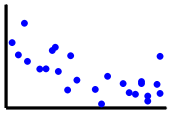


TOPIC D1: STATISTICS, DATA, AND VARIABLES

SPOTLIGHT: THE U.S. CENSUS BUREAU

The U.S. Census Bureau is the leading provider of data regarding the people and economy of the United States. Following independence in 1776, there was an immediate need to count the number of people in the entire country. Secretary of State Thomas Jefferson was the overseer of the first census taken in 1790 that counted 3.9 million inhabitants. It was clear in the early history of the United States that there was a need to collect statistics to help people understand the current status of the country and its growth. The content of the census has changed over time. In 1810, information was collected regarding the manufacturing and quality of products, in 1840 questions were added on fisheries, and in 1850 data were collected on taxation, churches and crime.

The Census Bureau oversees a census of the U.S. population every ten years. In addition, censuses are conducted on economic activity and state and local governments every five years, and every year the Census Bureau conducts over 100 other surveys.

Much of the data collected by the Census Bureau is publicly available on its website at www.census.gov. On the *American Factfinder* page of this Census website, one can find a wide variety of data regarding the population, economic activity, and geography of the United States. There is information about age, education, income and race of different states. There are tables and graphs regarding home ownership, including home values and mortgages, and a wealth of information regarding different types of businesses and industries.

In this topic, we will access data available in the Education section of the Census website. There are extensive data on school enrollment at different levels for all states. Data have been collected to learn about children's academic achievement, differences in achievement due to gender or race, the relative number of children attending private and public schools, and the availability of classes for gifted students. In addition, data are available documenting the educational attainment of adults and the relationship between

educational attainment and job earnings. The site, through its data tables, describes the differences in educational attainment between age groups and between the different regions of the country. The material posted on this website provides a great opportunity to learn about the education of Americans.

PREVIEW

In this opening topic, we get introduced to the science of statistics. Briefly, statistics is the process of formulating one or more questions about our world, collecting relevant data, and organizing and summarizing the data to answer the questions of interest. When we collect data, we will see that there are different ways that variables can be measured, and these different measurement types affect how we work with the data to draw conclusions. In this topic, we will get experience reading statistical studies and graphs in the media, and see how these reports reflect the different parts of a statistical investigation.

In this topic your learning objectives are:

- To understand statistics as a science of learning about our world.
- To understand that the science of statistics has four basic components common to any investigation that collects data to learn about our world.
- To understand the different measurement types of variables.
- To begin to have a critical view towards statistical studies reported in the media.

Topic D1

NCTM Standards

✓In Grades 9-12, all students should understand the meaning of measurement data and categorical data, and of the term variable.

✓In Grades 9-12, all students should evaluate published reports that are based on data by examining the design of the study, the appropriateness of the data analysis, and the validity of conclusions.

WARM-UP ACTIVITY: GETTING TO KNOW YOU

What do you know about your fellow students? Suppose you are interested in learning about the group of students that attend your school. In particular, you might be interested in learning about various physical traits of students such as gender and hair length and about social characteristics such as students' interest in movies, their diets, and their sleeping habits. We can obtain a convenient and interesting data set by asking questions from ourselves. Below is a list of questions that can be answered by every student in the class. In addition, through a class discussion, other questions can be proposed that will be answered by all of the students.

Your instructor will prepare a data set containing the student responses to all of the questions. This data set will be used to illustrate various methods for graphing and summarizing data. Using these descriptive methods, we will be able to draw some general conclusions about the students that attend our school.

QUESTIONS:

1. What is your height in inches? _____

2. What is your gender? _____

3. How many pairs of shoes do you own? _____

4. Choose a number between 1 and 10. _____

5. How many movie DVDs do you own? _____

6. What time (to the nearest half-hour) did you go to bed last night?
   _____

7. What time (to the nearest half-hour) did you wake up this morning?

   _____

8. How much did you spend on your last haircut (including the tip)?

   _____

9. How many hours do you plan to work on a job per week this semester?

   _____

10. For an evening meal, do you prefer water, soda (pop), or milk?

    _____

QUESTIONS GENERATED FROM CLASS DISCUSSION

11. _____

12. _____

13. _____

14. _____

15. _____

Statistics and Data

There is a general confusion about the meaning of "statistics." If we look at the *American Heritage Dictionary of the English Language*, we find two very different definitions of statistics.

1. *Statistics* is a collection of numerical data.
2. *Statistics* is the mathematics of the collection, organization, and interpretation of numerical data.

Let's focus first on the first definition. To many people, statistics are simply numbers like

61, 73, .406

But these are not statistics, they are simply numbers. Statistics are ***numbers with a context*** or underlying meaning. The author is a baseball fan and the above three numbers have interesting contexts:

- 61 is the number of home runs hit by Roger Maris in 1961
- 73 is the number of home runs hit by Barry Bonds in 2001
- .406 is the batting of Ted Williams in 1941 (the last player to have a batting average of over .400 for a single season)

A baseball fan cares about these statistics since they make him or her think about outstanding baseball players and their accomplishments.

We will refer to information collected from people or objects as ***data***. Actually, data is the plural form and a single piece of information is called ***datum***, although data is now commonly used to represent one or more pieces of information.

Why Do We Collect Data?

Why do we care about data? We collect data to help us learn about our world. We start off with some questions about something we wish to learn about, and we find appropriate data to help us answer these questions.

The Second Definition of Statistics

We've seen that statistics are numerical data that we work with. But statistics has a second deeper meaning -- it's the science of using these data to learn about our world and make conclusions.

The science of statistics has four basic components:

- **FORMULATING QUESTIONS**: First, state some questions or problems that we would like to address by collecting relevant data.
- **COLLECTING DATA:** Second, specify effective ways of collecting data that are useful in answering the questions of interest.
- **ORGANIZING & SUMMARIZING**: Next, organize and summarize the collected data to learn about its general features.
- **MAKING CONCLUSIONS**: Last, use the data to make conclusions. (It turns out that probability or chance plays an important role in decision-making.)

Any statistical study reported in the media will have these four components. At the beginning, there will be some questions that motivated the researcher to study a problem. If there were no questions, then there would be no reason to proceed further into a statistical study. Second, the researcher will collect data that he or she believes will be useful in answering the question. We will see that data can be collected or found from many sources. Next the researcher organizes the data in some useful way and make graphs and or calculations that are helpful in answering the main questions. Finally, the researcher has to use the graphs and calculations to address the questions of interest. It is possible that the data are insufficient or inconclusive on answering the questions and perhaps a new statistical study will be undertaken.

In this topic, we illustrate the four components of a statistical study for several examples. We begin with a description of our own study on learning about the educational achievement of Americans and then we look at statistical studies reported in the media.

Example: Educational Attainment

Formulating questions

Let's get started by looking at some educational data. All of us completed high school and most of us will be completing college. We take these accomplishments for granted, but we realize that not everyone in the United States completes high school and

certainly many students don't go to college. There is a current budget crisis in the state of Ohio and the legislature has to prioritize its spending. Should the money go to four-year state universities, or should it go to community colleges that specialize in two-year technical programs? In the discussion about this budget crisis, some people have said that one problem is that a relatively low number of Ohio adults are college educated. These adults are more familiar with technical education and may not appreciate the added value of a four-year college degree in preparing students for a variety of careers.

This brief discussion raises the following questions:

1. Is it common for a person in the United States to complete high school?
2. Are there differences in high school graduation rates between states? If so, which states have higher rates, and which states have lower rates?
3. How likely is it for a United States adult to complete college? How does the college completion rate vary between states?
4. How does Ohio compare with other states with respect to the college completion rate?
5. Is there a connection between a state's high school graduation rate and its college graduation rate?

Collecting Data

Where do we find data? Many books, such as almanacs, have data on population and demographic information on states in the United States. A quick way to get data is through the Internet.

Use your favorite search engine on the Internet and do a search for

educational attainment state

One of the sites it might pick up is from the U.S. Census Bureau:
http://www.census.gov/compendia/statab/cats/education/educational_attainment.html
This file contains a link to an Excel file with the title "Educational Attainment by State." In the footer of the file, we find the source for the data in the table is the 1990 and 2000 Census of Population and the 2008 American Community Survey.

We focus on the data for the most recent year in the table (2008). Here are the first few rows of the table.

State	High school graduate or more	Bachelor's degree or more	Advanced degree or more
Alabama	81.9	22.0	7.7
Alaska	91.6	27.3	9.7
Arizona	83.8	25.1	9.2

For each of the 50 states plus the District of Columbia, this table gives

- the name of the state or district
- the percent of these adults that completed high school
- the percent that obtained a bachelor's degree
- the percent that obtained an advanced degree (masters or above)

How did the U.S. Census Bureau get these data? One possibility is that the bureau asked every single adult in the United States their educational status. Actually, the government does attempt to collect data from all adults every 10 years -- this is called the census. But a census is very expensive and rarely done. These reported high school and college percentages are actually numbers (or statistics) computed based on a sample of adults taken from each state.

How did the Census Bureau take their samples? That's a good question and it should be asked whenever there is sampling reported in an article in the media. We'll talk about basic principles for taking "good" samples in a later topic.

Organizing and summarizing; drawing conclusions

Now that we have listed some questions of interest and have found relevant data, we next have to organize and summarize the data and then use our work to answer our questions. In topics D2 and D3, we will revisit this example and introduce methods of

graphing and summarizing data that will help us learn about the graduation rates of Americans.

Variables and Variable Types

In this example, we have collected different types of information from each state. The object that we are collecting information from is called the ***observational unit***. In this case, the observational unit is the state. The different types of information we collect for each state are called ***variables***. Here some variables are the name of the state, the percent of that state's adults that completed high school and the percent of adults that have a bachelor's degree. There are two distinct types of variables depending on how the variable is recorded. The name of the state is an example of a ***categorical variable*** -- this is in which its values can be grouped into different categories. The percentage of adults that graduate from high school is a ***quantitative variable*** -- this is a variable where the values are numerical and refer to the quantity or size of something.

As a second example, suppose we record the current grade point average, the hair color, and the number of music cds owned for 30 students in a class. Here the student would be the observational unit. Hair color would be a categorical variable, and the grade point average and the number of cds owned would be quantitative variables.

SPECIAL NOTE: Sometimes it can be difficult to tell if a collected "number" is a categorical or a quantitative variable. For example, is a person's social security number quantitative or categorical? For a variable to be of the quantitative type, it must make sense to add, subtract, multiply or divide these values. Is it meaningful to subtract two social security numbers, say 123 55 005 and 222 44 2121? The answer is "no" and that indicates that a person's social security number is an example of a categorical variable.

When we collect data, it is important to recognize if a given data value represents a categorical or quantitative variable. Our exploration of data will depend on its type. The way we explore categorical data will be fundamentally different from our treatment of quantitative data.

PRACTICE: GETTING TO KNOW YOU

In the "Getting to Know You" activity, the students in your class were asked several questions. For each question,

- explain why you (or someone else) might be interested in the answers to this question for a group of students
- state if the answer is a quantitative or a categorical variable

(The first two questions have been filled in for you.)

(a) Question: What is your height in inches?

Why? If a particular student was 70 inches tall, she might want to know how her height compares to the heights of other women at her school.

Type of variable: quantitative

(b) What is your gender?

Why? You might be interested in the relative numbers of men and women at your school.

Type of variable: categorical

(c) How much money do you have with you right now, in change only?

Why?:

Type of variable:

(d) Choose a number between 1 and 10. _____

Why?:

Type of variable:

(e) How many audio CDs do you own? _____

Why?:

Type of variable:

(f) What time did you go to bed last night? _____

Why?:

Type of variable:

(g) What time did you wake up this morning? _____

Why?:

Type of variable:

(h) How much did you spend on your last haircut (including the tip)?

Why?

Type of variable:

(i) How many hours do you plan to work on a job per week this semester?

Why?

Type of variable:

(j) How many cups of coffee did you drink yesterday?

Why?

Type of variable:

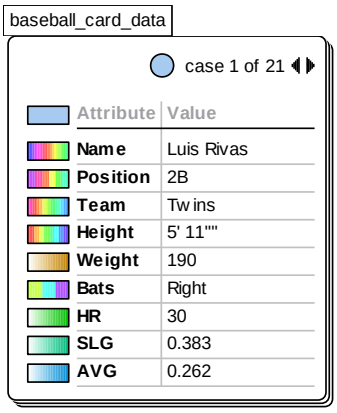
REFLECTION. Think of three additional pieces of information that you would like to learn about your fellow students. For each piece of information, explain why you are interested in this information, state the question you would ask, and describe if the answer to the question would be a quantitative or categorical variable.

Organizing and summarizing data: some initial thoughts

How does one organize and summarize data? To give some initial thoughts about how one gets started on this task, it is helpful to recall an episode from the author's youth. He enjoyed baseball and liked to collect baseball cards. He would buy a number of packs of cards and then spread all of the cards on the rug in my room. Each baseball card contained some data about a particular player, such as his age, height, weight, and statistics of his batting or pitching performance for recent seasons.

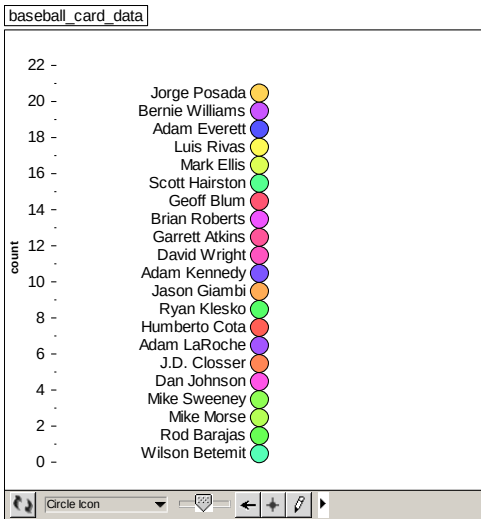
When your author looked at this collection of cards on the rug, he wanted to manipulate them in some way to get a better understanding of the statistics on the cards. In a similar manner, we wish to perform different operations on our dataset to learn about its basic features. We will discuss more formally in Topics D2 and D3 how to graph and summarize a single batch of data. But here we describe some basic operations that might be helpful in organizing data.

It can be fun to revisit one's youth, so the author recently purchased several packs of baseball cards for the 2006 season. There is a variety of data printed on each baseball – the data card below for Luis Rivas illustrates some of the variables collected.

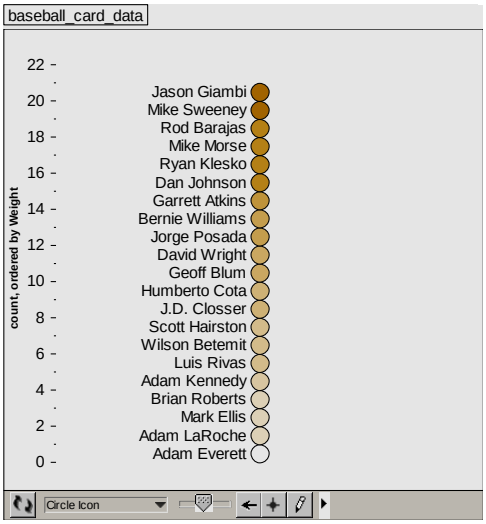


For this card, there are several categorical variables measured such as the player's Name, his fielding Position, the Team that he plays for, and the side of the plate that he bats (variable Bats). There are also quantitative variables available such as the player's height, his weight, the number of home runs (HR) hit in his major league career, and his career slugging percentage (SLG) and his career batting average (AVG).

Suppose one collects all of the 21 baseball cards of players who are not pitchers and have played at least one season in the Major League. We list the names of the players on the 21 cards as shown below.

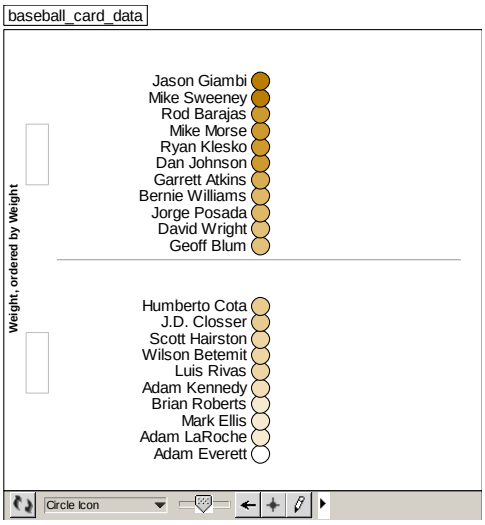


One way of organizing data is to *arrange* or sort the values on the basis of one variable. For example, suppose the author is interested in the weights of the baseball players and so he sorts the cards from the heaviest to the lighter player.



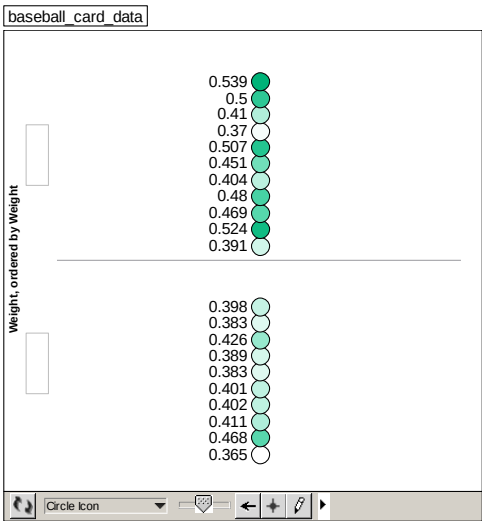
By doing this, we can identify the players who are unusually heavy (such as Jason Giambi) or light (such as Adam Everett) in this group. David Wright is in the middle of this sorted list, so he appears to have an "average" weight in this player group.

Another way to organize data is to *divide* them into two or more groups by the values of one variable. In this baseball example, we might be interested in breaking the players into the "light" players and the "heavy" players, where "heavy" is defined to be a player who weighs at least 200 pounds. In the below figure, we use a horizontal line to break the data into the two groups.



Once the data has been divided into groups by one variable, we might look at the relationship between the variable and a second variable. In our example, we might wonder if the light players and the heavy players differ with respect to another variable. One might think that a bigger player is more likely to hit extra-base hits such as doubles, triples, or homeruns. One measure of the ability of a baseball hitter to get extra-base hits is the slugging percentage (SLG). Do the light and heavy players differ with respect to slugging percentage?

To answer this question, the slugging percentage for each of our 21 players is recorded next to each player.



Looking at values of SLG for the two groups, it does appear that the heavy players have some of the largest values. To see if this first impression is correct, we summarize the slugging percentages in each group. A basic summary is an average such as a mean that is found by summing the SLG for each group and dividing by the number of players. If we do this, we get the following table.

Group	Average SLG
Heavy players	.458
Light players	.403

It does appear that heavy players tend to have a higher slugging percentage – by the table, it seems that heavy players, on average, have a .458 - .403 = .055 higher SLG than light players.

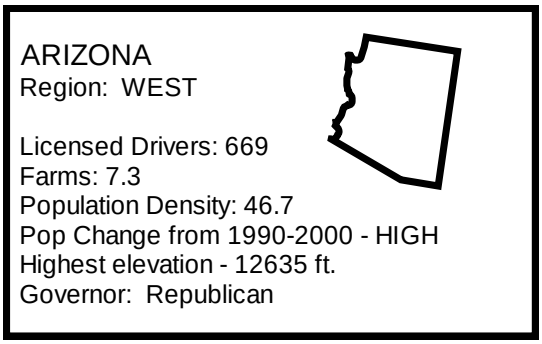
In the following activity, we will use the "arrange, divide, and summarize" data operations to learn about characteristics of different states in the U.S.

ACTIVITY: MEET THE STATES DATA

The United States is a diverse country in many ways. One will find significant differences between the states with respect to population density, geography, climate, employment, and cost of living. By exploring data collected from various states in this activity, we will begin to appreciate the diversity of the U.S. and start thinking about ways of effectively organizing, graphing, and summarizing these data to draw conclusions about this diversity. It is not important that you use a particular type of graph or summarize the data by the "right" statistic. Instead, you should choose methods that seem helpful in answering your questions.

This activity will be done in pairs. Each pair of you will be given a pack of State Cards, where one sample State Card is shown below. You are supposed to pose questions about several variables and then work with the data in various ways to answer these questions. (If these State Cards are not available, this activity can be done using packs of baseball or other sports cards that readily available in stores.)

MATERIALS NEEDED: One pack of special State Cards.



On a particular card, there are recorded:

- STATE: The name of the state.
- REGION: The location of the state in the United States.
- DRIVERS: The number of licensed drivers per 1000 residents in August 2002.
- FARMS: The number of farms (in thousands) in 2001.
- DENSITY: The population per square mile of land area (2000).
- POP CHANGE: The percentage change in population between 1990 and 2000 (HIGH OR LOW).
- ELEVATION: The highest point in the state (measured as feet above sea level).
- GOVERNOR: The political party of the state governor in 2002.

Your assignment is:

(1) Choose one quantitative variable that you are interested in and formulate a few questions about the variable.

My variable is _____

Questions about my variable:

(a)

(b)

(c)

(2) Arrange the cards from low to high with respect to the variable you chose in (1).

(3) Find the state with the lowest value, the state with the highest value, and the state with the "middle" value with respect to your variable.

	State	Value
Lowest value		
Highest value		
"Middle" value		

(4) Think of a second quantitative variable that you believe is related or associated with your first variable.

My second variable is _____

(5) Break the states into two groups of approximately equal size – the states that are low with respect to your first variable and the states that are high with respect to the first variable.

(6) For each group, find the "average" value of the second quantitative variable. (A simple average is the mean found by summing the values of the variable and dividing by the number of states.) Summarize your work below.

For states that were low in_____, the average of _____ was _____

For states that were high in _____, the average of _____ was _____

(7) Based on your work above, do you believe there is a relationship between your two variables? Explain.

EXTENSIONS: An almanac is a good source of data for different states. Make a list of 20 states and use the almanac to find an interesting variable for each state. Examples of interesting variables might be (1) birth rate, (2) percentage of population not covered by health insurance, (3) average temperature in July, (4) land area, (5) the mean SAT score, and (6) the percentage of people that voted for the Republican candidate in the most recent Presidential election. Answer the seven questions above using your collected data.

Reading Articles in the Media

Everyday we can read articles describing the results of statistical studies in the newspaper or the Internet. To illustrate, I looked at articles published in *USA Today* for a particular week in April, 2005. Here are some headlines of some relevant articles:

- "Alcohol's role in health not clear." An earlier study had suggested that there was evidence that a few alcoholic drinks a day was good for the heart. But this article discusses a more recent study that is inconclusive about the impact of drinking in reducing the risk of heart disease.
- "Perchance, to dream – of a good night's sleep." This article investigates the sleeping patterns of business travelers. By the use of several surveys, the article concludes that these "road warriors" are getting insufficient sleep.
- "Study: Fewer than expected dying from obesity." There is a general concern about the impact of obesity on death rates among Americans. But it is difficult to precisely estimate the number of deaths that are due to this health risk and this article describes how the conclusions from different studies vary.
- "2030 Forecast: Mostly gray." The Census Bureau recently projects, on the basis of current data, how the elderly population will grow faster than the total population.
- "Fewer high schoolers use Ecstasy, study finds." The drug Ecstasy was recently a popular drug among certain teens and young adults. But this article describes statistical evidence that this drug is losing popularity among this particular group of Americans.

Whenever we read an article such as these from *USA Today*, we should ask ourselves what questions were asked, how the data were collected, and if the conclusions drawn from the data appear valid based on the information that is provided. Unfortunately, many of the articles are brief, and it can be difficult to determine if the conclusions make sense due to the incomplete description of the study. But even if the article seems incomplete, you can think about what needs to be described to make the article more complete.

To illustrate this critical view, consider the following article about the possible benefits in eating grapefruit.

Grapefruit Lowers Weight, Fights Cancer
Studies find benefits to eating the citrus

By Kathleen Doheny

WEDNESDAY, Aug. 25 (HealthDayNews) -- A grapefruit or two a day, along with a healthy diet, could help shrink widening waistlines.

This finding comes from one of several studies on the benefits of citrus fruits presented Wednesday at the annual meeting of the American Chemical Society in Philadelphia.

The so-called grapefruit diet -- which advocates mostly eating grapefruit with some protein -- has been popular on and off for weight loss for years, said Dr. Ken Fujioka, director of nutrition and metabolism research at the Scripps Clinic in San Diego and lead author of a study evaluating grapefruit for weight loss. Most nutrition experts have deemed the grapefruit-and-protein regimen unhealthy, and Fujioka is not advocating any return to such a strict diet. However, his findings do suggest that a grapefruit or two each day, added to a balanced diet, might help the weight-conscious stay svelte.

In the study, Fujioka and his colleagues assigned 100 men and women who were obese to one of four groups. One group received grapefruit extract, another drank grapefruit juice with each meal, another ate half a grapefruit with each meal, while the fourth group received a placebo. "They weren't trying to diet," he said. "To make everyone even [on activity], all were asked to walk 30 minutes three times a week."

At the end of 12 weeks the placebo group lost on average just under half a pound, the extract group 2.4 pounds, the grapefruit juice group 3.3 pounds, and the fresh grapefruit group 3.5 pounds.

"In this study they had one and a half grapefruits a day," he noted. "That's not easy to do." And participants ate the fruit more like an orange: "They cut it in half, then into four sections, then separated the fruit from the skin." Eating grapefruit this way is thought to yield more beneficial compounds, he explained.

Exactly how grapefruit might spur weight loss isn't known, Fujioka said, but "it appears to help insulin resistance," which develops as people become obese.

The weight loss associated with eating grapefruit isn't surprising to another expert familiar with the study. "Eat fruit before any meal and you will lose weight," said Julie Upton, an American Dietetic Association spokeswoman. "The fiber fills you up, and fruit has fewer calories than other foods."

One half of a grapefruit has 60 calories, no fat, and six grams of fiber.

In analyzing this study, we ask the following questions:

1. What were the main questions addressed by the statistical study?

Here the investigators were trying to learn about the benefit of eating grapefruit towards the goal of losing weight. There was some support for a grapefruit-only diet in the past, and the scientists wished to learn if a moderate consumption of grapefruit would help to lose weight.

2. What data were collected to answer the questions?

The scientists measured the number of pounds lost in a 12-week period for each person in the experiment.

3. How did they collect the data?

Initially 100 obese people were selected to participate in the study. These people were placed into four groups where each group had a different type of grapefruit diet.

4. What variables were measured in the data? Label each variable collected as quantitative or categorical.

For each person there are two relevant variables: the group or diet plan and the number of pounds lost in 12 weeks. Group would be a categorical variable and pounds lost would be a quantitative variable.

5. What were the conclusions drawn from this statistical study? Do you believe that the conclusions are valid based on the information provided?

The groups that had grapefruit in their diet lost more weight, on average, than the group that didn't have any grapefruit in their diet. Based on the limited information in the article, it is difficult to say that eating grapefruit will help any obese person to lose weight. The last comment by the American Dietetic Association spokeswoman indicates that eating any fruit before a meal may help a person lose weight.

Graphs in the Media

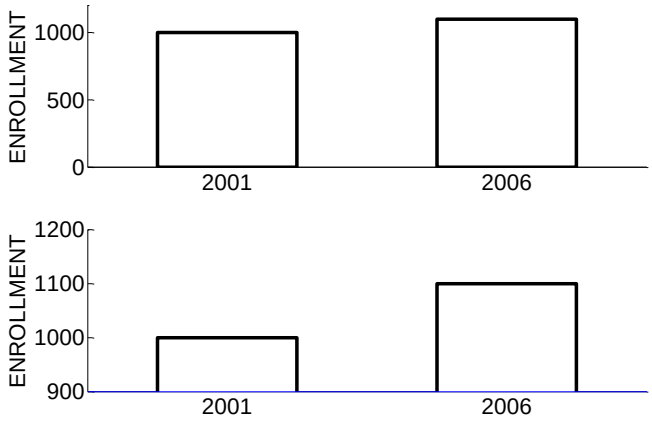
In newspapers, we will often see graphical displays that are used to convey numerical information. In some cases, the primary purpose of a graphical display is not to communicate information, but rather to entertain the reader. When we read a graph, we should think about the information that is being displayed and decide if the picture is an accurate representation of the information. Here are some specific questions to ask.

1. What is the main message of the graph?

2. What is the source of the data from which the numerical information has been computed?

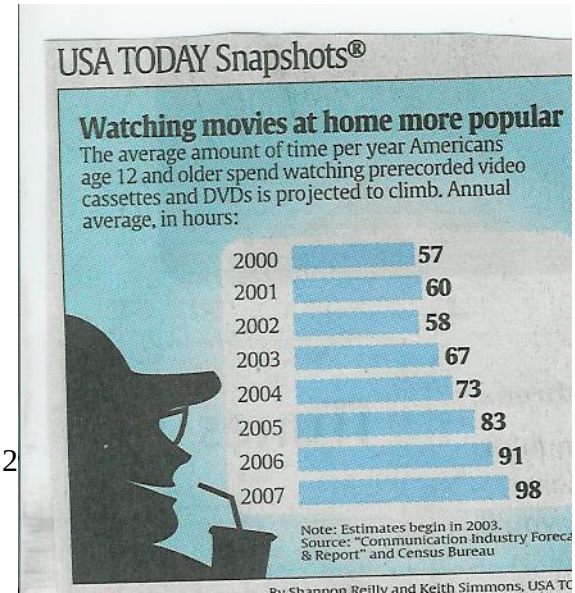
3. In a typical graph, there will be a primary display that shows the numerical information and other material (sometimes called chart junk) that is included to make the graph more attractive. Does the extra material distract the reader from seeing the primary graph?

4. Does the graph accurately represent the data? It is important that a graph follows the *area principle* where the area of the bar or figure should correspond to the number that one wishes to represent. For a simple illustration of this principle, suppose a university president wishes to construct a graph to show how the enrollment at the university has climbed in the last five years. The enrollment has increased from 1000 students in the year 2001 to 1100 students in the year 2006, a 10% increase. The figure below shows two sets of bar charts that could be used to display the enrollment numbers. The top graph obeys the area principle – the area of the bar for the year 2005 is 10% larger than the area of the bar for the year 2001 that does correspond to the actual increase in enrollment. In contrast, the bottom graph does not obey the area principle. Since the area of the 2006 bar is approximately twice the area of the 2001 bar, one gets the misleading impression that enrollment has doubled in the five-year period. The problem with this bottom graph is that the baseline enrollment in this graph has been set to 900. To give an accurate representation for bar graphs such as these, one always should use a baseline of zero instead of some arbitrary positive number.



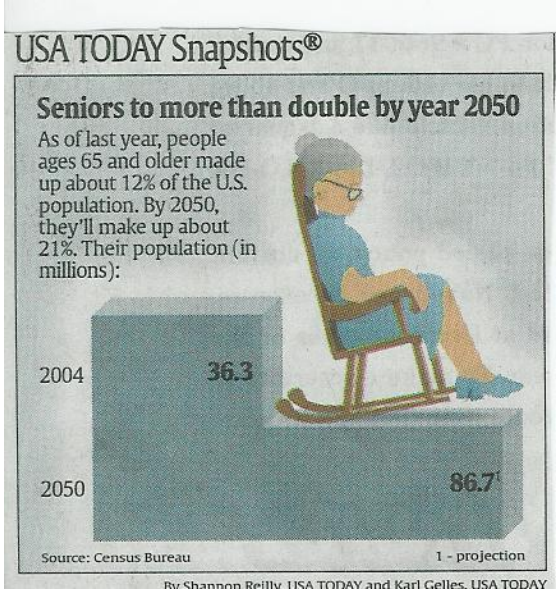
Let's illustrate this critical perspective through several graphs that recently appeared in USA Today.

"Watching movies at home more popular"



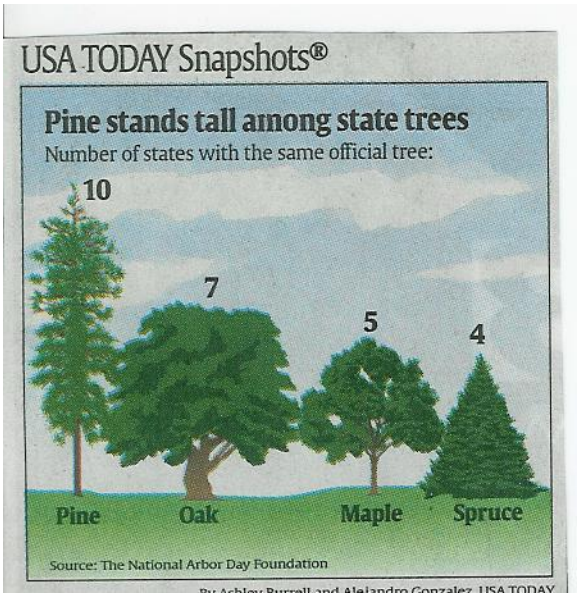
This graph shows how Americans' watching of video cassettes and DVDs has increased over time. For each year from 2000 to 2007, this graph shows the average number of hours watching movies per year by Americans age 12 or older. Blue bars are used to display these averages. To be an accurate representation of these data, the lengths of the bars should be proportional to the data values. For example, since 91 hours is roughly 50% longer than 60 hours, the length of the bar for 91 hours should be approximately 50% longer than the length of the bar for 60 hours. This seems to be generally true, so this display seems to be an accurate picture of the information.

"Seniors to more than double by year 2050"



This graph is used to show visually how the percentage of seniors (those ages 65 or older) will dramatically increase in the coming years. The population sizes (in millions) of seniors in the years 2004 and 2050 are displayed using bars. The length of the top horizontal bar corresponds to 36.3 million, the population of seniors in 2004 and the length of the bottom horizontal bar corresponds to 86.7 million, the population of seniors that is projected in 2050. The lengths of the bars look reasonable – the length of the bottom bar is over twice as long as the length of the top bar that reflects the data. But this display is hard to read. One focuses on the graphic of the woman in the rocking year that is irrelevant to the message of the graphic. Also the two horizontal bars have been merged into a single grey step and one might be tempted to look at the vertical heights rather than the horizontal heights.

"Pine stands tall among state trees"



Every state in the United States has an official "state tree" and the point of this graphic is to show what trees are most popular among the state

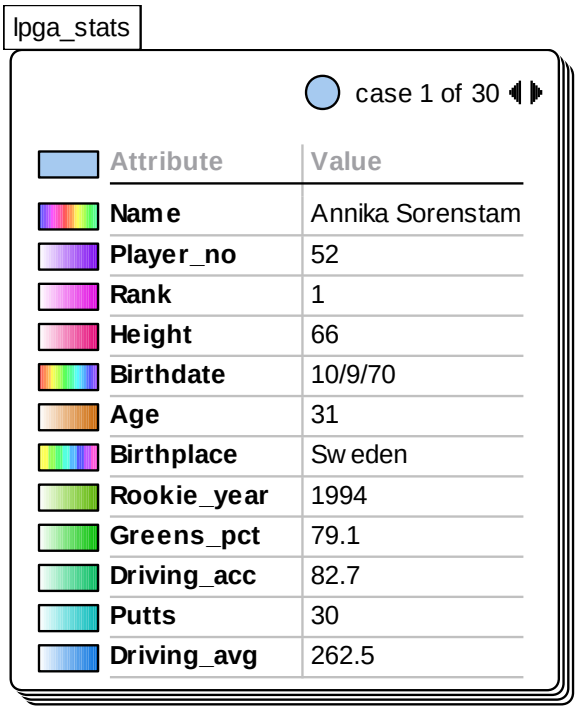
trees. We see that the pine tree is the most popular state tree (for 10 states), followed by the oak (7 states), maple (5 states), and spruce (4 states). Is this a good graphic of these data? The heights of the four tree diagrams do appear to correspond to the numbers – the height of the pine tree (10 states) is twice the height of the maple tree (5 states). But note that there are substantial differences in the widths of the diagrams. The area of the oak display is actually larger than the area of the pine display, which gives the impression the number of oak-tree states exceeds the number of pine-tree states. This graph does not obey the area principle.

TECHNOLOGY ACTIVITY: INTRODUCTION TO *TINKERPLOTS*

DESCRIPTION: This activity provides an introduction to the graphing package *Tinkerplots.* The author has collected data from the top 30 players in the Ladies Professional Golf Association in a recent year. For each golfer, the dataset contains

- NAME: the golfer's name
- PLAYER_NO: the number of the player as recorded on the LPGA database
- RANK: her rank in terms of money won (a rank of 1 corresponds to the golfer who has won the most)
- HEIGHT: her height in inches
- BIRTHDATE: her birth date
- AGE: her age in years when this data were collected
- COUNTRY: country of her birthplace
- ROOKIE: the year she was a rookie on the LPGA tour
- GREEN_PCT: green accuracy (the percentage of greens hit in regulation)
- DRIVING_ACC: driving accuracy (the percentage of drives that landed in the fairway)
- PUTTS: the average number of putts per round
- DRIVING_AVG: the average length of a drive (in yards)

Start the program *Tinkerplots*. Import the data into *Tinkerplots* from the file lpga_stats.txt. You should see the Data Card for the first woman golfer, Annika Sorenstam:



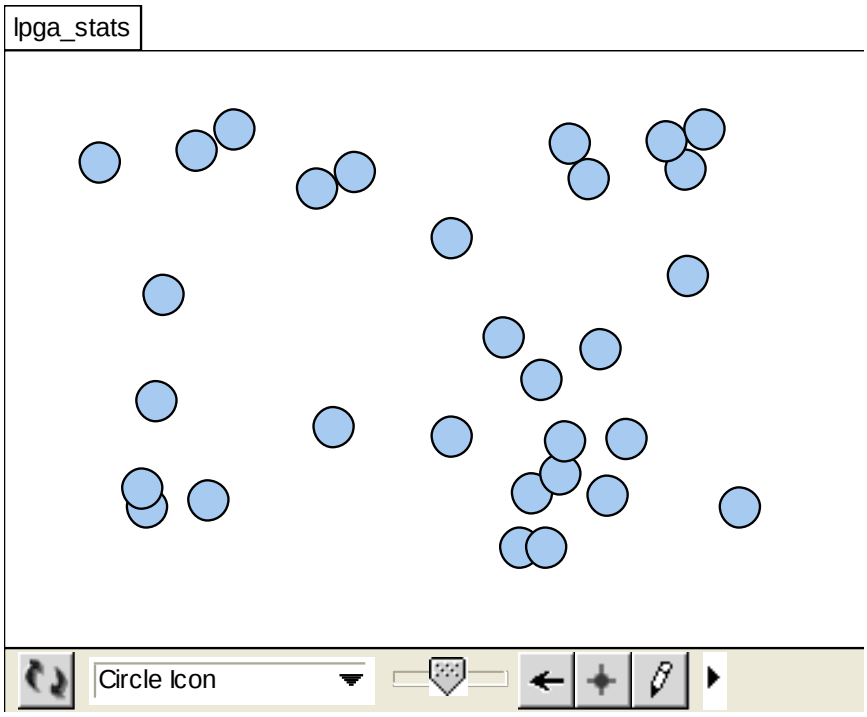
You can look at other Data Cards by clicking on the arrows at the top of the card.

An alternative representation of the data is by a Case Table, where the players appear as rows, and the variables appear as columns in the table. You can get this representation in *Tinkerplots* by dragging down the Table icon.

lpga_stats

	Name	Player_no	Rank	Height
1	Annika Sorenstam	52	1	66
2	Se Ri Pak	49	2	66
3	Juli Inkster	44	3	67
4	Mi Hyun Kim	31363	4	61
5	Karrie Webb	53	5	66
6	Laura Diaz	31460	6	68
7	Rachel Teske	31326	7	65
8	Grace Park	31452	8	67
9	Hee-Won Han	31318	9	67
10	Michele Redman	31475	10	68

If you drag down a Graph object, you see the 30 golfers represented as a random collection of case icons or dots. This representation is analogous to taking the stack of "golf cards" and scattering them on the floor.



Tinkerplots allows one to organize these case icons in different ways.

- One can *separate* the icons in two or more groups by use of any of the variables.
- One can *order* the icons from highest to lowest on some variable.
- When the icons are separated into groups, one can *stack* the icons in each group.

Use one or more *Tinkerplots* graphs to answer the following questions. For each question, you should show your graph by either printing and pasting it into your book, or electronically copying the graph in a word processing document.

1. Look at the Birthplace variable. What countries are represented by these women golfers? Are these golfers predominately from particular countries? Are Americans well-represented in this group?

2. Look at the Age variable. Who are the youngest and oldest golfers in this group? What is a typical age among these golfers? Explain how you found this typical golfer.

3. Look at the Rookie Year variable. (This is the first year that the golfer played as a professional.) Which golfer has played the longest on the tour? Which golfer has played the shortest? On average, how many years have these golfers played on the LPGA tour. (Explain how you computed the average.)

4. Look at one of the golf statistic variables (GREEN_PCT, DRIVING_ACC: PUTTS: or DRIVING_AVG). It is desirable for a golfer to have …

- a *high* percentage of greens hit in regulation (GREEN_PCT)
- a *high* percentage of drives hit in the fairway (DRIVING_ACC)
- a *low* average number of putts (PUTTS)
- a *high* average driving length (DRIVING_ACG)

Can you pick out a couple of the more successful golfers and the less successful golfers with respect to this variable? Are the best golfers (the ones ranked from 1 to 5) good with respect to this particular golf statistic?

WRAP-UP

In this opening topic, we have been interested to the science of *statistics*. In a statistical investigation, we begin with some questions of interest, and then we collect relevant data that we think will be helpful in answering the questions. *Data* is the general term for information that we collect about individuals and *statistics* is a term used to describe numerical and categorical data. Generally, several variables are collected from each individual, and these variables differ by how they are *measured*. It is important to distinguish *quantitative* variables from *categorical* variables – the distinction is important in how we graph and summarize data. Whenever a statistical study is reported in the media, it is important to judge if the collected data are useful in addressing the main questions of the study. We were introduced to the use of graphical displays in the media. When we view a graph in a newspaper, we should think if the graphical display is a clear and accurate representation of the quantitative or categorical information.

Classroom Capsule: "Children's Well-Being"

Overview: The students will be introduced to an interesting UNICEF study on the well-being of children from 21 industrialized countries.

Objectives: The students will start to think about the meaning of "well-being of a child". After they list several attributes of a child's well-being, they read an article on a UNICEF student from the Baltimore Sun. From the article, they identify the relevant data that are collected and the variables that are studied.

Description: Ask the students the following questions. (They can answer these questions orally or by writing.)

1. Consider an average child who is brought up in the United States and an average child brought up in France. Who do you think would be better off? Why?
2. What do you think are the positive aspects of a child brought up in the U.S.?
3. What do you think are the negative aspects of a child brought up in the U.S.?
4. How would you measure a child's material well-being?

5. How would you measure a child's health and safety?
6. How would you measure a child's educational well-being?
7. How would you quantify a child's family and peer relationships?
8. How would you measure the health and risk behavior of a child?
9. How would you measure a child's subjective sense of well-being?

Next, read the following article from the *Baltimore Sun* that compares the well-being of children from the United States and other industrialized countries.

From the Baltimore Sun (May 4, 2007)

U.S. scores low in study on children's well-being By Julie Deardorff

America is one of the richest countries in the world. It's also one of the worst industrialized places for kids to grow up and has a greater percentage of depressed people than impoverished, war-torn nations do, according to two major studies.

The first unflattering finding comes from a recent UNICEF child-welfare study that measured everything from the number of books in the home to infant-mortality rates, drinking and drug use and the percentage of children who eat meals with their families.

Of 21 wealthy nations surveyed, the United States ranked second to last. Only Britain was worse. Child well-being was highest in the Netherlands, Sweden, Denmark and Finland, places that invest heavily in their children.

The problem isn't just that, compared with the European countries, the United States lacks day-care services and has poorer health and preventive-care coverage, which has left 9 million children without health insurance.

America finished dead last in terms of infant-mortality rates, vaccinations, the percentage of newborns with low birth weights and deaths from accidental injuries. We finished second to last when the researchers assessed a child's diet, physical activity and weight, exposure to violence and bullying and the number of 15-year-olds who smoke, drink and have sex.

And, in what could explain why we're among the most depressed people on Earth, according to a study of 14 nations conducted jointly by the World Health Organization

and Harvard Medical School, we finished second to last when researchers examined relationships with family members and friends and family structure.

American children often don't eat the main meal of the day with their parents. Children say they don't spend time "just talking" to their parents. And they generally don't find their peers "kind and helpful," according to the study.

It shouldn't really be a surprise, then, that 9.6 percent of Americans suffer from depression or bipolar disorder, according to the WHO/Harvard study; that binge eating or drinking is up; or that children are heavily medicated for depression and attention-deficit disorder.

In material goods, American children have it all. But to make them feel loved, cherished and supported, they need family, community, a higher sense of purpose and meaningful cultural traditions - all things money can't buy.

Questions from reading the article:

1. What were the main questions addressed by the statistical studies described in the article?
2. What data were collected to answer the questions?
3. How did they collect the data?
4. What variables were measured in the data? Label each variable collected as quantitative or categorical.
5. What were the conclusions drawn from this statistical study? Do you believe the conclusions were valid based on the information provided?

Share and Summarize: Here are some important items that should be discussed for this example.

(a) There are many ways to describe a child's well-being. A child will have a good upbringing if he or she has sufficient material possessions, is healthy and safe, has sufficient education, good family and peer relationships, does not engage in bad behavior or risky situations, and has a positive self-image.

(b) Although we can agree that all six aspects are important, it can be difficult to measure the corresponding attributes. For example, how do you measure the quality of a child's family relationships?

(c) If we can find a suitable way of measuring a particular attribute, say self-worth, the next problem is how to collect data from a representative subset of the population of interest. For example, if we wish to compare the self-worth of American and French children, how do we take samples of the two populations?

(d) A newspaper article will typically focus on the conclusions of a statistical study and say little about the details of a study such as the manner in which the data were collected.

Application or Extension: Each of the students could be asked to find another newspaper article that compares the well-being of children of the United States and other countries. The student could discuss the article using the same questions given above. What information is given in your article that was not contained in the *Baltimore Sun* article?

EXERCISES

1. **Reading Articles**

Two articles that recently appeared in the media are copied below. Each article discusses the results of a statistical study. For each article, write a paragraph describing the different parts of the study. In particular, answer the following questions:

1. What were the main questions addressed by the statistical study?
2. What data were collected to answer the questions?
3. How did they collect the data?
4. What variables were measured in the data? Label each variable collected as quantitative or categorical.
5. What were the conclusions drawn from this statistical study? Do you believe that the conclusions are valid based on the information provided?

ARTICLE 1

Americans' Costly Health Care Not Better

By Jennifer Warner WebMD Medical News

May 5, 2004 -- Americans may pay more for health care than other countries, but they may not necessarily be getting the best medical care.

A new study shows the U.S. health care system ranks near the top on some but not all major health indicators and could take a lesson from the superior performance of other countries on several key areas, such as asthma and transplant surgery survival.

Researchers say the findings show there may be little evidence to back up the mantra often cited by policymakers, "Americans have the best medical care in the world."

"It is well known that the United States spends much more on health care per capita than other countries, and it is commonly assumed that we have the best health care system in the world, " says researcher Peter S. Hussey of the John Hopkins Bloomberg School of Public Health in Baltimore, Md., in a news release. "However, the results of our study show that the United States performs better than other countries in only a few areas, while performing worse in others."

"This raises the question of what Americans receive for all of the money devoted to health care," says Hussey.

Comparing Health Care Systems

Researchers say the study, which appears in the May/June issue of *Health Affairs*, is the first to use a universal set of standards to compare the quality of health care in five countries: Australia, Canada, England, New Zealand, and the U.S.

An international group of researchers collected and examined data on 21 health indicators that reflect the quality of medical care, including:

- 5-year cancer survival rates
- Cancer screening rates
- Avoidable events, such as suicide, asthma deaths, and smoking prevalence
- Vaccination rates
- Transplant survival rates

The study showed that no one country consistently scored among the best or worst overall.

For example, the U.S. had the highest breast cancer survival rate, cervical cancer screening rate, and lowest smoking (tied with Canada). But the U.S. performed among the worst in other areas, such as asthma-related deaths and survival after kidney and liver transplants. In fact the U.S. was the only country in which asthma-related deaths were increasing rather than decreasing.

The U.K. scored best in five of the eight avoidable event indicators, including pertussis and hepatitis B, but scored the lowest in five of the nine survival rate indictors and had the lowest cancer survival rate of the five countries studied.

"Each country in our study has areas of care where it can learn from the other countries and areas where it could teach others," says Hussey. "That tells us that there are opportunities for improvement in the quality of health care in all five countries."

SOURCES: Hussey, P.*Health Affairs*, May/June 2004; vol 23: pp 89-99. News release, John Hopkins Bloomberg School of Public Health.

ARTICLE 2:

Mass. to track all traffic stops after study finds profiling

The Northeastern University study, released Tuesday, was commissioned four years ago by the Legislature and included 366 departments — from cities and towns and the state police, to university, state transit and Amtrak police agencies. Just 92 got a passing grade.

Public Safety Secretary Edward A. Flynn warned against condemning departments that failed until more information can be gathered. The study caused the state to order 249 departments to collect a year's worth of data on all traffic stops.

"We are not today finding any agency guilty of having engaged in racial profiling," he said. "Data collection is not punishment."

Flynn said requiring agencies to collect more data will provide a clearer picture of racial profiling in Massachusetts.

"Every community deserves an explanation from its police department on how it uses its authority," Flynn said.

Northeastern used four statistical tests in analyzing 1.6 million traffic citations issued between April 1, 2001, and June 30, 2003: Ticketing resident minorities disproportionately more than whites; ticketing all minorities disproportionately more than whites; searching minorities more often than whites; and issuing warnings to whites more often than minorities.

According to the study, 15 police departments failed all four tests, 42 failed three tests, 87 failed two tests and 105 failed one. Among those that failed all four were Boston, Springfield and Worcester.

The Executive Office of Public Safety will use $1 million in grant money over the next six months to set up a uniform system for all police departments to report traffic stops, including those that do not result in any citations or written warnings. That information will be gathered over another year, then analyzed again.

Jack Collins, general counsel for the Massachusetts Chiefs of Police Association, said the added paperwork is a "witch hunt" and unnecessary.

Bishop Filipe Teixeira, a Catholic bishop from Brockton, said he's heard complaints in his community about minorities being targeted.

"We do have bad apples in the police departments," Teixeira said. "We have enough data. Let's get into action."

---

2. In a local or national newspaper, find three articles that describe conclusions from a statistical study. Summarize the information in each article by answering the questions given in Exercise 1.

3. **Graphical Displays in the Media**

Four graphical displays from USA Today are shown below. For each display, answer the following questions:

a. What is the source of the information? (That is, where did the information come from?)

b. How is the information portrayed in the graphical display?

c. What is the basic message communicated in the graphic?

d. Is the graphical display an accurate representation of the data? In particular, does the graphical display follow the area principle, where the areas of the bars or shape are proportional to the data values?