STATISTICS AND GRAPHS

This unit is about data and how it is displayed in graphs. The measures of central tendency will be reviewed. Data will be examined through frequency tables, scatter plots, and various types of graphs. Data will be compared and interpreted through graphs. A misuse of graphs will be examined through misleading reports of data in graphs.

Central Tendency of Data

Frequency Tables and Histograms

Displaying Data in Graphs

Pictographs

Constructing a Circle Graph

Scatter Plots

Same Data in Different Graphs

Misleading Statistics

Central Tendency of Data

There are several ways to describe the central tendency of a group of data. We will discuss three: mean, median, and mode. Let's use the example of a typical classroom in high school and describe the age of the occupants in the room.

Example: Listed below are the ages of the students in the classroom along with the teacher's age.

16, 16, 15, 15, 15, 16, 15, 14, 18, 47

Put the data in order from least to greatest to analyze it.

14, 15, 15, 15, 15, 16, 16, 16, 18, 47

Mean is the average of the data.

$$\text{Mean}: \quad \frac{14+15+15+15+15+16+16+16+18+47}{10} = 18.7$$

Mean = 18.7 years

Median is the middle number of the data when it is in order from least to greatest or vice versa.

*When there is an odd number of numbers in the data (listed in order), the middle number is the median. If there is an even number of numbers, the median is the average of the two middle numbers of the data.

14, 15, 15, 15, 15, 16, 16, 16, 18, 47

$$\text{Median}: \quad \frac{15+16}{2} = 15.5$$

Median = 15.5 years

*In this example, there is an even number of numbers in the data set, so the median is the average of the two middle numbers.

Mode is the number that occurs most often in the data.

14, 15, 15, 15, 15, 16, 16, 16, 18, 47

Mode = 15 years

*The number that occurs most often (four times) in this data set is 15.

Now let's see what happens when the teacher's age is removed.

14, 15, 15, 15, 15, 16, 16, 16, 18, ~~47~~

Mean: $\frac{14+15+15+15+15+16+16+16+18}{9} \approx 15.6$

*15.6 is rounded to the nearest tenth.

Median: 14, 15, 15, 15, 15, 16, 16, 16, 18 $= 15$

Mode: 14, 15, 15, 15, 15, 16, 16, 16, 18, 47 $= 15$

Mean = 15.6 years Median = 15 years Mode = 15 years

View the chart below and compare how the mean, median, and mode were affected by removing the "large data point" that was the teacher's age.

Measure of Central Tendency	With Teacher's Age	Without Teacher's Age
Mean	18.7	15.6
Median	15.5	15
Mode	15	15

How was the mean affected? The mean was affected greatly. After the removal of the teacher's age, the mean became more representative of the entire group of data. The mean dropped from 18.7 to 15.6.

How was the median affected? The median was slightly affected and became a little lower after the teacher's age was removed. The median dropped from 15.5 to 15.

How was the mode affected? There was no change in the modes.

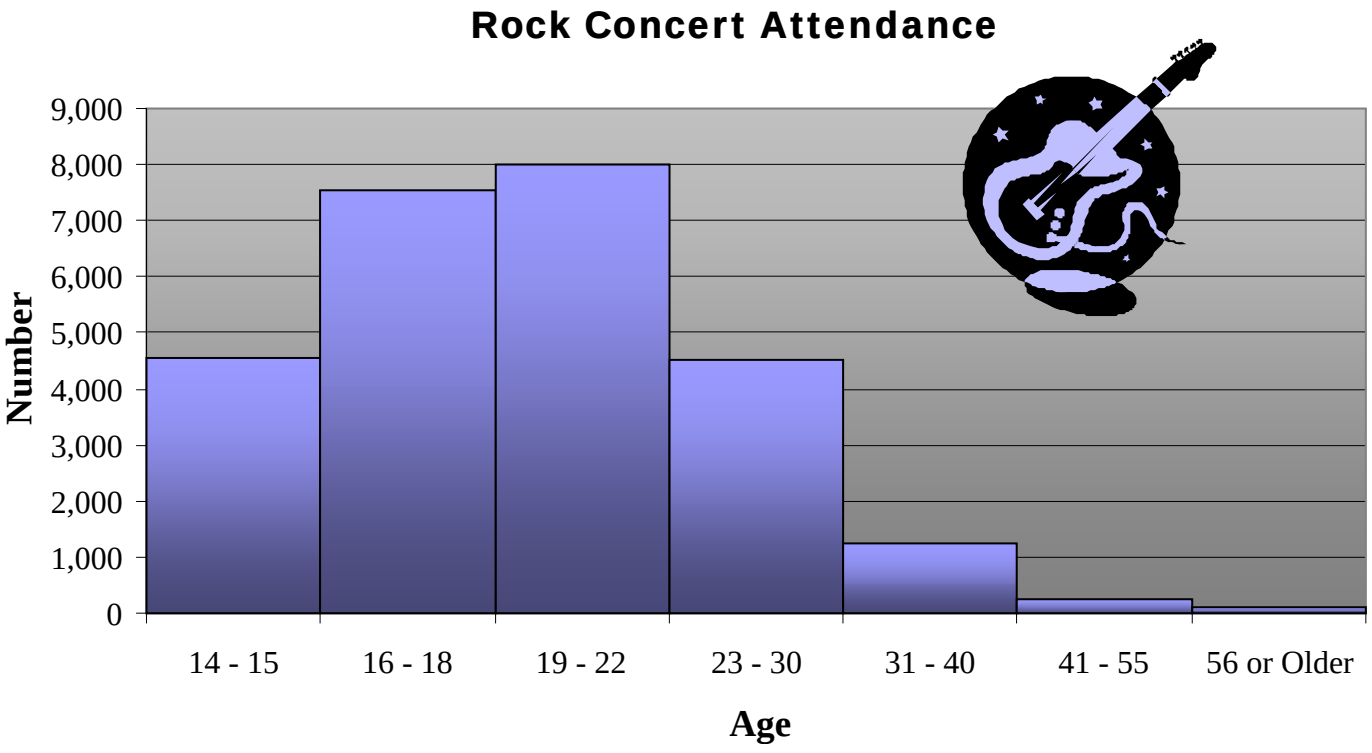
*Notice that the median and the mode were less affected by a single large (or small) data point in comparison to the mean. When reporting data, sometimes the median is given rather than the mean because it is more representative of the whole group of data.

Be careful when making decisions about reporting the results of calculated measures of central tendency. Choose the measure that demonstrates an **authentic center** of the data.

Frequency Tables and Histograms

Darin wanted to attend the rock concert being held at the local civic center. His mom agreed that he could go, but only if she went also. The data displayed below in a frequency table depicts the ages of the persons attending the rock concert. It is also displayed in a histogram. Histograms are used to display ranges of data.

Frequency Table	
Ages	**Frequency**
14 – 15 (Young Teens)	4538
16 – 18 (Teens)	7536
19 – 22 (Young Adults)	7999
23 – 30 (Adults)	4520
31 – 40 (Middle-Aged Adults)	1236
41 – 55 (Older Adults)	233
56 or Older (Seniors)	120



What age group had the highest attendance? The 19-22 year olds

What age group had the lowest attendance? 56 or Older
In a histogram, a display of data which possesses a **symmetric** distribution is one in which the two "halves" of the data appear as mirror-images to one another.

A **skewed** (non-symmetric) distribution of data is a distribution in which the data appears at more diverse values. Distributions of data are skewed if one of the tails of a histogram (the part that stretches out from the peak) is longer than the other.

A "**skewed right**" distribution is one in which the right tail is longer.

A "**skewed left**" distribution is one in which the left tail is longer.

How would the distribution of the data of the rock concert attendance be described?

The data is "**skewed right**". Most of the rock concert attendees were younger than 30 years old.

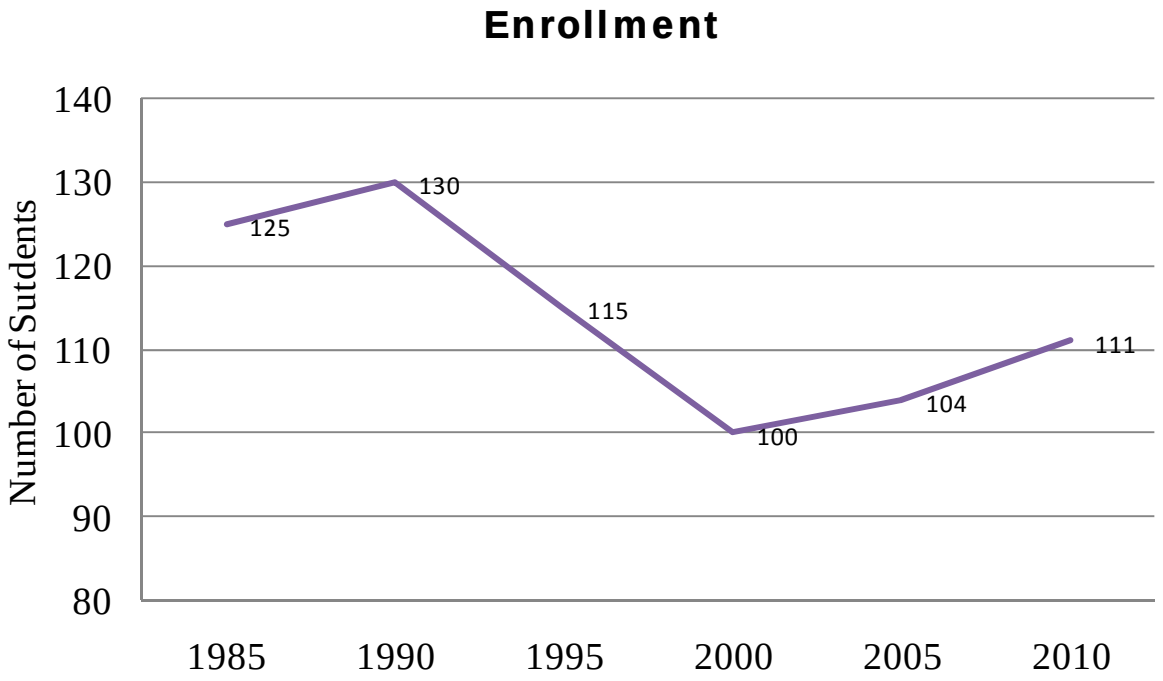
Display Data in Graphs

Let's take a look at some data and the resulting graphs.

Line Graph

Beth is examining the student enrollment in the school's ninth grade class. She started with the year 1985 and made a line graph to look for a trend in the enrollment.

Year	1985	1990	1995	2000	2005	2010
Number	125	130	115	100	104	111

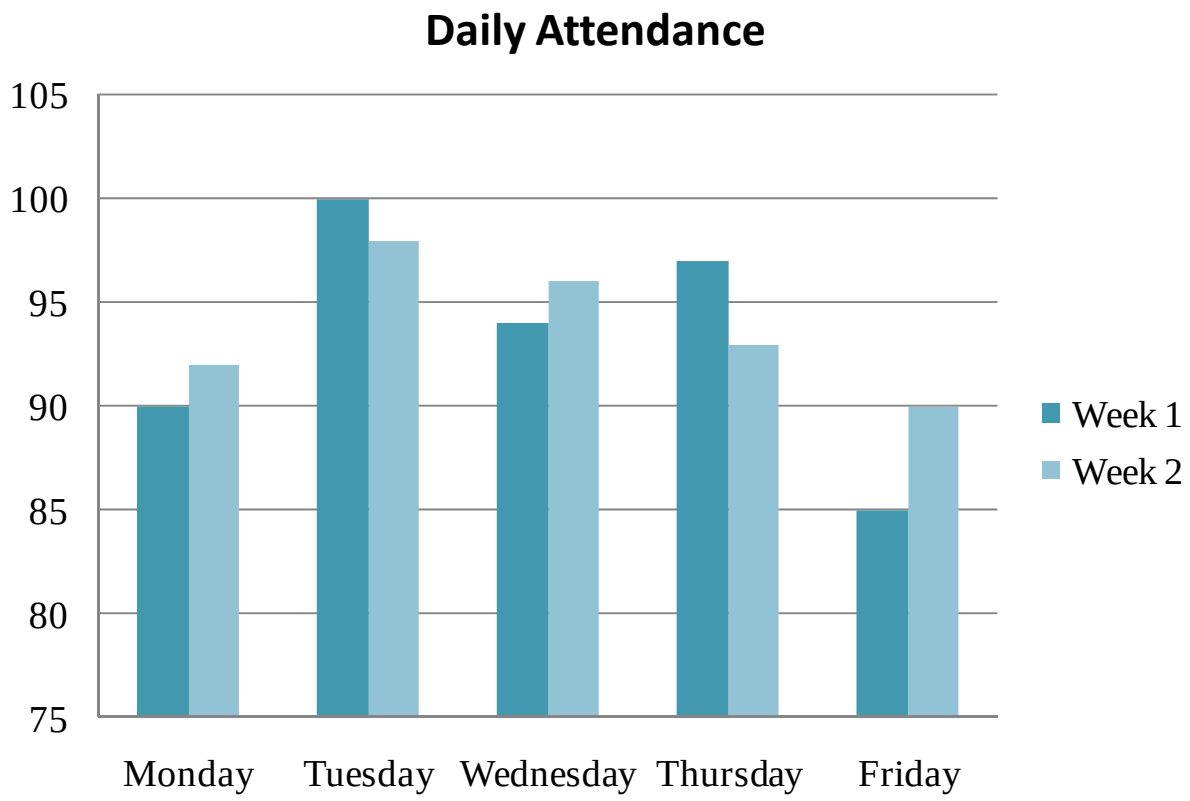


The line graph shows that there was a decline in enrollment during the 1990's, but more recently, the enrollment is increasing.

Bar Graph

Beth is examining the daily attendance of the ninth grade class. She selected a bar graph to compare the data for two weeks to show which day of the week most students were at school.

Daily Attendance of 9th Grade Class					
	Monday	Tuesday	Wednesday	Thursday	Friday
Week 1	90	100	94	97	85
Week 2	92	98	96	93	90

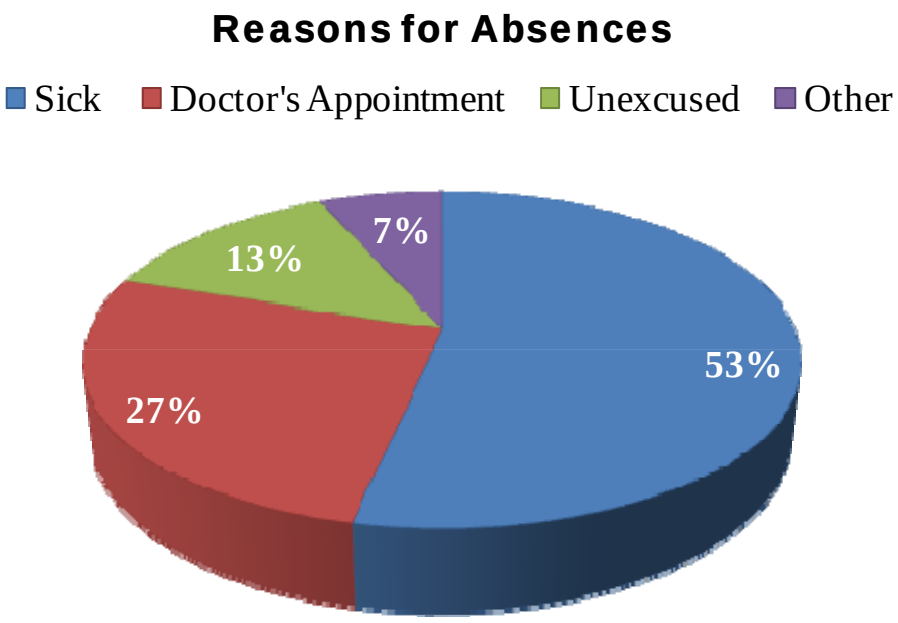


Beth found that the daily attendance is best on Tuesday.

Circle Graph

Beth is examining the reasons for student absences on Friday. She decided to display the information in a circle graph to show the absences based on percentages of the whole group that was absent.

Reasons for Absence	
Sick	8
Doctor's Appointment	4
Unexcused	2
Other	1



More than half of the students were reported off as sick on Fridays.

Pictographs

A **pictograph** is a graph that shows numerical information by using picture symbols to represent the data sets. Examine the scenario below to understand the construction and interpretation of pictographs.

California is the world's largest producer of strawberries. Mr. Bradford, a county extension agent, wanted to prepare a presentation on the number of harvested acres used in California to produce strawberries. He chose to present the data in a pictograph.

Look over the data collected and draw a pictograph to summarize the information in the table.

Number of Acreage Used for Strawberry Production in California	
Year	**Acreage**
1998	24,200
1999	28,500
2000	27,600
2001	26,400
2002	28,500
2003	29,600
2004	33,200

http://www.usda.gov/nass/

Method:

Step 1: Label the graph with a title and write the scale on the vertical axis.

Step 2: Choose a symbol to represent the data. Write a definition of the symbol at the bottom of the graph.

Step 3: Determine how many symbols will be used for each item by dividing.

Step 4: Draw the symbols.

*Since the data ranges from about 24,000 through 36,000, one strawberry symbol could represent 4,000 acres. We will proceed with constructing the pictograph based on the premise that one strawberry in the graph represents 4000 acres.

1998

Divide 24,200 by 4,000. The result is 6.1 rounded to the nearest tenth.

$$\frac{242\cancel{00}}{40\cancel{00}} \rightarrow \quad 40\overline{)242.00} \quad \begin{array}{r} 6.05 \\ \end{array} \approx 6.1$$

Draw 6.1 symbols next to 1998.

1999

Divide 28,500 by 4,000. The result is 7.1 rounded to the nearest tenth.

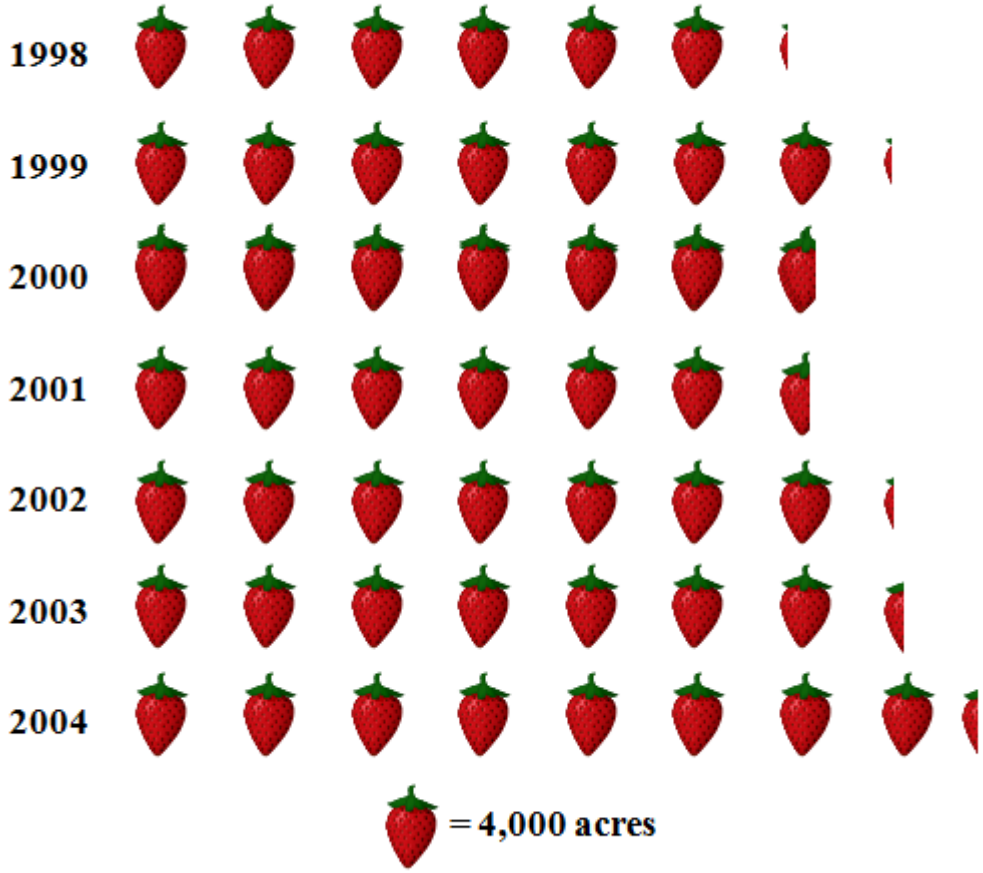
$$\frac{285\cancel{00}}{40\cancel{00}} \rightarrow \quad 40\overline{)2850.000} \quad \begin{array}{r} 7.125 \\ \end{array} \approx 7.1$$

Draw 7.1 symbols next to 1999.

Repeat the above process for the years 2000 through 2004.

2000	2001	2002	2003	2004
$\frac{276\cancel{00}}{40\cancel{00}} = 6.9$	$\frac{264\cancel{00}}{40\cancel{00}} = 6.6$	$\frac{285\cancel{00}}{40\cancel{00}} \approx 7.1$	$\frac{296\cancel{00}}{40\cancel{00}} = 7.4$	$\frac{332\cancel{00}}{40\cancel{00}} = 8.3$

California Acreage Used in Strawberry Production



Practice reading the pictograph and/or the chart of data to answer each of the following questions. (*The answers are provided below the questions.)*

1. Each symbol represents how many acres?

2. In what year was the most acreage used for strawberry production in California?

3. In what two years was the same amount of acreage used for strawberry production in California?

4. Refer to the pictograph to answer and explain the solution to the following question: About how many more acres were used in strawberry production in 2004 than in 1998?

5. Refer to the table to determine exactly how many more acres were used in strawberry production in 2004 than in 1998?

6. Use the pictograph to compare the strawberry production in years 2000 and 2001. In which year was more acreage used for strawberry production?

7. Refer to the table to determine exactly how many more acres were used in strawberry production in 2000 than in 2001?

8. In the instances above, when was it best to refer to the pictograph to answer the questions and why?

9. In the instances above, when was it best to refer to the table of data to answer the questions and why?

Answers

1. Each symbol represents 4,000 acres.

2. 2004

3. 1999 and 2002

4. About 8000 acres. In 1998 there are 6 whole berries and less than half of another berry. In 2004 there are 8 whole berries and less than half of another berry. Therefore 8 – 6 = 2 berries. 2 x 4000 = 8000 acres.

5. 9000 acres (33,200 – 24,300 = 9000)

6. 2000

7. 1200 acres (27,600 – 26,400 = 1200)

8. It is best to use a pictograph when making a quick comparison or an estimate.

9. It is best to use the table of data when determining an exact answer.

Constructing a Circle Graph

In a survey Madison determined her classmates' favorite colors based on the following choices: blue, green, yellow, red. She decided to display the results in a **circle graph**.

First she made a **tally** chart to record the responses.

She then made another chart to **organize** her calculations for drawing the graph.

Madison made fractions based on 20 (total responses) and changed them to percent. She then multiplied the percent by 360 since there are $360°$ in a circle to determine the central angle to represent each color.

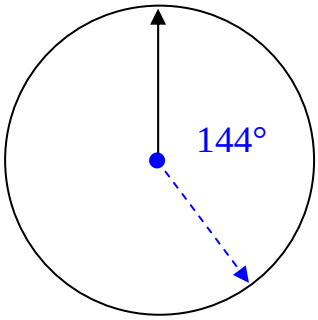
Tally of Favorite Color		
Blue	𝍸 |||	**8**
Green	𝍸	**5**
Yellow	|	**1**
Red	𝍸 |	**6**

Favorite Color		
Blue	$\frac{8}{20}$ = **40%**	$40\% \times 360 = 144°$
Green	$\frac{5}{20}$ = **25%**	$25\% \times 360 = 90°$
Yellow	$\frac{1}{20}$ = **5%**	$5\% \times 360 = 18°$
Red	$\frac{6}{20}$ = **30%**	$30\% \times 360 = 108°$
Check	$\frac{20}{20}$ = **100%**	**360°**

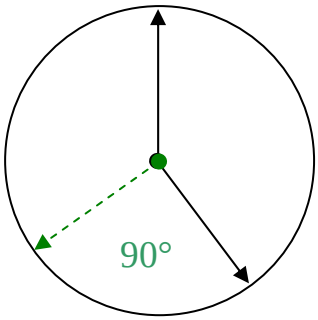
To check the calculations, Madison added the results. The fractions total **20/20**, the percents total **100%**, and the degrees total **360˚** which makes **1 whole circle**.

Now Madison is ready to draw the circle graph.

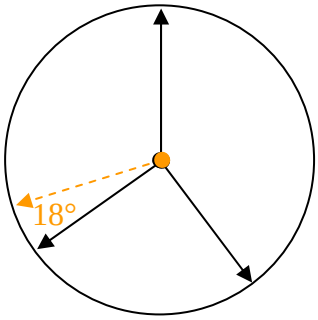
Step 1: Draw a circle with one radius. Place a protractor on the radius, using the center of the circle as the vertex, and draw the first angle (144° - blue).



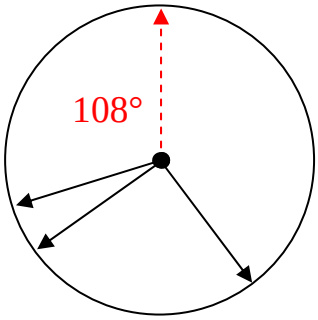
Step 2: Move the protractor to rest on the ray just drawn and proceed to make the second angle (90° -green).



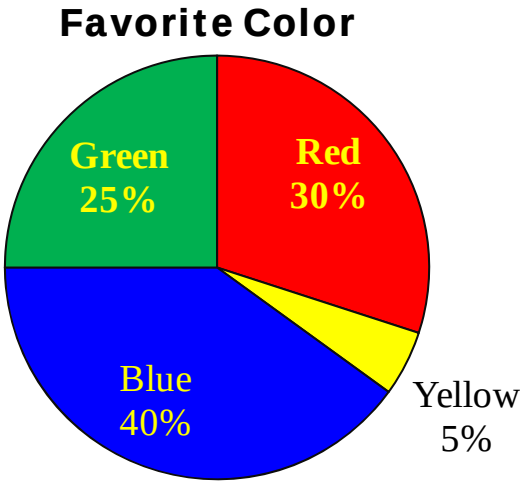
Step 3: Then move the protractor to rest on the ray she drew and proceed to make the third angle (18° - yellow).



Step 4: Madison will not have to draw the last angle, but will measure it. It should measure 108° (red).



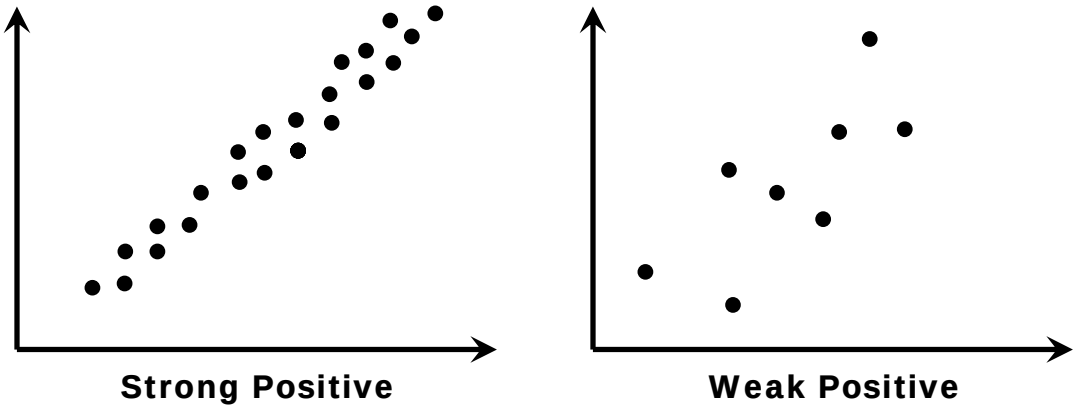
Now the circle has been divided proportionally to the percent for each category. Complete the graph by adding color, category labels, and a title.



Scatter Plots

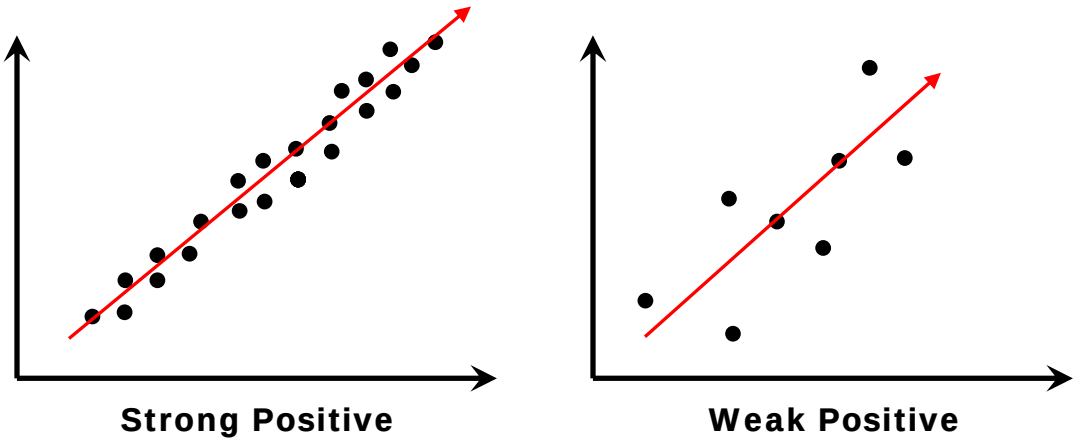
A **scatter plot** shows relationships between two sets of data in clusters of data with dots or symbols.

Positive Correlation of Data: If the points appear to suggest a line that slants upward and to the right, there is a positive relationship.



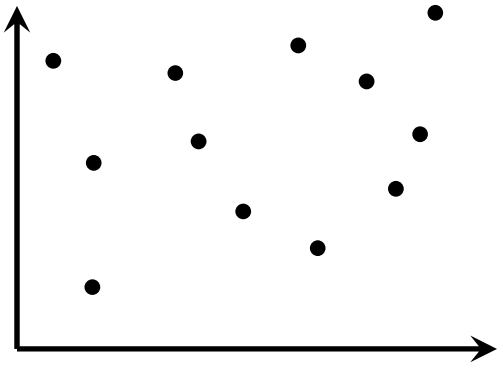
Positive correlation: Both data sets increase together.

Line of Best Fit: The straight line that is located as close as possible to most of the data points is the best-fitting line. Notice the positive slope of each straight line added to the graphs below.



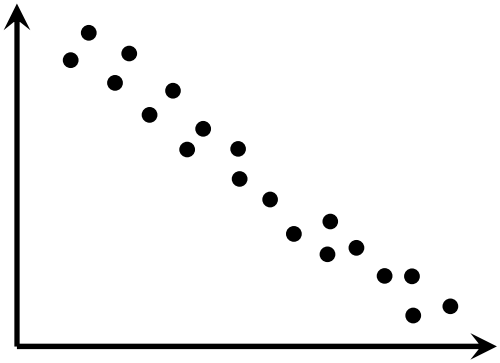
*NOTE: The best-fitting line does not need to pass through or even near all the points.

No Correlation of Data*:* If the points seem to be randomly located, then there is no relationship.

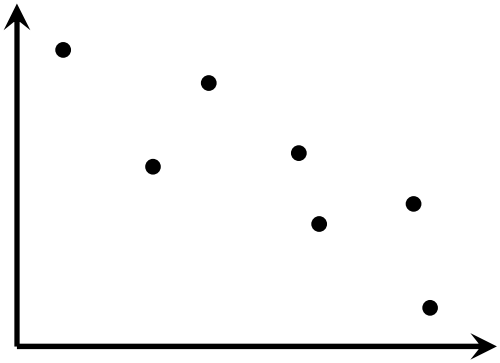


No correlation: The changes in one data set do not affect the other data set.

Negative Correlation of Data*:* If the points appear to suggest a line that slants downward and to the right, there is a negative relationship.



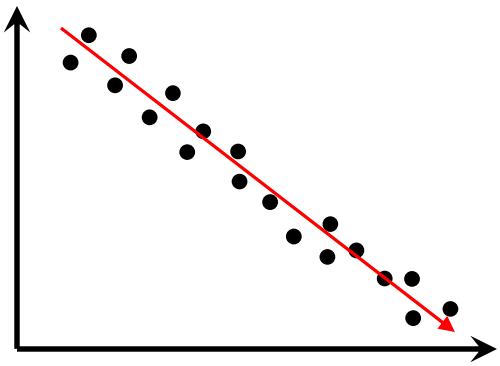
Strong Negative



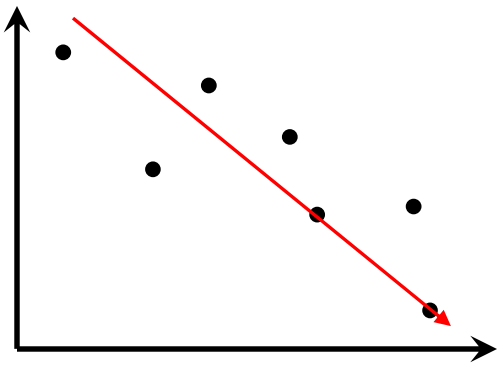
Weak Negative

Negative correlation: As one data set increases, the other decreases.

Line of Best Fit: The straight line that is located as close as possible to most of the data points is the best-fitting line. Notice the negative slope of each straight line.



Strong Negative



Weak Negative

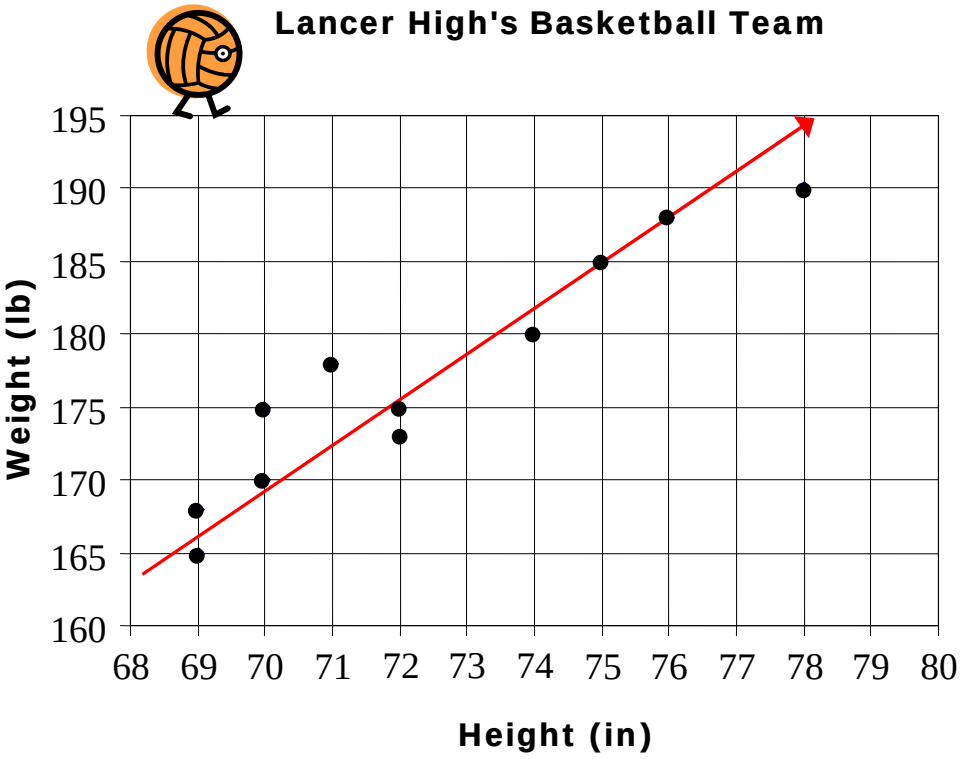
Example: Use the given data to make a scatter plot of the weight and height of each member of Lancer High's basketball team.

Height (in)	Weight (lb)
70	170
69	165
72	175
74	180
75	185
70	175
69	168
72	173
71	178
78	190
76	188
69	165

Step 1: Make a scatter plot of the data pairs. The points on the scatter plot are (70, 170), (69, 165), (72, 175), (74, 185), (75, 185), (70, 175), (69, 168), (72, 173), (71, 178), (78, 190), (76, 188), and (69, 165).

Step 2: Draw the line that appears to best fit the data points. There should be about the same number of points above the line as below it. The line does not have to pass through any of the data points.

What kind of correlation exists between the data sets? The line of best fit slants upward and to the right.



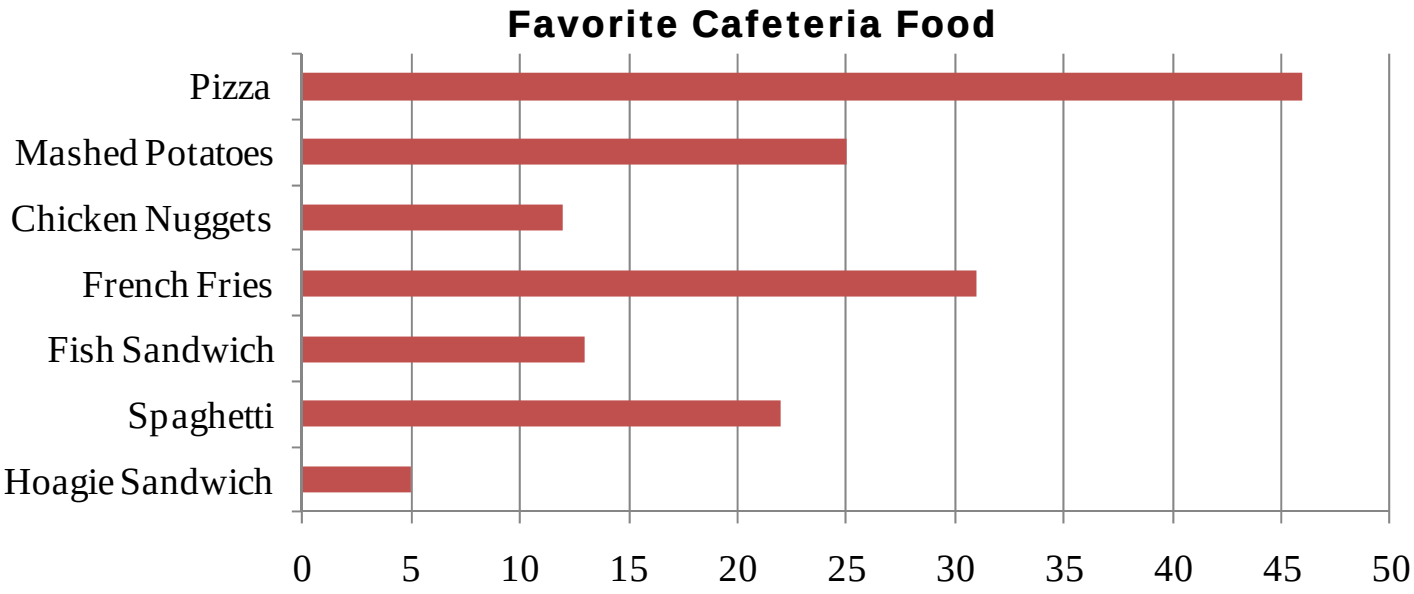
There is a positive correlation between the height of the team member and his weight.

Same Data in Different Graphs

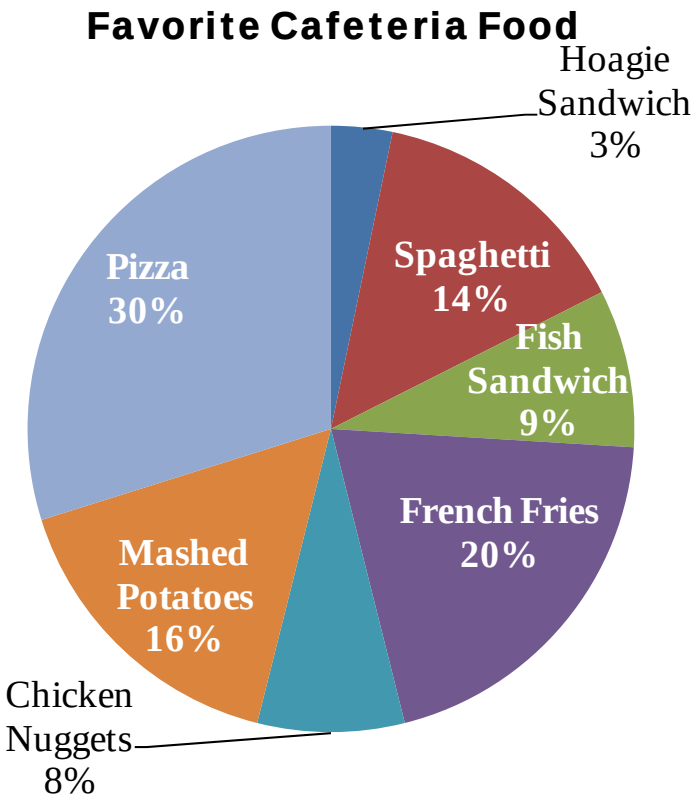
Jennifer surveyed the ninth grade students to determine the favorite food in the school cafeteria with these results: Pizza (46), Mashed Potatoes (25), Chicken Nuggets (12), French Fries (31), Fish Sandwich (13), Spaghetti (22), Hoagie Sandwich (5). She made a table to display her results.

Jennifer then decided to compare her results in two different types of graphs, a bar graph and a circle graph.

Which graph would you use to display the results to the student body?



In the **bar graph**, the **actual number** of students surveyed can be determined be reading the number represented by each bar.



In the **circle graph**, the **percent of students** is shown rather than the number of students.

In **both** graphs, **pizza** is clearly the favorite choice of the students.

Misleading Statistics

Statistics and graphs are very valuable tools for determining, processing, and evaluating data. However, some persons may misuse statistics purposely to promote an outcome they desire.

Example 1: Timothy purchased five raffle tickets that cost $1 each, based on the sales promotion that the top prize was $400 and the average award amount was $20. What the salesperson failed to mention to Timothy is that the rest of the prize awards were for $1.

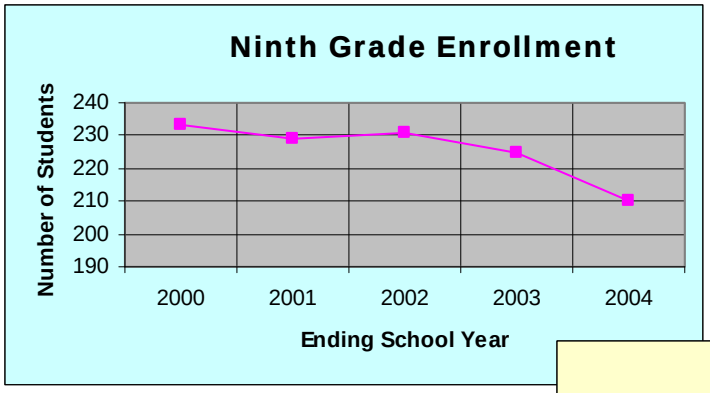
This promotion was mathematically correct, but if Timothy won, most likely his prize would be $1, not $400.

The Math:

1 prize of $400 = 400
20 prizes of $1 = 20
---
21 prizes = 420

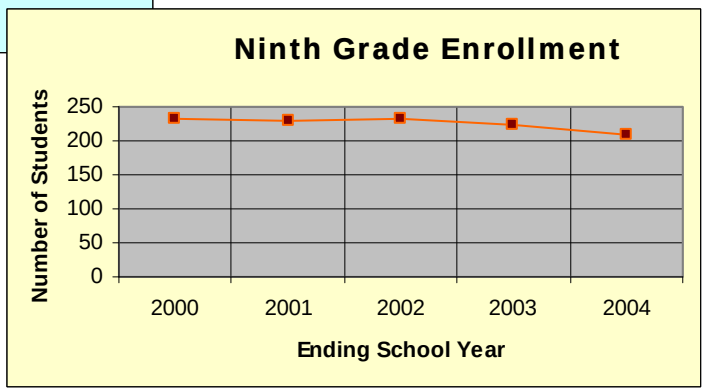
Average prize amount is $\frac{420}{21} = \$20$

Example 2: Graphs may be constructed in a way to emphasize certain data presented. Here are two graphs that show the enrollment of a ninth grade class over the school years 2000 through 2004.



In this graph, an overall decline in enrollment is shown with a sharp decline in the year 2004.

In this graph, the enrollment looks pretty steady with a small decline in the year 2004.



The different appearance of these two graphs was simply achieved by making changes in the beginning and ending values of the vertical axis and also by making changes in the amount of increments in the scale of the vertical axis.