Chapter Project

Swimming in an Ocean of Data

Data and information are everywhere. Data comes to us from the visible and audible world around us, as well as from the things we can physically touch. Our brains do an amazing job of processing this type of data. There are other types of data, large sets of numbers, graphs, and other information that we get from the news media, websites, and social media. Our brains do not handle those types of data as well without some instruction. Because we now live in a world where computers are constantly collecting data, we need more and more people to analyze and interpret that data.

Part 1: Investigate current career trends.

1. Go to the Bureau of Labor Statistics website to identify the fastest growing occupations, currently. https://www.bls.gov/emp/tables/fastest-growing-occupations.htm

   Identify which of these jobs focus on data analysis.
2. Do a quick web search for the U.S. News and World Report of the best jobs for this year. Looking at the top 25, list those jobs that focus on data analysis. https://money.usnews.com/careers/best-jobs/rankings/the-100-best-jobs
3. Choose one of the jobs listed in #1 or #2 above that focuses on data analysis and research more about it. Write a one paragraph synopsis of your research. Make sure to identify the average salary, what level of education is required, and if the positions for this occupation are expected to grow or decline. Cite any sources you use.

Part 2: Identify statistics in the media.

Find two examples of statistics in the media. You may use any news source, magazine, newspaper, or televised commercial.

For each example:

1. Print a copy or provide a link to the example.
2. Write a short synopsis of what you think the graphic or statistic is trying to illustrate.
3. Identify any portion of the graphic or provided statistic that you think may be misleading or inaccurate. Can you think of another way to illustrate or express this information that might make the authors purpose clearer?

Part 3: Research and identify other sources of Big Data.

Find two sources of Big Data you find interesting that are not mentioned in the textbook.

1. Cite the source where you found the data.
2. Write a brief explanation of why you find this data interesting and describe the usefulness of the information that could be obtained from analysis of this data.

Chapter Project

Birds of a Feather Evolve Together

"Over the past four decades, evolutionary biologists Rosemary and Peter Grant have documented the evolution of the famous Galápagos finches by tracking changes in body traits directly tied to survival, such as beak length, and identified behavioral characteristics that prevent different species from breeding with one another. Their pioneering studies have revealed clues as to how 13 distinct finch species arose from a single ancestral population that migrated from the mainland 2 million to 3 million years ago."

A video detailing the changes in these species that have evolved can be found at either of the following websites. View the video about this research prior to beginning the project.
http://media.hhmi.org/biointeractive/interactivevideo/finchquiz/?_ga=2.36754469.465667716.1528204380-868970706.1528204380 or https://www.youtube.com/watch?v=mcM23M-CCog

As you learn more about analyzing data, you will find that the type of data presented will dictate which types of graphs you will create, which types of statistics you can calculate, and which types of inference can be done. You will use a sample of the finch data, which can be found on stat.hawkeslearning.com, to identify types of variables. Definitions of the variables can be found on the next page.

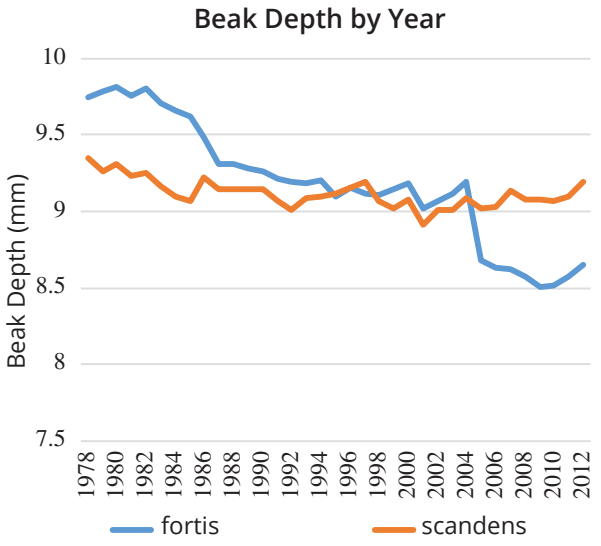
Data

The data can be found at stat.hawkeslearning.com **Data Sets > Finch Data.**

1. Were the Grants conducting an experiment or observational study on these finches?
2. If we were trying to determine the relationship between sex and weight, which variable would be the explanatory variable and which variable would be the response variable?
3. Identify each variable in the data set as:
   a. Qualitative or quantitative
   b. Continuous, discrete, or neither
   c. Nominal, ordinal, interval, or ratio
4. A second set of data the Grants collected is various averages per year (Finch by Year Data). This data can be used to show changes in the finches over time. The following graph highlights beak depth for two species of finches on the Galapagos from 1978 to 2012.

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Finch by Year Data.**



a. Does this graph represent time-series or cross-sectional data?

b. Is there a noticeable trend in the change of beak depth from 1978 to 2012 for the fortis species?

c. Is there a noticeable trend in the change of beak depth from 1978 to 2012 for the scandens species?

Source: https://www.hhmi.org/biointeractive/the-origin-of-species-the-beak-of-the-finch

Band refers to an individual's identity, more specifically, the number on a metal leg band it was given.

Species name is Geospiza fortis, which is the medium ground finch.

Sex is indicated as male, female, or unknown. The reason for the "unknown" category is that males start their lives looking like females. After one or more years they molt into a plumage with some black feathering that indicates they are males.

The variable **First adult year** refers to the year after the individual hatched from an egg.

The variable **Last year** refers to the last year of that individual's life. Fifty individuals did not survive beyond 1977, the year of the drought, whereas 50 survived to 1978 and later years.

The next six columns provide the morphological measurements of individuals in the group that died in 1977 and in the group that survived.

Weight is in grams; the other measurements are in millimeters.

Tarsus is a part of the leg.

Chapter Project

The Breakfast of Champions

A nutritionist at the Food and Drug Administration is studying the effects of cereal marketing on family meal choices. In particular, she would like to understand how cereal manufacturers market their products in grocery stores. She became interested in doing this study after noticing how the cereal was being restocked one day in her local grocery store. The store personnel were restocking the cereal shelves based on a reference sheet that told them where everything was to be placed. The placement of each cereal brand seemed very deliberate.

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Cereal Data.**

To gather data for her study, the nutritionist goes to the local grocery store and records data about cereal nutritional claims and shelf location for 77 cereals.

1. Identify the population and the sample in this scenario.
2. Consider the variables in the data set. Identify the variables that are qualitative and those that are quantitative.
3. Consider the variable *Shelf*. This variable is the shelf position of the cereal (bottom, middle, top) starting from the floor up. Based on your experience at the grocery store, do you think that the shelf position is related to the nutritional content of the cereal? Why or why not?

To see whether the shelf position is associated with one measure of nutritive value, the amount of sugar, look at the data for the variable *Sugars*. Compare the sugar content of cereals on each shelf by making a separate frequency table and histogram for the sugar content of the cereals on each shelf: a total of three frequency tables and three histograms. Use the sugar content values as they are – do not factor in the serving size. (The data for one of the cereals, Quaker Oatmeal, is missing. Just continue with what is available. That's the way it is in real life – values are missing, files are incomplete, etc.)

4. Make a frequency table for each shelf. (Hint: It might help to sort the variable *Sugars* by shelf location.)
   a. Use the same classes for all shelves.
   b. Use 6 classes.
5. Make a histogram for each shelf.
   a. Use graph paper and work neatly or use your calculator.
   b. Use the same scales for your histograms so you can compare the data easily.
   c. Title each histogram and label the axes.
6. Briefly describe the distribution in each histogram with respect to shape. Based on your histograms, which shelf position has cereals with the most sugar?
7. Consider your histograms for sugar content. Is the shelf position of a cereal related to its nutritive value as measured by sugar content? Explain your reasoning. What kinds of cereals are on each shelf?
8. What further data gathering would you recommend?
9. Does your analysis of the data generate any other questions?
10. Can you suggest how you might obtain data to answer these questions?

Source: https://www.kaggle.com/crawford/80-cereals/data

Chapter Project

A Day in the Life of a College Student

College students' physical, emotional, and mental health are at the forefront of many national discussions and statistical studies. Many factors can influence students' overall health, including diet, sleep, exercise, etc. In this project, you will look at real data gathered from 30 college freshmen and sophomores and determine related descriptive and inferential statistics.

The students surveyed were asked the following questions:

On a typical weekday,
How many hours of sleep do you get?
How many hours do you study?
How many calories do you intake?
How many hours do you exercise?
How many hours do you spend on social media?

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Student Life.**

Plan

1. Using the Student Life data set provided, what measures of central tendency (mean, median, mode) or dispersion (standard deviation, range, or variance) would you use to describe the data for each question?
   Choose at least two measures of centrality and two measures of dispersion. Explain the reasoning behind your choices.

Measures of Centrality	Why did you choose this measure?
Choice #1	
Choice #2	
Measures of Dispersion	Why did you choose this measure?
Choice #1	
Choice #2	

Explore

2. Using technology and the given data, calculate your preferred measures of central tendency and dispersion and record in the following table.

Measures of Centrality	Descriptive Statistics
Choice #1	Hours Sleeping: Hours Studying: Calorie Intake: Hours Exercising: Hours on Social Media:
Choice #2	Hours Sleeping: Hours Studying: Calorie Intake: Hours Exercising: Hours on Social Media:
Measures of Dispersion	Descriptive Statistics
Choice #1	Hours Sleeping: Hours Studying: Calorie Intake: Hours Exercising: Hours on Social Media:
Choice #2	Hours Sleeping: Hours Studying: Calorie Intake: Hours Exercising: Hours on Social Media:

**Extra Challenge: Pair up with another individual (or group) that is using a different type of technology to determine the measures. Compare and contrast the efficiency of both types of technology, sharing any helpful hints.

Understand

3. Record any inferential statements based on your descriptive statistics calculated above.
4. Based on your results from the previous question and your own personal experience, what can college students do to promote overall wellness and health?
5. Create a graphical display to share your opinion with others. Make sure to include descriptive and inferential statistics to support your claim.

Extension

6. If you were to do a follow-up statistical study, what question(s) would you like to explore? Why?

Chapter Project

Home Sweet Home: Using Linear Regression to Analyze and Predict Home Prices

An important problem in real estate is determining how to price homes to be sold. There are so many factors—size, age, and style of the home; number of bedrooms and bathrooms; size of the lot; and so on—which makes setting a price a challenging task. In this project, we will investigate the relationships among typical characteristics of homes and home prices, identify key variables related to pricing, and build linear regression models to predict prices based on property characteristics. Our analysis will be based on the Mount Pleasant Real Estate Data (available on stat.hawkeslearning.com). This data set includes information about 245 properties for sale in three communities in the suburban town of Mount Pleasant, South Carolina, in 2017.

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Mount Pleasant Real Estate Data.**

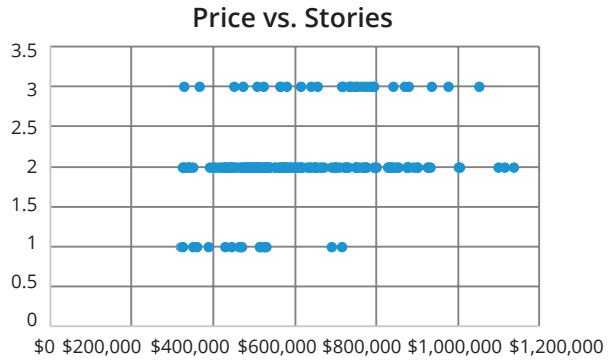
Phase 1: Data Preparation.

1. Download the Mount Pleasant Real Estate Data from stat.hawkeslearning.com and open it with Microsoft Excel.
2. To ensure the data contains comparable properties, eliminate duplexes and properties whose prices are outliers. What limitations does this impose on our analysis?
3. The statistical tools from the current chapter focus on numeric data, so eliminate non-numeric variables from the data. Does this remove potentially useful information?
4. Are there any redundant variables we could eliminate?

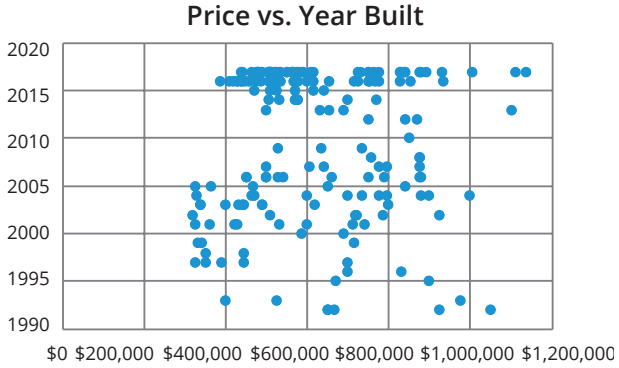
Phase 2: Discovering Relationships

5. How strongly does each remaining variable correlate to the price?
6. Which variable correlates most strongly with price?
7. Are any variables weakly correlated with price? Practically speaking, why do you think this is true?
8. Do scatter plots reveal any nonlinear pattern between price and the weakly correlated variables?

a.



b.

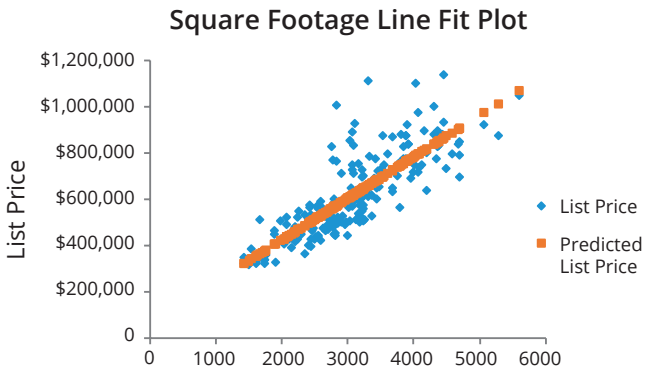


Phase 3: Constructing Predictive Models.

Enable the Analysis ToolPak add-in to Excel.
The regression tool will be used.

9. Find the regression line $\hat{y} = b_0 + b_1 x$ predicting home price by the variable most highly correlated to it. Assess the fit of the line in terms of error and the proportion of variation explained by the model.

10. For which properties do the model's predictions have the greatest errors? What is an intuitive reason for this?



Chapter Project

Should You Play the Hand You're Dealt?

This project is designed to be completed by groups, but an individual may complete it. Each group should have one standard deck of cards to use as directed.

Rules for Poker

Using a standard deck of 52 cards, the players are dealt cards depending on the exact type of poker game being played. For the purpose of this project, players will be dealt 5 cards. Card values and scoring can be found at www.bicyclecards.com/how-to-play/basics-of-poker/.

Part 1: Understanding the deck.

a. What is the probability of getting a queen?

b. What is the probability of getting a four?

c. What is the probability of getting a five?

d. What is the probability of getting a ten?

e. Does it matter which card value that is picked?

f. What does this tell you about the probability of drawing any particular card value from the deck?

g. Now consider how many different 5-card hands that can be dealt. Does order matter in getting these cards?

h. Remember that when order matters use a permutation and when order does not matter use a combination. So, would you use a permutation or a combination to determine the number of 5-card hands that can be dealt from a deck of 52 cards?

i. Based on your answer to the previous question, determine the number of 5-card hands that can be dealt from a deck of 52 cards.

Part 2: What is the probability of getting more than one of the same card in a 5-card hand?

a. What is the probability of getting a pair of some kind, such as a pair of fours, out of five cards

b. What is the probability of getting three of a kind, such as three fours, out of five cards?

c. What is the probability of getting four of a kind, such as four fours, out of five cards?

Part 3: Classroom Activity

Create small groups of four or less. Take a standard deck of cards and shuffle it well. Deal each person five cards.

a. How many people got a pair?

b. How many people got three of a kind?

c. Did anyone get four of a kind?

d. How does this help your group understand the probabilities from Part 2?

Part 4: Let's look at the probability of getting two pairs. This is two different numbers paired up, such as 3, 3, 7, 7, 9.

a. Would you use a permutation or a combination to determine the probability of getting a 5-card hand that contains two pairs?

b. Determine the probability of getting two (different) pairs in a 5-card hand.

Part 5: Let's look at the probability of getting a full house. A full house is three of one card and a pair of another card such as 8, 8, 8, 4, 4. Determine the probability of getting a full house.

Part 6: Let's look at a very special hand, the Royal Flush, which is getting an A, K, Q, J, and 10 all of one suit.

a. How many ways can you get a Royal Flush?

b. What is the probability of getting a Royal Flush?

Part 7: When you have all 5 cards from the same suit in a sequence, such as 3, 4, 5, 6, 7, this is called a straight flush.

a. How many ways can you get a straight flush in a 5-card hand?

b. What is the probability of getting a straight flush?

Part 8: Let's look at a flush, which is getting all five of one suit such as 2, 4, 6, 7, 10, of any one of the four suits. Now we want to just look at flushes, not flushes that happen to also be straights since a straight flush is itself a certain kind of hand.

a. What is the probability of getting a flush?

b. Let's look at a straight that is not also a flush, which is getting all five in a row of any suit such as 4, 5, 6, 7, and 8 no matter what suit they came from.

c. What is the probability of getting a straight?

Part 9: Conclusions

a. So, what have you learned from this project about playing poker with a 5-card hand and will it influence how you bet in the future when playing?

b. Is the probability in your favor of getting the hand you want?

Source: https://www.bicyclecards.com/how-to-play/basics-of-poker/

Chapter Project

Take Me Out to the Ball Game!

Use the Moneyball data set which contains selected statistics for Major League Baseball teams from 1962–2012. You can see a full description of this data set in Appendix B.

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Moneyball Data.**

1. Select the *Number of wins variable*, *W*, and compare the distribution of *W* for the American League (AL) with that of the National League (NL). Use side-by-side boxplots as described in Chapter 4.
2. Identify the outliers in both leagues (i.e., the teams that have a total number of wins far from the rest of the teams in their league).
3. Compare the distribution of the *Number of wins, W,* for NYM and TEX using a side-by-side boxplot and by investigating the numerical summaries of each. (Compare the shapes, means, medians, and the variability).
4. Discuss why the discrepancy in variability between the performance of NYM and the performance of TEX didn't cause a similar discrepancy in their respective leagues.
5. Based on historical data, the probability that in a given year the NYM will make the playoffs is $p = 7/47 = 0.149$. Let X be the discrete random variable that gives the total number of Playoffs made by NYM in the last 20 years, i.e., from 1993 to 2012.
   a. Assume that the outcomes for the NYM in these years are unknown for us. Also assume that the outcome in any of the years is independent of the outcome in any other year. Under these assumptions, what would be the distribution of X? Why?
   b. What is the probability that the total number of playoffs made by NYM during this 20-year period is exactly three?
   c. What is the probability that the total number of playoffs made by NYM is at most 3?
   d. What is the probability that the total number of playoffs made by NYM is at most 18?
   e. What is the probability that the total number of playoffs made by NYM is at least 15?
   f. What is the expected number of playoffs that NYM will make in this 20-year period?
   g. Find the variance of the number of playoffs that NYM is expected to make in this 20–year period?
   h. Can we use the Poisson distribution with λ=2.98 to model the number of playoffs that NYM will make? Why?

Source: https://www.baseball-reference.com/

Chapter Project

Darts and the Normal Distribution

Darts, like any game of skill, involves some degree of randomness. In this project, we will try to analyze the chances of earning various point totals in individual dart tosses. A typical dartboard is displayed in Figure 1.

Figure 1

We may consider two random variables in the position of each dart toss (see Figure 2):

1. The angle, θ, from the horizontal of the line (in red) on which the dart lands.
2. The position along this line, X, centered at the middle of the bullseye.

If the location of the dart is the black dot in Figure 2, we see that X and θ define the location of the dart, with a positive X above the horizontal and a negative X below it.

If a player aims a dart at the bullseye in the center of the dartboard and tosses it, of course he/she may or may not hit the target with some probability, but we can say more if we know something about the pattern in the randomness.

In particular, a highly skilled player when aiming at the bullseye, will hit the center of the bullseye on average, but with decreasing probability, the dart will land further away from the target, so a normal distribution centered at 0 is reasonable for X. A highly skilled player also shows no particular angular tendency, so we may model θ as a uniformly distributed random variable that is independent of X.

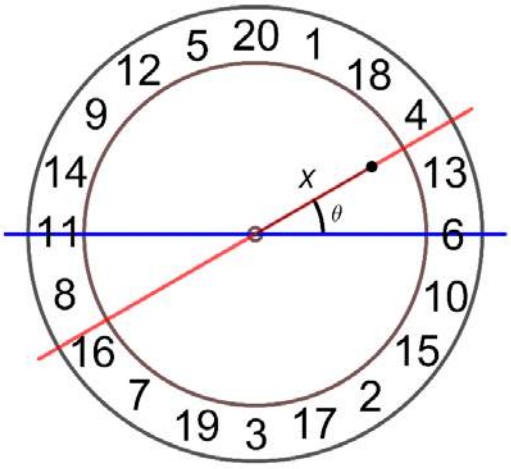


Figure 2

Common rules of darts have the following scoring scheme[1]:

Where dart lands	Point value
Inner Bullseye	50
Outer Bullseye	25
Single Region (black/white area within Double ring)	Score given on outside ring (1-20)
Triple Ring (inner red/green ring)	Triple the score given on the outside ring
Double Ring (outer red/green ring)	Double the score given on the outside ring
Outside the Double Ring	0

Questions (Assume all throws are aimed at the center of the bullseye.)

1. What is the range of the random variable θ?
2. For a skilled player, what is the mean of the normal random variable X?
3. Should the normal random variable X have a high or low σ for the most skilled players?
4. Find the probability of the dart landing in each region for a player with $\sigma = 50$ mm and for a player with $\sigma = 100$ mm given the following specifications for the dartboard. Each region corresponds to the given ranges of (absolute) distances from the center of the inner bullseye.[1]

Region	Distance is more than...	Distance is less than...
Inner Bullseye	0 mm	6.35 mm
Outer Bullseye	6.35 mm	15.9 mm
Inner Single Region	15.9 mm	95.3 mm
Triple Ring	95.3 mm	104.8 mm
Outer Single Region	104.8 mm	158.8 mm
Double Ring	158.8 mm	168.3 mm
Beyond	168.3 mm	

5. What is the probability of landing in a single scoring region (shown in yellow in Figure 3) for each player?
6. What is the probability of landing in the wedge of the dartboard marked with 20 if $\sigma = 50$ mm (i.e., in the yellow region or beyond)? (See Figure 4.)
7. What is the probability of scoring 7 points on one dart throw if $\sigma = 50$ mm?
8. What is the probability of scoring 8 points on one dart throw if $\sigma = 50$ mm?
9. What is the probability of scoring 6 points on one dart throw if $\sigma = 50$ mm?
10. What are the highest and lowest possible scores from a single dart throw? What is the probability of each if $\sigma = 50$ mm?
11. What is the expected value of the score for a dart toss by a skilled player if $\sigma = 50$ mm?

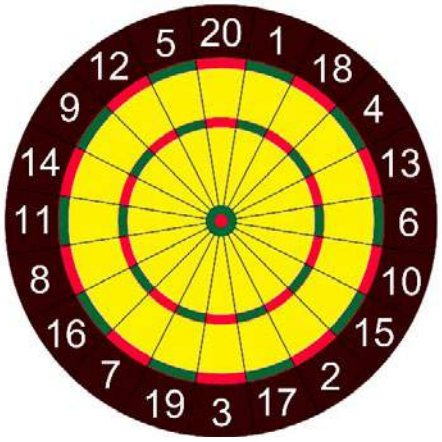


Figure 3

Figure 4

References

[1] World Darts Federation (2018). Playing and Tournament Rules. 20th edition. https://www.dartswdf.com/rules/

Chapter Project

A Random Sampling of College Life

Everyone has heard about the "Freshmen 15", referring to the number of pounds gained by a typical freshman during their first year of college. Recent studies show that students on average gain 3 to 10 pounds during their first 2 years of college, with most of that weight gain occurring during the first semester of their freshman year.

If a study conducted by a local university showed that on average a college student gains 10 pounds with a standard deviation of 2 pounds during their freshmen year, and student weight gain is normally distributed, answer the following questions.

1. What is the probability that a randomly selected student will gain between 6 and 8 pounds?

2. If 25 students are randomly selected, what is the probability that the average weight gain of the students is between 6 and 8 pounds?

Another study on college students says that the average college student credit card debt is around $3200 with a standard deviation of $500. Assuming that credit card indebtedness is normally distributed, answer the following questions.

3. What is the probability that a randomly selected student owes more than $3500 in credit card debt?

4. If 25 students are randomly selected, what is the probability that the average credit card indebtedness is more than $3500?

A recent news article stated that only 17% of college students between the ages of 18 to 24 years old voted in the last presidential election.

5. Assuming the voting rate stays the same, what is the probability that from a random sample of 500 college students from a local university, at least 20% will vote in the next presidential election?

Suppose you are taking a statistics class and you must conduct a survey of the students at your college as part of an assignment. Describe how you would go about sampling students using each of the sampling methods listed below.

6. Cluster sample

7. Stratified sample

8. Systematic sample

9. Random sample

10. Convenience sample

Chapter Project

Home Sweet Home: Using Confidence Intervals to Analyze and Compare Home Prices

One of the biggest purchases we make in our lives is a home. As we buy a home we ask ourselves many questions such as:

How much should I spend for a home?

How many bathrooms are there?

What is the cost per square foot?

Data

The data can be found at stat.hawkeslearning.com
Data Sets > Mount Pleasant Real Estate Data

Suppose you are looking for a house near Charleston in Mount Pleasant, SC, and you have narrowed your search to three subdivisions: Carolina Park, Dunes West, and Park West.

a. Download the Mount Pleasant Real Estate data set.

b. Import the data into Minitab, Excel or other statistical software.

c. For the variable *List Price,* calculate the sample mean, the sample standard deviation, and the sample size for the three different subdivisions. Put the calculations in a table and round to the nearest dollar for the sample standard deviation and the mean.

d. Based on the data set and the information we have, which confidence interval should we use here, a *z* or a *t* interval? Why?

e. Find the critical value for a 95% confidence level for each subdivision for the variable *List Price*.

f. Construct an interval to estimate the true average *List Price* for each subdivision with 95% confidence. Based on these confidence intervals, is it possible that Carolina Park and Dunes West have the same average *List Price*. Discuss.

g. Do you think a *List Price* of $520,000 is a reasonable value for the Carolina Park subdivision?

h. Do you think a *List Price* of $670,000 is a reasonable value for the Dunes West subdivision?

i. Do you think a *List Price* of $568,000 is a reasonable value for both the Carolina Park and Park West subdivisions?

Chapter Project

Employee Satisfaction: Statistical Inference Using Hypothesis Testing

Part 1

If the level of employee satisfaction drops below 0.60 overall, then there is a belief that there may be a serious problem with morale in that department. There have been rumors that the Human Resources department (**hr** in the data file) may be having just such issues. Using a statistical package, test to determine if the mean employee satisfaction level in the Human Resources department is less than 0.60.

1. Is *satisfaction_level* a qualitative or quantitative variable?
2. Graph the employee satisfaction level for the Human Resources department with an appropriate graph and calculate statistics appropriate for this type of data.
3. Conduct the appropriate hypothesis test using the following steps.
   a. Determine the null and alternative hypotheses.
   b. Use a significance level of $\alpha = 0.05$.
   c. Validate the assumptions of the hypothesis test, identify the appropriate test statistic, and compute its value.
   d. Determine the *P-value*.
   e. Make a decision to reject or fail to reject the null hypothesis, H_0.
   f. State the conclusion in terms of the original problem.
4. Based on our conclusion from the previous step, what type of error could we have just made (Type I or Type II)? State the practical implications of this error.
5. Would it be appropriate to compare this test to a confidence interval for the mean? Why or why not?

Part 2

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Employee Satisfaction.**

According to the US Bureau of Labor Statistics, there was a 12.4% incidence of workplace injury in 2016 in the private sector. If the workplace accident rate is above 12.4%, the company will increase inspections and implement additional safety training. For this data, is there statistical support to increase inspections and implement additional safety training? Use the entire dataset and the *work_accident* variable where 0 = no accident and 1 = an accident occurred.

1. Is *work_accident* a qualitative or quantitative variable?
2. Graph the accident at work data with an appropriate graph and calculate statistics appropriate for this type of data.
3. Conduct the appropriate hypothesis test using the following steps.
   a. Determine the null and alternative hypotheses.
   b. Use a significance level of $\alpha = 0.05$.
   c. Validate the assumptions of the hypothesis test, identify the appropriate test statistic, and compute its value.
   d. Determine the *P-value*.
   e. Make a decision to reject or fail to reject H_0.
   f. State the conclusion in terms of the original problem.
4. Based on our conclusion from the previous step, what type of error could we have just made (Type I or Type II)? State the practical implications of this error

Source: https://www.bls.gov/iif/oshwc/osh/os/ostb4741.pdf

Chapter Project

Pales in Comparison

Part 1

Different types of beers have different ingredients, flavors, and alcohol amounts. Two popular beers in the US are the American Pale Ale and the American IPA. The IPAs tend to have a stronger flavor and come in a variety of colors whereas the Pale Ales tend to be lighter in flavor and in color. Using a statistical package, test to determine if the mean alcohol by volume for the American IPA is the same as the mean alcohol by volume for the American Pale Ale.

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Beers and Breweries.**

1. Graph the alcohol volumes for the two types of beers using appropriate graphs and calculate statistics appropriate for this type of data.
2. Conduct the appropriate hypothesis test using the following steps.
   a. Determine the null and alternative hypotheses.
   b. Use a significance level of $\alpha = 0.05$.
   c. Validate the assumptions of the hypothesis test, identify the appropriate test statistic, and compute its value. Using the graphs created, determine if you should be conducting a two-sample test of the mean with equal or unequal variances.
   d. Determine the *P*-value.
   e. Make a decision to reject or fail to reject the null hypothesis, H_0.
   f. State the conclusion in terms of the original question.
3. Calculate the 95% confidence interval for the difference between the two means. Does this confidence interval support your results from the hypothesis test? Why or why not?

Part 2

Colorado and California are huge producers of beers with many microbreweries in each state. Both states produce a variety of different types of beers as well. Is the proportion of American IPA's compared to all other types of beers the same in both California and Colorado?

1. Graph the American IPAs as compared to all other types of beer for each state using appropriate graphs and calculate statistics appropriate for this type of data.
2. Conduct the appropriate hypothesis test using the following steps.
   a. Determine the null and alternative hypotheses.
   b. Use a significance level of $\alpha = 0.05$.
   c. Validate the assumptions of the hypothesis test, identify the appropriate test statistic, and compute its value.
   d. Determine the *P*-value.
   e. Make a decision to reject or fail to reject the null hypothesis, H_0.
   f. State the conclusion in terms of the original question.
3. Calculate the 95% confidence interval for the difference between the two proportions. Does this confidence interval support your results from the hypothesis test? Why or why not?

Chapter Project

Home Sweet Home: Transforming Data Prior to Regression Analysis

Use the data set named Mount Pleasant Real Estate Data which contains information about properties for sale in three subdivisons of Mount Pleasant, South Carolina in the year 2017.

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Mount Pleasant Real Estate Data.**

1. Classify the three variables *List Price*, *Squarefootage*, and *Subdivision* as qualitative or quantatitive and provide the level of measurement as described in Section 2.2.
2. Which of the quantitative variable(s) in the previous question should be considered the dependent variable (as described in Section 5.2)? Why?
3. Use technology to make a histogram for *List Price* and describe the distribution.
4. Create a scatterplot of *List Price* vs. *Squarefootage*. That is let y = *List Price* and x = *Squarefootage* and describe any patterns.
5. Can we use Pearson's correlation coefficient to measure the strength of the linear association between *List Price* and *Squarefootage*? Why or why not?
6. Is the simple linear regression model appropriate for predicting List Price based on *Squarefootage*? Explain your answer by investigating the model assumptions.
7. Take the natural log of List Price and call it *LnPrice*. Make a histogram for this variable
8. Describe the distribution of *LnPrice*.
9. Create a scatterplot of *LnPrice* vs. *Squarefootage* and describe any patterns.
10. Calculate Pearson's correlation coefficient between *LnPrice* and *Squarefootage* and interpret it.
11. Using technology, build a simple linear model to predict *LnPrice* from *Squarefootage*.
12. Use α=0.05 and test to see if the model is significant.
13. Predict the price of a 3045 square foot house.
14. Find the residual of the predicted value of the house in the previous question if the house is actually the house from the data set with ID = 27 in Carolina Park with a price of $575,000.

Chapter Project

Home Sweet Home: Using Multiple Regression to Analyze and Predict Home Prices

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Mount Pleasant Real Estate Data.**

An important problem in real estate is determining how to price homes to be sold. There are so many factors—size, age, and style of the home; number of bedrooms and bathrooms; size of the lot; and so on—which makes setting a price a challenging task. In this project, we will try to help realtors in this task by determining how different characteristics of homes relate to home prices, identifying the key variables in pricing, and building multiple-variable regression models to predict prices based on property characteristics. Our analysis will be based on the Mount Pleasant Real Estate Data (available on stat.hawkeslearning.com). This data set includes information about 245 properties for sale in three communities in the suburban town of Mount Pleasant, South Carolina, in 2017.

Phase 1: Data Preparation.

1. Download the Mount Pleasant Real Estate Data from stat.hawkeslearning.com and open it with Microsoft Excel.

2. To ensure the data contains comparable properties, eliminate duplexes and properties whose prices are outliers. What limitations does this impose on our analysis? Consider the following variables associated with each property.

x_1 = number of bedrooms	x_5 = has pool?
x_2 = number of bathrooms	x_6 = has dock?
x_3 = number of stories	x_7 = fenced yard?
x_4 = square footage	x_8 = golf course?

3. Are any of the variables qualitative? Adjust this data in a reasonable quantitative way for use in a regression analysis.

Phase 2: Constructing Predictive Models

Enable the Analysis ToolPak add-in to Excel. The regression tool will be used.

The idea of linear regression can be easily extended to the case where there are multiple independent variables that are used to predict the dependent variable. The linear regression equation will look like

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m$$

where each b_i is a coefficient and each x_i is an independent variable. In the context of real estate pricing, $\hat{y}$ = predicted home price. Excel can calculate regression models with multiple variables via the same regression tool that it does for single-variable regression models by simply using more columns of data for the X inputs. Intuitively, this should be more realistic for real estate pricing as there may be several variables that contribute to property values.

4. Construct the multiple regression equation with input variables x_1, x_2, ..., x_8.

5. What is the adjusted coefficient of determination, R_a^2, of the regression model? Explain the meaning of this value and how it differs from R^2.

6. Perform a hypothesis test to determine if the model is useful for predicting home values at a significance level of $\alpha = 0.05$. State the *P*-value and interpret its meaning.

7. Are any variables not useful predictors of home price at a significance level of $\alpha = 0.05$? State the *P*-values of these variables. Intuitively, what does this mean with respect to pricing properties?

8. Construct the multiple regression model with only the input variables whose coefficients are significant in the eight-variable regression?

9. How does the adjusted coefficient of determination of the new model in Problem 8 relate to the adjusted coefficient of determination from Problem 5? What conclusion can you draw from this?

Phase 3: Applying and Interpreting the Model

10. Suppose you own a 2000 ft^2 2-story house in one of the communities in the data set with 3 bedrooms, 2.5 baths, a pool, and it is located on a golf course, but has no dock or fenced yard. What does the model from Problem 4 predict the price of your house to be?

11. A common term in real estate is "comparables," or "comps" for short, which are properties that have similar characteristics. It is common for realtors to look up "comps" for a certain property to get an idea of how to price it. Locate the "comps" for your home in the data set. Create a box plot of the "comps" and estimate a price range for your house on this basis.

12. What advantages and disadvantages does this approach have to the multiple regression model above?

Chapter Project

Analysis of California DDS Expenditures

A government lobbying firm is interested in getting more money directed to people who have special needs. To convince state legislators that more money needs to be diverted to the California Department of Developmental Services, they first must determine which groups of people are in most need of those funds. This firm has hired you to conduct the analysis to answer the questions below.

Data

The data can be found at stat.hawkeslearning.com **Data Sets > California DDS Expenditures.**

Part 1

Determine if there is a difference in the average expenditures by gender. Conduct the appropriate hypothesis test using the statistical package of your choice.

a. Determine the null and alternative hypotheses.

b. Use a significance level of $\alpha = 0.05$.

c. Validate the assumptions of the hypothesis test, identify the appropriate test statistic, and compute its value. Include any graphs you use to validate assumptions.

d. Determine the *P*-value.

e. Make a decision to reject or fail to reject H_0.

f. State the conclusion in terms of the original question.

Part 2

After seeing your results, the lobbying firm has now decided that they also need to see the breakdown by both gender and age group (the pre-defined age range the consumer falls into).

1. Would this be a Two-Way ANOVA Randomized Block Design or a Two-Way ANOVA Factorial Design?
2. Conduct the appropriate hypothesis tests to see if there is a difference in the average expenditures by both gender and age group and, if appropriate, interaction between these variables.

   a. Determine the null and alternative hypotheses for each test.

   b. Use a significance level of $\alpha = 0.05$.

   c. Assume all assumptions are satisfied for each test, identify the appropriate test statistics, and compute their values.

   d. Determine the *P*-value for each test.

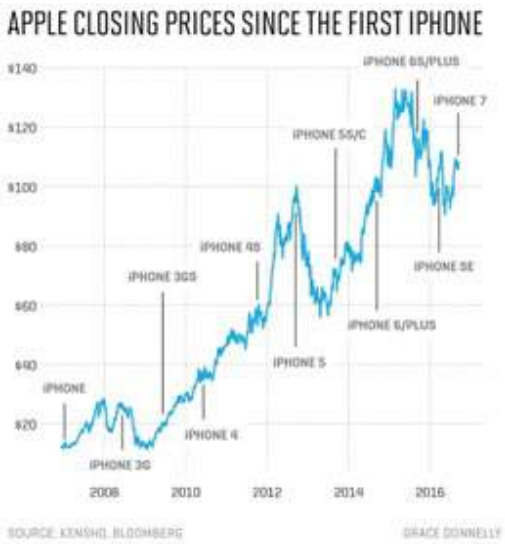
   e. Make decisions to reject or fail to reject H_0 as appropriate.

   f. State the conclusions in terms of the original question.

Chapter Project

Individual Stocks vs. Index-Matching Investments

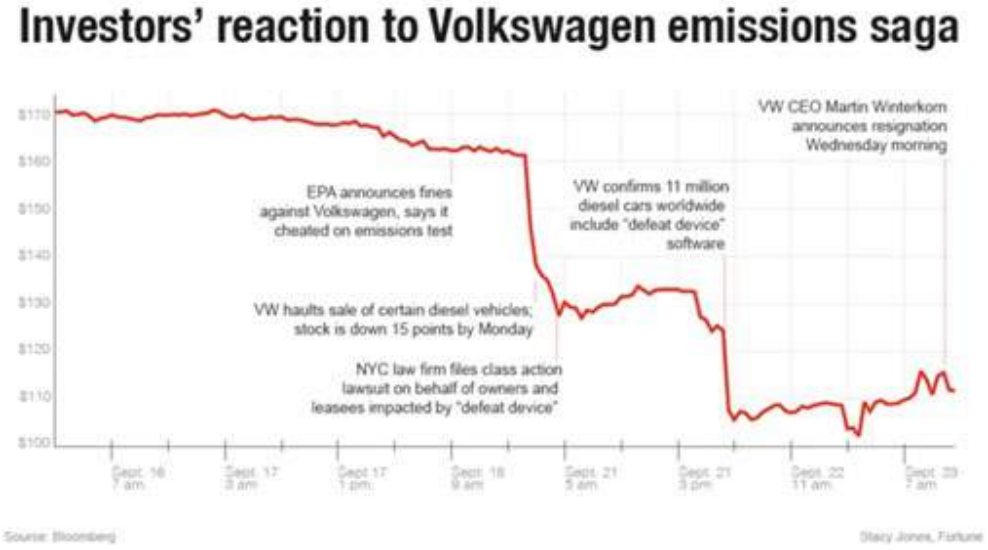
In stock market investing, the traditional approach is to buy and sell stocks for individual companies, but this is sometimes risky as it is very difficult to predict when large changes to prices may occur. Large upswings or downswings in stock prices may happen for many reasons. Some positive examples leading to large stock price increases include the following.



- Retail sales on Amazon have frequently exceeded expectations.[1]
- Apple introduced its wildly popular iPhone, and continued to introduce new versions.[2]
- Netflix exploded in popularity; a growth of 7.41 million subscribers in the first quarter of 2018 (up from the expected 6.35 million) led to a stock increase of 60% in that quarter.[3]

Some negative events leading to plummeting stock prices include the following.

- Microsoft lost a federal antitrust lawsuit in 1999 and its stock price dropped 14% in a single day.[4]
- In 2015, the Environmental Protection Agency (EPA) discovered Volkswagen had intentionally programmed certain diesel engines to activate emission controls for certain nitrous oxides only during testing, which made It appear as though the vehicles released less than the legal limit of the polluting gases, but the vehicles actually released 40 times the legal limit! In the wake of the scandal, Volkswagen stock prices dropped from $162 to $105 in just three days.[5]



As we see, certain events may cause large fluctuations in stock prices, which may be good or bad for the investor, but in either case, it results in a volatile and sometimes risky investment.

Index-matching funds are portfolios (groups of multiple stocks) structured so that they match a market index, such as the Standard & Poor's 500 Index (S&P 500).[6] Gains or losses to the investment will be (proportionally) the same as that of the whole S&P 500. Such an index tends to be less volatile than investing in individual stocks, therefore index-matching funds are considered less risky investments. These funds are often used as parts of retirement funds or other long-term investments. But, how true is this assumption? Is an S&P 500 index-matching fund actually a safer investment? Our goal for this project is to test that assumption.

1 Daily closing stock prices can be found on the web site http://www.macrotrends.net/stocks/charts.

The closing price data for three stocks (Amazon (AMZN), Starbucks (SBUX), and Coca-Cola (KO)), along with the S&P 500 index fund (SPY) from the beginning of 2000 to the end of 2017, can be found on our web site.

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Stock Comparison Data.**

2. We want to determine how the value of certain stocks and the S&P 500 compare over time, so notice that a new column called Price Change was created measuring the daily change in stock price for each stock, i.e.,
Day 2 Price – Day 1 Price
Day 3 Price – Day 2 Price
Day 4 Price – Day 3 Price
etc.

It does not make sense to simply compare stock prices of the different stocks because the prices are on very different scales. For example, if a $100 stock drops by $1, it is less consequential than if a $3 stock drops by $1. A reasonable adjustment is to consider a price change as a percentage of the previous day's price, so the example above would be a 1% drop compared to a 33.3% drop, which allows us to do a more reasonable comparison. A new column for each stock and the S&P 500 index has been created with these percentages and is labelled as *Return*.

3. Can you think of any descriptive statistics that would compare the volatility of the returns?
4. Create box plots of the returns for each stock and the S&P 500 index.
5. Do you notice any patterns in the box plots that suggest a difference in the stocks and the S&P 500 index? Explain their practical significance, if any.
6. Delete the outliers in the data for each stock and the S&P 500 index and create histograms for each on the same scale. [*Hint*. Excel's Data ToolPak add-in allows for easy histograms where we can specify the bins.]
7. Are any differences between the stocks and the S&P 500 index apparent from the histograms? Explain their practical significance, if any.
8. Use a test for goodness of fit to test whether the distributions of the non-outlying stock returns and index returns are different (at least over the range of the S&P 500 returns) at a significance level of $\alpha = 0.2$.

References
1 https://www.cnbc.com/2017/10/26/amazon-earnings-q3-2017.html
2 http://fortune.com/2016/09/09/apple-stock-iphone-launches/
3 https://www.thestreet.com/investing/stocks/netflix-shares-rise-after-beating-estimates-14557098
4 https://www.nytimes.com/2000/04/04/business/us-vs-microsoft-overview-us-judge-says-microsoft-violated-antitrust-laws-with.html
5 http://fortune.com/2015/09/23/volkswagen-stock-drop/
6 https://www.investopedia.com/terms/i/indexfund.asp

Chapter Project

Home Sweet Home: Using Nonparametric Tests to Compare Home Prices

Use the **Mount Pleasant Real Estate Data** which contains information about properties for sale in three subdivisons of Mount Pleasant, South Carolina in the year 2017.

Data

The data can be found at stat.hawkeslearning.com **Data Sets > Mount Pleasant Real Estate Data.**

1. Download the **Mount Pleasant Real Estate Data** into a statistical software package like Excel or Minitab.
2. Classify the three variables *List Price*, *Square Footage*, and *Subdivision* as qualitative or quantatitive and provide the level of measurement (nominal, ordinal, interval, or ratio).
3. Which of the quantitative variable(s) should be considered as the dependent variable? Why?
4. Use statistical software to make a histogram for *List Price* and describe the distribution.
5. Can we use the *t*-test to see if the **mean** home price is more than $500,000? Why or why not?
6. Since the underlying distribution is not normal, we have an opportunity to use nonparametric methods to analyze the data. Can we conclude that the **median** *List Price* in Mount Pleasant in 2017 is more than half a million dollars? State your hypotheses and perform a Sign Test using $\alpha = 0.05$.
7. Create side-by-side boxplots of *List Price* for the three Mount Pleasant subdivisions: Carolina Park, Dunes West and Park West. Describe the distributions of the three subdivisions and comment about their variability.
8. Use the Wilcoxon Rank-Sum Test to see if the distribution of *List Price* in Park West in 2017 is to the left of that in Dunes West.