Probability and Statistics Teacher's Edition - Enrichment

CK-12 Foundation

December 29, 2009

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.

flexbook
next generation textbooks

Contents

1 Probability and Statistics TE - Enrichment **5**

1.1 An Introduction to Analyzing Statistical Data 5

1.2 Visualizations of Data 9

1.3 An Introduction to Probability 11

1.4 Discrete Probability Distributions 16

1.5 Normal Distribution 20

1.6 Planning and Conducting an Experiment or Study 22

1.7 Sampling Distributions and Estimations 24

1.8 Hypothesis Testing 28

1.9 Regression and Correlation 31

1.10 Chi-Square 34

1.11 Analysis of Variance and the F-Distribution 36

1.12 Non-Parametric Statistics 38

Chapter 1

Probability and Statistics TE - Enrichment

1.1 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.

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:

1. Each student should measure the length of their right ring finger in both centimeters and inches.
2. 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.
3. 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.

4. 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 figure 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:

1. 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?
2. Find examples of experiments done on animal, other than humans, that may be considered unethical. How did humanity benefit from these experiments?
3. 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:

1. Compare and contrast the different philosophies on ethical experiments. How have these opinions evolved over time?
2. 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:

1. 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.
2. Make a dot-plot of the data set. Label and title the plot.
3. Calculate the mean and median of the data set.
4. Mark the mean and median on the plot.
5. Write an analysis of the work you have done that addresses the following topics.

- 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:

1. 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.

2. The table should have three columns titled like those shown in the text of the lesson.
3. 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.
4. 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.
5. The second row of the third column will square the values in the second column.
6. 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:

1. 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.
2. Make a dot-plot of each data set. Label and title the plots.
3. Calculate the standard deviation of each data set using a table. Check your answer with your calculator.
4. 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.

1.2 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:

1. 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.
2. 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.
3. 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:

1. 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 form 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:

1. 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.)
2. Describe the shape, center, and spread of the data.
3. For each stem-and-leaf plot make a box-and-whiskers plot of the same data.
4. Does seeing the data displayed in a different way make you want to change your description of the data's shape, center, and spread?
5. 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.

1. Calculate the interquartile range for the data. Remove the outlier of 34 from the set and calculate the interquartile range again.
2. Calculate the standard deviation of this sample. Remove the outlier and calculate the standard deviation again.
3. Calculate the percent change for each measure of spread.
4. 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:

1. $11, 10$
2. $15.28, 7.54$
3. $9\%, 51\%$
4. 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

1.3 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:

1. 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 an number less than three} = $\{1, 2\}$

Exercises:

1. The experiment of drawing one card from a standard deck has 52 outcomes. Define two simple and two compound events.
2. 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.
3. 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. $A \cap B = \{\} = \text{the empty set} = \varnothing, A \cup B = \{\text{M, T, W, H, F, S, U}\} = \text{the sample space}$

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

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:

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

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

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

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.

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 ask their opinion on the new plan.

Table 1.1

	In Favor	Against	
Female	30	90	120
Male	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:

1. What does this imply about the healthcare plan?
2. If the probabilities in the respective definitions were approximately equal, would the events be almost independent? How close to equal do probabilities calculated form 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:

1. How would you take a simple random sample that represents this class? The school?
2. What sampling frames could be used to get s 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:

1. 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 form the school office for the students. This can be assigned to small groups who report their method and findings to the class.
2. 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.

1.4 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 accustom 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.

1. What are the possible values of the discrete random variable?
2. Calculate the probability of each possible result.
3. 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.
4. Calculate the area of each rectangular bar in the graph of the probability distribution. What is the sum of the areas?
5. What is the total area over the random variable values of two and three? What is the probability of getting two or three heads?
6. What is the relationship between probabilities and area in a probability distribution?

Answers:

1. $\{0, 1, 2, 3, 4\}$
2. $P(0) = \frac{1}{16}, P(1) = \frac{4}{16}, P(2) = \frac{6}{16}, P(3) = \frac{4}{16}, P(4) = \frac{1}{16}$

3. .
4. $\frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16} = \frac{16}{16} = 1$
5. $\frac{6}{16} + \frac{4}{16} = \frac{10}{16}, P(2 \text{ or } 3) = P(2) + P(3) - P(2 \text{ and } 3) = \frac{6}{16} + \frac{4}{16} - 0 = \frac{10}{16}$
6. 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:

1. Design a probability experiment that produces values of a discrete random variable.
2. Perform the experiment at least 20 times and record the results.
3. Use the collected data to calculate the probability of each value of the random variable.
4. Make a table and a graph for the probability distribution.
5. Calculate the mean, variance, and standard deviation for the probability distribution.
6. 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 to be constant form 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

1. Why isn't X a binomial random variable?

2. Use classical probability, combinations, and the Multiplicative Rule of Counting to calculate the probability that exactly two of the presenters are your close friends.

3. 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.

4. Use classical probability, combinations, and the Multiplicative Rule of Counting to calculate the probability that exactly two of the four close friends are chosen.

5. 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:

1. 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.
2. $\frac{({}_4C_2 * {}_{26}C_1)}{{}_{30}C_3} = \frac{156}{4060} \approx 0.0384$
3. ${}_3C_2 * (.2)^2(.8)^1 = .096$, not a good estimate
4. $\frac{({}_4C_2 * {}_{1196}C_1)}{{}_{1200}C_3} \approx 0.0000250$
5. ${}_3C_2 * \left(\frac{4}{1200}\right)^2 \left(\frac{1196}{1200}\right)^1 \approx 0.0000322$, this estimate is much better

The Poisson Probability Distribution

Practice and Extend: Changing Units for the Poisson Distribution The parameter λ 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, λ could be the average number of accidents at a given intersection in a year, but the probability to the calculated is for three accidents in a given month. It is an easy fix; λ 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:

1. 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?
2. 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?

3. 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:

1. (a) mean = standard deviation = $\frac{1}{3}$
   (b) $p(1) \approx 0.2388$
2. (a) $\lambda = 140, P(X > 150) = 1 - P(X \leq 150) \approx 1 - 0.8134 = .1866$
   (b) $\lambda = 140, P(X > 200) = 1 - P(X \leq 200) \approx 1 - 0.9999992528 \approx 0$
3. $\lambda = 60, P(X = 70) \approx 0.02160$

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.

1. 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?
2. 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?
3. 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?
4. 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:

1. Geometric, $p = \frac{2}{3}, P(X > 2) = 1 - P(X \leq 2) \approx 1 - 0.8889 = 0.1111$
2. Binomial, $p = .6, n = 10, P(X < 4) \approx .0548$
3. 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.
4. Poisson, $\lambda = 22, P(X > 30) = 1 - P(X \leq 30) \approx 0.0405$

1.5 Normal Distribution

The Standard Normal Probability Distribution

Project: The History of Distributions

When students are able to place what they are leaning 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 where made?
- How was the distribution discovered? Was their a particular need or topic of inquiry that lead 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:

1. Think of a data set that is approximately normally distributed and for which you can gather at least 30 elements.
2. Enter the data into a spreadsheet and calculate the mean and standard deviation of the data set.

3. Calculate the $z-$score for each element of the data set and plot the $z-$score against the data values.
4. Make a histogram of the data. Use small bin widths so the histogram appears somewhat smooth.
5. Draw in an approximation of the density curve. Mark the mean and inflection points.
6. 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:

1. 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.
2. Use the normal distribution to find the probability of randomly selecting a boy between 145.5 cm and 155.5 cm. Her we are allowing a tolerance of 0.5 cm.
3. 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:

1. 0.009
2. 0.045

3. 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.

1.6 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:

1. Give a penny to each student in the class.
2. 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.
3. 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?
4. 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 leaned from this section to collect a sample that represents your school and conduct an unbiased survey.

Procedure:

1. 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.
2. Administer a survey to all the members of your sample.
3. Record the data you collected in a spreadsheet. Calculate appropriate summery statistics for the data.
4. 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 leaned 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.

1.7 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:

1. Calculate the mean of each sampling distribution.
2. Calculate the standard deviation of each sampling distribution, otherwise known as the sampling error, using the formula $s = \sqrt{\frac{PQ}{n}}$.
3. How do the shape, center, and spread of the sampling distributions change as the sample size increases?
4. 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:

1. Use the calculator to graph a normal distribution with mean 100 and standard deviation 35.
2. 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.
3. Ask the students to compare the graphs.
4. Calculate the probability of randomly selecting a member of the population with a value less than seventy.
5. Calculate the probability of randomly selecting a sample of three with a sample mean less than seventy.
6. Calculate the probability of randomly selecting a sample of six with a sample mean less than seventy.
7. 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 at 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. 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 Sundays game. $Y \sim B(4, .235)$

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

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 is 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.

1.8 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:

1. $H_0 : \mu = a$
   $Ha : \mu \neq a$
2. $H_0 : \mu = a$
   $H_a : \mu > a$
3. $H_0 : \mu \leq a$
   $H_a : \mu > a$
4. $H_0 : \mu \geq a$
   $Ha : \mu < a$

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:

1. Write a null and alternative hypothesis with a proportion about the students at your school.
2. Take a sample. Be sure to use proper sampling techniques to get a sample that represents the population. Review chapter six if necessary.
3. Calculate the sample proportion.
4. Test the hypothesis at probability level 0.05.
5. 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

1. 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.
2. 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.
3. Provide complete solutions to these three exercises.

Part Two: Research and Apply

1. 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.
2. Write out the null and alterative hypotheses. Conduct the test, and interpret the results.
3. 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:

1. 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.
2. 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.

1. 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?
2. Are the results of your experiment in the form of a mean or a proportion?
3. Are the two groups dependent or independent?
4. Did you be use the t distribution or the normal distribution? Why? What is the critical value?
5. State the null and alternative hypotheses.
6. Calculate the standard error of the difference between the two groups. Show clear, organized work. Carefully chose the appropriate method.
7. Calculate the test statistic. Show clear, organized work.
8. Will you reject or fail to reject the null hypothesis? Why?
9. Interpret the results of the test in the context of the experiment. What have you learned from the experiment and test?

1.9 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 curviliniear 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.

1. Describe the strength of each relationship. Is it what you expected it would be?
2. 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.
3. 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 predication, 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: $s_{y*x} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$, 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.

$b \pm tS_b$, 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?

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.

1.10 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:

1. 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.
2. Calculate the expected frequency of each fist digit. One would expect the first digits to be evenly distributed among the numbers 1 to 9.
3. State the null and alternative hypotheses for your data.
4. Use the chi-square distribution table to write a rule for rejecting the null hypothesis at the 0.05 significance level.

5. Calculate the chi-square statistic to compare the observed and expected frequencies.
6. Determine whether to reject or to fail to reject the null hypothesis.
7. 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.

1.11 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.

1.12 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 texts 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 asses 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 asses 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:

1. Perform the Kruskal-Wallis test on the normally distributed data and the approximately normally distributed data collect in the enrichment activity for the one-way ANOVA test.
2. Take two samples from a population that you know does not have a normal distribution.
3. Perform the $F-$test on the data collected in step two.
4. Perform the one-way ANOVA test on the data collected in step two.
5. Perform the Kruskal-Wallis test on the data collected in step two.
6. Compare the results of the three tests performed in the previous steps.
7. 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?