much money do Washington DC residents spend on groceries per month? To answer this question, I take a random sample of 200 individuals and ask them their monthly expenses on groceries. I calculate the mean of their expenses and I come up with $125.67. This value is known as a point estimate for the monthly cost of groceries for residents of Washington DC. However, as we saw in the last chapter, I can get a different point estimate every time I pull a different sample. For instance, what would happen if in the second sample I pulled, I found a family that spent $2MM/month on groceries? This might make the point estimate well over $2,000. The question begs, what is the point of estimating if we get a different answer each time we pull a sample?

What we want to be able to do is account for this variability in point estimates. If we can do that, then we can at least have some sort of measure of reliability for our estimate. To generate this reliability measure, first recall the Central Limit Theorem (CLT) from the last chapter. The CLT taught us that if the population distribution has a mean μ and a standard deviation σ, then for sufficiently large n, the sampling distribution of x- bar has mean μ and standard deviation of

$$\frac{\sigma}{\sqrt{n}}$$

Then, as n gets large, the sampling distribution of x-bar becomes approximately normal. Recall that our sample size was 200. So even though we don't know what the true distribution is, we know that the sampling distribution of x-bar is approximately normally distributed. And if we go even further back, to say chapter 3, we know that ~95% of the observations will fall within 2 standards of the mean for our sampling distribution (for a mound shaped distribution). Recall that the standard deviation of a sampling distribution is

$$\frac{\sigma}{\sqrt{n}}$$

If we take a large number of sample means, then 95% of the time, the unknown population mean is between

$$\overline{X} - 2 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2 \cdot \frac{\sigma}{\sqrt{n}}$$

The form above is known as a confidence interval. The thought behind the confidence interval is that instead of estimating the parameter by a point estimate, a whole interval of likely estimates is given.

Definition

Confidence Interval: A confidence interval for a population parameter is an interval estimate with an associated probability p that is generated from a random sample of an underlying population such that if the sampling was repeated numerous times and the confidence interval recalculated from each sample according to the same method, a proportion p of the confidence intervals would contain the population parameter in question. Confidence levels are often calculated at 90%, 95%, or 99%.

The confidence coefficient, known as α, is the probability that a randomly selected confidence interval encloses the population parameter – that is, the relative frequency with which similarly constructed intervals enclose the population parameter when the estimator is used repeatedly a very large number of times. The confidence level is the confidence coefficient expressed as a percentage. In general to calculate alpha:

α=1-.90=.1 (90% confidence)

α=1-.95=.05 (95% confidence)

α=1-.99=.01 (99% confidence)

So, for a given sample, if we have the mean, the standard deviation, the sample size, and a specified alpha level, we can calculate a confidence interval. We can do this in Excel using the ***confidence*** function. The function is as follows:

$$=\text{confidence}(\alpha,\sigma,n)$$

The above gives us what is known as the *margin of error*. We add and subtract this from the sample mean to get the confidence interval.

Example

You are a Q/C for Gallo. The standard deviation for a two liter bottle of wine is .05 liters. A random sample of 100 bottles shows that the average amount poured into a bottle is 1.99 liters. What is the 90% confidence interval estimate of the true mean amount in 2-liter bottles?

We know the mean amount of wine is 1.99 liters. Our standard deviation is .05 and we want a 90% confidence interval. Use the confidence function in Excel to find the margin of error.

=confidence (.1,.05,100) =0.0979

Therefore, our confidence interval is

1.99±0.0979

The interpretation of the above is that if we repeated our experiment a number of times and compute the confidence intervals each time, the confidence intervals you get will include the population mean 90% of the time. In a sense, there is a 90% chance that any specific confidence interval (including the one we calculated) will contain the population mean.

Example

A random sample of 225 1st year statistics tutorials was selected from the past 5 years and the number of students absent from each one recorded. The mean number of absences was 11.6 and the standard deviation was 4.1 absences. Find a 90% confidence interval for the mean number of absences.

=confidence (.1, .4.1,225) =0.445

11.6±0.445

The interpretation is that if repeated samples were taken and the 90% confidence intervals were computed each time, 90% of the intervals would contain the true population parameter.

In summary, confidence intervals measure a population parameter. They consist of a point estimate with some measure of reliability. The take the form

$$\text{estimate} \pm \text{margin of error}$$

where we used Excel to calculate the margin of error. A confidence level gives the probability that the interval will capture the true parameter value in related samples. The proper interpretation of a confidence interval must include the idea that the numbers were arrived at by a process that gives the correct result 95% of the time. The key to the interpretation is the idea of the repeated process. You do not know whether our sample is one of the 95% (for instance) for which the interval we calculated captured our true mean or if it is one of the unlucky 5%.

7.1 Small Sample Confidence Intervals

When the sample size is small (say, below 30), you have less information on which to base your conclusions about the mean. In addition, the Central Limit Theorem cannot kick in, so you can't rely on the normal distribution to make estimates about the margin of error from. When situations like these arise you use a different distribution, the t-distribution.

The t-distribution has many of the same properties as the normal distribution – it's just a flatter version of it. The thought behind the t-distribution is that since you have less information with which to work with, the distribution is less forgiving than the normal. It is less forgiving by having heavier tails than the normal. These heavier tails mean that there is more opportunity to producing values that are further from the mean than they would be for a normal distribution. Each different sample size has its own t-distribution, because the smaller the sample size, the heavier the tail. Put another way, the less information you have, the heavier and flatter the t-distribution is, which as you can imagine, leads to producing values far from the mean.

Each t-distribution is distinguished by its "degrees of freedom". If your sample size is n, the degrees of freedom for the corresponding t-distribution is n-1. For example, if your sample size is 11, you would work with a t-distribution with degrees of freedom equal to 10.

From there, the formula for a confidence interval is the same as with the large sample confidence interval.

$$\text{Estimate} \pm \text{Margin of Error}$$

The excel formula for calculating a margin of error with the t-distribution is

$$=\text{confidence.t}(\alpha,\sigma,n)$$

Example

A sample of 15 test tubes tested for number of times they can be heated on Bunsen burner before they cracked gave average of 1,230 and s=270. Construct a 99% confidence interval for mu.

=confidence.t(.01,270,15) = 207.54

1230±207.54

7.2 Chapter 7 Problems

1) If 100 randomly selected runners from a race can complete a mile run with an average time of 9 minutes with a standard deviation of 30 seconds, find a 95% confidence interval for the average time on all the runners in the race.
2) Based on a sample of 100 participants, a 95% confidence interval for the average SAT score for individuals who took a prep course is 11.4 points plus or minus .51. Interpret this confidence interval.
3) Suppose survey results say that 69% of adult females from a sample taken of shoppers use a credit card when making purchases. The margin of error for the survey that produced these results was +/- three percent. Interpret these values.
4) A sample size of 100 produced the sample mean of 16 with a standard deviation of 3. Compute a 95% confidence interval for the population mean.
5) A sample of 30 individuals has a mean score on a test of 58 with a standard deviation of 3.2. Compute a 99% confidence interval based on this sample data.
6) A sample of data from car lot had the following consumption of gasoline in gallons. 32, 16, 25, 22, 27, 19, 20. Calculate a 90% confidence interval for the mean daily use of gasoline.
7) Suppose you have selected a sample of data with values: 4,6,3,5,9,3. Construct a 90% confidence interval for the population mean.
8) A random sample of 90 observations produced a mean of 25.9 and a standard deviation of 2.7. Find a 95% confidence interval for the population mean.
9) A random sample of 90 observations produced a mean of 25.9 and a standard deviation of 2.7. Find a 90% confidence interval for the population mean.
10) A random sample of 90 observations produced a mean of 25.9 and a standard deviation of 2.7. Find a 99% confidence interval for the population mean. How does this interval differ from the two you calculated in problems 8 and 9?

8 Hypothesis Testing

A statistical hypothesis is a belief about a population parameter such as its mean, proportion or variance. A hypothesis test is a procedure which is designed to test this belief. Put another way, you have a belief about a population parameter and to confirm this belief, you run a hypothesis test to see if this belief is true or not. The calculation steps of a hypothesis test are similar to what was done for a confidence interval (as we will see when we get to that point) but there is a large difference in what we are trying to accomplish with a hypothesis test over a confidence interval. In a confidence interval we have no idea what the parameter is and we use the data to estimate it. In a hypothesis test we use the data to confirm an a priori belief about it. In this chapter we will walk through the steps to run a hypothesis test for large and small sample sizes. There are several components that go into a hypothesis test and we will go through each of them first before putting them together to run a hypothesis test.

8.1 The Null and Alternative Hypothesis

Every hypothesis test contains two hypothesis: the null hypothesis and the alternative hypothesis. The null hypothesis is a statement about the population parameter being equal to some value. For example, if you believe that the average weight loss for a population of dieters is equal to 2 lbs in one week, you have H_o:μ=2. The null hypothesis is defined as follows:

- The null hypothesis is what is being tested.
- Designated as H_0 (pronounced h-nought).
- Specified as H_0 = some numeric value (or ≤, ≥)

The third bullet is the key one to remember when setting up your null hypothesis. The equality sign is always in the null hypothesis as it is what your belief is and what you want to test.

Along with the null hypothesis, every hypothesis test also has an alternative hypothesis, denoted H_a If H_0 turns out to be false, then one accepts the alternative hypothesis is true. The alternative hypothesis is defined as follows:

- It is the opposite of the null hypothesis.
- Designated as H_a.
- Always has an inequality sign in it: <, >, or ≠.

The third bullet in our definition for the alternative hypothesis is very important. Which sign you choose when setting up your alternative hypothesis depends on what you want to conclude should you have enough evidence to refute the null. If you want to claim that the average weight loss is greater than 2 lbs, then you write H_o:μ>2. If you do not have any directional preference about the alternative, you write H_o:μ≠2.

Example

Set up the null and alternative hypothesis for the following question:

Q: Is the population average amount of TV viewing 12 hours per week?

A: Here the question asks if the average amount of TV viewing is 12 hours per week or not? The question does not differentiate more or less than 12 hours, just whether or not it isn't. Clearly our null hypothesis is

H_o:μ=12

The alternative hypothesis, again, since we do not care whether or not the TV is less than 12 or greater than 12:

H_A:μ≠2.

Example

Set up the null and alternative hypothesis for the following question:

Q: Is the average amount spent in the local bookstore greater than $25?

A: Here we are interested in if the spending is more than $25 so we set up the alternative hypothesis to be only greater than 25.

H_o:μ=25

H_A:μ>25

8.2 One or Two Sided Hypothesis

The type of hypothesis that we are interested in testing sets up a either one or two sided test. If you have no a priori expectation of the outcome of your test (or you are not concerned if the alternative is higher or lower) you have a two sided test, and you use the inequality sign ≠ and you have a two sided test. On the other hand if you want to test whether it is higher or lower, you use the < or > in the alternative and have a one sided test.

Example

A pharmaceutical company is interested in testing whether or not their new drug relieves pain more than their current drug. They run a statistical hypothesis test with the outcome being the mean score on a pain relief scale. Thus we test:

$H_0 : \mu_{current} = \mu_{new}$ $H_a : \mu_{cirrent} < \mu_{new}$ one-sided

8.2.1 The Test Statistic

Another important component to hypothesis testing is the test statistic. The test statistic is based on the statistic that estimates the parameter. Because normal calculation require standardized variables, we use our test statistic the standardized sample mean.

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

This random variable has the standard Normal distribution N(0,1).

8.2.2 The P-value

The last concept that is important for our understanding of hypothesis testing is calculating (with Excel) and interpreting the p-value. Tests of hypothesis quantify the chance of obtaining a particular random sample result if the null hypothesis were true. It is a way of assessing the accuracy or how believable the null hypothesis is given the evidence provided by the sample mean. A small p-value implies that random variation due to sampling is not likely to account for the observed difference. When we have a small p-value, we reject the null hypothesis. But how small is small? Oftentimes, a p-value of 0.05 or less is considered significant. The phenomenon observed is unlikely to be entirely due to chance event from the random sampling. The 0.05 is known as the significance level of the test. It is decided before you conduct the test. Written formally,

- If the p-value is equal to or less than the significance alpha, then we reject H_o in favor of H_A
- If the p-value is greater than the significance level then we fail to reject H_o.

To test the hypothesis of a mean based on a simple random sample of size n from a normal population with unknown mean μ and known standard deviation σ we rely on the properties of the sampling distribution that we studied in the previous chapter. The p-value is the area under the sampling distribution for values at least as extreme in the direction of the alternative hypothesis as that of our random sample. First we use the test statistic from above to calculate a z-value and then use excel to find the p-value associated with the value. The Excel formula for calculating a p-value is below:

=NORMDIST(X,mean,stdev,cumulative)

where **X** is the value of the test statistic

mean is 0 (since we are using a standard normal curve)

stdev 1 (again, since we are using the standard normal curve)

and **cumulative** is a true/false argument that should always be set to true for the exercises in this chapter.

Example

Perform the following test: $H_0: \mu = 227$ g versus $H_a: \mu \neq 227$ g where the mean is 222g, the standard deviation is 5g, and the sample size is 30.

Our test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{222 - 227}{5/\sqrt{30}} = -5.477$$

plugging this into our Excel formula

=normdist(-5.477,0,1,true)

=0.000002

Since this value is clearly below 0.05, we would reject the null hypothesis.

The steps for conducting a hypothesis test are summarized below:

1. State the null and alternative hypothesis.
2. State a significance level.
3. Calculate the test statistic.
4. Find the p-value for the observed data.
5. State a conclusion based on comparing the significance level to the p-value.

A couple of things to keep in mind about hypothesis testing when it comes to using them in real life work. There is a difference between practical significance and statistical significance. Statistical significance, such as we do in hypothesis testing, only shows that the observed effect is due to random sampling or not. This does not mean that it is practically important. With a large enough sample size, significance can be reached even for the tiniest effect. Only field experts can determine whether an observed effect is practically significant. There is no rule of thumb or general consensus on how big an effect size needs to be for it to be practically significant. This is because effect size significance will vary across fields. Always think about the context of the problem that you are working in to gain insight on effect sizes. Effects could be too small to be relevant.

8.3 Chapter 8 Problems

1) A simple random sample of 10 people from a certain population has a mean age of 27. Can we conclude that the mean age of the population is not 30? The variance is 20. Test at the 0.05 level. Is this a one or two tail test?
2) For the same set up as above, can we conclude that the mean age of the population is less than 30? Is this a one or a two tailed test?
3) An office building uses light bulbs that have a mean life of 900 hours. A new manufacturer claims that his light bulbs last longer than 900 hours. They ask you to conduct a test to determine the validity of their claim. You take a sample of 64 bulbs and calculate a mean of 920 hours with a standard deviation of 80 hours. Test this at .05 level.
4) A restaurant tells its customers that the average cost of a dinner there is $52 with a standard deviation of $4.50. A group of concerned customers thinks that the average cost is higher. In order to test the restaurants claim, 100 customers purchase a dinner at the store and find the mean price is $52.80. Perform a hypothesis test at .05 significance level and state the decision.
5) Perform the same test as above, but assume only 15 customers purchased dinner. All the other parameters are the same. How does this affect the results of the test?
6) A sample of 40 sales receipts from an electronics store has mean $137 and a standard deviation of $30.2. Use the value to test whether or not the mean sales at the electronics store are different from $150. Test at the .01 level.
7) If you get a p-value of .001, do you reject the null hypothesis at the 0.05 level? The 0.01 level?
8) You believe that a population mean is equal to 6.2. To test this belief, you carry out a hypothesis test at the .05 level based on a sample of data that you collected:
   a) Mean = 5.9
   b) Standard deviation is 4.1
   c) Sample size =42

9 Correlation and Regression

In this chapter, we move away from the one sample tests that we did in the previous chapters and take a look at some different ways that we study the relationship between two variables. Like what we did in chapter 3, we can study the relationship between two variables using graphical and quantitative techniques. Statistical studies involving more than one variable attempt to answer questions such as:

- Do my running shoes have an impact on the speed of my 10K?
- If I drink a protein shake can I lift more weight?
- If I spend more money on tutoring will my test scores improve?
- How much will my test scores if I spend more money on tutoring?

The most basic way to study relationships between two variables is to use a graph and try to look at the overall pattern and study the deviations from the pattern. A scatterplot is one of the most effective ways of studying the relationship between the graphs. It shows the relationship between two quantitative variables measured on the same individuals. Recall the idea of the dependent and independent variables from chapter 3.

Definition

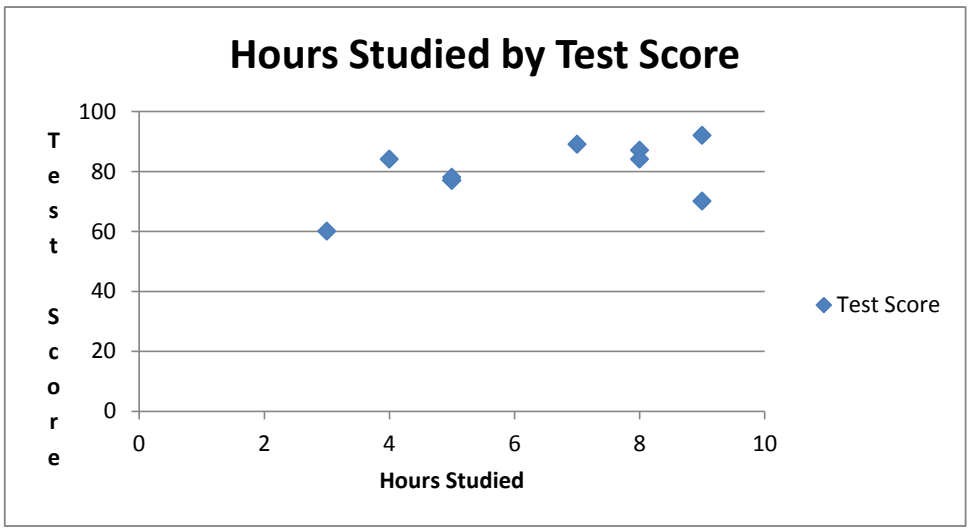
Dependent Variable: The variable which represents the effect that is being tested

Independent Variable: The variable that represents the inputs to the dependent variable, or the variable that can be manipulated to see if they are the cause.

9.1 Scatterplots

In a scatterplot the independent variable is plotted on the x-axis and the dependent variable is plotted on the y axis. Every individual in the data appears as a point in the plot. Take the most basic example of height and weight. Below is a dataset that contains the hours spent studying and test scores ten individuals. We are interested in the relationship between the two variables. How is one affected by changes in the other?

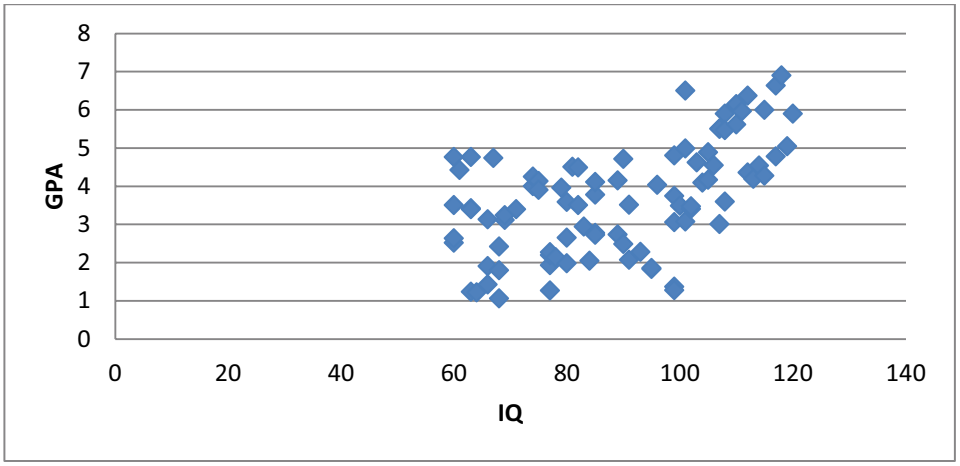
Student	Hours Studying	Test Score
Jan	9	70
Fred	7	89
James	4	84
William	5	77
Penny	5	77
Pricsilla	8	87
Carrie	8	84
Scott	9	92
Thomas	5	78
Benjamin	3	60



Scatterplot of Hours Spent Studying by Score

By looking at the graph, you can see that there is a slight upward trend in the data. That is, those that study more tend to do better. Scatterplots are useful for identifying general trends in the data as well as identifying points that do not follow the patterns as those data points are always interesting too. For instance, what might one say about Jan, who studied for 9 hours but only received a 70 on the exam? That would be a data point worth further exploring.

Scatterplots are usually studied on three dimensions, the form, direction, and the strength of the association. For instance is the form linear or curved? Or is there no pattern at all? In the above scatterplot, there is a slight linear pattern. The direction of the plot is also important. Is it positive, negative or is there no direction at all? Our plot above shows a slight positive direction. Lastly, the scatterplot measures the strength of the relationship. How closely do the points fit the "form". Let us take a different example and identify the form, direction, and strength of the relationship between two other measures: IQ and grade point average.



The form of the relationship is linear for the most part. There are no curves in the plot and there is a clear pattern. The direction is positive; as IQ scores go up, the GPA tends to rise as well. The strength of the relationship is quite strong.

9.2 Correlation

Quite strong however is a qualitative term that we have selected to describe the strength of the relationship based on our judgment. What would be ideal is to have a statistic, or some quantitative measure, to measure the strength of the relationship. The correlation coefficient, denoted as "r" is the most widely accepted measure of association between two variables.

Definition

Correlation: A measure of the direction and strength of a linear relationship.

Before we get into the formula for calculating the correlation, there are several facts about correlation worth pointing out:

- The correlation measures the strength and direction of a linear relationship between two quantitative variables.
- The correlation is always between -1 and +1.
- The negative value of the correlation means association between the variables is negative; similarly positive values of r means the association between the two variables is positive.

The Excel formula for calculating the correlation is

=correl (x,y)

where x, y are the arrays of the data you wish to calculate the correlation on. Since the correlation does not concern itself with the dependent and independent variables, you do not need to make this distinction in Excel. For our above example the correlation is 0.57. The relationship is positive (as one value grows, the other value grows as well) and moderately strong, as indicated by the correlation.

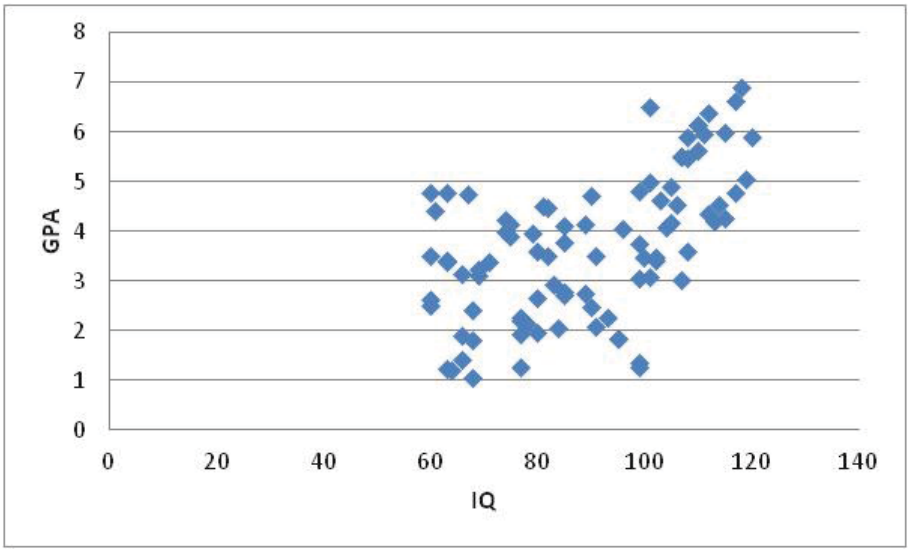
9.3 Simple Linear Regression

While correlation tells us about the strength and direction of the linear relationship between two variables we often would like to understand of how both variables vary together. For instance, is one variable increasing faster than the other? What is the rate of decrease of one variable if the other is increasing? Linear regression is one way that we can quantify this relationship.

A regression line is a straight line that describes how a dependent variable changes with respect to an independent variable. We use this line to do two things. The first, which we just mentioned, is to explain the change in a dependent variable in terms of an independent variable. We also use a regression line to predict the value of the independent variable for a given dependent variable.

That being said, it is clear that the distinction between the independent variable and the dependent variable is very important. In correlation this distinction was not important since correlation measured the relationship between two variables. Since in regression we measure how a dependent variable changes as an independent variable changes, making this distinction is key. For example, we found that the correlation between IQ and GPA was moderately strong. The correlation told us that as one measure increased, the other measure increased as well. However, this does not tell us anything about how the GPA increases (the dependent variable) as a person's IQ increases (the independent variable). Regression can do this for is. It might tell us that a one unit increase in IQ can lead to a 0.5 increase in GPA.

What is the intuition behind the regression line? To illustrate this, let's revisit our IQ vs. GPA scatterplot from our study of correlation.



The scatterplot clearly shows the moderate relationship between IQ and GPA but it clearly it is not a perfect relationship. There are some high IQ's with low GPAs and vice versa. Linear regression attempts to quantify the relationship. Linear regression fits a straight line through the data in such a way that the line minimizes the sum of the vertical distances between the points of the data set and the fitted line. Clearly, several lines could be fit to the data. The best line is the one that minimizes these distances. The least squares regression line is the line

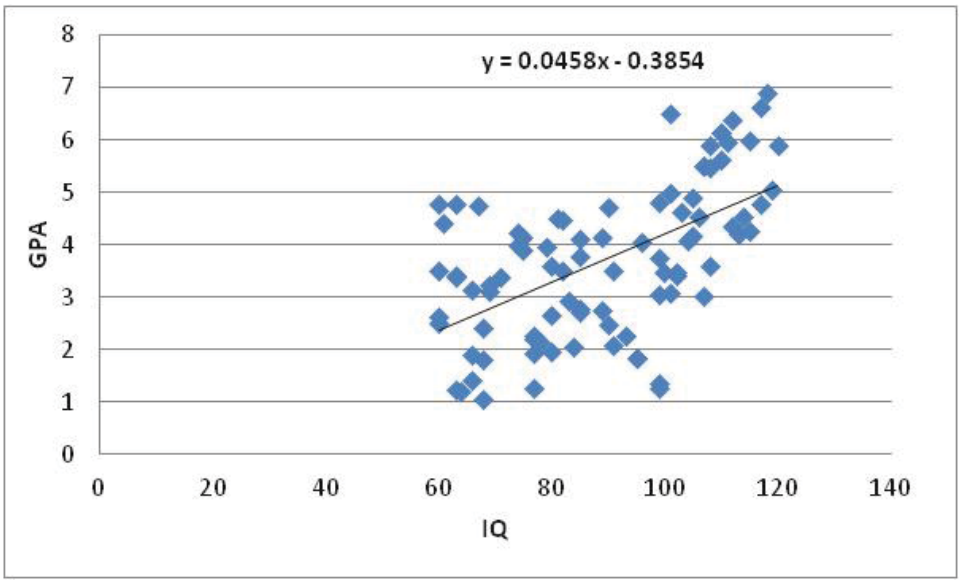
$$\hat{y} = b_0 + b_1 x$$

where y-hat is the predicted response for any x

b_o is the intercept of the line

b_1 is the slope of the line.

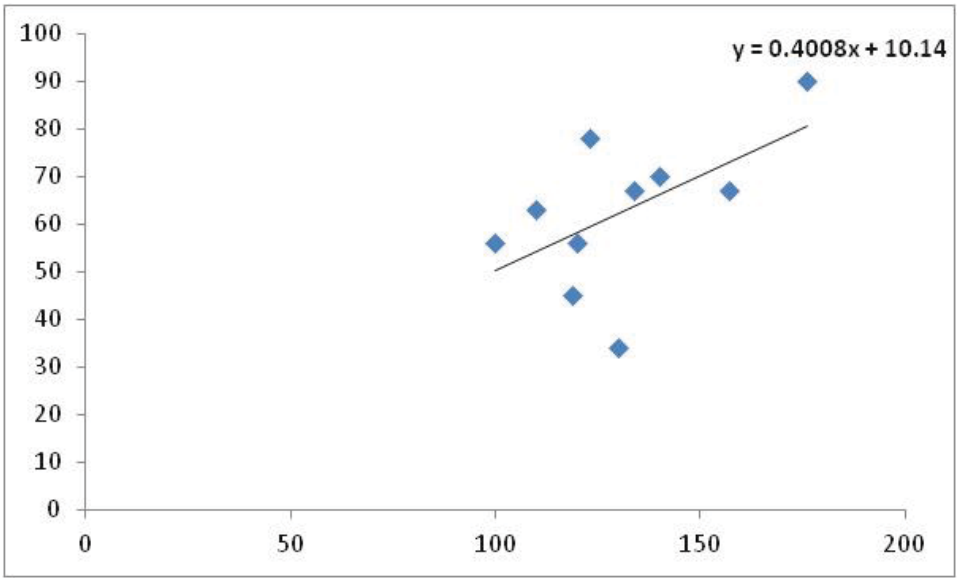
The calculations for the line can be done using a software package. The key to regression is proper interpretation of the slope and the intercept of the line. Lets revisit our IQ vs. GPA example, this time with the fitted regression line. The line is in the upper right hand side of the graph.



The line is y = 0.0458x – 0.3854. The slope is 0.0458. The slope of a regression line tells us that for every one unit change in the independent variable leads to the dependent variable to change by however much the slope value is. So in this example, a unit increase in IQ leads to a 0.048 increase in GPA. The interpretation of the intercept is what the value of the dependent variable would be if there was no effect of the independent variable. In this case, the interpretation of the intercept does not really make sense as it is impossible to have an IQ of 0 (or a negative GPA).

Example

Below is a scatterplot looking at the relationship between advertising dollars and sales of a product. The regression is in the top right hand corner.



The regression for this equation is y=0.4008x + 10.14. Thus, our interpretation of the slope is that for every dollar spent on advertising our sales will increase by 0.40 dollars (or 40 cents). The intercept interpretation is that if we spent no money on advertising, we would have made $10.14 in sales.

Regression lines can also predict values for the independent variable that are not in the data set. To do this, plug the value of interest where the "x" is in the independent variable. In the above example if we wanted to know how much sales we would have if we spent $100 on advertising, the equation would look like:

$$y=0.4008(\$100)+10.14$$

$$= \$50.22$$

Not a great return on our advertising dollars.

To calculate the regression coefficients in Excel, you need to use two formulas: one for the slope and one for the intercept. The formulas are below:

=slope(range of y, range of x)

=intercept(range of y, range of x)

Example

The regression equation for selling price(dependent variable) vs. square footage of homes (independent variable) is y = 4795 + 92.8x.

Q: What is the interpretation for the slope of the line?

A: For every one unit increase in square footage, home price increases by $92.8.

Q: Is the intercept meaningful?

A: No. There is no such thing as a house with 0 square footage.

Q: If the price of square footage of a house is 2500 what do we predict the as the selling price?

A: $236,795

9.4 Chapter 9 Problems

1) You are given the following data set:

Y	X
38	4
42	3
29	11
31	5
28	9
15	6
24	14

   a) Compute the correlation between these two variables
   b) Fit the linear regression line for these two variables. (where Y is the dependent variable and X is the independent variable)
   c) Interpret the slope of the linear regression line.
   d) Interpret the intercept for the linear regression line.
   e) What would be the predicted value for Y if X is 10.

2) Articulate the difference between correlation and simple linear regression.
3) Interpret the following correlation coefficient in terms of strength and direction: 0.843
4) You fit a regression line to data that measures the a dependent variable weight, on an independent variable, calories eaten. The line is as follows: y=145+.005x.
   a) Interpret the slope of the line
   b) Interpret the intercept of the line. Does the intercept interpretation make sense in this problem?

10 Endnotes

1. The Atkins Diet is a strict diet that eliminates carbohydrates from the diet.
2. www.data.gov is a website that consolidates several different government sponsored data sources.
3. A person's weight could be any value within a range of human weights, not just fixed weights.