

# Advanced Probability and Statistics Teacher's Edition (Being Reviewed)



# Advanced Probability and Statistics Teacher's Edition

---

Jared Prolo, (JaredP)  
Teresa Zwack, (TeresaZ)

**Say Thanks to the Authors**

Click <http://www.ck12.org/saythanks>

*(No sign in required)*



To access a customizable version of this book, as well as other interactive content, visit [www.ck12.org](http://www.ck12.org)

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the **FlexBook®**, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the **FlexBook Platform®**.

Copyright © 2012 CK-12 Foundation, [www.ck12.org](http://www.ck12.org)

The names “CK-12” and “CK12” and associated logos and the terms “**FlexBook®**”, and “**FlexBook Platform®**”, (collectively “CK-12 Marks”) are trademarks and service marks of CK-12 Foundation and are protected by federal, state and international laws.

Any form of reproduction of this book in any format or medium, in whole or in sections must include the referral attribution link <http://www.ck12.org/saythanks> (placed in a visible location) in addition to the following terms.

Except as otherwise noted, all CK-12 Content (including CK-12 Curriculum Material) is made available to Users in accordance with the Creative Commons Attribution/Non-Commercial/Share Alike 3.0 Unported (CC-by-NC-SA) License (<http://creativecommons.org/licenses/by-nc-sa/3.0/>), as amended and updated by Creative Commons from time to time (the “CC License”), which is incorporated herein by this reference.

Complete terms can be found at <http://www.ck12.org/terms>.

Printed: April 15, 2012



## AUTHORS

Jared Prolo, (JaredP)

Teresa Zwack, (TeresaZ)

<b>1</b>	<b>Probability and Statistics TE - Teaching Tips</b>	<b>1</b>
1.1	An Introduction to Analyzing Statistical Data . . . . .	2
1.2	Visualizations of Data . . . . .	5
1.3	An Introduction to Probability . . . . .	7
1.4	Discrete Probability Distributions . . . . .	10
1.5	Normal Distribution . . . . .	12
1.6	Planning and Conducting an Experiment or Study . . . . .	14
1.7	Sampling Distributions and Estimations . . . . .	16
1.8	Hypothesis Testing . . . . .	19
1.9	Regression and Correlation . . . . .	21
1.10	Chi-Square . . . . .	23
1.11	Analysis of Variance and the F-Distribution . . . . .	24
1.12	Non-Parametric Statistics . . . . .	26
<b>2</b>	<b>Probability and Statistics TE - Enrichment</b>	<b>28</b>
2.1	An Introduction to Analyzing Statistical Data . . . . .	29
2.2	Visualizations of Data . . . . .	33
2.3	An Introduction to Probability . . . . .	36
2.4	Discrete Probability Distributions . . . . .	41
2.5	Normal Distribution . . . . .	45
2.6	Planning and Conducting an Experiment or Study . . . . .	47
2.7	Sampling Distributions and Estimations . . . . .	50
2.8	Hypothesis Testing . . . . .	54
2.9	Regression and Correlation . . . . .	57
2.10	Chi-Square . . . . .	60
2.11	Analysis of Variance and the F-Distribution . . . . .	62
2.12	Non-Parametric Statistics . . . . .	64

## CHAPTER

**1**

# Probability and Statistics TE - Teaching Tips

## CHAPTER OUTLINE

---

- 1.1 AN INTRODUCTION TO ANALYZING STATISTICAL DATA**
  - 1.2 VISUALIZATIONS OF DATA**
  - 1.3 AN INTRODUCTION TO PROBABILITY**
  - 1.4 DISCRETE PROBABILITY DISTRIBUTIONS**
  - 1.5 NORMAL DISTRIBUTION**
  - 1.6 PLANNING AND CONDUCTING AN EXPERIMENT OR STUDY**
  - 1.7 SAMPLING DISTRIBUTIONS AND ESTIMATIONS**
  - 1.8 HYPOTHESIS TESTING**
  - 1.9 REGRESSION AND CORRELATION**
  - 1.10 CHI-SQUARE**
  - 1.11 ANALYSIS OF VARIANCE AND THE F-DISTRIBUTION**
  - 1.12 NON-PARAMETRIC STATISTICS**
-

---

# 1.1 An Introduction to Analyzing Statistical Data

---

## Definitions of Statistical Terminology

This Probability and Statistics Teaching Tips FlexBook is one of seven Teacher's Edition FlexBooks that accompany the CK-12 Foundation's Probability and Statistics Student Edition.

To receive information regarding upcoming FlexBooks or to receive the available Assessment and Solution Key FlexBooks for this program please write to us at [teacher-requests@ck12.org](mailto:teacher-requests@ck12.org).

Probability and Statistics is by far the most text based subject in mathematics. Students new to statistics get a real baptism by total immersion with the early sections as we very quickly attempt to bring them up to speed on all of the terminology that will be used. Because of the significant difference between traditional math classes and stats, some care is going to be needed to teach how to read the textbook, as well as how to read problems.

If your school has a method of taking notes, you should first consider using that method, if you aren't already. If your school doesn't have a prescribed method, then it's worth checking in with the humanities and possibly the science teachers to find out if your students are being asked to use, or had been asked to use, a particular method then that would also be a good choice. If your students have not been asked to learn or use a note taking method, it will be worth your time to teach one, and enforce the use of it. Likely there will be some resistance; AP stats students are likely to be very skilled and resistant to note taking systems. However, even very skilled students can become bogged down and lost quickly with how much language and vocabulary is needed for understanding.

My favorite method is Cornell Notes, or a slightly modified version thereof. Many of my past students have hated it at first, but began to believe in the value when they were able to reference and study from a concise source when the topics became more challenging. I don't always have my students attempt to recall entire lectures or pages of reading from their notes, but the recording of key vocabulary and topics and then the summary written at a later time are very important. A PDF from the university of Cornell can be found at: [http://lsc.sas.cornell.edu/Sidebars/Study\\_Skills\\_Resources/cornellsystem.pdf](http://lsc.sas.cornell.edu/Sidebars/Study_Skills_Resources/cornellsystem.pdf), as well as many other sources. Other note taking systems can be found in various academic literacy guides and texts.

Another key is being very careful about the language used in class, and making sure that words that have special meaning are reserved. As mathematicians we frequently are sloppy with our language, letting the context define if we mean the special meaning of the word or a more general meaning. One big one is variable. With continuous, discrete, quantitative, qualitative, random and all kinds of variables being very important in statistics, students can get confused quickly, especially as the treatment of these variables is slightly different than it may have been in the past.

---

## An Overview of Data

With the exception of a few theoretical situations, all of statistics is based around data. However, as students will see later, the importance of clear appropriately gathered data can't be overstated. The first step is seemingly always designing the method of gathering the data. Most of the topics presented here are going to be examined in depth in later sections, so there is no need to spend lots of time teaching about the specifics of each at this time.

### 1.1. AN INTRODUCTION TO ANALYZING STATISTICAL DATA



This is a good time to look at some of the terms and relate them to students' experiences. Especially in contemporary times, studies and experiments are a large part of the news and popular culture. Have students find news articles citing studies and bring them in for discussion. It will especially be powerful if students choose items that have some pertinence to their lives. Studies about school violence, teen health, standardized testing and similar studies will have more meaning for students than ones on heart disease, home prices and other favorites of the media. Current events can be a huge part of the stats classroom, and it helps to make it a rich and memorable class. News articles are wonderful resources because they usually include incomplete or incorrect information when it comes to the math involved. Students can be led in a discussion of each study, what they have included, what they have omitted, and what the omitted details likely are. The examination of sample studies in the text provide the template for such examinations.

---

## Measures of Center

The measures of center are frequently the only statistics that a vast majority of the population will ever use. However, the treatment of these in previous classes is very incomplete, often beginning and ending with mean, median and mode. This is a great place to begin with students, as the task of a first year statistics teacher is frequently to show how common perceptions are not always meaningful in statistics. The association with the "average" being the mean is easily discredited, as is discussed in the text. A fun exercise is to look at alumni lists for either a high school or a university. If there is a super-star athlete, or a top executive on that list, the mean and median income of graduates will be quite different, and outlines clearly why it is sometimes advantageous to use poor choices in statistics to promote a particular idea, in this case saying that the "average" alumnus of a school can expect to earn well above what is realistic.

There are many more measures of center presented and each will have its appropriate place. The text give a brief nod to "It depends" as the mantra of statisticians, but at some point students are likely to struggle with the apparently arbitrary nature of statistics. The purpose of statistics is rarely to nail down truths (although with careful practice, stats can yield surprisingly close results), but to inform and give a clear picture for trends when the data set is frequently too complex to use directly. The different choices about the mean, weighted means and trimmed means shows that statisticians have choices, and there is frequently no clear direction on which to use and that is OK. In the early stages students need to stay calm and just roll with it, as much of the confusion is cleared up with practice and experience.

One key topic here is the difference between the Population Mean and Sample Mean. There isn't much more to say about it at this point, but when we get into continuous distributions the difference between the two becomes really important. Make sure students are aware of the difference and are careful about using the correct label and terms for each.

---

## Measures of Spread

The second of the two major base topics for stats is Spread (the other being Center as discussed in the previous section). These two topics will return over and over again as we begin to look at different distributions. As such, these sections should not be glossed over quickly. Trying to go back and understand where variance comes from is far more difficult when you are also trying to learn about the motivation behind continuous distributions, like the normal distribution.

At the top of page 32 there is a little mention of an important idea:

Even though we are doing easy calculations, statistics is never about meaningless arithmetic and you should always be thinking about what a particular statistical measure means in the real context of the data.

Highlight it! Make a poster of it! Frequently AP stats is derided as not being as tough, or as mathematically intensive as it's brethren AP Calculus. I am not sure where exactly this prejudice comes from, but I suspect it has a lot to do with the fact that very, very few preliminary skills from algebra and geometry are needed for success in statistics, while Calculus will expose every hole in one's high school math experience. This does not, however, directly relate to an easier time in a mathematical sense. Statistics requires a level of attentiveness that other classes do not. There are plenty of "cookie-cutter" problems in calculus that once the correct method is chosen the following steps follow a reliable pattern and it is only a matter of following the algorithm that has been used a hundred times before. Statistics requires a new level of understanding for the nature of the question, the process and then the results. The AP exam rewards such understanding. The free response sections are graded specifically to reward conceptual understanding and deemphasize rote algorithmic procedures. IF a student were to make a small arithmetic mistake, get an incorrect value for the standard deviation, and then interpret the value correctly, the penalty is minimal. However, if a student were to make a small arithmetic mistake and get a probability of 1.17 and the student left that answer as correct the penalty is substantial. The arithmetic mistake is not so bad, but not understanding that a probability greater than 1 necessarily shows an error belies a fundamental lack of understanding of statistics.

You may notice that both this guide, and the text, are constantly mentioning "but later". This is bad form, in my opinion, but speaks to the ease of overlooking details now with the stiff penalty of how complex topics are later. Sometimes having a former student speak at the start of a school year is helpful, not only for alerting students to challenges ahead, but also for time management, formatting of the test and other topics useful to be successful in an advanced placement class. Students will likely get tired of hearing it all from the teacher, but by the time they realize they should have been paying closer attention, it will be too late. I also tend to be ruthless with grading at this point, creating the expectation of extreme attention to detail. Later on I will relax a bit.



---

## 1.2 Visualizations of Data

---

### Histograms and Frequency Distributions

Statistics classes are dangerous places for students who try to outsmart math problems. You probably know the type, the kid who will always bring up some fact of why a particular study is nonsense, or provides data, or worse anecdotes, that put the class discussion off track. For example, the bottled water discussion will likely have students bringing up all kinds of strange numbers, reasoning and justification or bottled water use, or lack thereof. This is also a regional aspect to this, as my former students living in unincorporated locations of the Santa Cruz mountains with untreated well water, as well as my former students from New Orleans where the Mississippi River is the source, have different considerations than my former students in San Francisco (which pipes in amazing water from Yosemite National Park). Decisions have to be made as an instructor as to how much “digression” is going to be allowed. Sometimes there are teachable moments regarding using the data presented, not bringing in personal bias, or sometimes students have valid questions about how data is collected. Students, especially those who have had success in school, really want to participate. Often in the early chapters they don’t have much to contribute to problems as it is a new subject for them. Simply dismissing students’ contributions can leave some students with a sour attitude, but certainly not everyone’s slightly off topic contributions can be entertained. I tend to try to listen, and then if the comment is not productive to my lesson, I try to let the class know what I would like to hear, and what isn’t going to help us with the ultimate goal. This becomes less of a problem as the year progresses, students learn more stats, and the topics are tougher.

Most of these topics are going to be review from previous classes. I wouldn’t spend a ton of time in class, but would use these as warm-ups. Give students a topic, like hours of TV watched daily, and have them collect the data and chart it in the first 5 – 10 min of class. Making clear and accurate graphical representations is a skill to be practiced without a ton of content. It is also OK to be critical of “artistic” skills here, as presenting data for use by others requires clear charts. Making the histograms and other charts look good is part of the job.

---

### Common Graphs and Data Plots

This is again a chapter to review, practice and make sure that everyone is on the same page. I would make the decision ranking the most important representations of data. This is subject to some debate. I deemphasize pie charts and stem and leaf plots. It is true that the general public loves pie charts, and while I find them easy to read, they are of limited use for extended work. Stem and leaf plots are really difficult to read, and in many cases tough to teach. The idea of using different place values to categorize and split the numbers up seems to confuse students, and is tough to read as an end user of the chart. Scatterplots for bivariate data and dot plots are more useful, especially for later units. I make sure my students are comfortable and proficient with making these charts.

---

### Box and Whisker Plot

Box and whisker plots are great, especially for comparing two or more sets of data in a graphical manner. Students will have the most success with creating these graphs by following a somewhat algorithmic process. The steps

outlined in the text is exactly how I work these problems, and would teach and stress the same. The biggest problem that I observe students having is attempting to draw the graph and then add numbers, or not using an even scale for the graph. The best part about the box and whisker plot is the clear representations of center and spread. If an even scale is not used then no quality information can be pulled from the graph. I stress that the number line must be drawn and labeled before any part of the box and whisker graph is drawn. This is even more important when trying to compare two sets of data.

The idea of the middle 50% is an important one. It is worth spending some focused time on, as not only is it useful for finding outliers, but is also an important statistic on its own. One thing students may be unsure of is if the 5 number summary is re-calculated after outliers are “thrown out”. The answer is no, as the summary is resistant to outliers. The changes would not be more descriptive; the graph is the only thing changed for clarity’s sake. For this reason, and some others, I never say “throw out the outliers” as it implies that they aren’t an important part of the data set. Outliers are still important, and have to be treated carefully rather than simply discarded. This is most evident, for example, in analysis of data with an important, but strangely spread data, such as air pollutants by California counties. Los Angeles county is going to be way far out there, but you can’t accurately represent the climate situation in the state without considering Los Angeles. List it, but it is most helpful to show it as a point outside of the whisker, because it’s important to show how far away that single county is, not because it’s not important.

## 1.3 An Introduction to Probability

### Events, Sample Spaces and Probability

Finding the sample space is the key for student success in computing probabilities. Frequently problems arise when possible outcomes are missed, or double counted, or otherwise. That is to say, there are lots of less than intuitive possibilities so tools to support students finding correct answers are helpful. One of those tools is listed in the text, where a table is made with the different outcomes. Lists are also good, especially for sample spaces that only involve one item, like a single die. There is also no shame in drawing pictures or working with manipulative. I frequently visualize a deck of cards in my mind when working with those probabilities, but as a bridge player I'm comfortable with cards, while a student might not be. Don't hesitate to give students cards to work with.

I encourage my students to never work in percentages. I realize that it's kind of picky, but I find it useful to drive home that probabilities must be between zero and one inclusive. The strong restriction is very useful as it frequently provides feedback for mistakes. If you are computing any probability and it ends up outside of that interval, then something is wrong. It's also a great tool to use to eliminate answers on a multiple choice test. I like to have a mantra to really drive it home. Between 0 and 1 and all possibilities add to 1.

### Compound Events

Elementary set operations are a fundamental part of mathematics, but one that is taught in inconsistent places. I've always marveled at the fact that nearly all of my college textbooks, including my graduate level texts, start with a preliminary chapter on set theory. This has always led me to believe that set theory is either not taught, or is given an incomplete treatment in a lot of classes. There really isn't any reason to give a deep treatment to it here. Sometimes sticking too much to the strict notation is going to cause more problems than it's worth, since the ideas here are intuitive and students have worked with them in venn diagrams and other problems in the past.

An important thing to remember is that set operations are binary operators. That is, even if there are more symbols, only two sets can be operated on at a time. Due to the fact that the combination of unions and intersections is not associative (that is to say:  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ ) I always include parenthesis if there is more than one operator, even if they are all the same operation, just to avoid confusion. This is, and matrices, are some of the first structures students encounter that do not follow all the classic rules of real numbers that students are used to. It can be an opportunity to push a talented class, but it is really an extra topic that has limited utility to the ultimate goal for a stats class.

### The Complement of an Event

The classic example of value here is the birthday problem: In a given room, what is the probability that at least a pair of people have the same birthday? This is a great problem, as it shows how working smarter using some principles of probability makes a seemingly tough problem easy. It also has a result that is fairly counter intuitive; the probability of at least a match is much higher than one would presume. Both are key ideas to drive across to students studying

probability for the first time. Asking for “at least one pair” means that if you were to directly calculate the probability it would take a very, very long time. There are just too many ways to get a match once the number of people in the room exceeds 4 or 5. However, asking the question “what is the probability that no two people share the same birthday” is logically equivalent to the first question and much easier to compute. Subtracting this quantity from 1 (finding the complement) then yields the answer.

The key here might be in re-writing the question in terms of the complement. Students probably don’t see what the big deal is in calculating complements, and that is really simple. However, making the complement work for you requires seeing where to apply it, and then what exactly you are looking for. The key hints are when a question is asking for a probability where multiple situations are possible. In the birthday problem, asking for exactly a single pair should be calculated directly, but at least a pair dictates that the complement is easier. Students should practice identifying and re-writing questions to make it so that time isn’t wasted attempting monumental calculations.

---

## Conditional Probability

Students are going to get tripped up with the order of conditional probabilities. The more intuitive way of thinking of conditional situations is “if then” as opposed to “given”. While I am nearly always in favor of understanding the concept and avoiding formulae, in this case the formula is great. Because the order is a little strange, and frequently is mixed up, this is one of the few formulae that I put on a poster and ask students to commit to memory. Application is easy once the spaces are put in their proper place.

---

## Additive and Multiplicative Rules

This is probably student’s first exposure to the dreaded problem of double counting. In this case, in contrast to the previous section, I don’t emphasize the formula here. Finding values or probabilities that are double counted is a huge part of statistics, and one that most stats students have memories of sitting in a group, having problems with getting the correct answer, pulling hairs out, only to then have someone say “double counting!” For this reason, I really try hard to get my students to understand where double counting occurs and try to train them to always be aware of where the error is likely to occur.

---

## Basic Counting Rules

Combinatorics are the foundation of lots of basic probability questions. Cards are a great way to do look at problems, especially with the recent attention paid to poker. The television broadcasts will show probabilities for winning in “real time”. This can be used to practice and find those same percentages, which are relatively simple to compute. The foundation of each probability is finding the different cards that will allow for a win, against the total number of possible cards left. This is a nice extension of finding general probabilities for each type of poker hand.

Whenever combinatorics and probability comes up gambling is not far behind. This can cause problems in some cases considering the ethics of teaching typical gambling games in a classroom to students who are not legally able to wager bets. A couple of thoughts on the subject, and a general defense. First, it may be useful to make the distinction that while games are being taught and talked about, at no point is gambling going to occur. This is similar to a health class where effects of drugs are being discussed, but clearly no endorsement, or use, of drugs is happening. There is also historical context, as the earliest theories and work on probability was, in fact, motivated by gambling. The legacy remains, even in situations where gambling is no longer associated. For example, a hand with no high card points in bridge is called a Yarborough, named in honor of a lord who would offer his opponents

### 1.3. AN INTRODUCTION TO PROBABILITY

a 1000 : 1 payout if no points were dealt (odds of getting a yarborough are 1827 : 1, so the Earl made quite a profit on this). Second, I would promote the idea that gambling institutions profit from a lack of knowledge of probability. Like the Earl of Yarborough, the idea of casinos is to present a situation that looks favorable to the gambler where in actuality the advantage is firmly in the direction of the house. Knowing exactly how much of an advantage is disheartening to a gambler, and in many cases will cause a loss of interest in gaming.

While it is useful to compute a couple of combinations and permutations by hand to get a sense for how they work, I quickly pull out the calculators. In fact permutations and combinations might be the functions I use most often on the calculator right behind the trig functions. There is no benefit to spending any extra time with the tedium of not using the calculator functions.

## 1.4 Discrete Probability Distributions

### Two Types of Random Variables

Random variables are sometimes hard for students to understand the nature of the random variable. Because it is a new idea, this may take lots of time to make sure students understand what exactly is going on. I have found that discrete random variables are slightly easier than continuous ones. The problem is that the random variable can take on a number of unknown values, which also tend to be represented as variables. Random variables can be thought of as bins, that hold any number of other variables. This is helpful when talking about distributions later on.

### Mean and Standard Deviation of Discrete Random Variables

This is where you need to connect back to the early lesson when we were pretty strict about calling the mean and the standard deviation parameters. When talking about distributions of a random variable, the most important thing is the mean and the standard deviation, and once the method of finding each is established most texts will then give a table of the common ones.

Up to this point there has been lots of theory. This section begins the statistical portion in earnest, and deserves some extra time for practice in finding means and variances. This is especially important before we start to define special distributions where means and variances will frequently be calculated by the rule for each individual distributions.

### The Binomial Distribution

This the big one for discrete distributions. The formula is one that needs to be memorize, as so many problems come down to the binomial distribution. The idea of a probability of pass or fail applies in many circumstances. One place that can be fun to look at for students (especially those who are sick of medical studies...) is to look at some of the probabilities that are published about sports. For example, currently Accuscore is famous for running many simulations of each game a team plays to come up with probabilities that each team will make it to the playoffs. Presented in popular media, the mechanics of what goes on behind the scenes is unknown. (Gambling disclaimer! Accuscore attempts to profit off of selling their information to people who use it to place bets. They technically run a monte carlo program with using a binomial distribution as the probability distribution function (my assumption, but I'd be really surprised if I'm wrong.) They try to bamboozle people out of their money with fancy language and guarantees, another reason knowledge is power.) Students are more than capable of running their own simulations, for sporting events or other types of contests. The trick that students will probably discover is that the binomial distribution is great, but there is a huge catch to it. How to find  $p$ .

I try to get my students to come up with the questions on their own. Every question they are likely to encounter has something along the lines of “given that  $p$  is .6...”, so when students are given a more open ended problem, hopefully they will ask how to find the probability of success. They can be guided to begin thinking about it; asking “Does this answer make sense? How do you think they chose  $p$ ?” after solving some of the books problems. It isn't always easy, or clear, how to choose success. Sometimes it is determined by a specific probability space, but for



human studies, it can be a bit of a guessing game. It is an important consideration as students begin to learn about the job of statisticians and what is directly computed and what has to be assumed or chosen by the person doing the study.

---

## The Poisson Probability Distribution

The Poisson distribution, also sometimes called the exponential distribution, is very useful in practice. It is used frequently for risk management scenarios and other applied questions in physics and business. The topic will also be a good one to revisit when confidence intervals are studied, as the two frequently go hand in hand.

However, it is not a topic on the AP examination. In checking a number of textbooks for a first year probability and stats course at the university level, it seldom made an appearance in those as well. Therefore, how you use this section is dependent on what the goals of your class are, how skilled they are, and how badly you may need to speed things along before the examination. Again, it is a great topic, but may not fit with your plans for this course.

---

## The Geometric Probability Distribution

The geometric distribution has a very particular use. That is, how long before a single event will occur. The troubles students are likely to encounter are the geometric distribution's similarity to the binomial distribution and the circumstances that it is applied.

The forms of the two distributions simply must be memorized. There unfortunately no way around it. Something to focus on is what the variable in each distribution represents. In the binomial distribution the variable is how many successes out of a fixed total. The geometric distribution's variable is how many trials. Now if students are really clued in conceptually they will realize that the two can be confused because they are essentially the same. Look at a binomial probability of one success:

$$p(x) = np^1(1-p)^{n-1}$$

The only difference is the  $n$  multiplied at the front. This is because the binomial distribution does not care when you have the success. The Geometric distribution only give the probability for a success at a particular point, i.e. the probability of 1 success in 5 trials as opposed to having the first success on the 5<sup>th</sup> trial.

Because of this choosing when to apply the geometric distribution is sometimes tricky. The thing to focus on is "Before the first success" or "until first success". The number of successes is fixed in the geometric distribution, and that can be the clue pulled from the problem to choose correctly.

## 1.5 Normal Distribution

### The Standard Normal Probability Distribution

The big one. The book does an excellent job of introducing each key part of the normal distribution. I try to keep students away from the equation that describes the distribution for as long as I possibly can. By accessing students prior experiences with things that are normal, and then looking at the key parts of the distribution graph, students will have a better chance of a strong conceptual understanding that can be better applied for a variety of problems. Really, the equation of the distribution is mostly helpful only after calculus.

A key idea that must be stressed is the movement away from discrete distributions and into continuous distributions. This can be tough, because there are many things that we measure in a discrete manner, but model with the continuous curve. SAT scores is one, as there is no possibility of getting half of a question right (well, technically fractional credit is possible with the penalty for errors, but this still does not create a continuous score range) and no one gets half points on the scale after normalizing scores. The key here is the language that I used: we model behavior with the continuous distribution. Students will be well served to remember this little step to avoid thinking that normality requires continuity.

As a personal note, I really get personally confused by  $Z$ -scores. I suspect this is because I had many calculus classes before my first stats class, and therefore used calculus to solve the problems. The only reason why I bring this up is that reading a standard normal chart, finding  $Z$ -scores and relating normal distributions that might not be standard to the standard normal curve is a critical skill for a non-calculus based stats class. For my high school and university classes I would spend significant time on the algorithmic process of finding scores and relating them. It's one of the few times in a probability and stats class where a purely algorithmic skill is needed to be practiced (and practiced and practiced!) Plan extra time just for this skill.

### The Density Curve of the Normal Distribution

A key idea that I talk about when introducing the standard normal curve, empirical rules and the standard deviation is where the inflection point is placed. This helps with students drawing good curves as well. The inflection point is always going to be a single standard deviation from the mean in each direction. This means that the inflection is going to “pull in” or “spread out” proportionally with the rest of the curve. I often have my students practice drawing generic curves with the same mean, and placing the inflection point in the same place. I don't focus on it yet, but I also want them to understand that the areas (under the whole curve, between two inflection points) must remain equal, therefore establishing the connection between the height of the peak and the spread of the graph.

I'd like to take an additional moment to stress the importance of sketching the graph and the area you are inspecting when working problems involving normal probability density. Because the direction of the “tail” changes the way things are looked at, or the additional steps needed to find an area that is in the middle, or split on both ends, the sketch and shading is critical to staying on track. Another thing that I have my students get in the habit of is labeling the area once they find it. For instance, if they are looking at the area between one standard deviation and the mean, they would shade the area and then label it “.34”.

This is one of the key areas where calculators can really help with making things easy and fast. Because the calculator will find the continuous density, regardless of the location of the mean. This is a calculator skill I would make sure

all students are comfortable with.

It would be wise to get a copy of the tables that will be provided in the AP examination. I don't have my students' use the charts in the text, but copies of the tables that they have for the examination.

---

## Applications of the Normal Distribution

The lessons at the start of this section are designed primarily as an intermediate step. Rare is the occasion where those specific types of questions are asked. However, trying to solve questions in context requires both the mechanical understanding of how to find the missing information as well as the contextual understanding of what information is given, what is needed and how to set it all up. Trying to teach both at the same time might be too much all at once, hence these initial exercises. If these lessons are used, then I strongly recommend stopping before the section using real data, and having the students work problems 1-3 at the end of the section. Then after everyone is on the same page for those questions you can move on to the real data.

There are a couple of real data problems included in this section, but I would add a bunch more. Interpreting normal data is a huge part of what students will be asked to do. Fortunately, most of the data that is out there is normal, so finding data sets is pretty easy. I have included an appendix of sources from the internet at the end.

The big task here for students is making sure they know what quantity they are being asked to find. Most of the time this is straightforward because it is directly asked for. However, they might not be comfortable “translating” to the variable names that we have been using, like percentile, z-value, mean and standard deviation. For instance the last example asks “How rare is it that we find a female marine iguana...” This isn't using any term that we have before. Students, through the practice of doing many guided problems, need to develop a lexicon of the different ways that quantities can be asked for. In this case, how rare, what is the chance of, and other similar statements are looking for the percentage under the curve beyond a certain marker, in the case of this example below 400g.

---

## 1.6 Planning and Conducting an Experiment or Study

---

### Surveys and Sampling

This lesson is a bit on the thick side. It is reasonable to break it up into smaller parts and move them around as you see fit. Some items are interesting, but not always productive (like the discourse on randomness).

Something I like to do is find more of the bad sampling practices (or bad conclusions) and present them to students and ask them to figure out what probably went wrong. Some classic examples are the Dewey v. Truman presidential election (complete with incorrect newspaper headline!), why Yao Ming has made the NBA all-star game nearly every year, even when he is hurt and not playing, and others. Frequently these are not so cut and dry that one factor can be singled out, but often many things combined to make a bad survey or poll. This is partially where students begin to understand the cliché “You can make statistics say anything you want...” which is frequently used by the undereducated to discredit well done surveys. The educated student realizes that there is plenty of opportunity for errors, and because of that, and other factors, absolute certainty is impossible, but a well done survey is a very solid source of information. Students need not memorize each type of bias by name. This is only useful for cocktail parties to stress that you know what you are talking about in terms of studies and surveys.

The randomness idea is lots of fun, but maybe beyond the scope of this class. Since we have not yet looked at the Uniform Distribution, which is how most computers generate random numbers (hence the TI calculators always returning a number between 0 and 1), we really don't have the impetus to work with computer generated randomness. An easy way to show the book's point about seeded generators is to reset the calculator. Don't do this if you have lots of programs in memory. A reset all, will set each calculator's seed to the same number, and therefore every TI calculator will return the same “random” number after a reset. Even better, because the seed is incremented the same way for each calculator, the entire sequence of random numbers will all be identical until the seed is changed.

---

### Experimental Design

It may not be critical to the AP examination, but one of the mantras in my classroom, whether it be a stats class or not, appears at the bottom of p235. “Correlation is not causation.” For my students who are not going to pursue a future in math or science this is a critical idea. While any student at the end of a year of probability and stats can understand why, the more people who understand, and can communicate the idea, the more intelligent and informed decisions can be made as a community. It's one of the few chances we have as math instructors to teach for social justice.

Experiment design is one of the hardest things to “master” as a professional. Teachers are very familiar with this and I encourage you to share your experiences with educational research. It's tough for any study to give clear results due to the impossibility of true control, many, many confounding variables and the reality that a classroom will never be a lab, nor a lab a classroom. A task that is usually taken on by most AP stats classes after the exam is to design and carry out a large scale study involving their school or peers. I've had some amazing work done by students, including some really salient studies about drug use by students that was in direct conflict with the findings for our site from the CA Healthy Kids survey. The discussions about why the two findings were different were some of the best moments in any of my classes ever. It's maybe good to have the students start thinking about their projects now,

if you plan to conduct one after the exam. It will help reinforce the ideas here, as well as set students up for success later on. With the way things go, the month and a half after the exam always seems like a ton of time, until you get there and it runs out in a hurry. Preparations must be made so that surveying can begin shortly after the test.

---

## 1.7 Sampling Distributions and Estimations

---

### Sampling Distribution

Sampling distributions are frequently tough for students. Most of what has been studied up to this point are fairly concrete, but now there is a level of abstraction that can be tough to follow. The problem is that we are now talking about sampling a random variable that is itself a function of other random variables. The best way to handle this is to try to present as many different kinds of explanations as possible, using different language. One thing that is nice is that it is very possible to carry out an experiment in the classroom just like the example that is outlined in the test. Another good practice is to make absolutely sure you have been, and continue to be consistent with the notation used, with  $\mu$  and  $\bar{x}$  both representing means, but the population and the sample respectively.

This is the beginning of some of the interesting facts about the normal distribution. There will be theorems later on that states why, but students should have their attention brought to all of the times that the normal distribution appears.

The formulae for sampling error, the mean and the standard deviation are ones that should be added to the “memorize this” list.

---

### The z-score and the Central Limit Theorem

Using the term  $z$ -score is unfortunately necessary. It's always preferable to not introduce new terms when they are not needed. In this case the  $z$ -score is really just the standard deviations away from the mean, with the small exception of negative or positive signs indicating direction below or above respectively. It is a term that is in common use, however, and will be referred to by that name in both the AP examination as well as later courses students may take. The only possible exception being a calculus based stats class for math majors. It's always a good idea to cycle through some problems regarding reading  $z$ -score tables and standardizing values even outside of the chapters that include the topic.

The central limit theorem might be the most important single idea of a first year class. It is critical to know the formulas for sample proportions and sample means, but knowledge of the mechanics of the central limit theorem are not needed. The text does not present a formal proof, nor do most texts for a first year stats class. Because of student's work with various previous problems, students should be familiar with the idea that so many natural occurrences are normal, so a sort of study of a large number of studies should also be normal.

---

### Binomial Distribution and Binomial Experiments

This is the beginning of the set of lessons where a common distribution is discussed, the mean and standard deviation is derived and then practiced. It is a common practice to not determine how to find the mean each time, but rather work from formulae once the distribution is determined. I will give my students a sheet of the common distributions and each of their means and variances. On a timed test it's a major advantage to be able to recall how to find each parameter, and move on with the problem.



Depending on what your technology resources, you may run a computer program that shows an animation of the binomial distribution sampling as  $n$  increases. There are instructions available for programs like *geometers sketchpad* and *fathom*, as well as some java applets, like [\[INSERT LINK\]](#). This can give a nice conceptual sense for how, even if the probability of success is off to the side, the sampling approaches the normal distribution.

---

## Confidence Intervals

Statistical inference is a large part of the *AP* examination. Confidence intervals is probably the key topic for the section on inference. It's also sometimes counter intuitive to students, so careful use of language is required.

I force students to use very precise language here. The correct language is "95% confidence level", not 95% chance, or 95% probability of anything else similar. It can be considered picky, but there is no probability here, and that's important. Especially as students are considering different distributions and samples of different distributions, clarity helps.

It's worth having a discussion with students about what they consider to be an appropriate level of confidence for their studies. Obviously 99% is very strong, but students should be aware of how much more "expensive" that level of confidence is. 90% is very easy, but doesn't make you feel very, well, confident. The common levels of confidence tend to be 90%, 95%, 97% and 99%. Students should work some problems and get a sense for what they feel the "sweet spot" is for keeping the sample size manageable, but gives a high enough confidence level.

---

## Sums and Differences of Independent Random Variables

I can get confusing for students when trying to figure out when to add probabilities and when to multiply. There are a couple of tricks to help students out. First, I always tell them that choosing one and then another, without replacement, is logically identical to choosing two at the same time. In the case of the book's example of miners, the question can be reformatted to choosing one miner, than another. This clearly implies a multiplication of the two individual probabilities to get the probability for the two together. However, if we then ask about the probability of getting at least one of the two having the illness, then adding the different probabilities is needed. The difference is that the multiplication happens on the "front end", multiple people, multiple coins, multiple dice and so on. The addition happens on the "back end", where a condition is set where multiple outcomes can meet the requirement.

I tend to define expected value for my students as "probability times payout." This isn't always strictly true, like for the television hook-up example, the number of TVs in the house is not directly a payout. Students seem to be able to make the connection fairly easily, and it applies most of the time as expected values are connected to wagering and business more often than not. I would shy away from the contribution to the mean language.

The linearity problem is the first time that students will see why the variance is used frequently in upper level courses. Stress that you can't use the standard deviation in the same way because of the rules of the square root, although I am sure you will come across at least one student who tries.

---

## Student's t Distribution

The story of the Student's  $t$  is kind of cool. The person responsible for the invention of the distribution is William Sealy Gosset. Gosset worked for Guinness who was, at the time (1900s), interested in scientifically boosting barley crop production. His work involved small sample sizes, so he had to find a distribution that would work in testing different plots against each other. Due to a previous employee publishing trade secrets of brewing in an academic

paper, all employees were barred from publishing, regardless of content. As a consequence, Gosset chose to publish under the name “Student”.

One thing that some statistics books will stress is that there is no proving the null hypothesis true. The book dances around this, but if you are aware of the rule then you will see that they specifically say “accept the null hypothesis” or “there is no evidence against the null hypothesis.” Strictly speaking, rejecting the null hypothesis is a stronger condition than not rejecting one, or accepting one. It’s kind of like how a single counterexample will disprove a theorem definitively, but no amount of examples in support will prove that it is true. The supporting examples will give hints to give you confidence that the theorem is probably true, and that is exactly how a non-rejection should be treated. Sometimes the null hypothesis is written in a manner where rejection is the goal, therefore showing the opposite has a high degree of certainty to be true. Therefore, students need not only the skill to compute the statistic, but writing good null hypotheses.

Degrees of freedom is one of those things that sounds fancy, but really isn’t. There are reasons for the name, but they are beyond the scope of the class, and really aren’t necessary. The beginning and end of what students need to know is that it’s  $n - 1$ .

---

## 1.8 Hypothesis Testing

---

### Hypothesis Testing and the p-value.

In this section, you will need to key your students into the precise language being used. While it may not always be completely fair, but the proper language of hypothesis testing is one way that tell the people who have a proper training in statistics apart from the people who are pretending. This is possibly an issue when writing summaries on the open response questions on the AP examination. The use of proper language goes a long ways to promoting the idea that you know what you are talking about. The key phrases are “There is no correlation...”, “There is no difference...” for the start of every null hypothesis, “ Does not reject”, “Does reject” in regards to the results of the test. It might seem strange to students, but this isn’t the place for creativity or fancy language (or good grammar... sort of; many standard math phrases that are used over and over again are grammatically suspect, but we plow ahead anyway).

A good tip for helping students to understand hypothesis testing is to look at a couple of examples for each idea. For instance, when you introduce two-tailed testing, give students some problems where the null hypothesis is already written and they only need to evaluate whether or not they are rejecting in based on the level of significance chosen. If the information is not broken up, then students easily get lost in the details.

Calculating error and knowing the different types may seem trivial. It also isn’t a large part of later classes and work in statistics. However, it is something that students are expected to know for the AP examination, so they should be made aware of that.

---

### Testing a Proportion Hypothesis

This chapter is about the time that students start having a really tough time keeping all of the test and sample statistics straight. It may be a good use of time to take a break from new information and have students study, re-copy or interact with each of these statistics so they can hopefully keep them straight. The AP examination can be pretty stressful, and if time isn’t taken to review throughout the year, it will be hard for students to recall what they need in the test.

Proportion hypothesis testing is a simple instance of hypothesis testing. There are a couple of things that work slightly differently, like calculating the standard deviation, that should be the focus of this unit.

---

### Testing a Mean Hypothesis

It is important for students to have an intuitive sense of why the procedures need to change for small samples. The test outlined here should be tested with the chi-squared distribution, but the idea is the same regardless. Everyone is aware that the theoretical probability for a coin flip is .5. Students can then perform experiments with small numbers of observations, say 1 to 10. Discussing all the experiences from the students, they will see that there is great variation in results even though everyone has a fair coin. If all results are combined on one graph, it should be clear, and intuitively so, that the more flips performed resulted in a better match to the theoretical mean. Going

one step further, it's simply impossible to be anywhere close to the theoretical mean with 1, 3 or 5 samples. Students should be asked "what are the implications of this?" Likely answers include: the need for more samples (which is not always possible), the problem with using binary or discreet outcomes (get used to it, rarely can we depend on continuous empirical results), and most importantly, the uncertainty that is inherent to small sample sizes. What would it take to be certain that a coin is fair with only 5 experiments? Most students will probably indicate that they could not reject the hypothesis of the coin being unfair regardless of outcome. Now bumping it up to 10, now there will probably be some outcomes that will not be accepted. This is the conceptual foundation for small population, or non-normal tests.

---

## Testing a Hypothesis for Dependent and Independent Samples

It is not standard practice among different texts to assign different symbols for the hypothesis depending on if it is a dependent or independent test. However, this is a good idea, especially as it is easy to get confused or forget a step. The key part of this section is understanding the procedure for testing two different populations, especially when testing for growth. This is the first instance where students will have a chance to see and understand how to show a difference outcome based on procedure. This has to take into account the baseline information for each party, making it a dependent sample. The most interesting consequence for students, as outlined in one of the examples, is the utility of this process to examine various practices in school. If students are working on a project for the end of the year, this may become an integral test for answering many of the questions that students might have about their school.

Students may become a little lost with all the tests. There is a chart in a later section outlining when each test is to be used. It is also advisable at this point to have the students create one of their own, with the creation being an exercise to help them remember.

---

## 1.9 Regression and Correlation

---

### Scatterplots and Linear Correlation

Graphical representations are going to be the primary focus of bivariate data in a first year class. Many of the techniques required for statistical analysis of multiple variables requires calculus, so a basic treatment is all students are equipped for at this point. The scatterplot is a very familiar structure to students at this point. In some ways your task is not going to be teaching the principles of a good scatterplot, but rather un-teaching the bad habits or misconceptions that students have developed over the years. Those bad habits are usually focused around poor scaling, sloppy labeling and a general lack of precision. If they are representing data for analytical use, they will need to take extra care in having a graph that is clear and accurate enough to be useful. Also, simply using a computer grapher is not sufficient, as the scaling can still induce poor conclusions.

For many complex measures, like the various correlation coefficients, a table where values are determined step by step, as on page 341, is very useful in keeping all of the variables and values in the correct place. (It's also a great tool for finding standard deviations by hand. If you are careful, you will realize that there is plenty conceptually in common between the standard deviation and correlation coefficients.)

---

### Least Squares Regression

This is a biggie. Not only for the class or the AP exam, but for understanding all kinds of statistical work in the future. If you have students designing statistical projects, some of them will likely need to use least squares regression for their work. Plan a touch of extra time to make sure the class understands this section.

There are two ways to go about presenting this chapter. One is to have students work out their own plan for finding the best fit line. They should be capable of developing a number of different ideas, with a little bit of guidance. I start out by having students, or groups of students, construct a line with a straightedge that they think is the best fit. Inevitably, there will be some students who have different opinions. I ask if anyone can think of a way to test to see analytically who's line is the closest fit for all of the data. Usually students get very close, if not exactly on the correct answer. The drawback to this method is that it takes time, and that it can be confusing for students in the long run. There are many valid methods of finding or confirming a line to fit data. Least squares is the most common in statistics, so students are expected to know it. Some may get confused if a number of ideas are presented.

The other method is to go straight for the table, like the text does for example 1. This is very quick, clear and usually results in a great rate of success for students working these problems. The drawback anytime a strict algorithmic method is applied is that students miss the concept behind the method. This is of lesser importance in this section as opposed to others.

Of additional information presented, all of it is good, but the only part that is really important for the AP examination is the part on calculating residuals. Students may also be asked in a free response question to perform a t-test on residuals to determine if the line is a good enough fit.

---

## Inferences about Regression

The AP examination will usually ask a question where students are required to make an inference about the correlation between two statistics. There are a number of steps to doing so, all of which have been covered somewhere in this text, but not all of them in this section. At this point, students should be reminded of all of the conditions that are required for this inference. Some of it may seem like it's tedious routine, but it is easy to apply rules and tests in statistics in places where they are not going to give meaningful results. Furthermore, the results will seem logical, taking away the logical check system that students usually have.

The sample must be random, as with nearly anything that we are going to be looking at. This is again taken from the design of the experiment as covered in earlier chapters. The errors must be normal, which can be checked through various plotting methods. This is one condition that can frequently be overlooked without consequence, but is technically a requirement. Residual errors must be centered around 0. This can be figured with a plot, or by finding the mean of the errors. Along with the mean of the errors, and another reason to make a residual plot, the standard deviation of errors should be the same for all  $X$ . The fact that the errors are independent can also be determined from this plot, provided no trends in plotting are popping up. Once all of these items are checked, then the process can proceed as stated in the examples in the text. Presenting the solution will be a good choice on the AP free response examination for clarity and to show that it is known what the requirements are for the test.

---

## Multiple Regression

Multiple variable regression is tough to visualize. This is compounded by the problem of making 3-D graphs. For this reason, and the fact that there is no limit as to how many variables can be used, it makes sense to show a single example with a graph, and then move onto making the computations without a visual. There are many instances in mathematics where a two variables, or even three, are used to graphically develop a rule that can be extended beyond what can be represented; linear programming is a classic example. There is nothing lost by having students simply follow steps for a solution now that they have experience with regression for two variables. Another good plan is to use technology in this section for ease of solution or visualizing results. The text mentions SAS and SPSS but there are many stats packages that will perform regression with multiple variables.



---

## 1.10 Chi-Square

---

### The Goodness of Fit Test

I have found that students often have an easier time with chi-squared tests than with many of the other tests. Also, the chi-squared distribution can be intuitively constructed from binomial trials. For those reasons, I choose to introduce the chi-squared distribution before the  $t$  or the  $F$  distributions. The other nice thing about working with the chi-squared test is that the experiments are quick and easy to conduct. After presenting the information I will have a chart up at the front of the class everyday polling the students. Students are now asked to formulate the null hypothesis (which I have already developed to choose the poll question) and then perform the chi-squared test and make a conclusion.

A bit of caution about something in the text. Students are going to want to hold on to a particular chi-squared value to apply to all situations, regardless of the degrees of freedom or level of confidence. The book doesn't help as it almost sounds like it declares that a particular chi-squared value is the threshold for rejection. I always make sure that students have to work different problems so they don't develop a habit.

---

### Test of Independence

It is debatable whether you should keep this topic separate from the previous, or if they should be merged. The purpose of merging the two topics is that they are really the same. The difference in the test for independence is the same as fit, except for the way the null hypothesis is written. A reason to give it its own treatment is because the AP examination treats it more in that manner. I tend to treat them as one section. Conceptually, the difference is in how the null hypothesis is written and correlation is shown the same way that independence is. Also, while other sections can be helped by being stretched out a little, students do not need as much practice with chi-squared tests as many of the others.

A fun activity I remember from university is that we ran chi-squared tests for homogeneity on various random number generators. Even with different seeds, the random generator on the calculator did not fare well in our experiment. I can be a very fun one for students, especially if human responses are added in.

---

### Testing one Variance

The text mentions that the  $F$ -test is sensitive to non-normality, which is only partially true. In this chapter it is, as the specific instance of the  $F$ -test being sensitive is when showing that variances are the same. The  $F$ -test is actually guarded pretty well against non-normality in other methods of testing, which is why it is a part of ANOVA, which is specifically mentioned as a robust test. I don't know how much I would bother students with these details, but since a good student is diligent about checking requisite conditions, it may be worth mentioning.

---

## 1.11 Analysis of Variance and the F-Distribution

---

### The F distribution and Testing Two Variances

Note that the  $F$  distribution and other tests of variance are not topics on the *AP* examination. This does not make them less valid, but they should be retained until after the examination at topics to fill out the remainder of the year.

This may be a little bit tough for students to understand the motivation behind these tests of variance. Another problem is that the  $F$  distribution is complex enough that first year students are not exposed to details behind the distribution. The best thing for students to remember is that while means are often easy to find or approximate, the variance is not. The key use of the  $F$  test is going to be when two populations are being examined, or when a comparison can be made to a known variance.

---

### The One-Way ANOVA Test

ANOVA tables are computed frequently in professional statistical analysis. They can be tedious, but in the real world they are nearly always done with the assistance of some computational software. I suspect that it is because of this that ANOVA tables are not a required topic for the *AP* examination. I think that I have completed a single ANOVA table by hand, ever. This is also what I have my students do and then move quickly on to learning how to use the software program of choice.

ANOVA is a difficult procedure to run on some software packages. At the very least, there is little consistency between any of the software programs. In fact, I was asked to present to the math department when I was an undergraduate the different ANOVA output from different software packages. It took me a very, very long time to prepare the presentation due to the difficulty of learning all of the different syntax and output. I had looked at Mathematica, MAPLE, SPSS, SAS, R, S+ and Excel. I did not have access to Minitab, MATLAB (I don't know for sure if the base package will do ANOVA) or Fathom. The most interesting finding was the extreme limitations of Excel, as well as some errors in calculation. Excel errors are well documented, and many have workarounds, but I can't ever recommended it as an accompaniment to the stats class. Maple, SPSS and I have heard Fathom, are pretty user friendly and are popular choices. Whatever your package is, make sure to block out time around these chapters to get comfortable with using the software for these computations.

---

### The Two-Way ANOVA Test

Since there is little difference between a one way and a two way ANOVA test, there shouldn't be much practice necessary. There are a couple benefits to this. First, it is a way to reinforce some of the topics learned in the one way lesson. Second, by contrasting some of the elements that are different about the two way test, the important parts of the one way test become apparent. Remember that the one of the toughest things about seeing a new topic for the first time is that everything seems of equal importance, and everything seems unique. This is especially true as the theory gets tougher and more and more examples are presented. Students will have trouble at times telling what

parts of the problem are specific to the example and which parts are always going to be part of the ANOVA table. Looking at two way tests can help with this.

Another consideration is how much time needs to be spent on setting up data for each program. I know that some programs require entry in cells, or imported from comma separated value files. Other programs require the data in the form of a matrix. Students need to be aware of what the requirements of their software is, but it will also help if they know a little about what possibilities are out there, in case they come across different software at a later date.

## 1.12 Non-Parametric Statistics

### Introduction to Non-Parametric Statistics

Non-parametric statistics are a more exotic test. The situations in which data is encountered that is not normal are few, and as a consequence this section is very rarely covered in any statistics course, let alone a high school level AP course. These sections are advisable only for students who are truly exceptional. Furthermore, these tests are almost never run by hand. Every statistics specific package will have tests, and they should be used.

An interesting note is that a common use of the runs test is to test for randomness. This makes sense, as testing for correlation is the opposite, and obviously random numbers will not be randomly distributed. Appendix A: Online Data Sources

There are a number of sources of real data that is pertinent and interesting for projects and assignments for students. Here is a short list of sources I have used.

Data.gov: <http://www.data.gov/catalog/raw>

The central clearinghouse for raw data for all of the federal government agencies. Data is available in XML or CSV files.

US Census: [http://www2.census.gov/census\\_2000/datasets/](http://www2.census.gov/census_2000/datasets/)

Raw data from the 2000 census. There are other tools on the census site, but most of them manipulate the data too much rather than just presenting numbers.

UN Data: <http://data.un.org/>

A clearinghouse for the data collected from various UN organizations including the WHO, WTO, UNICEF and UNESCO.

Baseball Reference: <http://www.baseball-reference.com/>

While there are other sports that collect statistics, none do it with the verve of baseball. This is the most complete resource for all kinds of baseball statistics.

#### Appendix B: Statistical Software Packages

This list is not exhaustive, but should give a brief overview as to what software is available to support classroom activities.

Fathom: <http://www.keypress.com/x5656.xml>

Fathom is published by a textbook publisher, so it has a wide following in schools. It is easy to use, powerful enough for most, and has many activities and lesson plans available for use. It also is one of the cheaper packages. Windows and MacOS

SPSS: <http://www.spss.com/>

SPSS has outgrown its acronym: Statistical Package for the Social Sciences. It is now a huge data mining and statistics suite owned by IBM. It is easy to get started with, as it uses a spreadsheet type interface for most of its data. Expensive. All platforms.

SAS: <http://www.sas.com/>

SAS is geared more towards enterprise uses. I have never been in a school lab with SAS installed, but it remains a popular choice for many, with lots of online community help and resources. Expensive. All platforms.

R: <http://www.r-project.org/>

R is a FREE open-source clone of the functionality from the legendary math and statistics package S and S+ from AT&T. It is huge, the documentation is hard to read, it has few graphical interface items, and is truthfully more of a programming language/environment. It's what I use. It does everything. Did I mention its free? All platforms.

Minitab: <http://www.minitab.com/en-US/default.aspx>

Minitab is, in my experience, somewhere between SAS and SPSS in usability. It is a favorite of many business schools. There is a significant group of AP teachers using Minitab in their classes, so lesson help should be easy to find. Expensive. Windows only.

Mathematica: <http://www.wolfram.com/>

An all around math software and programming environment. It's using a cannon to kill a fly for a stats class. Tough to use, but has incredible documentation. Very, very expensive, but massive discounts are available to teachers and students. All platforms.

Excel: <http://office.microsoft.com/en-us/excel/default.aspx>

You were waiting for this one... Well, the good news is you probably have it already, and everyone is familiar with at least the basics. The bad part is that the range of tests is very limited, and there are an incredible amount of documented errors with various distributions and functions. I can't recommend it, but it has worked for some in the past, and will continue to work for many in the future. You probably don't need to buy it. Windows and MacOS.

Random.org: <http://www.random.org/>

Not software, but a website that is committed to provide about as good of random numbers as is possible. There is also a ton of info on randomness, why it's elusive and some other techniques for better generating randomness. A great resource for teachers and students. Free!

## CHAPTER

**2****Probability and Statistics TE -  
Enrichment****CHAPTER OUTLINE**

---

- 2.1 AN INTRODUCTION TO ANALYZING STATISTICAL DATA**
  - 2.2 VISUALIZATIONS OF DATA**
  - 2.3 AN INTRODUCTION TO PROBABILITY**
  - 2.4 DISCRETE PROBABILITY DISTRIBUTIONS**
  - 2.5 NORMAL DISTRIBUTION**
  - 2.6 PLANNING AND CONDUCTING AN EXPERIMENT OR STUDY**
  - 2.7 SAMPLING DISTRIBUTIONS AND ESTIMATIONS**
  - 2.8 HYPOTHESIS TESTING**
  - 2.9 REGRESSION AND CORRELATION**
  - 2.10 CHI-SQUARE**
  - 2.11 ANALYSIS OF VARIANCE AND THE F-DISTRIBUTION**
  - 2.12 NON-PARAMETRIC STATISTICS**
-



## 2.1 An Introduction to Analyzing Statistical Data

This Probability and Statistics Enrichment FlexBook is one of seven Teacher's Edition FlexBooks that accompany the CK-12 Foundation's Probability and Statistics Student Edition.

To receive information regarding upcoming FlexBooks or to receive the available Assessment and Solution Key FlexBooks for this program please write to us at [teacher-requests@ck12.org](mailto:teacher-requests@ck12.org).

### Definitions of Statistical Terminology

#### Activity: The Effect of Units on Continuous Measurements

In this activity students will explore the effect units have on continuous variables. A continuous variable must be measured using some instrument and some unit. The distance between two cities could be measured in feet, meters, miles, or kilometers. The instrument could be the odometer of a car or a bike, a pedometer, satellite technology, or the distance could be calculated using a map. The method and units of measurement affect the data that is gathered.

**Materials:** a class set of rulers that have both inches and centimeters

#### Procedure:

- Each student should measure the length of their right ring finger in both centimeters and inches.
- The class data should be compiled. After the students have made the measurements they can write them on the board. One side of the board can be used for measurements made in inches and the other for those taken in centimeters. The instructor could also write the data on the board as the student read off their information, or a paper could be passed around the room and the data transferred to the board.
- Make a dot plot for each set of data. This will be a number line with measurements along the bottom. One graph will use inches, and the other, centimeters. Dots, one for each measurement, can then be placed above the number line.
- Analyze the data in a class discussion with the following questions.
  - Theoretically, is the length of a finger discrete or continuous?
  - Does the data displayed on the board look discrete or continuous?
  - How does rounding affect the data?
  - How do the units of measurement affect the data?

As the discussion progresses, students should realize that even though the length of a finger is a continuous variable, the data is effectively discrete. There is only so much accuracy that can be gotten from a ruler and measurements must be rounded. What value the number is rounded to depends on the units used. When inches are used, there may be fingers recorded as length  $2\frac{5}{8}$  inches, but when centimeters are used that same finger might be 6.7 centimeters. The practicalities of measurement make continuous variable an abstraction, but the smaller the units and the greater the degree of accuracy, the closer the variable comes to being truly continuous.

---

## An Overview of Data

### Research and Discuss: Experimental Ethics

Many fields depend on experimental data. Pharmaceutical companies use experiments when developing new drugs, as so do cosmetic companies when creating new beauty products. Psychologists are famous for conducting behavioral experiments. The subjects of these experiments are often animals, including people. There is a wide variety in the opinions on what is, or is not, ethical treatment of the subjects of these experiments, and when the knowledge gained by the experiment justifies the discomfort, trauma, or death of the participant.

#### Research Topics:

- Find examples of experiments done on people that are considered unethical by the current standards of our society. Include examples from the fields of psychology and medicine, and example from different time periods in history. How did humanity benefit from these experiments?
- Find examples of experiments done on animal, other than humans, that may be considered unethical. How did humanity benefit from these experiments?
- Different organizations have developed guidelines for what constitutes an ethical experiment. Find examples from a variety of groups including psychologists, medical doctors, pharmaceutical companies, animal rights groups, governments, and any other relevant group.

#### Discussion Topics:

- Compare and contrast the different philosophies on ethical experiments. How have these opinions evolved over time?
- When is the cost to the individual subjects of the experiments justified by the benefit of the results to society?

#### Procedure:

Assign individuals or groups of students to different research topics. Have them present what they found in class to stimulate discussion.

---

## Measures of Center

### Assignment: The Mean, Median, and the Data

The measures of center are important tools used to summarize and describe sets of data. Calculating the mean or median of a set of data will come easily for students at this level. The skill that needs to be developed at this stage is the ability to interpret what the mean and median convey about a specific data set. Students need to be able to get information about the data set by comparing these two measures of center, and be able to decide which is a better description of the data in a specific situation. They need experience with data sets that are familiar and of interest to them.

#### Guidelines:

- Find a data set with between 30 and 50 elements. You can collect this data yourself or get it from a reliable source. Cite the source of your data or describe your collection method.
- Make a dot-plot of the data set. Label and title the plot.
- Calculate the mean and median of the data set.
- Mark the mean and median on the plot.
- Write an analysis of the work you have done that addresses the following topics.

### 2.1. AN INTRODUCTION TO ANALYZING STATISTICAL DATA

- Does the set have outlier(s)? Is the shape of the graph symmetric?
- Are the mean and the median close in value? Why or why not?
- Which measure of center is closer to the outlier(s)?
- Which measure of center best describes a typical value in the data set?

Have students present their work to the class. Display the dot-plots on the walls of the classroom. The exposure to these sets, along with their measures of center, will help students develop an understanding and intuition for what the mean and median can tell them about a set of data.

---

## Measures of Spread

### Integrating Technology: Spreadsheets

Knowing how to use a spreadsheet is a valuable skill. Many college science classes require that data analysis for lab work be done on a spreadsheet. Students at the college level are expected to have basic knowledge of, and the ability to use this tool. A quick perusal of the requirements given in job descriptions for work in the fields of accounting and finance, as well as many other fields, would convince anyone that learning to use a spreadsheet is a worthwhile pursuit. Students are adept at picking up new technology, and before long will be showing you useful features of this program.

**Objective:** Calculate the standard deviation of a large set of data by making the usual table on a spreadsheet.

**Procedure:**

- Provide the students with a large set of data, one with at least a hundred elements. Another option is to have the students provide their own data set so that it will be of more interest to them. They can use sports statistics, measurements that indicate climate change, or anything. For grading purposes it will be easier to have everyone using the same set of data, especially for a first attempt.
- The table should have three columns titled like those shown in the text of the lesson.
- In the first column the value of the variable must be entered. If everyone in the class is using the same set of data, you can provide the spreadsheet to the students with the first column already filled.
- The second row of the second column will contain a formula for the deviation of each value from the mean. The students can be taught how to reference other cells and to fill down.
- The second row of the third column will square the values in the second column.
- Now students can get the sum of the third column and calculate the standard deviation.

**Notes:**

- This would be done more efficiently with two columns, but it would give the students less practice. A brief discussion of error magnification and an explanation of the three column requirement are appropriate.
- Excel is the most widely used and nicest spreadsheet, but use what is available to you.
- Students can find detailed explanations of how to use these programs with a quick internet search.

### Assignment: Picturing the Standard Deviation

The standard deviation provides vital information about a set of data. It is a key component of many of the calculations that are done in statistics. The mean of a data set is not particularly useful unless it is paired with the standard deviation. In the past students have had multiple exposures to the mean, but this is most likely the first time they have encountered the standard deviation. They will have a difficult time seeing where it is and what it measures. Experience with the standard deviation is the key to their understanding. This assignment will provide students with an opportunity to work with the standard deviation of data sets that are of interest to them.

**Guidelines:**

- a. Find two data sets each with at least 20 elements. Chose one data set with numbers that are fairly spread out and one where the values are all relatively close to the mean. You can collect this data yourself or get it from a reliable source. Cite the source of your data or describe your collection method.
- b. Make a dot-plot of each data set. Label and title the plots.
- c. Calculate the standard deviation of each data set using a table. Check your answer with your calculator.
- d. On your dot-plot, highlight all the values that are within one standard deviation of the mean in yellow, those between one and two standard deviation from the mean in pink, and those between three and four standard deviation in green.

Have students present their work to the class. Display the dot-plots on the walls of the classroom.

Discuss if and how the standard deviation would change if the measurements were made in different units.

The process of finding data sets with large and small standard deviations will make the students think about what the standard deviation tells them about the data.

## 2.2 Visualizations of Data

### Histograms and Frequency Distributions

#### Activity: The Effect of Bin Width on the Shape of a Histogram

Describing the shape of a histogram is not always a straightforward task. There is not always one correct answer. Usually, combinations of words must be used to obtain phrases like, “approximately normal with an outlier”. To make matters more ambiguous, how the data is grouped often affects the shape of the histogram.

**Materials:** The instructor needs a graphing calculator and an overhead display mechanism. Students can follow along with their graphing calculators for practice, but this will slow down the activity.

#### Procedure:

- Gather several sets of data from the students. Chose sets that you would expect to have different shapes. Their heights will be approximately normal, but the number of pets that they have will most likely be skewed with outliers. So as not to use class time collecting and entering data, students can provide the information the day before on index cards and the instructor can enter the data into the calculator before class. The lists can be transferred to the students’ calculator with a cord if the students are to follow along.
- Make a histogram with one of the data sets. Start with a set that will be easy to describe. Display the histogram and ask students to describe the shape. Display the same data set with different bin widths and compare the resulting histograms. Sometimes the shape appears to be quite different.
- Repeat with the other sets of data.

#### Additional Topics for Discussion:

- Find the mean, median, and standard deviation for each data set. Note that in the case of outliers and skewed data, the mean is pulled toward the outlier or tail. How is the standard deviation affected by outliers and skewed data?
- This is a good time to discuss how subjective statistics can be, and how data can be manipulated to seem to support various points of view.

### Common Graphs and Data Plots

#### Technology Project: The Right Graph for the Data (with Excel)

Learning to create these different graphs is not terribly challenging for students. Choosing the best graph or data plot to display a specific set of data is the most important skill students will take away from this lesson. This task also gives the students practice using the powerful, and commonly used tool, a spreadsheet.

#### Procedure:

- Find three sets of data, each with at least 20 elements. Collect this data yourself or get it from a reliable source. Cite the source of your data or describe your collection method.

- One set of data will be categorical, to be used in a bar and pie graph.

- One set of data will be bivariate, and will be used in a scatter plot. Choose two variables that they believe will have a fairly strong association.
- One set of data will be bivariate with the explanatory variable being time. This data will be displayed in a line plot.

2. Each data set will be entered into columns in a different page of a spreadsheet program. The first cell in each column should contain a title. Select the data and insert the proper graph of plot for each of the three data sets.

3. Write a paragraph describing the plots and graphs using vocabulary from this section of the text. What have you learned about the data sets from the visual display that you made?

This assignment reverses the typical situation. Here students are looking for data to fit a specific graph. It still gives the students the opportunity to match data sets with visual displays. If time allows have the students present their graphs and plots to the class. Orally describing their work to others will make it more meaningful for them.

---

## Box and Whiskers Plots

### Activity: Stem-and-Leaf Plot to Box-and-Whiskers Plot

Students familiarized themselves with stem-and-leaf plots in the previous section. A stem-and-leaf plot is basically a histogram made of the ones digits of the numbers in the data set. They are a good representation of small data sets because the actual values are retained, while also giving a visual representation of the data. Students have seen variations of this method for representing data many times before. They understand it well. The box-and-whiskers plot is a new concept; it is based on position instead of value. Students will need some experience with this type of display before they will be able to gain a good understanding the data from a box-and-whiskers plot.

#### Procedure:

- Select some stem-and-leaf plots with different shapes that you have made in the past or have seen in the text or elsewhere. (The instructor can make the selection or leave it up to the students.)
- Describe the shape, center, and spread of the data.
- For each stem-and-leaf plot make a box-and-whiskers plot of the same data.
- Does seeing the data displayed in a different way make you want to change your description of the data's shape, center, and spread?
- Which plot would be easier to make for a large data set? In what circumstances would you chose to use the stem-and-leaf plot? The box-and-whiskers plot?

This activity will give students practice reading the numbers from stem-and-leaf plots, and making box-and-whisker plots. Most importantly though, it will teach the students how to interpret box-and-whiskers plots and get them thinking about the strengths and weaknesses of the different types of visual displays of data that they have learned to make so they can chose the best method in any situation.

### Investigation: The Effect of an Outlier on Measures of Spread

The most important measure of spread used in statistics is by far the standard deviation/variance. Students need to realize that the standard deviation as well, as the mean, are not resistant to outliers or skewed data.

#### Procedure:

Use the reservoir data for California given in this lesson.

- Calculate the interquartile range for the data. Remove the outlier of 34 from the set and calculate the interquartile range again.
- Calculate the standard deviation of this sample. Remove the outlier and calculate the standard deviation again.

## 2.2. VISUALIZATIONS OF DATA

- c. Calculate the percent change for each measure of spread.
- d. Use the calculations made in (1) – (3) to evaluate how well the interquartile range and standard deviation represent the original data (before the outlier was removed).

**Answers:**

- a. 11, 10
- b. 15.28, 7.54
- c. 9%, 51%
- d. The interquartile range is the better representation of spread for this set of data. In the case of the standard deviation, it does not seem reasonable for one value to have such a large affect on a single summary statistic.

Many calculations in statistics can only be done with a standard deviation, so the standard deviation must be used even if the data is heavily skewed or there are significant outliers. In these situations statisticians may chose to trim the data set, or leave off outliers. This investigation will help student see why this is a reasonable

## 2.3 An Introduction to Probability

### Introduction

#### Assignment: Experiments, Events, and Outcomes

When a probability experiment is performed, there are many ways the outcomes can be divided into events. How these events are defined often determines how the probabilities are distributed and calculated. Students are not ready to know all the details this early in the course, but they should learn to be flexible and creative in defining events.

#### Definitions

A simple event is one that contains only one outcome, or in other words, can happen in only one way. A compound event contains more than one outcome.

#### Example:

- For the experiment of rolling a six-sided die, there are six possible outcomes. The sample space, the set of all possible outcomes, can be represented as  $S = \{1, 2, 3, 4, 5, 6\}$ . List two simple and two compound events for this experiment.

Answer: Simple Events:  $A = \{1\}$  or  $B = \{5\}$

Compound Events:

$C =$  rolling an even number  $= \{2, 4, 6\}$  or

$D =$  rolling a number less than three  $= \{1, 2\}$

#### Exercises:

- The experiment of drawing one card from a standard deck has 52 outcomes. Define two simple and two compound events.
- Buying a box of cereal is a probability experiment where every different type of cereal is a possible outcome. Define two simple and two complex events.
- Describe a probability experiment and the possible outcomes. Define two simple and two complex events.

Answer will vary.

### Compound Events

#### Assignment: More Practice with Unions and Intersections

Students can always use more practice finding the intersection and union of sets. A good guideline for students to keep in mind is that the intersection of two events is more restrictive and results in a set that is smaller than, or in some cases equal in size to the original sets, and that the union of two events is more inclusive and results in a set that is larger than, or in some cases equal in size to the original sets.



**Exercises:**

1. Experiment: drawing one card from a standard deck of cards

Events:  $A = \{\text{a red suite}\}$ ,  $B = \{\text{a face card}\} = \{\text{Jack, Queen, King}\}$

List the outcomes in  $A \cup B$ , and in  $A \cap B$ . If each card in the deck is equally likely to be chosen what is the probability of each compound event?

2. Experiment: Asking participants the question, “What is your favorite day of the week?”

Events:  $A = \{\text{their response is a weekday}\} = \{\text{M, T, W, H, F}\}$ ,

$B = \{\text{their response is a weekend}\} = \{\text{S, U}\}$

Find  $A \cup B$  and  $A \cap B$ .

3. Describe a probability experiment and events such that the intersection of the events is the empty set and the union of the events is the entire sample space.

Answers:

1.  $A \cap B = \{\text{JD, QD, KD, JH, QH, KH}\}$ ,  $P(A \cap B) = \frac{6}{52} \approx 11.5\%$

$A \cup B = \{\text{all 13 hearts, all 13 diamonds, JS, QS, KS, JC, QC, KC}\}$ .  $P(A \cup B) = \frac{32}{52} \approx 61.5\%$

2.  $A \cap B = \{\} = \text{the empty set} = \emptyset$ ,  $A \cup B = \{\text{M, T, W, H, F, S, U}\} = \text{the sample space}$

3. Answers will vary.

## The Complement of an Event

### Assignment: Just Two Possibilities

Being able to calculate a probability often depends on thinking of the situation in the right way, by correctly fitting it into a mold. It is frequently useful to take a sample space with many possible outcomes and simplify it into one event and its complement. This is a step toward applying a binomial distribution. Students will learn about distributions in the next chapter. Give them the opportunity to play with dividing the sample space into an event and its complement now.

**Example:**

1. Divide the sample space of the following experiment into one event and its complement.

Experiment: A shopper purchases a box of cereal.

Outcomes:  $S = \{\text{all existing types of cereal}\}$

Answer:  $A = \{\text{shopper buy cereal with less than 10 grams of sugar per serving}\}$  and

$A^c = \{\text{shopper buys cereal with 10 or more grams of sugar per serving}\}$

**Exercises:**

1. Find two more ways the sample space from the experiment described in the example can be divided into an event and its complement.

2. Find two ways in which to divide the sample space of the following experiment into an event and its complement.

Experiment: A card is drawn from a standard deck.

Outcomes:  $S = \{\text{all 52 possible cards}\}$

3. Find two ways in which to divide the sample space of the following experiment into an event and its complement.

Experiment: The weight category associated with the BMI (body mass index) of an adult participant.

Outcomes:  $S = \{\text{Underweight, Normal, Overweight, Obese}\}$

## Conditional Probability

### Extension: Using Conditional Probability to Calculate the Probability of Intersections

In this section students calculate conditional probabilities using the formula  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ ,  $P(B) \neq 0$ . Multiplying both sides of this formula by  $P(B)$  yields a way to find the probability of the intersection of two sets,  $P(A \cap B) = P(A|B) \cdot P(B)$ ,  $P(B) \neq 0$ . This is an extremely useful and widely used method for calculating probabilities.

#### Exercises:

1. If a certain study finds that the probability of having an accident each year given that the driver regularly speeds is 0.2, and that 60% of drivers regularly speed. What is the probability that a randomly selected driver regularly speeds and will be in an accident this year?

Answer: Let  $A$  be the event of having an accident sometime during the year, and  $S$  be the event of selected driver who regularly speeds, then  $P(A|S) = 0.2$ , and  $P(S) = 0.6$ . So the probability that a randomly selected driver speeds and is involved in an accident is

$$P(A \cap S) = P(A|S) * P(S) = 0.2 * 0.6 = 0.12$$

2. Teresa is having trouble deciding between two elective courses. She estimates that the probability of getting an  $A$  in Environmental Studies is  $\frac{3}{4}$  and the probability of getting an  $A$  in Psychology is  $\frac{2}{3}$ . If she decides which class to take by flipping a fair coin. What is the probability she finishes the year with an  $A$  in Psychology?

Answer:

$$\begin{aligned} P(\text{Psychology and } A) &= P(A \text{ given Psychology}) * P(\text{Psychology}) \\ &= \frac{2}{3} * \frac{1}{2} = \frac{1}{3} \end{aligned}$$

3. Use conditional probability to calculate the probability of drawing two diamonds from a standard deck of cards?

Answer: Let  $D_1$  be the event that the first card is a diamond and  $D_2$  be the event that the second card is a diamond, then the probability that both cards are diamonds is

$$P(D_1 \cap D_2) = P(D_1) * P(D_2|D_1) = \frac{13}{52} * \frac{12}{51} = \frac{156}{2652} \approx .0588$$

## Additive and Multiplicative Rules

### Extension: Tests for Independence

Many of the theorems and rules of probability apply only to independent events, and it is not always a simple matter to determine if two events are independent. There are two widely used tests for independence.

#### 2.3. AN INTRODUCTION TO PROBABILITY

Two events,  $A$  and  $B$ , are independent if  $P(A|B) = P(A)$  or  $P(B|A) = P(B)$ .

Two events,  $A$  and  $B$ , are independent if  $P(A \cap B) = P(A) * P(B)$

### Exercises:

A company of 200 employees is considering a new health care plan. The following distribution shows the responses of all 200 employees based on the variables gender and opinion when they are asked their opinion on the new plan.

**TABLE 2.1:**

	In Favor	Against	
<b>Female</b>	30	90	120
<b>Male</b>	8	72	80
	38	162	

1. Are the events,  $F$  = being female, and  $A$  = being against of the new health plan, independent? Justify your answer with both definitions of independence.

Answer: No, gender and opinion on the healthcare plan are dependent.

$$P(F|A) = \frac{90}{160} = 0.5625 \text{ and } P(F) = \frac{38}{200} = .19$$

$$P(F \cap A) = \frac{90}{200} = .45 \text{ and } P(F) * P(A) = \frac{38}{200} * \frac{162}{200} = .1539$$

### Topics for Discussion:

- What does this imply about the healthcare plan?
- If the probabilities in the respective definitions were approximately equal, would the events be almost independent? How close to equal do probabilities calculated from this type of data need to be for the events to be considered independent?

## Basic Counting Rules

### Discussion and Activity: Taking a Simple Random Sample

In theory, making random selections from a population to form a sample sounds quite simple, but in practice, designing a method where each member of a large group has an equally likely chance to be chosen is often quite difficult. The challenge is finding a good sampling frame. A sampling frame is the list of units from which the sample is drawn. It might be a telephone directory, but not every member of the population is listed. It is difficult to find a complete sampling frame.

### Generating Random Numbers

Once the frame is formed, each unit on the list can be assigned a number. Units will then be selected with a random number generator. Most calculators and computers can select random numbers. On the TI-84 for example, this can be done by selecting MATH, then moving over to PRB, the fifth option is randInt(. Three numbers are required in the argument of this function. The first two are the range in which the user would like the random number to be, and the third indicates how many numbers are required.

### Discussion Topics:

- How would you take a simple random sample that represents this class? The school?

- b. What sampling frames could be used to get a simple random sample of the residents of a city, college student, or mothers? What members of the population would not be included in the frame? How would the absence of these members affect the results of the survey?

**Activities:**

- a. Have the students take a simple random sample of some population in the school, perhaps the athletes, honor roll students, or drama participants. Each selected student should be asked a question or complete a short survey, just to make the process more interesting. The instructor may have to request information from the school office for the students. This can be assigned to small groups who report their method and findings to the class.
- b. The students can research sampling frames. What creative methods have been used? What are the strengths /drawbacks of these methods? What outstandingly bad methods have been used? What were the results? This can be assigned to small groups and the results presented in class.

One option is to assign the first activity to some groups and the second to others, or let the students choose which assignment they would like to complete.

## 2.4 Discrete Probability Distributions

### Introduction

#### Explore: Probability as Area

In a graph of a probability distribution, area is equal to the probability. This fact becomes especially important when calculating probabilities with a normal distribution in the next chapter or using integrals to find probabilities in later classes. It is helpful to get students accustomed to this concept now while working with the more straight forward case of discrete random variables.

#### Procedure:

Consider the probability experiment of flipping a fair coin four times with the discrete random variable,  $X$  = the number of heads in those four flips.

- What are the possible values of the discrete random variable?
- Calculate the probability of each possible result.
- Display the probability distribution for this experiment as a table and as a graph. Check your work by seeing if the two conditions that must be true for all probability distributions are satisfied.
- Calculate the area of each rectangular bar in the graph of the probability distribution. What is the sum of the areas?
- What is the total area over the random variable values of two and three? What is the probability of getting two or three heads?
- What is the relationship between probabilities and area in a probability distribution?

#### Answers:

- $\{0, 1, 2, 3, 4\}$
- $P(0) = \frac{1}{16}, P(1) = \frac{4}{16}, P(2) = \frac{6}{16}, P(3) = \frac{4}{16}, P(4) = \frac{1}{16}$
- .
- $\frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16} = \frac{16}{16} = 1$
- $\frac{6}{16} + \frac{4}{16} = \frac{10}{16}, P(2 \text{ or } 3) = P(2) + P(3) - P(2 \text{ and } 3) = \frac{6}{16} + \frac{4}{16} - 0 = \frac{10}{16}$
- The area is equal to the probability.

### Mean and Standard Deviation of Discrete Random Variables

#### Project: Collecting and Displaying Discrete Data

At this point in the class a project that pulls together many of the important aspects of what the students have been learning in the first few chapters will have substantial benefit. It provides an opportunity to review, and the added benefit of work with real data will improve students experience and intuition.

**Objective:** To design and perform an experiment that yields values of a discrete random variable and to display the resulting probability distribution as a table and as a bar graph to be analysed.

#### Procedure:

- Design a probability experiment that produces values of a discrete random variable.
- Perform the experiment at least 20 times and record the results.
- Use the collected data to calculate the probability of each value of the random variable.
- Make a table and a graph for the probability distribution.
- Calculate the mean, variance, and standard deviation for the probability distribution.
- Describe the shape of the distribution.

### Technology:

Encourage the students to use a spreadsheet to record data, and to make calculations, tables, and graphs. Increasing the amount of data the students must collect will make using technology more rewarding. A quick demonstration of how to use formulas and fill in columns will get students started. If they need additional help, or want to learn more about how to use this tool, additional information and tutorials can be found online.

---

## The Binomial Probability Distribution

### Explore: Selection Without Replacement

The characteristics of a binomial experiment require that the probability of success be constant from trial to trial, and that the trials are independent of each other. These conditions do not hold when selecting from a finite group without replacement, but if the group is large enough, and the number of trials is small enough, maybe we can get “close enough”.

#### Case One:

In a class of 30 students you have 4 good friends. The instructor is randomly selecting three students to present their project today. Let  $X$  = the number of your good friends selected

- Why isn't  $X$  a binomial random variable?
- Use classical probability, combinations, and the Multiplicative Rule of Counting to calculate the probability that exactly two of the presenters are your close friends.
- Use the binomial probability distribution with  $p = \frac{4}{20}$  to approximate the probability of having exactly two of your close friends chosen. What do you think of this approximation?

#### Case Two:

In a school of 1200 students, 3 will be randomly selected to complete a survey about the school lunch program. Let  $X$  = the number of those same 4 good friends that are selected for the survey.

- Use classical probability, combinations, and the Multiplicative Rule of Counting to calculate the probability that exactly two of the four close friends are chosen.
- Use the binomial probability distribution with  $p = \frac{4}{700}$  to approximate the probability of having exactly two of the close friends chosen. What do you think of this approximation?

### Answers:

- Once a student is selected they are removed from the pool of choices so the probability of success for the next trial changes. The trials are also not independent. If the first trial is a success, the chances of a success on the second trial are lower.
- $\frac{{}_4C_2 \cdot {}_{26}C_1}{{}_{30}C_3} = \frac{156}{4060} \approx 0.0384$
- ${}_3C_2 * (.2)^2 (.8)^1 = .096$ , not a good estimate
- $\frac{{}_4C_2 \cdot {}_{1196}C_1}{{}_{1200}C_3} \approx 0.0000250$
- ${}_3C_2 * \left(\frac{4}{1200}\right)^2 \left(\frac{1196}{1200}\right)^1 \approx 0.0000322$ , this estimate is much better

## 2.4. DISCRETE PROBABILITY DISTRIBUTIONS

## The Poisson Probability Distribution

**Practice and Extend: Changing Units for the Poisson Distribution** The parameter  $\lambda$  gives the mean number of events in a certain amount of time, distance, volume, or area. Sometimes though, it is not given in the desired units. For instance,  $\lambda$  could be the average number of accidents at a given intersection in a year, but the probability to be calculated is for three accidents in a given month. It is an easy fix;  $\lambda$  can just be divided by twelve. Of course, it won't be month specific, December probably has a higher average than May, and this method will not reflect that difference. In many situations, statisticians calculate the best value that they can, and then consider the inaccuracies.

### Exercises:

- a. On average there is one flaw found in every yard of sheetrock produced on a specific machine.
  - a. What are the mean and standard deviation of the distribution of flaws per foot?
  - b. What is the probability of finding a flaw in the first foot of sheetrock?
- b. A busy executive receives an average of 14 emails an hour.
  - a. What is the probability that she will receive more than 150 emails in a ten hour work day?
  - b. What is the probability that she will receive more than 200 emails in a ten hour work day?
- c. After being given a free trial, two out of three participants will enroll in a certain telephone service. What is the probability that exactly 70 out of 90 participants will enroll?

### Answers:

- a.
  - a. mean = standard deviation =  $\frac{1}{3}$
  - b.  $p(1) \approx 0.2388$
- a.  $\lambda = 140, P(X > 150) = 1 - P(X \leq 150) \approx 1 - 0.8134 = .1866$
- b.  $\lambda = 140, P(X > 200) = 1 - P(X \leq 200) \approx 1 - 0.999992528 \approx 0$
- b.  $\lambda = 60, P(X = 70) \approx 0.02160$

## The Geometric Probability Distribution

### Practice: Identifying Discrete Probability Distributions

Now that students know the basics about the Binomial, Poisson, and Geometric distributions it is time for them to work on identifying which distribution can be used in a given situation.

### Exercises:

Identify which, if any, of the discrete probability distributions can be used in the following situations. Give the value(s) of the parameter(s) for the distribution, and find the indicated probability.

- a. In a class of 27, 18 students know the answer to question number three on the last exam. If the instructor randomly chooses students, what is the probability that he will have to call on more than two students before he is given the correct answer?
- b. In early August 2009, approximately 60% of Americans were in favor of a public option for health care. In a random sample of 10 Americans, what is the probability that less than 4 are in favor of a public plan?
- c. A bowl with four pieces of black liquorish flavored candy and eight pieces of cherry flavored candy is passed around a party. If it is not possible to distinguish between the types of candy when selecting, what is the probability that the first four candies taken are all black liquorish?

- d. A student band sells songs on its website to raise money for their favorite charity. They sell an average of 22 songs each month. What is the probability of more than 30 songs being sold next month?

**Answers:**

- a. Geometric,  $p = \frac{2}{3}$ ,  $P(X > 2) = 1 - P(X \leq 2) \approx 1 - 0.8889 = 0.1111$   
b. Binomial,  $p = .6$ ,  $n = 10$ ,  $P(X < 4) \approx .0548$   
c. None of the distributions learned in this chapter can be used. Each trial does not have the same probability of success and the trials are not independent.  
d. Poisson,  $\lambda = 22$ ,  $P(X > 30) = 1 - P(X \leq 30) \approx 0.0405$



---

## 2.5 Normal Distribution

---

### The Standard Normal Probability Distribution

#### Project: The History of Distributions

When students are able to place what they are learning in mathematics in historical context, the material becomes much more interesting. Students like to hear stories about people and places. The following research topics will help them relate to the mathematics and mathematicians. The information they gather can be written up in a report, presented to the class with visual aids like PowerPoint, or both.

#### Assignment One:

Each student, or group of students, should be assigned one of the distributions studied in class, and research the following topics.

- What mathematician(s) or statistician(s) developed the distribution? Give a short biography of their life or lives.
- Describe the time period when the distribution was developed. What historic events happened around that time? What other scientific or mathematical discoveries were made?
- How was the distribution discovered? Was there a particular need or topic of inquiry that led to its discovery?
- How is the distribution related to other distributions?

#### Assignment Two:

There are an amazing number of different distributions in use. Students could also find a completely new distribution to research. In this case they can include the following topics as well as the ones mentioned above.

- What are the characteristics and parameters of the distribution?
- Give some examples of situations where the distribution is used to calculate probabilities.
- What is the shape of the distribution? Display graphs of the distribution with different values of the parameter(s).

---

### The Density Curve of the Normal Distribution

#### Project: Analyzing Normally Distributed Data (with Excel)

This project makes use of important concepts and skills that have been developed in the previous chapters. It gives the students a chance to bring many topics together, use technology, and work with real data.

#### Procedure:

- a. Think of a data set that is approximately normally distributed and for which you can gather at least 30 elements.
- b. Enter the data into a spreadsheet and calculate the mean and standard deviation of the data set.
- c. Calculate the  $z$ -score for each element of the data set and plot the  $z$ -score against the data values.
- d. Make a histogram of the data. Use small bin widths so the histogram appears somewhat smooth.

- e. Draw in an approximation of the density curve. Mark the mean and inflection points.
- f. Calculate the percent of the data that lies within one standard deviation of the mean. Repeat the process for two and three standard deviations.

**Analysis:** Address the following topics in writing.

- Describe the shape of the distribution. Hopefully it is approximately normally distributed, but describe where it deviates from the idealized normal curve. Is it a bit skewed? Are there any outliers? Where is the symmetry off? ...
- Use the normal probability plot to analyze how well the distribution approximates a normal distribution.
- Are the inflection points of the distributions located one standard deviation away from the mean? Use the data to explain any discrepancies.
- Compare the percents calculated in step six to the Empirical Rule. Use the data to explain any discrepancies.

---

## Applications of the Normal Distribution

### Explore: Area Under a Continuous Distribution

The probability of attaining a certain value of a discrete random variable is equal to the area of the rectangle over that value in the probability distribution. With a continuous distribution, an interval is substituted for the discrete value. This creates an interesting anomaly when finding the probability of one exact value. Take example four in the text concerning the height of twelve year old boys in Britton. What is the probability of randomly selecting a boy with height exactly 155 cm? Here the *exactly* is taken very seriously. It is 155 cm, not 154.99999999 cm or 155.0001 cm. This probability would correspond to an area with height given by the normal density function and width zero, so the probability would be zero. There must be some interval or range of acceptable heights in order to calculate a probability other than zero.

### Procedure:

- a. Use the normal distribution to find the probability of randomly selecting a boy between 154.9 cm and 155.1 cm. Here we are allowing a tolerance of 0.1 cm.
- b. Use the normal distribution to find the probability of randomly selecting a boy between 145.5 cm and 155.5 cm. Here we are allowing a tolerance of 0.5 cm.
- c. Use the normpdf function on your calculator to find the height of the normal density function at  $x = 155$  cm. Compare this result with your answer in #2. Explain this relationship.

### Answers:

- a. 0.009
- b. 0.045
- c. 0.046, The numbers are extremely close since the width of the area is one. The slope of the normal curve is not decreasing at a constant rate though, so the probabilities are not exactly the same.

The interval becomes more important when using integrals to calculate probabilities with other continuous distributions in more advanced probability classes. Similar modifications are made when using the normal approximation to the binomial distribution with the correction for continuity. Introducing student to the concept now will give them an advantage later.

## 2.6 Planning and Conducting an Experiment or Study

### Surveys and Sampling

#### Activity: The Affect of Sample Size on Sampling Error

Students will intuitively understand that increasing the sample size will produce a sample that better represents the population. This activity will help them to quantify this relationship, and prepare them to learn about the role sample size takes in determining the reliability of results in later chapters.

#### Procedure:

- Give a penny to each student in the class.
- Have each student flip the penny ten times and record the number of heads. After they are finished they can write the number on the board.
- Discuss the variation in the sampling errors. What was the largest error? How many students got the expected value of five heads? If the expected value was not known, how could the sampling error be determined?
- Combine all of the results into one total and calculate the proportion of heads for the class. How close is this to the expected value of 0.5?

#### Discussion and Research:

This would be a good opportunity to discuss the difference between classical and empirical probability, and how they are related in the Law of Large Numbers. In this case, the classical probability states that 50% of the flips of a fair coin will land heads, and the empirical data is the actual numbers of heads produced when the coin is flipped. The Law of Large Numbers states that the more times the coin is flipped the closer the empirical probability will come to the classical probability. An interesting extension is to apply this logic to gambling. What is the classical probability of winning any game in a casino? What does this say about your chances of winning?

#### Project: Survey a Representative Sample of the School

By putting to use what they have learned in a safe, familiar environment, students can solidify and gain insight into their new knowledge. This will be a fun project that allows students to learn about their classmates and themselves.

**Objective:** Use the techniques learned from this section to collect a sample that represents your school and conduct an unbiased survey.

#### Procedure:

- Choose a sampling method described in this section that will produce a representative sample of your school. The sample size,  $n$ , should be approximately five percent of the population.
- Administer a survey to all the members of your sample.
- Record the data you collected in a spreadsheet. Calculate appropriate summary statistics for the data.
- Make a graph that provides a good visual representation of the data.

**Analysis:** Address the following topics in writing.

- Describe the sampling method you used and why you chose it? What was challenging about taking the sample? How did you get the sampling frame? Why do you think this sample represents your school?
- Describe the survey you administered. How did you avoid the different types of bias described in this section?
- Describe the shape, center, and spread of the data set you collected. What conclusions can be made from the work that you did?

---

## Experimental Design

### Project: Conduct an Experiment

Designing and conducting an experiment will give the students practical knowledge in the field. The goal is for them to make their experiment as close as possible to a randomized clinical trial. It is essential that the treatments are randomly assigned. Repetition may not be possible.

**Objective:** Use the techniques learned from this section to conduct an experiment where a treatment is randomly assigned to participants so that a cause and effect relationship can be determined.

**Procedure:**

1. Chose a cause and effect relationship that you can practically and ethically test with an experiment.
2. Randomly assign the treatment(s) to different groups of participants.

Note: Participants do not need to be randomly selected to be in the experiment, they just need to be randomly divided into treatment groups.

3. Apply the treatments. Use a placebo to make the experiment blind. Make it double blind if possible. Use blocking when needed.
4. Record the results in a spread sheet. Calculate appropriate summary statistics.
5. Make a graph that provides a good visual representation of the data.

**Analysis:** Address the following topics in writing.

- Describe the possible effects of any confounding or lurking variables. How did you minimize, or eliminate these effects.
- Did you get the results you expected? Were the differences between the treatment groups large enough to be significant?

---

## Experimental Design

### Research and Discuss: Experimental Ethics

Many fields depend on experimental data. Pharmaceutical companies use experiments when developing new drugs, as so do cosmetic companies when creating new beauty products. Psychologists are famous for conducting behavioral experiments. The subjects of these experiments are often animals, including people. There is a wide variety in the opinions on what is, or is not, ethical treatment of the subjects of these experiments, and when the knowledge gained by the experiment justifies the discomfort, trauma, or death of the participant.

**Research Topics:**

4. Find examples of experiments done on people that are considered unethical by the current standards of our society. Include examples from the fields of psychology and medicine, and example from different time periods in history. How did humanity benefit from these experiments?

5. Find examples of experiments done on animal, other than humans, that may be considered unethical. How did humanity benefit from these experiments?
6. Different organizations have developed guidelines for what constitutes an ethical experiment. Find examples from a variety of groups including psychologists, medical doctors, pharmaceutical companies, animal rights groups, governments, and any other relevant group.

**Discussion Topics:**

3. Compare and contrast the different philosophies on ethical experiments. How have these opinions evolved over time?
4. When is the cost to the individual subjects of the experiments justified by the benefit of the results to society?

**Procedure:**

Assign individuals or groups of students to different research topics. Have them present what they found in class to stimulate discussion.

## 2.7 Sampling Distributions and Estimations

### Sampling Distribution

#### Activity: Making a Sampling Distribution

In this activity students will create several sampling distributions. The concept of a sampling distribution typically takes students awhile to grasp. Stress that the distribution includes *all* possible samples of a specific size.

#### Procedure:

1. Dived the class into groups of five or six students. Each group represents a population to be studied. The parameter of interest could be the average number of siblings of each student in the group.

Now the student will make sampling distributions for different sample sizes.

2. First make a sampling distribution with sample size one. This will be a dot-plot with six values. One for the number of siblings of each student in the group.

3. Now use a sample size of two. Look at all possible combinations of two in the group. Find the average number of siblings of each combination of two and graph it on a separate dot plot.

4. Continue by making a new dot plot with samples of size three, then four, and so on until there is just one sample that contains all the members of the group. Use the combination formula to make sure you have found all the possible samples of each size.

#### Analysis:

- Calculate the mean of each sampling distribution.
- Calculate the standard deviation of each sampling distribution, otherwise known as the sampling error, using the formula  $s = \sqrt{\frac{PQ}{n}}$ .
- How do the shape, center, and spread of the sampling distributions change as the sample size increases?
- If just one sample where to be taken randomly from each distribution, which one would most likely have mean closest to the true population mean.

### The z-score and the Central Limit Theorem

#### Discuss and Explore: The Relationship between Sample Size, Sampling Error, and Probability

This process will give students the opportunity to visually and quantitatively see the effect of sample size on sampling distributions and see the effect of sample size on the reliability of estimates of population parameters taken from samples.

#### Discussion:

Ask the students to consider the formula  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ . What happens to  $\sigma_{\bar{x}}$  as  $n$  gets large?

#### Explore:

- Use the calculator to graph a normal distribution with mean 100 and standard deviation 35.
- Then graph sampling distributions for this normal distribution with sample size 3, 6, and 12. The sampling distribution will also be a normal distribution with the same mean as the original distribution and standard deviation given by the above formula.
- Ask the students to compare the graphs.
- Calculate the probability of randomly selecting a member of the population with a value less than seventy.
- Calculate the probability of randomly selecting a sample of three with a sample mean less than seventy.
- Calculate the probability of randomly selecting a sample of six with a sample mean less than seventy.
- Calculate the probability of randomly selecting a sample of twelve with a sample mean less than seventy.

### Analysis:

Ask the student to explain the affect sample size has on the shape, center, and spread of the sampling distribution and its affect on the probability that the sample mean is close to the population mean.

## Binomial Distributions and Binomial Experiments

### Extension: The Normal Approximation to the Binomial Distribution

The binomial distribution is discrete; the probability for each possible value must be calculated separately. A complex event that contains many outcomes would be tedious to calculate manually using the binomial formula. It could be done with technology as shown in the text, or with the normal approximation to the binomial distribution. This will be a brief introduction to the latter method.

### Conditions and Formulas:

- The binomial distribution is only close enough to a normal distribution if both  $np$  and  $nq$  are greater than 10.
- If both of these conditions are met, a  $z$ -score can be calculate with  $\mu = np$ , and  $\sigma = \sqrt{np(1-p)}$ .
- A correction for continuity must be made on the value of  $x$  so that the entire discrete bar can be included or excluded as the situation requires. This will usually involve adding or subtracting 0.5 from the  $x$ -value.

### Example:

In early August 2009, approximately 60% of Americans were in favor of a public option for health care. In a random sample of 500 Americans, what is the probability that less than 200 are in favor of a public option?

Answer:

Step One – Check

$500 * 0.6 > 10$  and  $500 * 0.4 > 10$  It is appropriate to use the normal approximation because both these statements are true.

Step Two

$$\mu = 500(.6) = 300, \sigma = \sqrt{500(.6)(.4)} \approx 11, x = 200 - 0.5 = 199.5$$

Step Three

$$z = \frac{200 - 300}{11} = -1.82$$

## Step Four

$$P(x < 30) \approx \text{normcdf}(-999, -1.82) \approx .0344$$

## Confidence Intervals

### Project: Confidence Intervals in the News

Students become much more interested in a topic and motivated to learn about it when they see applications for the topic outside of the classroom. Statistics is prevalent in our daily lives, and many examples can be found in the news.

**Objective:** Collect and analyze examples of confidence intervals used in the reporting of news stories.

**Guidelines:**

- Look for examples of confidence intervals in newspapers, news magazines, television broadcasts, and in news stories covered online. Cite your sources.
- Find examples from a variety of areas. Include science, politics, weather, updates on the war, or other topics of interest to you.
- Identify the confidence level, the margin of error, and interpret the meaning of the confidence interval for the given situation.

**Note:**

This assignment should extend over a large period of time. Ideally, students will spend the entire length of the assignment on the lookout for confidence intervals. It will also take some time to get confidence intervals in a variety of subject areas.

This could be a written report, a presentation made to the class, or could take the form of a poster that will decorate the walls of the classroom. If time is available, the presentations are preferable since they are easy to grade and allow all the students to benefit from the work of their peers.

## Sums and Differences of Independent Random Variables

### Practice and Extend: Sums of Independent Random Variables from Normal Distributions and from Binomial Distributions

Students have already learned how to use the normal distribution and the binomial distribution to calculate probabilities. These exercises review these old skills, and give the students the opportunity to see more examples of sums of independent random variables.

**Exercises:**

1. A college gives an entrance exam with both a math and writing section. The math scores are normally distributed with a mean of 500 and a standard deviation of 35. The writing scores are also normally distributed with a mean of 485 and a standard deviation of 50. If the scores are independent and a student is randomly selected, what is the probability that the sum of her math and reading score is higher than 1020?

Answer: The sum of two independent random variables, both from normal distributions, also has a normal distribution with  $\mu_{X+Y} = \mu_X + \mu_Y = 500 + 485 = 985$  and standard deviation  $\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{35^2 + 50^2} \approx 61$ . Therefore,  $P(X + Y > 1020) = \text{normcdf}(1020, 100,000,000,000, 985, 61) \approx 0.2831$

### 2.7. SAMPLING DISTRIBUTIONS AND ESTIMATIONS



2. Oscar plays little league baseball, and has a batting average of 0.235. This means he gets a hit 23.5% of the times he is at bat. Oscar has two games this weekend. If he is at bat 5 times in Saturday's game, and 4 times in Sunday's game, what is the probability that he will get more than three hits this weekend? Assume the number of hits he gets in each game is independent.

Answer: The sum of two independent random variables, both with binomial distributions, also has a binomial distribution.

Let  $X$  be the number of hits in Saturday's game.  $X \sim B(5, .235)$

$Y$  be the number of hits in Sunday's game.  $Y \sim B(4, .235)$

Then  $X + Y \sim B(9, .235)$ . Therefore  $P(X + Y > 3) = 1 - P(X + Y \leq 3) = 1 - (P(0) + P(1) + P(2)) \approx 0.3573$

---

## Student's t Distribution

### Practice: t Distribution, Standard Normal Distribution, or Neither

For each of the following situations determine if a confidence interval could be calculated using the standard normal distribution, the t distribution, or if the requirements of neither are met. Explain your reasoning.

1. A sample of 50 tomatoes is taken from a field. The sample had a mean weight of 120 grams with a standard deviation of 20 grams.

Answer: The standard normal distribution should be used because the sample size is large. With a large sample the shape of the population distribution and the fact that the population standard deviation is not known is irrelevant.

2. A college instructor analyzes all the midterm scores for biology 101. The scores are normally distributed with a mean of 72 and a standard deviation of 13. One class of twelve scored exceptionally high with an average of 92. He will treat this small class as a sample.

Answer: The standard normal distribution should be used. Even though the sample size is small, the population is normally distributed and the population standard deviation is known.

3. The time a dog spends in a shelter before being adopted is approximately normally distributed. The Mountain View shelter found homes for 5 dogs last month. The mean time these dogs spent in the shelter was 4 months, with a standard deviation of one month.

Answer: The  $t$  distribution should be used. The sample is small, the population standard deviation is not known, and the population is approximately normally distributed.

4. Five students compare their recent test scores. Their scores have an average of 81, and a standard deviation of 6 percentage points.

Answer: Neither of the distributions can be used. The shape of the population distribution is not known, nor is the population standard deviation.

## 2.8 Hypothesis Testing

### Hypothesis Testing and the P-Value

#### Extend: A Null and Alternative Hypotheses for Any Situation

The nature of the situations that is being tested determines if a right-tailed, left-tailed, or two-tailed hypothesis test is used. These exercises allow the student to consider these types of tests in a different, creative way. They will look at the types of test available, and think of situations where each can be used.

**Directions:** For each set of null and alternative hypotheses chose a number for the variable and describe a situation where this type of test would be appropriate.

#### Example:

1.  $H_0 : \mu = a$

$H_a : \mu < a$

Solution: Let  $a = 12$  oz. This null and alternative hypothesis can be used for testing the mean amount of soda found in all soft-drink cans produced in a specific factory. The consumer is interested in finding out if they are receiving the full amount of soda that they paid for, or if the soda producers are cheating them.

#### Exercises:

a.  $H_0 : \mu = a$

$H_a : \mu \neq a$

b.  $H_0 : \mu = a$

$H_a : \mu > a$

c.  $H_0 : \mu \leq a$

$H_a : \mu > a$

d.  $H_0 : \mu \geq a$

$H_a : \mu < a$

Note: The null hypothesis must have the “equal” part of the two choices.

### Testing a Proportion Hypothesis

#### Project: Hypothesize About Your School

Students will put their new knowledge to work by applying it to their high school. They will also review, and put to use, sampling techniques learned previously in the class. This project will be a fun way to make statistics real, and to make the material relevant, active, and long lasting.

**Objective:** Test, and construct a confidence interval for, a hypothesis about a proportion that describes the population of your school.

#### Procedure:

#### 2.8. HYPOTHESIS TESTING

- Write a null and alternative hypothesis with a proportion about the students at your school.
- Take a sample. Be sure to use proper sampling techniques to get a sample that represents the population. Review chapter six if necessary.
- Calculate the sample proportion.
- Test the hypothesis at probability level 0.05.
- Construct the 95% confidence interval for the population proportion using the sample proportion that you found.

**Analysis and Conclusion:**

Write a report and/or prepare a presentation for the class that explains your null and alternative hypotheses, your sampling method, and the results of your test.

Analyze the results: Were you correct? Why or why not?

How confident are you in the method you used to gather your sample?

If you were to do this over again, what would you change?

How could you improve your accuracy?

Is the hypothesis test or the confidence interval more useful for your purposes?

Is there anything else to take into consideration when interpreting your results?

---

## Testing a Mean Hypothesis

**Extension: Comparing a Subgroup to the Whole**

The text has addressed two uses for hypothesis tests. The first is to test a claim. A statement is made about the mean of some measurement made on a population, and then that claim is tested by comparing it to a sample from that population. The second is to see if a significant difference can be found between the mean of a subgroup and the mean of the group as a whole. In this second application, both means are known, and are compared with the hypothesis test. In this assignment, students will explore the latter case by writing exercises, and by doing research so they can perform this test with real data.

**Part One: Writing Exercises**

- Think of three situations where it would be useful to compare a subgroup to the whole with a hypothesis test. Each exercise should be from a different context. Think about science, politics, education, social justice, sports, and other areas. Be creative.
- Write out an exercise, like what would be presented to a student, for each situation. Choose reasonable numbers for the problem. Be sure to include all the necessary information to complete the test.
- Provide complete solutions to these three exercises.

**Part Two: Research and Apply**

- Choose a situation where you can find real data to compare a subgroup to the whole with a hypothesis test. You may have to research a few possibilities to find one where you can get all the necessary data. Make a list of the quantities you will need.
- Write out the null and alternative hypotheses. Conduct the test, and interpret the results.
- Create a report and/or presentation for the class. Be sure to cite the source(s) of your data.

Include the following:

How did you choose the significance level?

Were these the results you were expecting?

How reliable is the data you collected?

What other tests could be made to expand on what you discovered?

---

## Testing a Hypothesis for Dependent and Independent Samples

### Project: Test for Significance in Experiment Results

For this project, students will perform an experiment, thereby actively reviewing proper experimental technique, and analyze the results with a hypothesis test. Because they are working with data that they produced, from a topic that is of interest to them, the learning will be deep and long lasting.

**Objective:** Conduct a controlled experiment and determine if the results are statistically significant.

**Procedure:**

- a. Plan, and execute a controlled experiment. Review chapter six to ensure that your experiment meets all the guideline of a clinical trial, except for repetition.
- b. Conduct the proper hypothesis test at the 0.05 significance level to determine if the results of the experiment are statistically significant.

**Analysis:** Include the following in a written report and/or presentation.

- a. Describe your experiment and how it meets the standards of a clinical trial (without repetition). Was your experiment blind or double blind? Did you use a placebo?
- b. Are the results of your experiment in the form of a mean or a proportion?
- c. Are the two groups dependent or independent?
- d. Did you use the t distribution or the normal distribution? Why? What is the critical value?
- e. State the null and alternative hypotheses.
- f. Calculate the standard error of the difference between the two groups. Show clear, organized work. Carefully chose the appropriate method.
- g. Calculate the test statistic. Show clear, organized work.
- h. Will you reject or fail to reject the null hypothesis? Why?
- i. Interpret the results of the test in the context of the experiment. What have you learned from the experiment and test?

## 2.9 Regression and Correlation

### Scatter Plots and Linear Correlation

#### Project: Finding Sets of Bivariate Data with Different Types of Relationships

In this project students will explore bivariate data. They will consider possible relationships, visual representations of the data, and gain experience with correlation coefficients including their strengths and weaknesses as describers of the relationship. The students will also practice using Excel, an extremely powerful and useful tool for analyzing data.

**Objective:** To find four sets of bivariate data with different relationships, graph each set of data, and calculate each correlation coefficient.

#### Procedure:

1. Find or collect four sets of bivariate data each with at least fifteen pairs. Find one set with each type of relationship listed below. Cite the source of your data or describe the collection method.

- a positive linear correlation
- a negative linear correlation
- no correlation (or very close to none)
- a curvilinear relationship

2. Use Excel to make a scatter plot for each set of data. Label the axes and title each plot.

3. Use Excel to calculate the correlation coefficient for each set of data.

**Analysis:** Include the following analysis with the work you did on Excel in a written report and/or PowerPoint presentation.

- a. Describe the strength of each relationship. Is it what you expected it would be?
- b. For which sets of data is the correlation coefficient an accurate measure of the strength? Discuss why it would, or would not be an accurate descriptor in each case.
- c. Do you believe there is a causal relationship between the two variables in each set of data? Why or why not?

### Least-Squares Regression

#### Project: Calculating and Analyzing the Least-Squares Regression Line (continued from project for the previous section)

**Objective:** To find, analyze, and use the regression line for the data sets found in the previous project.

#### Procedure:

1. Use Excel to calculate the slope and y–intercept of the least-squares regression line for the set of data with positive correlation in the project for the previous lesson.

2. Graph the least-squares regression line over the scatter plot made in the previous project.

3. Use Excel to calculate the residual for each point. Find the sum of the residuals. Is it what you expected it to be?
4. Make a residual scatter plot on Excel and use to identify outliers.
5. Decide if you would like to eliminate any outliers from your set, and recalculate the equation of the least-squares regression line if necessary.
6. Use the least-squares regression line to make three predictions.
  - For the first prediction, use a value of the predictor variable that is inside the range of data you collected, but for which you have no value. This is called interpolation.
  - For the second prediction, use a value of the predictor variable that is above the range of data you collected. This is called extrapolation.
  - For the third prediction, use a value of the predictor variable that is below the range of data you collected. This is also called extrapolation.

Do you think that interpolation or extrapolation is more accurate? Why?

7. Repeat the process for the set of data with negative correlation in the project for the previous lesson.
8. Consider the data with the curvilinear relationship of the previous project. Is it possible to apply a transformation to achieve linearity? Play around with the data and see what you can do.

---

## Inferences about Regression

**Project: Hypothesis Testing and Confidence Intervals for the Regression Coefficient (continued from projects for the previous two sections)**

**Objective:** To analyze the reliability of the regression coefficient calculated in the previous project with a hypothesis test, and to make a confidence interval for the regression coefficient.

**Procedure:**

1. Use the positively correlated data from the previous project and conduct a hypothesis test on the regression coefficient at the 0.05 significance level. Follow the steps below.
  - State the null and alternative hypothesis.
  - Calculate the test statistic using Excel. Recall that the standard error of estimate is calculated as follows:
 
$$s_{y \cdot x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$
 where  $\hat{y}$  is the value of  $y$  predicted by the least-squares regression line for each value of  $x$ , and  $y$  is the actual value in the data.
  - Find the critical value using the  $t$  distribution and  $n - 2$  degrees of freedom.
  - State the conclusion of the test and interpret the results in the context of the data.
2. Repeat the process using the negatively correlated data, and then again using the data with little to no correlation.
3. Construct a 95% confidence interval for the regression coefficient of the least-squares regression line for the positively correlated data by using the following formula.
 
$$b \pm t S_b$$
 where  $b$  is the regression coefficient calculated in the previous project, and  $t$  is obtained from the  $t$  distribution table for  $\frac{\alpha}{2}$  area in the right tail of the  $t$  distribution and  $n - 2$  degrees of freedom
4. Repeat using the negatively correlated data, and then again using the data with little to no correlation.
5. Analyze the results. Are these the outcomes you expected? Do they make sense? Why or why not?

## 2.9. REGRESSION AND CORRELATION

---

## Multiple Regression

### **Project: Calculating and Analyzing the Multiple Regression Equation for Student Collected Data Using Excel**

**Objective:** To use Excel to calculate the multiple regression equation for data you have collected and to analyze the contribution of each variable to the relationship.

#### **Procedure:**

1. Think of a relationship for which you can gather data where one variable is determined by at least four predictor variables. Use a sample with at least fifteen ordered pairs. Cite your source or describe your collection method.
2. Enter your data into Excel and use the Data Analysis tools to calculate the regression statistics.

#### **Calculate the Multiple Regression Equation**

3. Write the regression model and interpret the regression coefficients.
4. Use the test statistic for the for each predictor variable to decide if it should be used in the regression equation. Eliminate variables that do not significantly contribute to the variance of the outcome variable, and recalculate the equation if necessary.

#### **Hypothesis Testing**

5. State the null and alternative hypothesis for the  $R$  value.
6. What is the  $F$  – statistic and the associated probability for your data?
7. State and interpret the results of your test.

#### **Confidence Interval**

8. Find the 95% confidence interval for each variable still in the regression equation.

#### **Predict**

9. Use your final regression equation to make some relevant predictions.

## 2.10 Chi-Square

### The Goodness-of-Fit Test

#### Project: The Frequency of First Digits

In this project students will see the results of Benford's Law. Benford's Law is a popular mathematical curiosity that the students will enjoy. It states that the first digits of numbers are not uniformly distributed. There are far more ones than nines. What is happening is that it is the logarithm of the first digits that is distributed uniformly. This project will give students the opportunity to use the chi-square goodness-of-fit test, and it will expose them to a counterintuitive mathematical law that may catch their interest and prompt them to explore some mathematics independently.

**Objective:** Perform a goodness-of-fit chi-square hypothesis test on a set of data to determine if the first digits are uniformly distributed over the set  $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ .

#### Procedure:

- Collect data and record the frequency of each possible first digit 1 to 9. Any source of data with a large range can be used. An atlas or almanac will make a good source of data. Populations of countries or cities, lengths of rivers, atomic weights, addresses, or every number in an addition of the newspaper are all good examples of sets of data that can be used. Collect a data set with at least 100 elements. Cite the source of your data.
- Calculate the expected frequency of each first digit. One would expect the first digits to be evenly distributed among the numbers 1 to 9.
- State the null and alternative hypotheses for your data.
- Use the chi-square distribution table to write a rule for rejecting the null hypothesis at the 0.05 significance level.
- Calculate the chi-square statistic to compare the observed and expected frequencies.
- Determine whether to reject or to fail to reject the null hypothesis.
- Write a summary statement based on the results of your test.

### Test of Independence

#### Extension: Statistics in the Social Sciences

The chi-square test of independence is frequently used in the social sciences to access if two factors are related. Being exposed to some of the many useful and widespread applications of this test and to applications of statistics in general, motivates students to learn and remember the material in this course.

#### Research:

Students can look for examples of research done in the social sciences that make use of the chi-square test of independence. These can be found in advanced text books, scientific journal, or by searching the internet. The students can focus on the social science that most interests them. Some of the possible areas that can be explored are psychology, politics, education, or cultural studies.



Many college majors require a basic statistics course. Statistics is, in fact, a more common requirement than calculus. Students can choose a college or university that interests them and look at which degrees and majors require a basic statistics class and which require more advanced work in statistics. This could be done in conjunction with the counseling department. Students can make a display or short presentation to share what they found with the other students at their school. This will encourage career planning and goal setting in the student population and increase student interest in the school's statistics program.

**Application:**

Students can design and conduct an experiment or survey in a social science area and use the chi-square test of independence to analyze the results. Psychological experiments are often of interest to students, but any area of social science can be used. Proper experimental and survey techniques can be reviewed in chapter six of this text. Excel can be used to perform calculations and display data. Students can write a report and/or give a presentation to the class explaining their experiment or survey and interpreting what the results of the chi-squared test of independence implies for their particular topic of inquiry.

---

## Test One Variance

**Extension: The Chi-Squared Tests in Biology**

The three chi-squared tests studied in this chapter are often used in biological studies. This would be a good time to do a cross-curricular project with the biology department. Biostatistics is an emerging field of study. Here are some ideas of how statistics and biology could be studied together.

**Research:**

Students can look for examples of biological studies that make use of any of these chi-squared tests. These might be found in an advance biology or biostatistics texts, scientific journals, or by searching the internet. The students could write a short analysis of the study and/or bring it in to share with the rest of the class.

Students can find universities that offer degrees in biostatistics and examine the programs that they offer. They should look at the following areas:

- Who are the professors and what are their areas of expertise?
- What classes are required for the degree? What are some of the electives?
- What are the prerequisites of the program?
- What are the typical standardized test scores and grade point averages of students admitted to the program?
- What are students that completed the degree doing now?

**Application:**

Students can design and conduct a biological study that makes use of some, or all, of the chi-squared tests studied in this chapter. This would be best done in conjunction with a biology class. Usually students currently taking statistics have already completed a basic biology class, but may currently be in an advanced or AP biology course.

## 2.11 Analysis of Variance and the F-Distribution

### The F-Distribution and Testing Two Variances

#### Project: The History of Statistics

Some students can happily work in the abstract world of numbers and symbols indefinitely, but most need and appreciate seeing the human side of statistics along with its theory and applications. Statistics has a rich history with many interesting characters for students to explore. Linking a statistical method to a memorable story or person will improve the student's ability to retain what they have learned in the course.

**Objective:** Find and report on the history of the development of the  $t$  distribution and the  $F$  distribution.

**Procedure:** Answer the following questions for each distribution in a written report and/or a class presentation.

- When was the distribution first used?
- Who discovered the distribution?
- What motivated the distributions discovery?
- How did the distribution get its name?
- What other notable work was taking place in science and mathematics at the time the distribution was first used?
- What was happening historically, politically, and in the art world at the time the distribution was first used?
- What developments have been made concerning the distribution since the time of its first use and who made these developments?
- What new uses have been found for the distribution since the time of its first use and who found these applications?

### The One-Way ANOVA Test

#### Explore: Assessing Variance with the F-test and ANOVA When Data is Approximately Normally Distributed

It is sometimes difficult for students to remember all the requirements and sensitivities of a particular test. Students have learned two tests that compare the variances of two sets of data. The  $F$ -test is only accurate when the data is normally distributed, but the ANOVA test is not as sensitive to small deviations from normality. By applying both of these tests to data that meets the normal requirement and to data that does not meet the strict normal requirement, students will be able to experience the magnitude of the discrepancy in the results. The experience will help them to remember the normal requirement for the  $F$ -test.

**Objective:** To compare the results of the  $F$ -test and the ANOVA test when analyzing the variance of data that is normally distributed and when analyzing data that is approximately normally distributed.

#### Procedure:

Part One: Normally Distributed Data

1. Take two samples from a population you know to be normally distributed. For example, you can use the heights of women and the heights of men since you know height to be normally distributed.
2. Perform an  $F$ –test on the variances of the two data sets.
3. Perform an ANOVA test on the variances of the two data sets.
4. Compare the results.

Part Two: Data that is approximately normally distributed.

5. Take two samples from a population that you know has an approximate normal distribution.
6. Perform an  $F$ –test on the variances of the two data sets.
7. Perform an ANOVA test on the variances of the two data sets.
8. Compare the results.

Analysis:

- (9) Analyze the normality requirement of the  $F$ –test using what you have learned from the tests just conducted.

---

## The Two-Way ANOVA Test

### Project: Technology Handbook

Throughout the course, students have used technology to perform statistical calculations. Excel and the TI-84 calculator have been used most frequently. In this assignment, student will gather and solidify their knowledge for their own future use and for the use of other students.

**Objective:** To write a clear, precise handbook that could be used by a person with no prior knowledge of the technology to perform statistical calculations with Excel and with the TI-84 calculator.

### Procedure:

1. Make an index.
  - Review what you have done with technology in this class and decide what should be included in your handbook.
  - Find a user friendly way to organize the information. Do you want to have an Excel section and a TI-84 section, or do you want to categorize by statistical topic? Maybe you have a different method to organize the material.
2. Write clear, concise directions so that each statistical calculation can be preformed with each technological tool.
  - Include pictures.
  - Ask someone who is not knowledgeable in this area to read over your work and identify areas that are unclear.

This project can be done in conjunction with the English department. Technical writing of this sort is a valuable skill for students to develop.

## 2.12 Non-Parametric Statistics

### Introduction to Nonparametric Statistics

#### Extension: Statistics in Literature

As the course comes to a close, it is nice to have a fun assignment that is a change of pace from what the students have been doing in class so far. The novel Bringing Down the House is an exciting story where what students have learned in this course figure prominently in the plot. It also makes statistics, math, mathematicians, and being smart look cool. This assignment will also satisfy any cross-curricular requirement of your school. Students can get credit for the essay in both their statistics and their English classes. The statistics teacher can grade the essay for content, and the English teacher can grade the same essay for the quality of the writing. Work with the English teacher(s) beforehand to come up with an assignment everybody likes.

**Assignment:** Read Bringing Down the House, by Ben Mezrich.

Write a 750 to 1000 word essay addressing one of the following topics.

- Is card counting cheating according to the members of the team, to the casinos, to the law?
- How do the players use math and stereotypes to count cards?
- How does card counting affect the lives of the members of the team? Is it an addiction?
- How does competition/greed affect the teams?
- If you can think of another good topic to address, go for it.

Include lots of detail and examples from throughout the book. Make sure I can tell you read the entire book carefully. Put page numbers in parenthesis at the end of quotes.

### The Rank Sum Test and Rank Correlation

#### Project: Analyzing Nominal and Ordinal Data

In the first chapter of this text students learned about the levels of measurement. Since then, the focus of their studies has been on data measured at the interval or ratio level. This last chapter returns to the lower levels of measurement, nominal and ordinal. Before students start working with the nonparametric tests in this chapter, it would be beneficial for them to review the levels of measurement found in the second section of Chapter One of this text.

**Objective:** To perform hypothesis tests on data measured at the nominal and ordinal levels that is collected by the students.

#### Procedure:

Part One: Data Measured at the Nominal Level

1. Take a random sample and divide the results into two categories.
2. Use a sign test at the 0.05 significance level to determine if one categorical variable is really “more” than the other.

- State the null and alternative hypotheses.
- Determine the critical value. Will you use the normal or  $t$ -distribution chart?
- Calculate the test statistic.
- Determine and interpret the results.

#### Part Two: Data Measured at the Ordinal Level

3. Collect two sets of data that you wish compared and rank the results.
4. Perform a rank sum test. Include all the hypothesis testing steps as above.
5. Interpret the results.

---

## The Kruskal-Wallis Test and the Runs Test

### Explore: Assessing Variance with the Kruskal-Wallis Test

#### Continued from the Enrichment Activity for the One-Way ANOVA Test

In the enrichment activity for the one-way ANOVA test, students used the  $F$ -test and the one-way ANOVA test to assess the variances of two data sets. They did each test twice, once with normally distributed data and once with approximately normally distributed data. In this activity, students will repeat the process with data that is not normally distributed at all, and also perform the Kruskal-Wallis test on all three data sets. By comparing the results of the tests, the students will be able to see the importance of considering the distribution of the data when deciding which test should be used in a particular situation.

**Objective:** To assess the variance of samples taken from three different populations with different distributions in three different ways in order to compare the results in terms of the robustness of the tests.

#### Procedure:

- a. Perform the Kruskal-Wallis test on the normally distributed data and the approximately normally distributed data collected in the enrichment activity for the one-way ANOVA test.
- b. Take two samples from a population that you know does not have a normal distribution.
- c. Perform the  $F$ -test on the data collected in step two.
- d. Perform the one-way ANOVA test on the data collected in step two.
- e. Perform the Kruskal-Wallis test on the data collected in step two.
- f. Compare the results of the three tests performed in the previous steps.
- g. Of all the tests done in both enrichment activities, which are giving dependable results and which are not because the test criteria have not been met?

