BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Video 1: Statistics Help Decision Making Under Uncertainty
Chapter 20 (Time Series Analysis)
Chapter 22 (Decision Making and Risk)

Video 2: Variation is Everywhere
Chapter 1 (Statistics and Variation)
Chapter 6 (Displaying and Describing Quantitative Data)

Video 3: Decision Making Error Probability
Chapter 11 (Testing Hypotheses about Proportions)

Video 4: Models Are Not Reality (But Can Be Incredibly Useful)
Chapter 18 (Multiple Regression)
Chapter 20 (Time Series Analysis)

Video 5: Graphing and Exploring Data Can Reveal "Eurekas"
Chapter 6 (Displaying and Describing Quantitative Data)
Chapter 25 (Quality Control)

Video 6: Quantifying Uncertainty in Making Estimates
Chapter 12 (Confidence Intervals and Hypothesis Tests for Means)
Chapter 10 (Confidence Intervals for Proportions)

Video 7: Statistics Turn Data Into Information
Chapter 2 (Data)
Chapter 4 (Displaying and Describing Categorical Data)
Chapter 6 (Displaying and Describing Quantitative Data)

Video 8: One Outlier Can Change the Results of a Statistical Analysis
Chapter 7 (Scatterplots, Association, and Correlation)
Chapter 8 (Linear Regression)

Video 9: Collecting Good Data is as Important as Analyzing It
Chapter 3 (Surveys and Sampling)
Chapter 2 (Data)

Video 10: Checking the Assumptions and Conditions of a Model is Crucial
Chapter 14 (Paired Samples and Blocks)
Chapter 22 (Design and Analysis of Experiments and Observational Studies)

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 1: Statistics Help Decision Making Under Uncertainty

Summary

This video illustrates how *Deckers Outdoor Corporation* uses statistical models to predict seasonal demand for its various lines of footwear. Based on historical data and past trends, as well as current economic activity and customer behavior, Deckers not only forecasts demand for its large retail customers but also for its online store that carries the entire line for each of its brands. These forecasts drive many decisions, including those dealing with inventory, production scheduling, and even materials acquisition further up Deckers' supply chain. While uncertainty exists, it is important for Deckers' to forecast demand as accurately as possible given the negative cost consequences associated with either over- or under- predicting demand. The main theme of the video is that Deckers' uses statistical models to improve decision making under uncertainty along its supply chain.

Video-Specific Questions

1. Deckers Outdoor Corporation, like many of its competitors, sells its footwear online. What types of data does Deckers Outdoor Corporation gather to forecast demand?

To forecast demand in future seasons, Deckers relies on historical sales data. However, this is dependent on the item being forecast. For example, this works well for carryover items (items carried in the past and will be carried in the future), as historical data are available. If an item is new but similar to an item carried in the past, historical data for the similar item is used to forecast demand for the new item. If the item being forecast is more fashion-forward with no comparable item carried in the past, then historical data from other fashion-forward items carried in the past are used. Information on factors that might influence demand are also gathered, such as consumer trends, competitor offerings, pricing, and general economic conditions.

2. What components may be present in historical sales data?

The four components of time series may be present in Deckers Outdoor Corporation historical sales data. These include trend (long term increase or decrease), seasonal (monthly), cyclical (tied to the economic cycle), and random components.

3. What are the consequences of forecasting errors in either direction (over or under) for Deckers Outdoor Corporation?

If Deckers carries too little inventory, the company faces stock-outs and unhappy customers who can't buy the products they want, potentially affecting customer goodwill.

If Deckers carries too much inventory, the company faces the disposal of unsold footwear at discounted prices. A wrong decision in either direction (too little or too much) has negative cost implications (loss of profit) for Deckers.

4. Part of the process that Deckers Outdoor Corporation follows in forecasting demand involves revisiting the past year's predictions to see how closely they matched actual sales. What are some measures of forecast accuracy that can be used to do this?

Some measures of forecast accuracy are the mean squared error (MSE), mean absolute deviation (MAD), and mean absolute percentage error (MAPE). The MSE penalizes large forecast errors because the errors are squared, but it is not in the same units as the data. The MAD is in the same units as the data. So, both MSE and MAD are dependent on the magnitude of the data values. The MAPE provides a percentage that relates the size of the errors to the magnitudes of the data values.

5. What other decisions, besides those dealing with inventory, depend on Deckers' **demand forecasts?**

Deckers' demand forecasts affect decision making up through its supply chain, such as production scheduling and raw materials acquisition.

Relevant Chapters in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 20 (Time Series Analysis)
Chapter 22 (Decision Making and Risk)

Concept-Centered Teaching Points

Payoff tables and decision trees structure decision making (Chapter 22).

Discuss the types of decisions that would have to be made for Deckers' online store. Ask students to think about how a payoff table could be constructed for one of these decisions (e.g., how much inventory to carry for one of its popular brands…UGGs). What are the possible actions? What are the states of nature that impact the decision? How would the outcomes or payoffs be determined? Who might be involved in the decision making process?

Probabilities quantify uncertainty in decision making (Chapter 22).

Extend the discussion started above by considering how probabilities are used in the decision making process. How might the probabilities for the various states of nature be determined? How does historical data on demand factor into the decision making process? Ask students to think about how additional information about the states of nature might be obtained to revise the probability estimates. You can also construct a payoff table with some values proposed by students and calculate the expected value of

each action as well as the standard deviation (which can be used to discuss the topic of risk).

Forecasting relies on past patterns to predict the future (Chapter 20).

The discussion can revolve around the components of time series data (such as those that may be present in the historical sales data used for forecasting at Deckers) such as trend, seasonality, and cycle. It should be noted that extrapolating any of these components into the future for forecasting relies on the assumption that past patterns will continue into the future. This is also a good opportunity to discuss difficulties associated in forecasting the business (or economic) cycle as well as how uncertainty is reflected by the irregular (or random) component of a time series.

Useful Links

http://cygnus-group.com/CIDM/index.html
(CIDM = Center for Informed Decision Making)

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 2: Variation is Everywhere

Summary

This video illustrates how *Southwest Airlines* must understand the sources of variation that impact flight schedules to better manage its operations. The potential sources of variation that are outside of the airlines' control are identified. In addition, the video shows how Southwest may choose to intentionally introduce variation into flight schedules take advantage of opportunities for its planes to depart the gate early. The main theme of the video is that Southwest Airlines must understand, quantify, and predict variation to assure consistent high-quality service in meeting its customers' needs and expectations.

Video-Specific Questions

1. What are some examples that illustrate the presence of variation?

Stock prices vary based on fluctuating market perceptions of performance, general state of the economy, news releases about the company
The taste of specialty coffee varies depending on the barista's skill, age of the beans, temperature of the roast
The time it takes to complete a transaction at a drive-through bank window varies based on the type of transaction, efficiency of the window teller, number of bank customers in line before you
Downloading speed of a website varies based on type of internet connection, type of computer, complexity of the webpage design

2. What factors contribute to variation in flight schedules for Southwest Airlines?

Many sources of variation that affect turnaround of an aircraft, and therefore flight schedules, are outside of Southwest's control. They include late arriving aircraft, weather conditions, the number of passengers on the aircraft, the number of aircraft on the ground at the same time, and variation in ground crew service and staffing.

3. How can the variation in data be measured?

When dealing with quantitative data, such as the turnaround time of an aircraft, several numerical measures can be used to express variability. These include the range, the interquartile range (IQR), the variance, standard deviation, and coefficient of variation. In addition, visual displays, such as boxplots and histograms, are useful in determining the spread in the data. Frequency and relative frequency tables, pie charts, and bar charts

are also useful in understanding the distribution of values when dealing with categorical data.

4. Can variation be eliminated? Explain.

Variation is inherent in any process. Not every factor affecting a process can be controlled, therefore variation cannot be eliminated. Even in what appears to be relatively well controlled processes, say in production facilities, variation cannot be reduced to zero (although it can be reduced to tolerable levels to assure quality). For example, an automatic process that fills cereal boxes will not fill every box with exactly the same amount. But the variation among contents is small enough to meet the amount specified in the label. Although variation cannot be eliminated, it can be managed. By studying a process, it may be possible to either eliminate sources of variation or reduce their effects. The latter is done at Southwest Airlines. They use the "five minute early push" to intentionally reduce the adverse effects from factors outside their control. When all goes right and turnaround times are faster than expected, flights can depart five minutes earlier than scheduled. Doing this early in the day provides a buffer against falling behind schedule later in the day due to uncontrollable factors such as poor weather conditions.

Relevant Chapters in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 1 (Statistics and Variation)
Chapter 6 (Displaying and Describing Quantitative Data)

Concept-Centered Teaching Points

Variation is inherent in every process (Chapter 1).

Discuss any process and the type of data that might be collected on the process to illustrate how measurements will exhibit variation. Because the theme of the video is transportation, one possibility is to simply ask students to think about the time it will take them to get to each statistics class during the semester. If they were to record these times, would they expect them to be exactly the same? Would the variability in times be the same for each student? If at the end of the semester we were to compare the average travel time among all students in the class, would we be able to conclude that the student with the lowest average travel time was always the first student to get to class? Why or why not?

Effective use of statistics requires understanding variation (Chapter 1).

Extend the discussion started above by visiting the Bureau of Transportation Statistics website (link provided below). As an example, go to the table found at http://www.bts.gov/publications/bts_special_report/2008_009/html/table_11.html titled *Wait Time for Passenger Screening Compared to Expectations 2006 vs. 2007*. Ask students to look at the statistics provided and discuss if there are real differences in passenger responses from 2006 to 2007. For each response category, the percentage and

its margin of error are reported. Without getting into specifics, you can discuss that drawing conclusions using statistics cannot be done by only comparing percentages, but that the margin of error (which is based on variation) must also be taken into account.

Variation can be measured and quantified (Chapter 6).

Even though it is early in the semester, many students are already familiar with some of the simpler measures of variation (e.g., range). This is a good opportunity to talk about how variation can be measured and quantified. As an example, go to the table found at http://www.bts.gov/publications/bts_special_report/2008_008/html/table_01.html. Ask students to compare the variability in average "taxi-in times" for 2007 to 2006. You can discuss the advantages and disadvantages of a measure such as the range, and perhaps give a general overview of other measures that are typically used in statistical analysis.

Useful Links

http://www.bts.gov

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 3: Decision Making Error Probability

Summary

This video illustrates how *Norchem Drug Testing* designs its multi-stage testing process to minimize the probability of errors. While two errors are possible in drug testing results (false positives and false negatives), Norchem takes special care to avoid false positives given their serious consequences for an individual. The main theme of the video is that Norchem must understand the types of errors possible in hypothesis testing and design its processes to minimize (or eliminate) their adverse effects in decision making.

Video-Specific Questions

1. What possible errors can occur at Norchem Drug Testing?

Two errors are possible at Norchem Drug Testing: a test result comes back positive when it is not (a false positive) or a test result comes back negative when it is not (a false negative). The goal of Norchem's multi-stage testing process is to minimize the probability of an error in either direction.

2. Which type of error is of greater concern to Norchem Drug Testing?

A false positive is of greater concern than a false negative to Norchem Drug Testing (or any other drug testing company) because of its serious consequences. A false positive drug test means that someone's liberties will be restricted when they shouldn't be (e.g., put in prison, under house arrest, etc.). A false positive leads to false testimony against an individual.

3. Norchem Drug Testing tries to ensure that the probability of a false positive is near zero. What type of probability is this? In the context of hypothesis testing, what would we call this probability?

This probability is a conditional probability. We can state it as the probability that the test will come back positive "given" that the true result is negative. If we assume as the null hypothesis that the individual is not using drugs, then rejecting this when it is true results in a false positive, and within the context of hypothesis a Type I error is committed. In statistics, the probability of making a Type I error is known as α.

4. Give a scenario other than drug testing that is designed to keep the probability of making a Type I error small.

The most well known scenario is the judicial system: presumed innocent until proven guilty. Penalizing someone who is innocent is of greater concern than freeing someone

who is guilty. By using the reasonable doubt threshold, the system is designed to minimize the chance that someone who is innocent is found guilty.

Relevant Chapter in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 11 (Testing Hypotheses about Proportions)

Concept-Centered Teaching Points

There are two errors possible in hypothesis testing.

Ask students why we cannot be 100% sure of our decision in hypothesis testing. This is a good opportunity to discuss statistical inference generally and to stress how sample data are used to draw conclusions about population parameters (the reason why we can never be 100% certain about our results). Discuss the two types of errors possible in hypothesis testing: Type I and Type II errors. A Type I error occurs when we mistakenly reject a null hypothesis that is true and a Type II error occurs when we mistakenly fail to reject a null hypothesis that is false.

The level of significance (α).

What is the level of significance (α) in hypothesis testing? Discuss how it is the probability of making a Type I error (rejecting the null hypothesis when it is true). In hypothesis testing we want to control the probability of making a Type I error, therefore we specify α as a very small value such as 0.01, 0.05, or 0.10. The typical α value is 0.05. This is a good opportunity to also discuss rejection regions, the sampling distribution of the test statistic under the assumption that the null hypothesis is true, and the unlikely possibility of obtaining a test statistic value in the rejection region when the null hypothesis is true.

The *p*-value.

What is the *p*-value in hypothesis testing? Stress that like α, the *p*-value is a probability, but that this probability is associated with the value of the test statistic obtained from the data while α is predetermined by the investigator. Specifically, the *p*-value shows how likely it is to observe the test statistic value given that the null hypothesis is true. Therefore, small values indicate that it is not very likely. Ask students how unlikely is unlikely enough for us to reject the null hypothesis. Then you can discuss how the decision is made by comparing the *p*-value to α.

Useful Links

https://www.norchemlab.com/main/drug_references/mythbusters1.pl

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 4: Models Are Not Reality (But Can Be Incredibly Useful)

Summary

This video illustrates how *Starwood Hotels and Resorts* must rely on complex statistical models to predict future redemptions of earned points in its Starwood Preferred Guest program. Although not perfect representations of reality, these models are used to forecast the worth of future redemptions for reporting to the SEC and IRS, as well as for planning at Starwood. The models take into account historical point redemption patterns by tier, travel trends, and changing economic conditions and are continuously adjusted to improve forecast accuracy. The main theme of the video is that these models for predicting future earned point redemptions, while not reality, are invaluable to Starwood's reporting and planning processes.

Video-Specific Questions

1. For what purpose does Starwood Hotels and Resorts use a statistical model?

Starwood Hotels and Resorts uses a statistical model to determine the value of earned points that will be redeemed in its Starwood Preferred Guest program. It must be listed as a liability on the company's balance sheet. The model is used to predict future redemption of points.

2. What statistical model(s) might be appropriate for Starwood's situation?

Multiple regression may be an appropriate methodology for developing Starwood's model if a number of independent variables are considered to affect point redemption. Another possibility is to use a forecasting method that extrapolates historical point redemption patterns into the future.

3. What might be plausible variables to include in the model? What may be some potential sources of uncertainty?

Historical data on point redemption is an important input to the model. Starwood keeps track of how members of the program redeem their points by tier because members in higher tiers tend to redeem more points than those in lower tiers. Consequently, dummy variables coded to represent different tier levels could be included. Potential sources of uncertainty are changing travel trends and general economic conditions.

4. Describe the process by which Starwood Hotels and Resorts seeks to improve the accuracy of the model.

The model is not reality. It is important that Starwood compare the model's predictions to actual point redemptions. As guests redeem their points, the liability on Starwood's financial statements is reduced and a pre-determined payment is made to the property as compensation for the point redemption. These data are collected and used to analyze the predictive accuracy of the model. Over time the model can be adjusted to improve accuracy.

Relevant Chapters in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 18 (Multiple Regression)
Chapter 20 (Time Series Analysis)

Concept-Centered Teaching Points

Multiple regression models are not deterministic (Chapter 18).

This is a good opportunity to discuss how regression (statistical) models are stochastic (probabilistic) rather than deterministic. Give an example of a deterministic model (such as determining a salesperson's commission) and contrast it with a multiple regression model that includes uncertainty (like the one that could be used for Starwood Hotels and Resorts). Ask students how uncertainty is represented in a regression model. Why can't a regression model explain 100% of the variation in the dependent variable? What goes into the "error" term?

Correctly interpreting multiple regression results (Chapter 18).

You can refer to standard multiple regression output for an example and ask students questions about interpreting the results. What measure is used to determine how much variability in the dependent variable is explained by the multiple regression model? What are reasonable values for R^2 in practice? How should multiple regression coefficients be interpreted? What does it mean when a multiple regression model is significant? What does it mean when the regression coefficient for an independent variable is significant?

Multiple regression-based forecasting models for time series (Chapter 20).

This is a good opportunity to discuss regression-based forecasting models. Ask students to think about how multiple regression models can be used with time series data. How can the model capture trend? How can seasonality be represented in the model? Discuss the coding of dummy (or indicator) variables. What are some potential problems of using multiple regression with time series data? Discuss the issue of autocorrelation and the use of the Durbin-Watson statistic.

Useful Links

http://forecasters.org/internet-sites.html

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 5: Graphing and Exploring Data Can Reveal "Eurekas"

Summary

This video illustrates how *Autoliv* uses graphs and charts to understand quality-related data collected from its production processes. Its goal, to assure quality for its customers, relies on identifying and correcting any process problems before defective parts are produced. By graphing data collected from samples of custom steering wheel parts produced in an injection-molding process, Autoliv discovers that one of the mold cavities yields a mean weight for these parts that is off the target specification. This "eureka" enables Autoliv to take corrective action on the process preventing any unacceptable parts from reaching its customers. The main theme of the video is that by using graphs to understand its process data, Autoliv is able to find and correct process problems that would have otherwise gone undetected and may have resulted in dissatisfied customers.

Video-Specific Questions

1. For what purpose does Autoliv **collect data? Give some of examples of the types of data collected.**

Autoliv collects data routinely on every aspect of its operation to assure quality. Like many other companies, Autoliv is interested in improving products and processes, streamlining operations, and reducing costs. The company wants to find deviations from target as soon as possible so that it can identify the problem in the process and correct it to prevent producing defective parts. Autoliv produces custom steering wheel parts using an injection-molding process. Autoliv collects both qualitative and quantitative data for compliance with design specifications. Gloss and color are examples of qualitative data; weight and tear seam thickness are examples of quantitative data.

2. Explain why Autoliv **must rely on sampling for data collection.**

Rather than testing every part manufactured, Autoliv relies on sampling. Sampling is obviously more time and cost effective. In addition, at Autoliv and in many other companies, the quality testing required may actually destroy the product. When destructive testing is involved, it is not feasible to test (and therefore scrap) every part produced in the process.

3. What types of graphs / charts can Autoliv **use to explore the data?**

Histograms, boxplots, and control charts can be used to understand Autoliv's process data.

4. Describe how the "eureka" revealed to Autoliv **in the film was discovered.**

Autoliv discovered that steering wheel parts from the injection molding process differed in weight depending on the cavity used for molding. By understanding its process, Autoliv was able to see that the mean weight of parts produced using one of the cavities was off the target. This also required knowledge of the spread (variability) in weight values. Charting made it possible to see the difference between the mean weights of parts produced by the two cavities.

Relevant Chapters in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 6 (Displaying and Describing Quantitative Data)
Chapter 25 (Quality Control)

Concept-Centered Teaching Points

The five-number summary and boxplot are used to explore data (Chapter 6).

Ask students to give some summary measures for quantitative data. Record those that are part of the five-number summary (minimum value, first quartile, median, third quartile, maximum value). Why are these measures useful for exploring data? What can we learn from these measures? Discuss how these five values can be displayed in graphical form on a boxplot. Discuss the usefulness of boxplots for seeing the spread of a data set, determining if the data set is symmetric, and identifying outliers. Note that it is easy to construct multiple boxplots on the same scale to compare different data sets.

Histograms are useful for understanding quantitative data (Chapter 6).

Histograms are one of the most popular tools for understanding data that are quantitative and continuous. Discuss how histograms are constructed. Ask students to think about why histograms are so useful. Stress that along with center and spread, we can use histograms to determine the shape of a distribution (e.g., unimodal, bimodal, symmetric, or skewed) as well as to detect outliers.

The structure and purpose of control charts (Chapter 25).

Control charts are arguably the most important tools in statistical process control. Discuss the basic construction of control charts (centerline, upper and lower control limits). While there are many types of control charts (attributes and variables), discuss the basic idea of using these charts to find "out of control" points. Using Autoliv to provide context, you can talk about how samples (subgroups) are selected periodically from a process and some quality characteristic (e.g., weight) is measured for each item in the sample. Then subgroups statistics are calculated (e.g., mean and standard deviation) and plotted on control charts. A subgroup statistic that falls outside the control limits is very unlikely, so it is considered an "out of control" point. Stress that the purpose of control charts is to both stabilize and improve processes.

Useful Links

http://www.asq.org/learn-about-quality/data-collection-analysis-tools/overview/histogram.html

http://www.asq.org/learn-about-quality/data-collection-analysis-tools/overview/control-chart.html

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 6: Quantifying Uncertainty in Making Estimates

Summary

This video illustrates how *McDonald's* collects data to estimate its drive-through service time. Not only is measuring variability key to quantifying the uncertainty surrounding this estimate, but it provides McDonald's with information about the consistency of its drive-through service and opportunities for improvement. By understanding the sources of variability in their drive-through service times, McDonald's can make changes to reduce variability and provide its customers with consistently high quality drive-through experiences. The main theme of the video is that McDonald's must be able to quantify uncertainty by measuring variability to obtain meaningful estimates of its drive-through service time.

Video-Specific Questions

1. McDonald's makes the claim that it takes 90 seconds or less from the time a car stops at the order point to delivery of an order through the window. How can McDonald's substantiate this claim when drive-through times vary?

Even though drive-through service times at McDonald's vary, they can back their claim by collecting data. Data allows McDonald's to be relatively confident that the drive-through service time is, on average, within a certain range. The appropriate statistical inference procedure for this scenario is to construct confidence intervals for the average drive-through service time.

2. How does McDonald's quantify the uncertainty surrounding its estimate of drive-through service time?

When constructing a confidence interval for a mean, the uncertainty surrounding the estimate depends on the level of confidence desired, the variability in drive-through service times, and sample size. The more data collected, that is the larger the sample size, the better the estimates. Also, better estimates can be obtained by reducing the variability in drive-through service times, say through quality improvement efforts.

3. How can the variability in drive-through service times be quantified? Why is it important to understand variability?

Measures of variability that can be used for drive-through service times are the range, variance and standard deviation. The standard deviation is used in constructing a confidence interval for the mean. It is important to measure (and monitor) the variability because even if the mean service time is on target, a large amount of variability means that too many customers will have a poor drive-through experience. In addition, by

understanding the sources of variability (inadequate staffing, inefficient layout, inappropriate equipment) efforts can be made to improve the process and reduce variability. Data collection after making improvements can determine whether improvement efforts were successful.

4. Why is it important for McDonald's to effectively communicate its statistical findings? What types of decisions depend on these data?

It is important to communicate statistical findings in everyday language for management to make decisions about staffing (e.g., using a wireless order taker), layout design (e.g., side-by-side drive-through), and capital investments.

Relevant Chapters in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 12 (Confidence Intervals and Hypothesis Tests for Means)
Chapter 10 (Confidence Intervals for Proportions)

Concept-Centered Teaching Points

Sample data are used to estimate population parameters (Chapter 10).

While the example in the video deals with estimating the mean drive-through service time at McDonald's, you can ask students to give examples of other parameters that may be estimated (e.g., the proportion). Ask them for specific examples. You can take one of these examples and give them a value for a "point" estimate and ask how they might evaluate the "accuracy" of such an estimate. This can lead to discussing the advantages of developing an interval estimate (confidence interval) rather than just using a point estimate for a population parameter.

The concept of margin of error (Chapter 10 or 12).

Extend the discussion by focusing on the margin of error. Most students will be familiar with poll results that report the margin of error in terms of percentage points (e.g., see http://www.usatoday.com/news/washington/2009-07-13-poll-health-care). Discuss what affects the margin of error (e.g., sample size, confidence level) and the tradeoff between certainty and precision. This is also a good time to ask students why it is not possible to be 100% confident that an interval estimate contains the true population parameter.

Confidence intervals should be interpreted correctly (Chapter 12).

Students often have difficulty with the concept of sampling distributions and the correct interpretation of confidence intervals. Consider using the example about McDonald's estimating its mean drive-through service time to provide context and develop a 95% confidence interval. Give students a variety of interpretations that are incorrect (e.g., the mean drive-through service time at McDonald's is 90 seconds 95% of the time). Discuss why it is easy to misinterpret a confidence interval and provide the correct interpretation.

Useful Links

http://www.gallup.com/Home.aspx

http://www.usatoday.com/news/washington/2009-07-13-poll-health-care_N.htm

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 7: Statistics Turns Data into Information

Summary

This video illustrates the types of data collected by *Southwest Airlines* to understand its own operations as well as to report to the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS). Southwest Airlines and the BTS use statistics to turn these data into meaningful information that can be used to track airline performance and help passengers make informed decisions. Along with collecting these data within a contextual perspective (Who, What, Why, When, Where), common descriptive statistics such as percentage tables and pie charts are used to organize and present data in a meaningful way. The main theme of the video is that statistics are needed to turn raw data into useful information.

Video-Specific Questions

1. What types of data are collected by Southwest Airlines and reported to the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS)?

Examples of the types of data collected by Southwest Airlines and reported to the BTS include on-time arrival data (tracking when aircraft depart the gate, when they arrive at the gate, when they land, when they take off), customer complaints, and how many bags are lost. These data can be qualitative or quantitative. For instance, on-time arrivals can be categorical (on time or not on time) or quantitative (number of minutes early or late relative to the scheduled arrival time).

2. Describe the data collected, maintained, and published by the Department of Transportation about airlines from a contextual perspective (*Who, What, Why, When, Where*).

Who: which flights
What: flight delays, consumer complaints, mishandled baggage
Why: extreme weather, security delays, other internal process problems
When: the date each occurred
Where: which airport was involved

3. How does the BTS turn raw data into meaningful information?

The BTS uses common descriptive statistics to turn raw data into meaningful information. The specific descriptive statistics used in the video are pie charts (showing the percentage of delays attributed to various causes for a given month across all airlines) and relative frequency (percentage) tables (showing the percentage of air carrier delays

due to circumstances within the airline's control, such as maintenance problems, for each airline).

4. Why is the information obtained from data collected about the airline industry valuable? What decisions are based on this information?

This information is valuable for several reasons. It can be used to track airline industry performance over time, it can be used for individual airlines to benchmark their own performance and find opportunities for improvement, it can be used to rank airlines, and it can help passengers make informed decisions about which airline to fly.

Relevant Chapters in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 2 (Data)
Chapter 4 (Displaying and Describing Categorical Data)
Chapter 6 (Displaying and Describing Quantitative Data)

Concept-Centered Teaching Points

There are two basic types of variables (categorical and quantitative) (Chapter 2).

Ask students to give examples of categorical variables. You can discuss how categorical variables can be either nominal or ordinal scaled. Ask students to give examples of quantitative variables. Here you can make the distinction between discrete and continuous variables as well as discuss how quantitative data can be either interval or ratio scaled.

Common descriptive statistics appropriate for categorical data (Chapter 4).

The key idea is that we cannot get meaningful information from raw data. Oftentimes students overlook the fact that tables and charts are also descriptive statistics. You can focus on the common tables and charts used to display categorical data (e.g., pie charts, bar charts, frequency tables). Staying with the theme of the video, you can visit the BTS website and find examples of these types of descriptive statistics (e.g., see http://www.bts.gov/publications/highlights_of_the_2001_national_household_travel_survey/html/figure_02.html).

Common descriptive statistics appropriate for quantitative data (Chapter 6).

Again, emphasize that quantitative data in their raw form are not very informative. You can choose to focus on any of the descriptive statistics appropriate for quantitative variables. If you are showing all of the videos in the series, note that histograms and boxplots were featured in Video 5; consequently you may want to focus on other methods for displaying quantitative data such as tables or time series plots (if time is of importance). Staying with the theme of the video, you can visit the BTS website and find quantitative data to use in illustrating these tables and plots (e.g., see

http://www.bts.gov/programs/economics_and_finance/air_travel_price_index/html/table_09.html).

Useful Links

http://www.bts.gov

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 8:
One Outlier Can Change the Results of a Statistical Analysis

Summary

This video illustrates how *Starwood Hotels and Resorts* identifies outliers and researches why they happened in its historical demand data to better forecast demand for its rooms in the future. Overlooking outliers can lead to incorrect forecasts that result in negative consequences such as improper staffing or loss of investor confidence. The importance of detecting and understanding the reasons for outliers is stressed. The main theme of the video is that an outlier can greatly affect the results of statistical analysis and that, like at Starwood Hotels and Resorts, every effort should be made to identify and investigate outliers.

Video-Specific Questions

1. How do managers at Starwood Hotels forecast demand for rooms? What goes into generating a good forecast?

In addition to considering future bookings from special events and current economic conditions, managers at Starwood Hotels forecast demand by looking for patterns in historical occupancy demand data.

2. What is an outlier? When outliers are found in the historical room occupancy data, how does Starwood Hotels deal with them?

Outliers are unusual observations. In the historical room occupancy data for Starwood Hotels, outliers are unusual one-time events that triggered a spike or drop in occupancy. For example, a sporting event like the Super Bowl would trigger a spike; an extreme weather event preventing guests from reaching their destinations would trigger a drop. When found, outliers are examined further to determine the cause and whether it is likely to happen again.

3. What statistical methods are particularly sensitive to outliers?

Many statistical measures and methods are sensitive to outliers. Methods relevant in this scenario for which the results may be affected greatly by outliers include trend analysis, regression analysis and correlation.

4. How can outliers be detected?

Graphical methods such as a scatterplots and time series plots can be used to detect outliers. Also, converting data values to z-scores will identify those values that are more than three standard deviations away from the mean.

Relevant Chapters in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 7 (Scatterplots, Association, and Correlation)
Chapter 8 (Linear Regression)

Concept-Centered Teaching Points

Use a scatterplot to detect outliers (Chapter 7).

Correlation and linear regression are particularly affected by outliers. It is a good opportunity to emphasize the importance of constructing the scatterplot as a first step not only to check the linearity condition but as a means for detecting any outliers before using these types of statistical analyses.

An outlier can distort a correlation (Chapter 7).

Discuss the ways in which an outlier can distort a correlation (e.g., Bozo the Clown effect: adding Bozo to the data inflates the correlation between shoe size and IQ). You can ask students to provide other examples.

Use residual plots to detect outliers in regression (Chapter 8).

Regression analysis results are also affected by extraordinary points. Here you can discuss how points in regression can be unusual in two ways: as outliers or influential points. This is also a good opportunity to introduce residual plots and how they can be used to detect outliers when regression is used for cross-sectional or time series data (like the historical demand data at Starwood Hotels and Resorts).

Useful Links

For an interesting article published in the *Journal of Statistics Education* on outliers see http://www.amstat.org/publications/jse/v13n1/datasets.hayden.html.

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 9: Collecting Good Data is as Important as Analyzing It

Summary

This video illustrates the importance *Deckers Outdoor Corporation* places on collecting good data. Good data are timely, accurate, and relevant. Good data are required to back up decisions both within Deckers Outdoor Corporation and along its supply chain. Data are collected to help understand the root cause of defects, to make decisions about raw materials, manufacturing processes, shipping, storage, inventory, and even where to locate stores. The main theme of the video is that the best statistical analyses cannot compensate for not having good data.

Video-Specific Questions

1. What factors affect footwear purchasing patterns?

While factors such as energy prices, consumer spending and economic conditions affect footwear purchasing patterns, companies like Deckers Outdoor Corporation realize the need to collect good data to capture their share of the market. They need to understand industry trends to design footwear that customers want to buy.

2. Give examples of the categorical data collected by Deckers Outdoor Corporation. What are the sources of its data?

Not all of the data Deckers collects is numeric. They collect data on the styles, colors, and sizes of footwear that sell. Sources for these data include feedback from their sales force, retail store accounts, and online retail sales managers. These data are useful in determining what designs to produce, in what quantities, and where to distribute them.

3. What types of "good" data are required in managing Deckers' supply chain?

To manage the supply chain, good data need to be collected on customer defects to get to the root cause of the problem. In this way the process can be changed to prevent rather than correct defects. Also, good data need to be collected to understand distribution center performance.

4. Give examples of how missing data can affect Deckers Outdoor Corporation.

Inventory in Deckers' online stores is based on demand forecasts. The online retail manager checks online sales daily, paying attention to how actual sales compare to forecasted sales. If actual sales are much higher than forecasted, she can intervene to secure more inventory so that sales demand can be met. If data are missing in the system, she is unable to intervene and may run short, resulting in dissatisfied customers.

Relevant Chapters in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 3 (Surveys and Sampling)
Chapter 2 (Data)

Concept-Centered Teaching Points

A sample should be representative of the population (Chapter 3).

This is a good opportunity to discuss concepts such as defining the population, how to select a sample that is representative of the population, the importance of randomization, and sample size issues. Ask students to relate these concepts back to Deckers Outdoor Corporation. For example, consider Deckers collecting data on preferences for shoe styles. What is the population? What is (are) the variables of interest? What methods could be used to select a representative sample? What issues are relevant in determining sample size? What are the problems with using a convenience sample (say, at a retail store location) or a voluntary response sample (e.g., broadcasting a survey in their online store)?

Careful wording of questions is key in collecting good survey data (Chapter 3).

Begin the discussion by noting that even if a sample is representative of the population, there are other sources of bias that must be considered. The importance of carefully wording questions on a survey so that they are understood as intended should be stressed. Ask students for examples of poorly worded questions or even questions that are intentionally misleading. You can also discuss how pilot pre-testing of a questionnaire can be done to elicit respondent feedback to help ensure a valid survey.

It is important to provide a contextual perspective for data (Chapter 2).

The main point here is that when data are collected a sufficient amount of information should be recorded to provide context. Data without context are useless. Again, using Deckers as an example, consider the data it collects on the styles, colors, and sizes of footwear it sells through various outlets (retail stores, online, etc.). Ask students to provide the *Who, What, When, Where,* and *Why* for this type of scenario.

Useful Links

http://www.aapor.org
(AAPOR = American Association for Public Opinion Research)

BUSINESS INSIGHT VIDEOS
A Guide for Use with *Business Statistics* by Sharpe, De Veaux, and Velleman

Business Insight Video 10: Checking the Assumptions and Conditions of a Model is Crucial

Summary

This video illustrates the steps involved in developing new products at *McDonald's.* From gathering data to track quick service restaurant trends and identify gaps in McDonald's menu offerings to carefully designed consumer taste tests of newly developed products, statistics plays a vital role in both the development process and successful launch of new products. Checking that the assumptions and conditions of the models used for analyzing the data is critical to making correct decisions about new creations and different product attributes. The main theme of the video is that successful new product launches at McDonald's, necessary for staying competitive in the quick service restaurant business, rely on the appropriate use of statistical models for analyzing customer data.

Video-Specific Questions

1. What is the first step in the process of developing new products at McDonald's?

The first step in McDonald's new product development process involves gathering data to track customer trends across the country, determine gaps that exist relative to these trends in the McDonald's menu, and determine what is on McDonald's competitors' menus that works for them.

2. What type of statistical methodology would be appropriate to analyze consumer panel taste tests?

Various types of experimental designs could be used depending on the number of different factors being considered and whether blocking is used. For instance, a randomized block design could be used in which each consumer tastes and rates all of the different new products. Similarly, a paired *t*-test could be used in the scenario described in which McDonald's is trying to decide if a product tastes better if made on one piece of equipment versus another (each consumer tastes both products made on each piece of equipment).

3. What sorts of conditions and assumptions might need to be checked?

The conditions and assumptions required for ANOVA would have to be checked. If blocking is used, the order in which consumers taste the products should be randomized. The ratings data should have similar variances for the different products. The ratings should also be normally distributed to allow statistical inference.

4. What types of data does McDonald's collect in the final steps before a new product is launched nationally?

Based on their carefully designed experiments, McDonald's continuously changes the product until the desired taste test ratings are achieved. If the new product does well in a few of their restaurants, then McDonald's begins an advertised sales test. They promote the new product in a few markets across the country and, based on the data collected, determine whether it is ready to be launched nationally.

Relevant Chapters in *Business Statistics* by Sharpe, De Veaux, and Velleman
Chapter 14 (Paired Samples and Blocks)
Chapter 22 (Design and Analysis of Experiments and Observational Studies)

Concept-Centered Teaching Points

There are benefits in using a paired (or blocked) design (Chapter 14).

This is a good opportunity to discuss one of the principles of experimental design (pairing or blocking). Because McDonald's is featured in the video, you can ask students to think about the best way to design a taste test. Should the same panelists taste all of the newly developed menu items? What are the benefits of doing so? How might not using the same panelists to taste all the menu items affect the results? Ask students to think of other scenarios (e.g., to evaluate weight loss programs) in which pairing (or blocking) is preferable to using independent groups.

Adhere to the basic principles of experimental design (Chapter 22).

In addition to blocking, you can discuss the other principles of experimental design: control, randomization, and replication. Ask students to think about how these principles apply to McDonald's taste testing. Ask students to suggest other scenarios where these principles apply (e.g., testing the efficacy of a new medical treatment). This is a good opportunity to stress the difference between experimental and observational studies.

Checking assumptions and conditions is critical for drawing valid conclusions from data (Chapter 22).

The key concept here is that model assumptions and conditions must be met to draw valid conclusions from data obtained in experimental studies. Stress that to analyze data using ANOVA methods, it is important to check that the following assumptions are satisfied: independence (achieved by randomization), equal variance, and normal population. Ask students how these assumptions can be checked Boxplots can be used to check that variances are similar across groups and normality can be checked using histograms or normal probability plots.

Useful Links

http://www.asq.org/learn-about-quality/data-collection-analysis-tools/overview/design-of-experiments.html

Chapter 2 – Data

SECTION EXERCISES

SECTION 2.1

1. a) Each row represents a different house that was recently sold. It is best described as a case.
b) Including the house identifier, there are seven variables in each row.

2. a) Each row represents a different transaction (not customer or book). It is best described as a case.
b) Including the transaction identifier, there are eight variables in each row.

SECTION 2.2

3. a) House _ID is an identifier (special type of categorical); Neighborhood is categorical (nominal); Mail_ZIP is categorical (nominal – ordinal in a sense, but only on a national level); YR_BUILT is quantitative (units – year), but could also be treated as categorical (ordinal); FULL_MARKET_VALUE is quantitative (units – dollars); SFLA is quantitative (units – sq. ft.).
b) These data are cross-sectional. Each row corresponds to a house that recently sold so at approximately the same fixed point in time.

4. a) Transaction ID is an identifier (special type of categorical); Customer ID is an identifier (special type of categorical); Date is categorical or may be treated as numerical if redefined as how many days ago the transaction took place; ISBN is an identifier (special type of categorical); Price is quantitative (units – dollars); Coupon is categorical (simply nominal); Gift is categorical (simply nominal); Quantity is quantitative (unit – counts).
b) These data are cross-sectional. Each row corresponds to a transaction at a fixed point in time. However the date of the transaction has been recorded. Consequently, since a time variable is included the data could be reconfigured as a time series.

SECTION 2.3

5. The real estate data of Exercise 1 are not from a designed survey or experiment. Rather, the real estate major's data set was derived from transactional data (on local home sales). The major concern with drawing conclusions from this data set is that we cannot be sure that the sample is representative of the population of interest (e.g., all recent local home sales or even all recent national home sales).

6. The student is using a secondary data source (from the Internet). The main concerns about using these data for drawing conclusions is that the data were collected for a different purpose (not necessarily for developing a stock investment strategy) and information about how, when, where and why these data were collected may not be available.

CHAPTER EXERCISES

7. **The news.** Answers will vary.

8. **The Internet.** Answers will vary.

9. **Sales.** The description of the study has to be broken down into its components in order to understand the study. *Who*– who or what was actually sampled–months at a major U.S. company; *What*–what is being measured–money spent on advertising ($ thousand) and sales ($ million); *When*–monthly from 2004–2006; *Where*–United States (assumed); *Why*–to compare money spent on advertising to sales; *How*–how was the study conducted–not specified; *Variables*–what is the variable being measured–there are 3 variables–the date, the amount of money spent on advertising which is quantitative, and sales which is quantitative; *Source* – data are not from a designed survey or experiment; *Type*–data are time series; *Concerns*–none.

10. **Food store.** *Who*– who or what was actually sampled–existing stores; *What*– what is being measured–weekly sales ($), town population (thousands), median age of town (years), median income of town($), and whether or not the stores sell beer/wine; *When*–not specified; *Where*–United States; *Why*–the food retailer is interested in understanding if there is an association amongst these variables to help determine where to open the next store; *How*–how was the study conducted–data collected from their stores;

Variables–what is the variable being measured– sales ($), town population (thousands), median age of town (years), median income of town($), which are all quantitative. Whether or not the stores sell beer/wine is categorical; *Source* – data are not from a designed survey or experiment; *Type* – weekly sales are time series, all other variables are cross-sectional; *Concerns*–none.

11. Sales II. *Who*– who or what was actually sampled–quarterly data from a major U.S. company; *What*–what is being measured–quarterly sales ($ million), unemployment rate (%), inflation rate (%); *When*–quarterly from 2004–2006; *Where*–United States; *Why*–to determine how sales are affected by the unemployment rate and inflation rate; *How*–how was the study conducted–not specified; *Variables*–what is the variable being measured–quarterly sales ($ million), unemployment rate (%), and inflation rate (%) which are quantitative; *Source* – data are not from a designed survey or experiment; *Type* – data are time series; *Concerns*–none.

12. Arby's menu. *Who*–Arby's sandwiches; *What*–type of meat, number of calories (in calories), and serving size (in ounces); *When*–not specified; *Where*–Arby's restaurants; *Why*–assess the nutritional value of the different sandwiches; *How*–information gathered on each of the sandwiches offered on the menu; *Variables*–the number of calories and serving size (ounces) are quantitative, and the type of meat which is categorical; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

13. MBA admissions. *Who*–MBA applicants; *What*–sex, age, whether or not accepted, whether or not they attended, and the reasons for not attending (if they did not accept); *When*–not specified; *Where*–a school in the northeastern United States; *Why*–the researchers wanted to investigate any patterns in female student acceptance and attendance in the MBA program; *How*–data obtained from the admissions office; *Variables*—sex, whether or not the students accepted, whether or not they attended, and the reasons for not attending if they did not accept (all categorical) and age (years) which is quantitative; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

14. MBA Admissions II. *Who*–MBA students; *What*–each student's standardized test scores and GPA in the MBA program; *When*–2000–2005; *Where*–London; *Why*–to investigate the association between standardized test scores and performance in the MBA program over five years (2000–2005); *How*–not specified; *Variables*—standardized test scores and GPA, both quantitative variables; *Source* – data are not from a designed survey or experiment, data are available from student records; *Type* – data are time series; *Concerns*–none.

15. Pharmaceutical firm. *Who*–experimental participants; *What*–herbal cold remedy or sugar solution, and cold severity; *When*–not specified; *Where*–major pharmaceutical firm; *Why*–scientists were testing the effectiveness of an herbal compound on the severity of the common cold; *How*–scientists conducted a controlled experiment; *Variables*—there are 2 variables. Type of treatment (herbal or sugar solution) is categorical, and severity rating is quantitative; *Source* – data come from an experiment; *Type* – data are cross-sectional; *Concerns*–the severity of a cold might be difficult to quantify (beneficial to add actual observations and measurements, such as body temperature). Also, scientists at a pharmaceutical firm could have a predisposed opinion about the herbal solution or may feel pressure to report negative findings about the herbal product.

16. Start-up company. *Who*–customers of a start-up company; *What*–customer name, ID number, region of the country, date of last purchased, amount of purchase ($), and item purchased; *When*–present day; *Where*–United States (assumed); *Why*–the company is building a database of customers and sales information; *How*–assumed that the company records the needed information from each new customer; *Variables*—there are 6 variables: name, ID number, region of the country, and item purchased which are categorical and date and amount of purchase ($) are quantitative; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–although region is coded as a number, it is still a categorical variable.

17. Vineyards. *Who*–vineyards; *What*–size (acres), number of years in existence, state, varieties of grapes grown, average case price ($), gross sales ($), and percent profit; *When*–not specified; *Where*–assume

United States as state is recorded; *Why*–business analysts hope to provide information that would be helpful to grape growers in the United States; *How*–not specified; *Variables*—size of vineyard (acres), number of years in existence, average case price ($), gross sales ($), and percent profit are 5 quantitative variables. State and variety of grapes grown are categorical variables; *Source* – data come from a designed survey; *Type* – data are cross-sectional; *Concerns*–none.

18. Gallup Poll. *Who*–1,180 American voters; *What*–region (Northeast, South, etc.), age (in years), party affiliation, whether or not the person owned any shares of stock, and their attitude (scale 1 to 5) toward unions; *When*–not specified; *Where*–United States; *Why*–the information was gathered as part of a Gallup public opinion poll; *How*–telephone survey; *Variables*— there are 5 variables. Region (Northeast, South, etc.), party affiliation, and whether or not the person owned any shares of stock are categorical variables. Age (in years), and their attitude (scale 1 to 5) toward unions are quantitative variables; *Source* – data come from a designed survey; *Type* – data are cross-sectional; *Concerns*–none.

19. EPA. *Who*–every model of automobile in the United States; *What*–vehicle manufacturer, vehicle type (car, SUV, etc.), weight (probably pounds), horsepower (units of horsepower), and gas mileage (miles per gallon) for city and highway driving; *When*–the information is currently collected; *Where*–United States; *Why*–the EPA uses the information to track fuel economy of vehicles; *How*– among the data EPA analysts collect from the automobile manufacturers are the name of the manufacturer (Ford, Toyota, etc.), vehicle type…."; *Variables*— there are 6 variables. Vehicle manufacturer and vehicle type (car, SUV, etc.) are categorical variables. Weight (probably pounds), horsepower (units of horsepower), and gas mileage (miles per gallon) for both city and highway driving are quantitative variables; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

20. Consumer Reports. *Who*–41 refrigerators; *What*–brand, cost (probably $), size (cu ft), type (such as top-freezer), estimated annual energy cost (probably $), overall rating (good, excellent, etc.), and repair history (in percent requiring repair over the past five years); *When*–2002; *Where*–United States; *Why*–the information was compiled to provide information to readers of Consumer Reports; *How*–not specified; *Variables*— there are 7 variables. Brand, type (such as top-freezer), and overall rating (good, excellent, etc.) are categorical variables. Cost (probably $), size (cu ft), estimated annual energy cost (probably $), and repair history (in percent requiring repair over the past five years) are quantitative variables; *Source* – some data (overall rating and repair history) likely come from a designed survey; *Type* – data are cross-sectional; *Concerns*–none.

21. Lotto. *Who*–states in the United States; *What*–state name, whether or not the state sponsors a lottery, the number of numbers in the lottery, the number of matches required to win, and the probability of holding a winning ticket; *When*–1998; *Where*–United States; *Why*–not specified but likely that the study was performed in order to compare the chances of winning the lottery in each state; *How*–not specified but data could be gathered from a number of different sources, such as the state lottery; *Variables*–– there are 5 variables. State name, whether or not the state sponsors a lottery are categorical variables. The number of numbers in the lottery, the number of matches required to win, and the probability of holding a winning tickets are quantitative variables; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

22. L.L. Bean. *Who*–LL Bean catalog recipients; *What*–number of catalogs mailed out, square inches in catalog, and sales ($ million) in 4 weeks following mailing; *When*–this information is currently reported; *Where*–United States; *Why*–to investigate association among catalog characteristics, timing, and sales; *How*–collect internal data; *Variables*— there are 3 variables. Number of catalogs, square inches in catalog, and sales ($ million) are all quantitative; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

23. Stock market. *Who*–students in an MBA statistics class; *What*–total personal investment in stock market ($), number of different stocks held, total invested in mutual funds ($), and the name of each mutual fund; *When*–not specified; *Where*–a business school in the northeast US; *Why*–the information was collected for use in classroom illustrations; *How*–an online survey was conducted, participation was probably required for all members of the class; *Variables*— there are 4 variables. Total personal investment in stock market

($), number of different stocks held, total invested in mutual funds ($) are quantitative variables. The name of each mutual fund is a categorical variable; *Source* – data come from a designed survey; *Type* – data are cross-sectional; *Concerns*–none.

24. Theme park sites. *Who*–potential theme park locations; *What*–country of site, estimated cost ($), potential population size (counts), size of site (hectares), whether or not mass transportation within 5 minutes of site; *When*–2008; *Where*–Europe; *Why*–to to present to potential developers on the feasibility of various sites; *How*–not specified; *Variables*— there are 5 variables. Country of site and whether or not mass transportation within 5 minutes of site are both categorical variables. Estimated cost (€), potential population size (counts) and size of site (hectares) are quantitative; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

25. Indy 2009. *Who*–Indy 500 races; *What*–year, winner, car model, time (hrs), speed (mph), and car number; *When*–1911-2009; *Where*–Indianapolis, Indiana; *Why*–examine trends in Indy 500 race winners; *How*–official statistics kept for each race every year; *Variables*— there are 6 variables. Winner, car model, and car number are categorical variables. Year, time (hrs) and speed (mph) are quantitative variables; *Source* – data are not from a designed survey or experiment; *Type* – data are time series; *Concerns*–none.

26. Kentucky Derby. *Who*–Kentucky Derby races; *What*–date, winner, winning margin (in lengths), jockey, winner's payoff ($), duration of the race (minutes and seconds), and track conditions; *When*–1875-2004; *Where*–Churchill Downs, Louisville, Kentucky; *Why*–examine trends in Kentucky Derby winners; *How*–official statistics kept for each race every year; *Variables*— there are 7 variables. Winner, winning jockey, and track conditions are categorical variables. Date, winning margin (in lengths), winner's payoff ($), and duration of the race (minutes and seconds) are quantitative variables; *Source* – data are not from a designed survey or experiment; *Type* – data are time series; *Concerns*–none.

27. Mortgages. Each row represents each individual mortgage loan. Headings of the columns would be: borrower name, mortgage amount.

28. Employee performance. Each row represents each individual employee. Headings of the columns would be: Employee ID Number (to identify the row instead of name), contract average ($), supervisor's rating (1-10), and years with the company.

29. Company performance. Each row represents a week. Headings of the columns would be: week number of the year (to identify each row), sales prediction ($), sales ($), and difference between predicted sales and realized sales ($).

30. Command performance. Each row represents a Broadway show. Headings of the columns would be: the show name (identifies the row), profit or loss ($), number of investors and investment total ($).

31. Car sales. Cross-sectional are data taken from situations that vary over time but measured at a single time instant. This problem focuses on data for September only which is a single time period. Therefore, the data are cross-sectional.

32. Motorcycle sales. Time series data are measured over time. Usually the time intervals are equally-spaced (e.g. every week, every quarter, or every year). This problem focuses on the number of motorcycles sold by the dealership in each month of 2008; therefore, the data are measured over a period of time and are time series data.

33. Cross sections. Time series data are measured over time. Usually the time intervals are equally-spaced (e.g. every week, every quarter, or every year). This problem focuses on the average diameter of trees brought to a saw mill in each week of a year; therefore, the data are measured over a period of time and are time series data.

34. Series. Cross-sectional are data taken from situations that vary over time but measured at a single time instant. This problem focuses on data for attendance of the third World Series game. Therefore, the data are cross-sectional.

Brief Case – Credit Card Bank

List the W's for these data:

Who – company cardholders
What – offer status (type of offer made to cardholder), credit card charges made by cardholder in August 2008, September 2008, and October 2008, marketing segment, industry segment, amount of spend lift after promotion, average spending on card pre- and post- promotion, whether or not cardholder is a retail customer or enrolled in the program and whether or not spend lift was positive.
Why – to determine what types of offers are most effective in increasing credit card spending
When – most likely in 2008
Where – although not specified, most likely national data collected in U.S.
How – demographic data most likely collected when credit card account was opened and spending data collected during transactions

Classify each variable as categorical or quantitative; if quantitative identify the units:

Variables:
Offer Status – categorical
Charges August 2008 – quantitative ($)
Charges September 2008 – quantitative ($)
Charges October 2008 – quantitative ($)
Marketing Segment – categorical
Industry Segment – categorical
Spend Lift after Promotion – quantitative ($)
Pre Promotion Avg. Spend – quantitative ($)
Post Promotion Avg. Spend – quantitative ($)
Retail Customer – categorical
Enrolled in Program – categorical
Spend Lift Positive – categorical

Chapter 3 – Surveys and Sampling

SECTION EXERCISES

SECTION 3.1

1.

a. False. Sampling error cannot be avoided, even when the sample is unbiased. Sampling error is always present when a sample statistic is used to estimate a population parameter.

b. True.

c. True.

d. False. Randomization will match the characteristics in a way that is unbiased. We can't possibly think of all the characteristics that might be important or match our sample to the population on all of them.

2.

a. False. The fraction isn't important. Although increasing the size of the sample reduces sampling error, it does not guarantee that the sample is representative of the population. If a sample is selected in a bad way (e.g., convenience) , even a very large sample will be biased.

b. False. Given modern methods, it is best to randomize.

c. False. A sample selected properly is also representative. For example, a simple random sample is representative of the population.

d. True.

SECTION 3.2

3.

a. Population–Professional food preparers in the U.S.

b. Sampling Frame–*Chef's Collaborative* membership listing.

c. Parameter–Proportion of professional food preparers in the U.S. who believe that food safety has improved.

d. Sampling method–Simple random sample (SRS).

4.

a. Population– All frequent flyer customers of the airline.

b. Sampling Frame–Airline database of frequent flier customers.

c. Parameter–Proportion who plan to use one of the airline's new hubs in the next 6 months.

d. Sampling method–Simple random sample (SRS).

SECTION 3.3

5.

a. No. It would be nearly impossible to get exactly 500 males and 500 females by selecting people from India at random.

b. Stratified sampling. This allows GfK Roper to divide the population of India into two groups based on gender (called strata) and then randomly select 500 individuals from each group. The resulting sample will consist of an equal number of males and females.

6.

a. No. It would be nearly impossible to get exactly 50 students from each class by selecting them at random from the student body.

b. Stratified sampling. This allows the Business students to divide the student body into four groups based on class (freshman, sophomore, junior and senior) and randomly select 50 students from each group (called strata). The resulting sample will consist of an equal number of students from each class.

7. The consumer advocacy group used systematic sampling. To ensure the sample is random, the consumer advocacy group should have started the systematic sampling by selecting the first member from the *Chef's Collaborative* list at random.

8. The airline used stratified sampling. The population of frequent fliers was divided into groups (strata) defined by tier level (silver, blue and red).

9.

a. Population—Human resources directors of Fortune 500 companies.
b. Parameter—Proportion who don't feel surveys intruded on their workday.
c. Sampling Frame—List of HR directors at Fortune 500 companies.
d. Sample—23% of HR directors don't feel that surveys intruded on their workday.
e. Method—Questionnaire mailed to all (nonrandom).
f. Bias—Nonresponse bias. Those that respond are obviously receptive to participating in surveys. Since who responds is related to the question itself, there is likely a difference between those who respond and those who don't respond to the survey with regard to the question.

10.

a. Population—unspecified.
b. Parameter—Proportion of the population who think businesses should pay for their employees' health insurance.
c. Sampling Frame—none since the population is unspecified.
d. Sample—Individuals who visited the website and responded.
e. Method—Voluntary response (no sampling method employed).
f. Bias—Voluntary response sample. Those who visit the website and respond may be predisposed to a particular answer. This is a potential source of bias.

SECTION 3.4

11.

a. The population of interest is professional food preparers in the U.S.
b. Members of the *Chef's Collaborative* who attended the recent symposium on "Food Safety in the 21st Century" that was held in Las Vegas.
c. The sampling frame is not necessarily representative of the entire group of food preparers. Those who attended the symposium may have different opinions from those who didn't. It is likely that those who attended the symposium have a special interest in the topic. Moreover, rather than selecting members to call in a random fashion, he started from the top of the list. The list was generated as members enrolled for the symposium; therefore those at the top of the list enrolled early (and perhaps were more enthusiastic about the issue). Consequently, there may be differences in opinion between members near the top of the list and those near the bottom. Finally, the script is biased (leading respondents to answer in a particular way) and may lead to an inflated estimate of the true population proportion who think food safety has improved.

12.

a. The population of interest is all frequent flier customers of the airline.
b. Customers who have recently registered for the "Win the trip to Miami" contest on the Internet.
c. The sampling frame may not be representative of all frequent fliers since those who are interested in a trip to Miami may be more likely to fly there. Also, only those with Internet access would have been able to register for the contest. Since customers without Internet access may have different flying preferences than those with access, there is potential bias due to undercoverage. The survey question is biased (encouraging customers to respond in a particular way) and may lead to an inflated estimate of the true proportion of airline customers who would consider traveling through the Miami hub.

13.

a. *Question 1* seems appropriately worded, although using the phrase "state-of-the-art" may give the impression that a higher price could be charged for the service. By noting that the monthly cost for the service is less than what it would cost to have a daily cup of cappuccino, *Question 2* predisposes the respondent to agree that $50 is a reasonable price. *Question 2* does not seem to be appropriately worded.

b. *Question 1* is more neutrally worded than *Question 2*. In addition, *Question 1* addresses the issue more directly: the willingness of customers to pay $50 per month for the new service.

14.

a. Both questions introduce bias by leading respondents to answer in a particular way. *Question 3* describes dial-up Internet connections as being "slow" in comparison with the high speed Internet service being offered and implies that speed affects the enjoyment of Web services. *Question 4* implies that high speed Internet service is important for children's education so it leads respondents to say that they would subscribe to the service.

b. *Question 3* is more appropriately worded than *Question 4*. However, there is room for improvement. For example, it could be reworded as: "Would high speed access improve your experience using the Internet?"

15.

a. True.

b. True

c. False. Measurement error refers to inaccurate responses. Sampling error refers to sample-to-sample variability and is always present when using a sample statistic to estimate a population parameter.

d. True

16.

a. False. This is a voluntary response sample and will likely be biased. Typically respondents to these types of invitations have strong opinions about the issue and won't be representative of the population.

b. False. Constructing a survey that has too many questions increases the time it takes for respondents to participate. A survey that is too long may reduce the response rate and introduce non-response bias.

c. False. A large sample does not ensure a valid survey. If the large sample is not representative of the population, the survey results will be biased.

d. True.

17.

a. This is a multistage design, with a cluster sample in the first stage and a simple random sample in the second stage. Churches are treated as clusters of the population, and three churches (clusters) are selected. Then a random sample of members is selected from each church.

b. Each church (cluster) is assumed to be representative of the population. This may not be the case if they are different (e.g., in terms of prominent ethnicity of its members such as Italian, Irish, etc.). If each church is not representative of the population, then bias will be introduced at the cluster stage.

18.

a. This is a multistage design involving cluster sampling and census. In the first stage, one day is selected at random. In the second stage, five boats (boats are treated as clusters) are selected within that day. In the final stage a census of the number and types of fish is taken for each boat.

b. If the one day that is selected is not representative of all fishing days, that will introduce bias. If the five boats selected within that day are not representative in terms of the numbers and types of fish typically caught by all fishing boats, then that will also introduce bias.

CHAPTER EXERCISES

19. Software licenses.

a. This sample was a voluntary response, not a random sample.

b. There is no confidence in the estimate sampled. Voluntary response samples are almost always biased, and so conclusions drawn from them are almost always wrong.

20. Drugs in baseball.

a. This sample was a cluster sample, with the teams selected being the cluster. A cluster is recognized by a random sample selection of teams within all of baseball. A sample is then taken from the selected teams.

b. It is a reasonable solution to the problem of randomly sampling players. You can sample an entire team at once relatively easily. You could select your random sample prior to showing up unannounced but it would be less efficient to search for the players randomly selected.

21. Gallup.

a. The population of interest is all adults in the United States aged 18 and older.

b. The sampling frame, a list of individuals from which the sample will be drawn, consists of U.S. adults with landline telephones, which are the only numbers available for a study like this.

c. An increasing number within the population (e.g., many college students and others with mobile only service) don't have landline phones, which could create a bias.

22. Defining the survey.

a. They are using a multistage design in which the countries selected are clusters, and then a random sample is drawn within each cluster. They don't specify how the random samples are taken.

b. The difference in population size has no effect on the precision of estimates from these surveys. Only the sample size matters and the sample chosen should be representative of the entire population.

23. Alternative medicine.

a. Population–All Consumer Union subscribers.

b. Parameter–Proportion of Consumer Union subscribers who have used and benefited from alternative medicine.

c. Sampling Frame–All Consumer Union subscribers.

d. Sample–Subscribers who responded.

e. Sampling method–Questionnaire to all subscribers.

f. Bias–Nonresponse. Those who respond could have strong feelings about the topic and affect the results.

24. Global warming.

a. Population–U.S. adults.

b. Parameter–Proportion of sample who believe that global warming has already begun and the proportion of sample who think global warming will never happen.

c. Sampling Frame–U.S. adults aged 18 and over.

d. Sample–1012 randomly selected adults.

e. Sampling method–Random selection method not specified.

f. Bias–Probably not biased. A large sample that was randomly selected was interviewed so it follows that the conclusions could be generalized.

25. At the bar.

a. Population–Adult bar patrons.

b. Parameter–Proportion of sample who thought drinking and driving was a serious problem.

c. Sampling Frame–All chosen bar patrons.

d. Sample–Every 10^{th} person leaving the bar.

e. Sampling method–Systematic sampling (every 10^{th} person).

f. Bias–Probably biased toward thinking drinking and driving is not a serious problem. The sample consisted of bar patrons leaving the bar. A large percentage of them had something to drink, most likely leading to a biased viewpoint. In addition, bar patrons don't reflect what all adults think about drinking and driving.

26. Election poll.

a. Population–City voters.

b. Parameter–Not clearly specified; likely, the proportion of voters who think certain issues are important or favor certain issues.

c. Sampling Frame–All city resident voters.

d. Sample–Every city resident voter in one block from each district.

e. Sampling method–Multistage design for a combination approach. A single block chosen from each district (not clear what method used in selection) represents a cluster. The sample represents all of the residents that could be found and willing to participate. This is a convenience sample.

f. Bias–Parameter(s) of interest not clearly specified. Convenience sampling within block clusters is not random and could produce biased results.

27. Toxic waste.

a. Population–Soil located near former waste dumps.

b. Parameter–Concentrations of toxic chemicals.

c. Sampling Frame–Any accessible soil surrounding a former waste dump.

d. Sample–Soil samples taken from 16 locations near a former waste dump.

e. Sampling method–Not specified how the sample locations were chosen.

f. Bias–Not specified how soil sample locations were chosen and therefore cannot assume they were chosen randomly, perhaps accessibility or some other factors. Unless this is known, it is possible that bias can affect the results if soil taken is more or less polluted than a random selection would produce.

28. Housing discrimination.

a. Population–Landlords in a particular area.

b. Parameter–Proportion of landlords illegally denying fair access to rental apartments.

c. Sampling Frame–All advertised apartments.

d. Sample–Apartments actually visited and inquired about.

e. Sampling method–Not specified how the apartments visited were chosen.

f. Bias–Likely to be a fair study as long as the apartments visited were randomly chosen and not all in one section of town.

29. Quality control.

a. Population–Snack food packages.

b. Parameter–Proportion of snack food packages passing inspection, weight of bags.

c. Sampling Frame–All snack food packages produced in a day.

d. Sample–Packages in 10 randomly selected cases, 1 bag from each case for inspection.

e. Sampling method–Multistage sampling due to a combination of methods. The selection of the 10 cases is a cluster and the sampling selection of an individual bag from each case is probably a random sample, although this is not specified.

f. Bias–Should be unbiased as long as the individual bag chosen is random. There could be differences in the first bag of a case versus the last bag.

30. Contaminated milk.

a. Population–Dairy farms.

b. Parameter–Not clearly specified although perhaps the proportion of dairy farms passing inspection.

c. Sampling Frame–All dairy farms although not specifically stated this way.

d. Sample–Not specified. Probably a random selection of farms and then a random selection of milk samples.

e. Sampling method–Multistage sampling due to a combination of methods. The selection of dairy farms is a cluster and the sampling selection of an individual sample(s) from each dairy is probably a random sample, although this is not specified.

f. Bias–Should be unbiased as long as the farms and the milk samples are randomly selected.

31. Instant poll. The station's faulty prediction is most likely the result of bias. Only people watching the local TV station news have the opportunity to respond. The responders who volunteered to participate may have different viewpoints than those of other voters, who either chose not to respond or didn't have the opportunity to participate (didn't see the news program).

32. Paper poll. The newspaper's faulty prediction is most likely the result of sampling error. The description of the stratified sampling method does suggest that the sample is representative of the voting population. However, it is unclear whether the percentages by party, age, etc. were accurate when compared to the entire voting population. Random selection of individuals within each strata means that the sample statistics will vary from the population parameter. In addition, no measure of a sampling error percentage was given for the result.

33. Cable company market research.

a. Sampling strategy is volunteer response. Bias is introduced because only those individuals who see the ad and feel strongly about the issue will respond. The opinions may not be representative of the rest of the public.

b. Sampling strategy is a cluster of one town selected to be sampled. Bias is introduced because one town may not be representative of all towns.

c. Sampling strategy is an attempted census, accessing all customers. Bias is introduced because of nonresponse to the mailing survey.

d. Sampling strategy is stratified by town, selecting 20 customers at random from each town to be surveyed, including follow up. This strategy should be unbiased and representative of the public opinion about the cable issue.

34. Cable company market research 2.

a. Sampling strategy is volunteer response. Bias is introduced because only those individuals who see the ad and feel strongly about the issue will respond. The opinions may not be representative of the rest of the public.

b. Sampling strategy is still volunteer response. Bias is introduced because only those who are strongly motivated to express their opinions will attend the meetings.

c. Multistage sampling, with a cluster sample within each town, consisting of those who live on a randomly selected street. Bias is introduced if there is a large percentage of residents on the selected street who do not participate or if the selected street is unrepresentative of the town as a whole.

d. Sampling strategy is systematic sampling. This strategy should be unbiased and fairly representative of the public opinion about the cable issue.

35. Amusement park riders.

a. This is a systematic sample (every 10^{th} person in line).

b. It is likely to be representative of all of those waiting in line to go on the roller coaster. It would be useful to compare those who have waited and are now at the front with those who are in the back of the line. Otherwise, survey every 10^{th} person about to board the roller coaster for a more consistent response.

c. The sampling frame consists of persons willing to wait in line for the roller coaster on a particular day within a given time frame.

36. Playground. The managers will get responses only from those who bring children to the park. It is very possible that parents and others who are dissatisfied with the playground's size and condition do not come to the playground.

37. Another ride. Biases exist because it could be that only those who think it is worth waiting for the roller coaster ride are likely to still be in line. Those who don't like roller coasters or don't want to stay in lines are not part of the sampling frame. Therefore, the poll won't get a fair picture of whether park patrons overall would favor more roller coasters.

38. Playground bias. The first sentence points out problems and issues that the respondent may not have noticed, and might lead them to feel they should disagree. The last phrase mentions higher fees which could make people reject the proposed improvements to the playground.

39. (Possibly) Biased questions.

a. This statement is biased because it leads the responder toward yes because of the word "pollute". The word "pollute" conjures up a negative image leading the responder to agree that companies should pay for this behavior. Another way to phrase it would be "Should companies be responsible for costs of environmental cleanup?"

b. This statement is biased because it leads the responder to no because of the words "enforce" and "strict" that conjure up images that could lead a responder to having negative reaction. Another way to phrase it would be "Should companies have dress codes?"

40. More possibly biased questions.

a. This statement seems unbiased, stating the question without extra leading phrases.

b. This statement is biased because it leads the responder to agreeing with space exploration because it has been a "great tradition" of the past. The responder is unlikely to disagree with an activity that is in line with a "great tradition" of the past. A better way to phrase the question would be "Do you favor continued funding for the space program?"

41. Phone surveys.

a. It would be difficult to achieve a random sample in this case because not everyone in the sampling frame has an equal chance of being chosen. People with unlisted phone numbers, without phones, and those at work or away from the home at the designated calling time cannot be contacted.

b. Another strategy would be to generate random numbers and call at random times or select random numbers from the phonebook and call at random times (this doesn't solve the unlisted phone number issue).

c. In the original plan, families that have one person at home are more likely to be included in the study. Using the second plan, more people are potentially included although people without phones or those not home when called are still not included.

d. This change does improve the chance of selected households being included in the study.

e. The random digit dialing does address all existing phone numbers, including unlisted numbers. However, there is still the issue of residents not being home at the time of the call. In addition, people without phones are still left out of the study.

42. Cell phone survey. Cell phones are not used equally in all demographic groups. Retired persons may not have cell phones. Individuals who cannot afford a cell phone or those who choose not to have one except for emergency purposes would also be left out of the study. As cell phones replace land lines and as they become more affordable and commonplace, this strategy would be a viable option.

43. Change.

a. Answers will vary

b. The parameter being estimated is the true mean amount of change that you carry daily just before lunch.

c. Population is now the amount of change carried by your friends. The average parameter estimates the mean of these amounts.

d. The 10 measurements in c) are more likely to be representative of your class (peer group with similar needs) but unlikely for larger groups outside of your circle of friends.

44. Fuel economy.

a. The mean gas mileage for your last 6 fill-ups (sample statistic).

b. The mean gas mileage for your vehicle (population statistic).

c. The results may not represent typical driving habits for an average driver of this car: speed, highway or city driving, aggressive driving, etc.

d. The EPA would be trying to estimate the mean gas mileage for all cars of this make and model.

45. Accounting.

a. Assign numbers 001 to 120 (3 digits required because the maximum number is 120) representing each order in a day. Use random numbers to select 10 transactions to check for accuracy.

b. Separate the transactions and sample each type (wholesale and retail) proportionately. This would be a stratified random sample.

46. Happy workers?

a. If all types of employees are sampled every month in relative proportions to their job types, the study should be without bias. It is not specified how the sample will be chosen. Because the company is doing the study on itself, there could be some inherent bias introduced due to self-interest.

b. The random sample approach assigns numbers to all employees, 001 to 439 (3 digits are required because the maximum number is 439). Then a random number table or random number generator software would be used to select the sample.

c. The random number approach does not recognize the separate labor groups where there are few managers and mostly laborers. A random sample would favor the largest group, the laborers.

d. A better solution would be to stratify by job type (proportionately to the numbers within each type).

e. Answers will vary. Assign numbers 01 to 14 to each person (2 digits are required because the maximum number is 14). Use a random number table or random number generator software to select 2 managers.

47. Quality control.

a. Randomly select 3 cases and then randomly select one jar from each case.

b. Assign numbers 01 to 20 to cases 07N61 to 07N80 respectively. Then generate three random numbers between 01 and 20 and select the appropriate case. Then assign random numbers 01 to 12 to each of the 12 jars within each case. For each case selected, generate a random number between 01 and 12 and select the corresponding jar within each case.

c. The method described involves two separate sampling methods and, therefore, it is multistage sampling.

48. Fish quality. There are some sources of biases for the results. The scientists asked fishermen to bring any fish they caught to the field station for inspection. They did not require it and, therefore, the fishermen that participated volunteered to do so and possibly do not represent the catch of all fishermen. Fishermen with discolored fish might be more likely to bring them in for inspection. In addition, it would have to be assumed that the fish caught and brought in for inspection are representative of all fish downstream of the chemical plant. It could be misleading to suggest that 18% of fish in the river have discolored scales.

49. Sampling methods.

a. Yellow pages may not include all doctor listings. If regular line listings are used, the list may include all doctors. If ads are used, not all doctors would be included and the ones with ad would not be typical of all doctors.

b. This sampling method is not appropriate. The cluster sample chosen (the randomly selected page) will only contain a handful of businesses and maybe only one or two business types.

50. More sampling methods.

a. It does not specify the method for calling local businesses and whether all or a random sample were called. Some people will say that they are willing to sign a petition when asked over then phone but later may not be willing to sign it. The mention of signing a petition may bias business owners to respond positively.

b. If the food court is the largest and perhaps only food court in the airport, then the results would be fairly representative. If travelers don't like the food available, they probably aren't eating there and made other choices.

Brief Case – Market Survey Research

While answers will vary, the questionnaire should begin with a description of the new product idea. If possible, an image of a prototype for the new product could be provided. Questions should address issues such as consumer behavior including habits and usage of such products, attitudes toward the new product and opinions about its various features, acceptance of the new product and likelihood of future purchase.

Some sample questions follow for a new smartphone:

Rate your level of agreement with the following statements on a 5 point Likert scale (1 = Completely disagree; 2 = Somewhat disagree; 3 = Neither disagree nor agree; 4 = Somewhat agree; 5 = Completely agree).

I currently own a smartphone.

I use my mobile phone to send text messages every day.

I use my mobile phone to check e-mail messages every day.

I need a smartphone to run advanced applications.

The new smartphone is superior to the mobile phone (or smartphone) I am using currently.

Below is a list of features available on the new smartphone design. Rate the importance of each on a 5 point scale (1 = Not important; 2 = Somewhat important; 3 = Important; 4 = Very Important; 5 = Extremely Important; 9 = Don't Know).

- 3G support
- Full touch screen …. etc.

Will you purchase the new smartphone when it becomes available on the market?
1 = No
2 = Maybe
3 = Probably
4 = Definitely

For this high technology product, conducting an online survey in conjunction with a cell phone service provider may be a reasonable approach. Current customers could be contacted via e-mail. Responses will help in decisions regarding the design of the new product (by determining customer expectations and preferences for various features) and launch (by predicting its success).

Brief Case – The GfK Roper Reports Worldwide Survey

What is the population of interest?
The population of interest is consumers worldwide. This can be further defined according to market or industry segment. For example, from the questions posed in this case we may define the population as worldwide retail consumers of food and personal care products.

Why might it be difficult to select an SRS from this population?
Obtaining a suitable sampling frame for the population of interest could be very difficult considering the global scale. A suitable sampling frame is required to select an SRS. A more reasonable approach might be to use a multistage sampling scheme that involves cluster sampling and SRS within each cluster.

What are some potential sources of bias?
Potential sources of bias may result from an incomplete sampling frame or undercoverage of portions of the population (i.e., countries excluded from GfK Roper Consulting studies). As with any study, attention should be paid to potential bias from poorly worded questions and/or bias due to nonresponse. In this study special care should be used in wording questions so that they are properly interpreted in light of cultural differences among countries.

Why might it be difficult to ensure a representative number of men and women and all age groups in some countries?
Countries vary in terms of the factors that may affect the ability to access or obtain valid responses from representative numbers of men, women and individuals in all age groups. Depending on the sampling scheme used, these factors may result in the over- or under- representation of certain segments of the population. For example, in some countries there may be significant differences between males and females in education level, literacy rate, working outside the home, or owning property (http://www.nytimes.com/2010/07/01/world/01iht-poll.html). In some cultures interviewing women may be prohibited (i.e., Islamic). With regard to age, technology use is generally more prevalent among younger individuals. The differences in some countries between younger and older segments of the population with regard to the use of cell phones and/or computers may be quite significant.

What might be a reasonable sampling frame?
Since GfK Roper Consulting monitors consumer attitudes and trends for several major market sectors, it most likely uses several means of contacting consumers including telephone (landline and mobile), surveys (online and mail) and in-home face-to-face interviews. Therefore reasonable sampling frames would be lists of telephone numbers or mailing lists of addresses.

Chapter 2
Data

What's It About?

In this chapter we introduce students to data. We talk about the importance of context (the W's), about variables, and make the distinction between categorical and quantitative data. We begin to introduce the vocabulary of Statistics.

Comments

It is valuable to get students involved with data from the start. We don't take a "big picture" approach at this time. There will be plenty of time to build models and draw inferences later. For now, let's just get our hands dirty with the data. When students have a good sense of what kinds of things data can say to us, they learn to expect to listen to the data. Throughout the course, we insist that no analysis of data is complete without telling what it means. This is where that understanding starts.

Rather than head directly for the "real purpose" of the course in the inference chapters, we prefer to emphasize the connection between our work with data and what they tell us about the world. No analysis is complete without a connection back to the real-world circumstances. Setting that stage is the underlying motivation for this chapter. We'll spend the next 4 chapters or so looking at and exploring data without making formal inferences.

Looking Ahead

You might have the students thumb through the book and read the opening of some chapters. Each one starts with a story about a company or business sector and proceed to analyses of related data, and most have additional stories and more data inside. Statistics is about the real world. Among other topics, we'll be discussing Keen footwear, MBNA and credit cards, Whole Foods Market, and even a small business. We need to get students thinking about the context of data and able to make the distinction between categorical and quantitative data. These are fundamental skills for everything that follows, and they'll be used throughout the course.

Class Do's

Get the class thinking about what the term "data" means. Students need to understand that data are not just numbers and that they must have a context (the W's). When data are quantitative, they should also have units. There are two ways we treat data: ***categorical*** and ***quantitative***. Don't get distracted by worrying about ratio, interval, and other distinctions. These are problematic and don't matter for the concepts and methods discussed in this book. Emphasize that the distinction between treating data as categorical or quantitative may be more about how *we* display and analyze data than it is about the variable itself. The variable "sex" is data, but just because we might label the males as 1 and the females as 0 doesn't mean that it's quantitative. On the other hand, taking the average of those 0's and 1's does give us the percentage of males. How about *age*? It is often quantitative, but could be categorical if broken down only into *child*, *adult*, and *senior*. Zip code is usually categorical, but if one business had an "average" zip code for their customers of 10000 while another had 90000, we'd know the latter had more customers in the western United States. Emphasize the importance of the context and the W's in summarizing these data.

Students should recognize that every discipline has its own vocabulary, and Statistics is no exception. They'll need to understand and use that vocabulary properly. Unfortunately, many Statistics words have a common everyday usage that's not quite the same. We'll be pointing those out as we go along.

Emphasize vocabulary words as they come up. One of the first should be *variable*. Make the point that it does not mean exactly the same thing as it did in Algebra. There, we call "*x*" a variable, but often that means that we just don't currently know its value. In Statistics a variable is an attribute or characteristic of an individual or object whose value varies from case to case.

A *statistic* is a numerical summary of data; one of the first you'll likely hear is that the class is x% male. Point out the difference between statistics and data. One comment that helps make the point: contrary to the advertisement that says "Don't be a statistic," you can't be a statistic, only a datum.

Some students will suggest pie charts or histograms. It's sufficient for now to point out that graphical displays are useful visual summaries of data.

Point out that summaries of data can be verbal, visual, and numerical. All are important. In fact, any complete analysis of data almost always includes all three of these.

After looking at the data from your class survey, some students may say things like, "The males are more conservative." Point out the difference between *univariate* and *bivariate* analysis. Note that bivariate is a lot more interesting.

Hope that someone objects to finding an overall average shoe size or to comparing men's and women's sizes—shoe sizes are inconsistent in terms of units. This adds emphasis to the importance of units and the W's.

The Importance of What You Don't Say

We are laying a foundation here. Stretching up to the attic at this point just makes everyone feel unsafe. Many fundamental Statistics terms are left unmentioned in this chapter. We've found it best to leave it that way. We'll get to them when the students have a safe place to file them along with their other knowledge. So we have an unusually long list of terms we recommend leaving for later in the course. In particular, avoid saying the following:

Hypothesis, Inference. These are certainly important in this course, but we have no background for discussing them honestly now, so they would just be confusing and intimidating.

Nominal, Ordinal, Interval, Ratio. "Nominal" is used by some software packages as a synonym for "categorical" as "continuous" is used for "quantitative." These distinctions arise from studies of measurement scales. But it isn't correct to claim that each variable falls into one of these categories. It is the use to which the data are put that determines what properties the variable must have. Ordinal categorical data may come up, but there are no special techniques for dealing with ordered categories in this course. And any differences between interval- and ratio-scaled data are commonly ignored in statistical analyses. If any of these terms were mentioned now, they'd never come up again anyway.[1]

Random, Probability, Correlation. Everyone has some intuitive sense of these terms, and we'll deal with them formally—but not for a while. Students may want to use these terms, but at this

[1] At least not until chapter 23's discussion of nonparametric methods.

early stage in the course, we don't need them. Without background and careful definition, they are likely to be misused.

Class Examples

1. Ask students to tell some things they learned about the class from inspecting the data collected in the opening day's survey. You can use that discussion to develop several of the important points of the chapter.
2. Consider 17, 21, 44, and 76. Are those data? Context is critical—they could be test scores, ages in a golf foursome, or uniform numbers of the starting backfield on the football team. In each case, our reaction changes.
3. Run through some other examples of data, asking about the W's, the variables (what are they, what type is each used as, and what are the units), and so on.

   •A Consumer Reports article on energy bars gave the brand name, flavor, price, number of calories, and grams of protein and fat.

   •A report on the Boston Marathon listed each runner's gender, country, age, and time.

   Solution:

   Consumer Reports
   Who: energy bars
   What: brand name, flavor, price, calories, protein, fat
   When: not specified
   Where: not specified
   How: not specified. Are data collected from the label? Are independent tests performed?
   Why: information for potential consumers
   Categorical variables: brand name, flavor
   Quantitative variables: price (US$), number of calories (calories), protein (grams), fat(grams)

   Boston Marathon
   Who: Boston Marathon runners
   What: gender, country, age, time
   When: not specified
   Where: Boston
   How: not specified. Presumably, the data were collected from registration information.
   Why: race result reporting
   Categorical variables: gender, country
   Quantitative variables: age (years), time (hours, minutes, seconds)

Resources

ActivStats[2]

• Start with Lesson 1 to let students familiarize themselves with the features of the software. Lesson 2 examines types of data and context.

[2] ActivStats (0-321-57719-1) can be purchased from Pearson at www.pearsonhighered.com or bundled with your textbook.

Web Links

• The Data and Story Library (DASL; http://lib.stat.cmu.edu/DASL/) is a source of data for student projects and classroom examples.

• The U.S. Census Bureau

Other

• Read polls, studies, or other reports in newspaper and magazine articles. It's always interesting to see how well (or poorly) they provide information about the W's.

If you have a computer and projection capabilities in class, you can find daily surveys at Gallup and other polling organizations. Current data are often particularly interesting to students. But don't use results of voluntary-response online surveys. We'll be making the point that these are fatally flawed—but we can't say that clearly without concepts and terms that we haven't developed yet.

Basic Exercises

1. The following data show responses to the question "What is your primary source for news?" from a sample of college students.

Internet	Newspaper	Internet	TV	Internet
Newspaper	TV	Internet	Internet	TV
Newspaper	TV	TV	Newspaper	TV
Internet	Internet	Internet	Internet	Internet
TV	Internet	Internet	TV	TV

   a. Prepare a frequency table for these data.
   b. Prepare a relative frequency table for these data.
   c. Based on the frequencies, construct a bar chart.
   d. Based on relative frequencies, construct a pie chart.

2. A cable company surveyed its customers and asked how likely they were to bundle other services, such as phone and Internet, with their cable TV. The following data show the responses.

Very Likely	Unlikely	Unlikely	Very Likely
Likely	Unlikely	Likely	Likely
Unlikely	Unlikely	Likely	Likely
Very Likely	Unlikely	Unlikely	Very Likely
Unlikely	Unlikely	Unlikely	Likely

   a. Prepare a frequency table for these data.
   b. Prepare a relative frequency table for these data.
   c. Based on frequencies, construct a bar chart.
   d. Based on relative frequencies, construct a pie chart.

3. A membership survey at a local gym asked whether weight loss or fitness was the primary goal for joining. Of 200 men surveyed, 150 responded fitness and the rest responded weight loss. Of 250 women surveyed, 175 responded weight loss and the rest responded fitness.

   a. Construct a contingency table.
   b. How many members have fitness as their primary goal for joining the gym?
   c. How many members have weight loss as their primary goal for joining the gym?
   d. Based on the results, should the owner of the gym emphasize one goal over the other? Explain.

4. The following contingency table shows students by major and home state for a small private school in the northeast U.S.

Major Program of Study

Home State	*Biology*	*Accounting*	*History*	*Education*
PA	80	65	55	100
NJ	50	40	65	95
NY	75	50	45	80
MD	65	55	40	40

   a. Give the marginal frequency distribution for home state.
   b. Give the marginal frequency distribution for major program of study.
   c. What percentage of students major in accounting and come from PA?
   d. What percentage of students major in education and come from NY?

5. The following contingency table shows students by major and home state for a small private school in the northeast U.S.

Major Program of Study

Home State	*Biology*	*Accounting*	*History*	*Education*
PA	80	65	55	100
NJ	50	40	65	95
NY	75	50	45	80
MD	65	55	40	40

   a. Find the conditional distribution (in percentages) of major distribution for the home state of NJ.
   b. Find the conditional distribution (in percentages) of major distribution for the home state of MD.
   c. Construct segmented bar charts for these two conditional distributions.
   d. What can you say about these two conditional distributions?

6. The following contingency table shows students by major and home state for a small private school in the northeast U.S.

Major Program of Study

Home State	*Biology*	*Accounting*	*History*	*Education*

PA	80	65	55	100
NJ	50	40	65	95
NY	75	50	45	80
MD	65	55	40	40

a. Find the conditional distribution (in percentages) of home state distribution for the biology major.
b. Find the conditional distribution (in percentages) of home state distribution for the education major.
c. Construct segmented bar charts for these two conditional distributions.
d. What can you say about these two conditional distributions?

ANSWERS

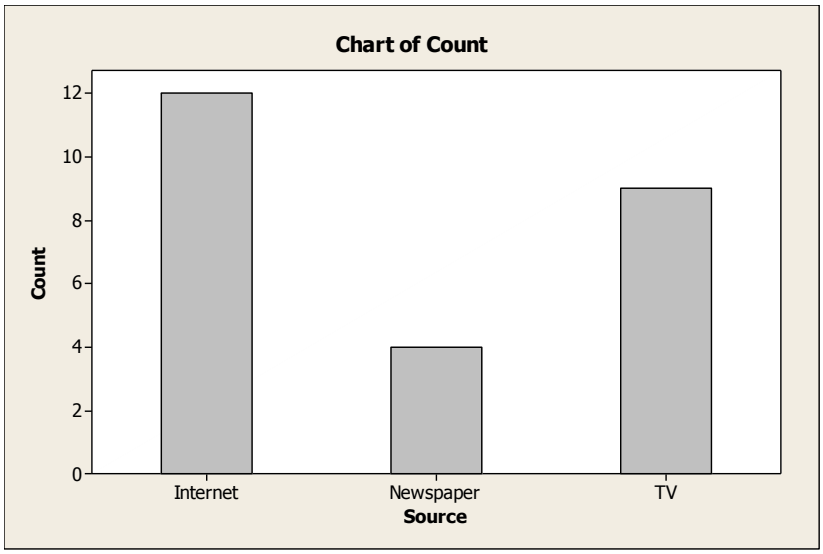
1. a.

News Source	*Number of Students*
Internet	12
Newspaper	4
TV	9

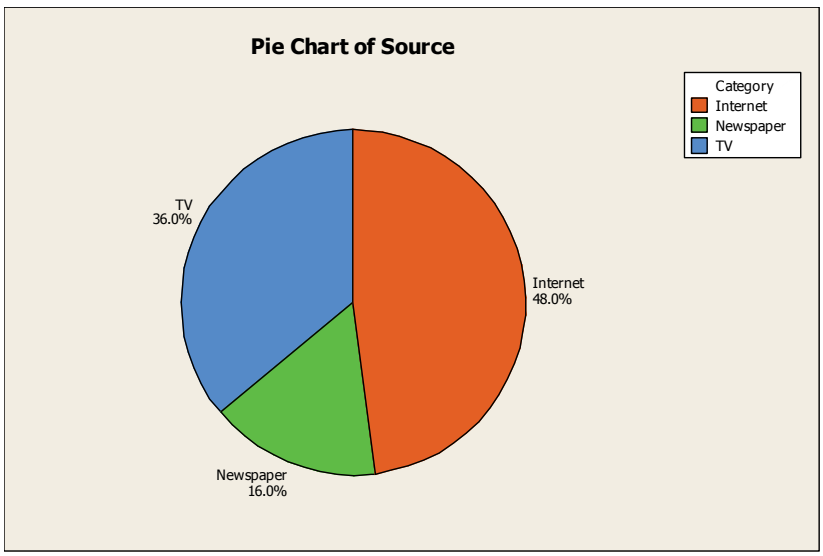
b.

News Source	*% of Students*
Internet	48 %
Newspaper	16 %
TV	36 %

c.



d.



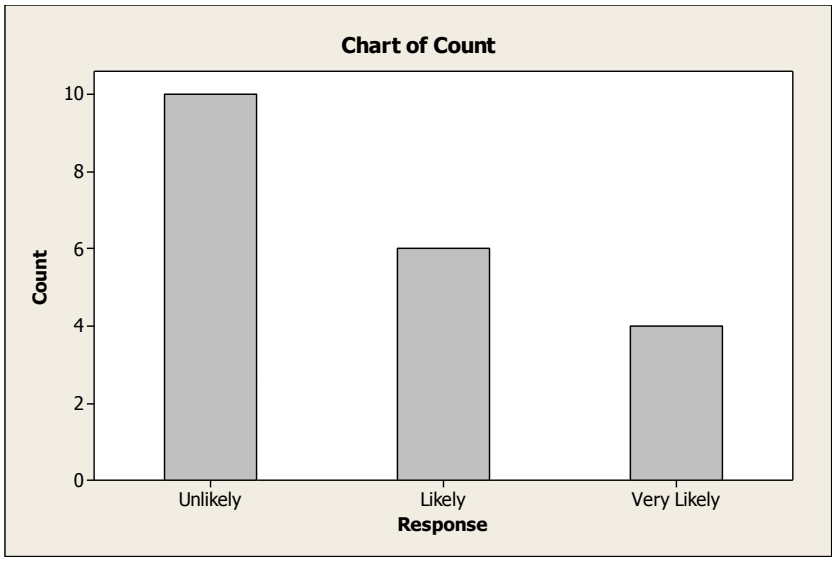
2. a.

Response	*Number of Consumers*
Unlikely	10
Likely	6
Very Likely	4

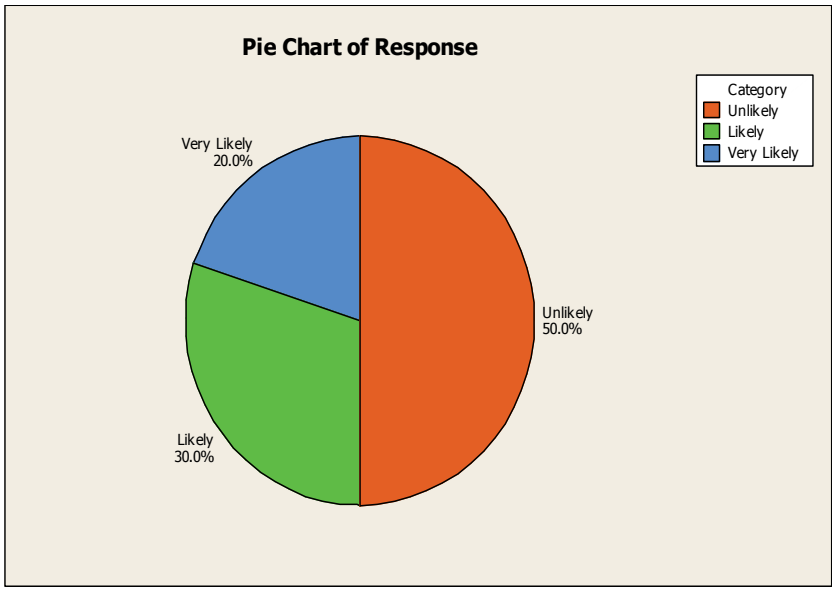
b.

Response	*% of Consumers*
Unlikely	50 %
Likely	30 %
Very Likely	20 %

c.



d.



3. a.

Goal for Gym Membership

Gender	Fitness	Weight Loss	Total
Men	150	50	200
Women	75	175	250
Total	225	225	450

b. 225
c. 225
d. No. 50% of the membership is pursuing each goal.

4. a.

Home State	***Number of Students***
PA	300
NJ	250
NY	250
MD	200

b.

Major	***Number of Students***
Biology	270
Accounting	210
History	205
Education	315

c. 6.5 %
d. 8 %

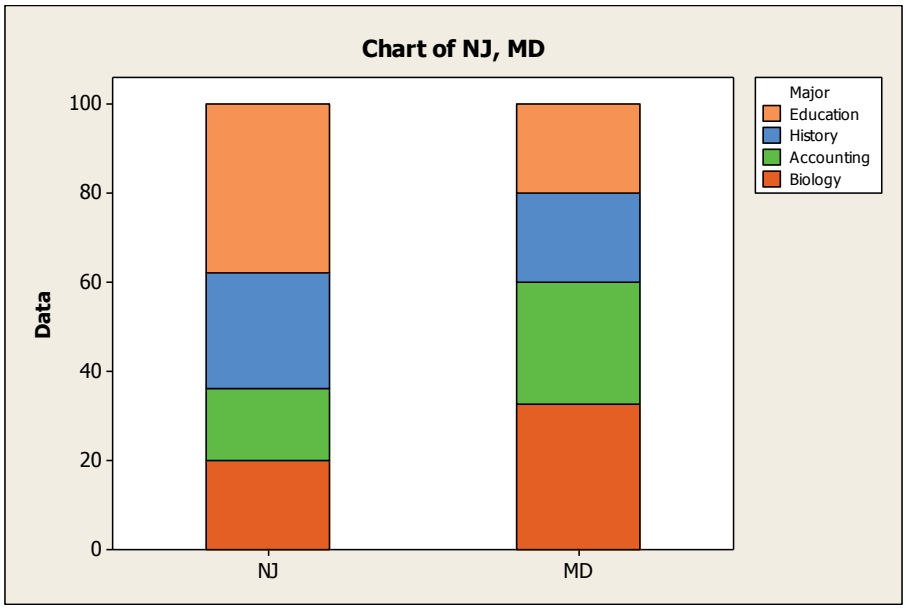
5. a.

Major	***Conditional for NJ***
Biology	20 %
Accounting	16 %
History	26 %
Education	38 %

b.

Major	***Conditional for MD***
Biology	32.5 %
Accounting	27.5 %
History	20 %
Education	20 %

c.



d. More biology and accounting majors come from MD compared to NJ.

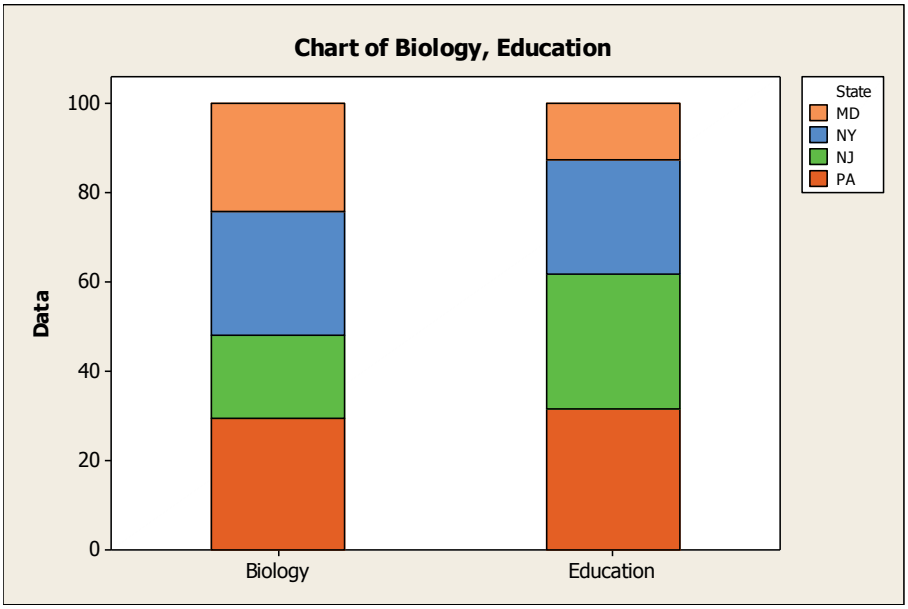
6. a.

Home State	***Conditional for Biology***
PA	29.6 %
NJ	18.5 %
NY	27.8 %
MD	24.1 %

b.

Home State	***Conditional for Education***
PA	31.7 %
NJ	30.2 %
NY	25.4 %
MD	12.7 %

c.



d. Fewer education majors are from MD and more are from NJ compared with biology majors.