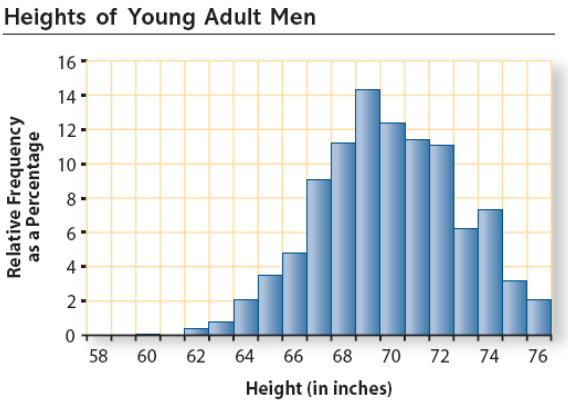
Exploring Distributions



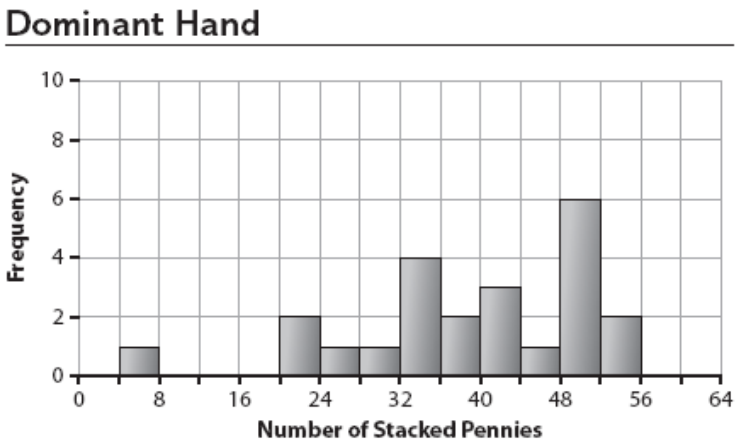
The statistical approach to problem solving includes refining the question you want to answer, designing a study, collecting the data, analyzing the data collected, and reporting your conclusions in the context of the original question. For example, consider the problem described below.
A Core-Plus Mathematics teacher in Traverse City, Michigan, was interested in whether eye-hand coordination is better when students use their dominant hand than when they use their non-dominant hand. She refined this problem to the specific question of whether students can stack more pennies when they use their dominant hand than when they use their non-dominant hand. In her first-hour class, she posed non-dominant hand. In her first-hour class, she posed

How many pennies can you stack using your dominant hand?

In her second-hour class, she posed this question:

How many pennies can you stack using your non-dominant hand?

In both classes, students were told: "You can touch pennies only with the one hand you are using; you have to place each penny on the stack without touching others; and once you let go of a penny, it cannot be moved. Your score is the number of pennies you had stacked before a penny falls."
Students in each class counted the number of pennies they stacked and prepared a plot of their data. The plot from the first-hour class is shown below. A value on the line between two bars (such as stacking 24 pennies) goes into the bar on the right.

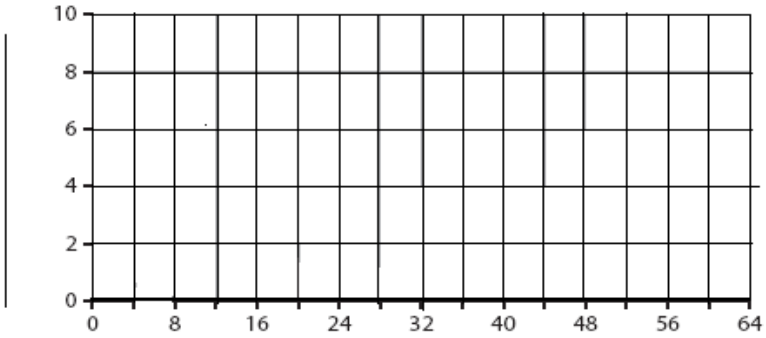


Examine the distribution of the number of pennies stacked by students in the first-hour class using their dominant hand.

a How many students were in the first-hour class? What percentage of the students stacked 40 or more pennies using their dominant hand?

b What do you think the plot for the second-hour class might look like?

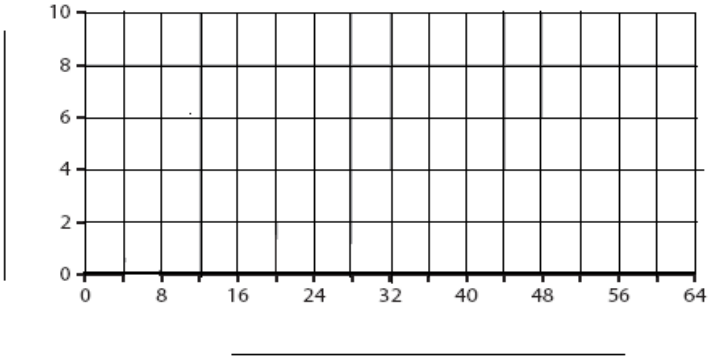
c. Check your conjecture in Part b by having your class stack pennies using your **non-dominant** hands. Make a plot of the numbers stacked by your class using the same scale as that for the dominant hand plot above.



d . Compare the shape, center, and spread of the plot from your class with the plot of the first-hour class on the previous page. What conclusions, if any, can you draw?

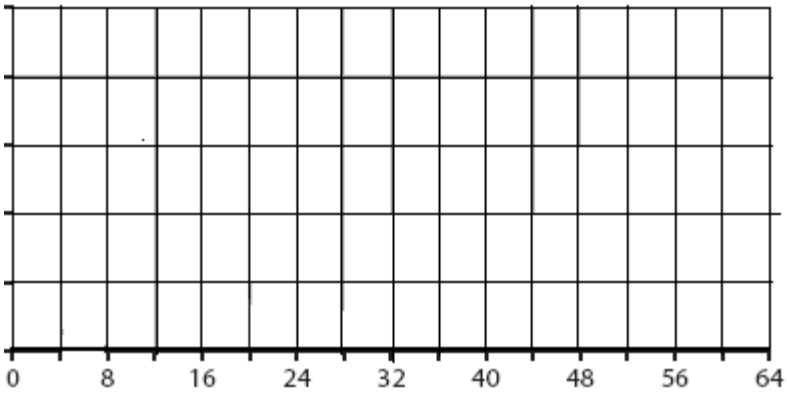
e. Why might comparing the results of first- and second-hour students not give a good answer to this teacher's question? Can you suggest a better design for her study?

f. i. Now make a histogram of the **dominant hand** data.

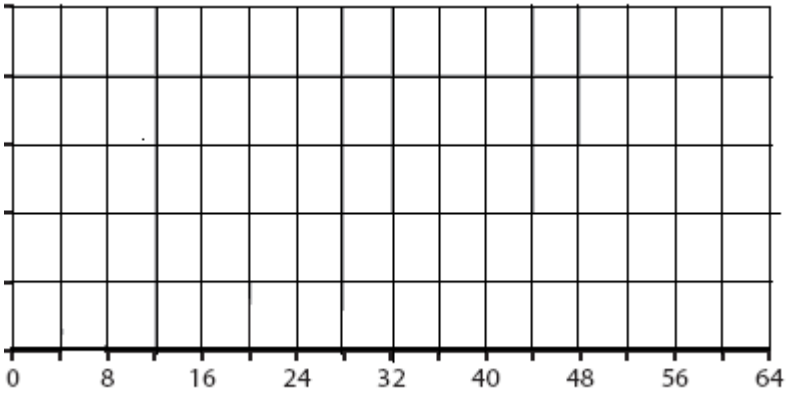


ii. Compare the shape, center, and spread of the plot from our classes dominant hand with the plot of our classes non-dominant hand. What conclusions, if any, can you draw?

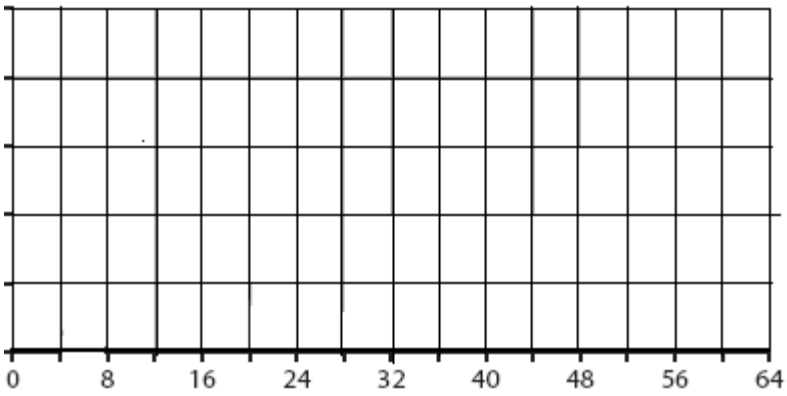
g. i. Now make a relative frequency histogram of the **non-dominant hand** data.



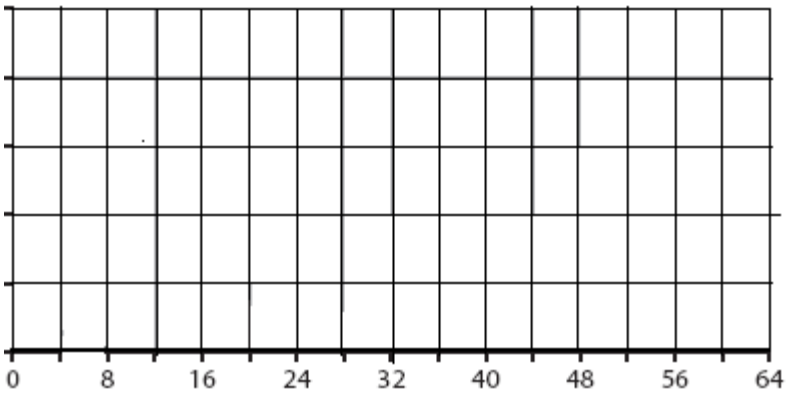
ii. Now make a relative frequency histogram of the **dominant** hand data



i. Now make a dot plot of the **non-dominant hand** data.



ii. Now make a dot plot of the **dominant** hand data

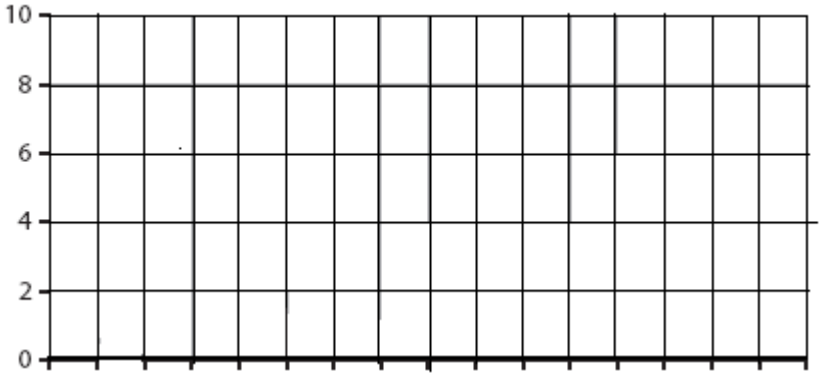


I. Now make a stem and leaf plot of the **non-dominant hand** data.

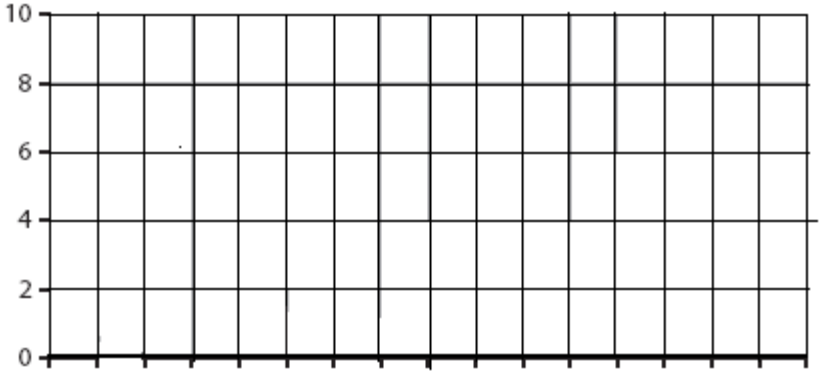
ii. Now make a stem and leaf plot of the **dominant** hand data

iii. Now make a back to back stem and leaf plot of the non-dominant and dominant hand data.

J. i. Now make a histogram using your calculator of the **non-dominant hand** data.



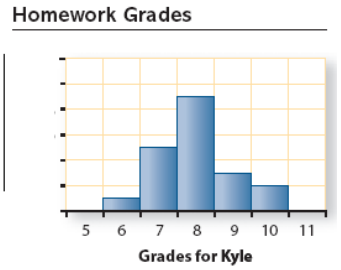
ii. Now make a histogram using your calculator of the **dominant** hand data.

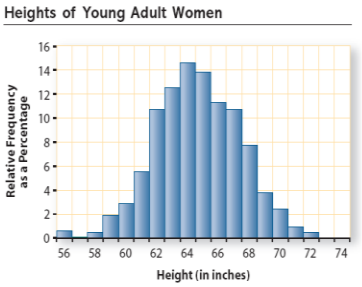


Making sense of data is important in everyday life and in most professions today. When describing a distribution, it is important to include information about its ***shape, center, range,*** **measure of *spread***, and to see if there are any ***outliers*.**

The Histograms below would be described as an **approximately normal** distribution.

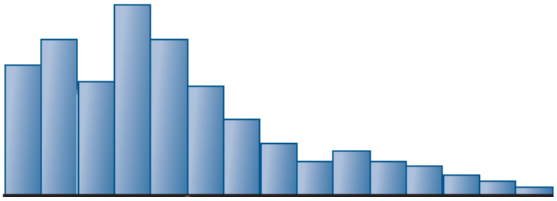
- The left half and the right half look like mirror images of each other.
- When describing a distribution that is **approximately normal** you should always give the **mean as the center** of the distribution and the **standard deviation** as the **measure of spread**

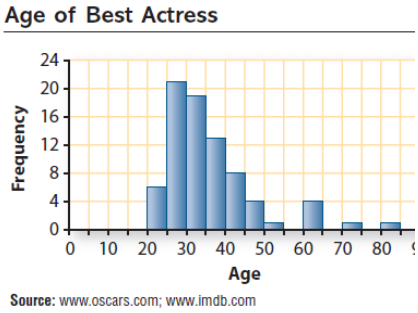




The Histograms below would be described as distributions that are **skewed right**.

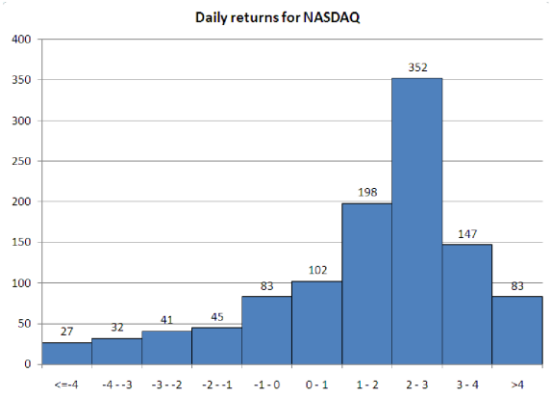
- These distributions are stretched toward the large values
- When describing a distribution that is **skewed right** you should always give the **median as the center** of the distribution and the **Interquartile Range** as the **measure of spread**.

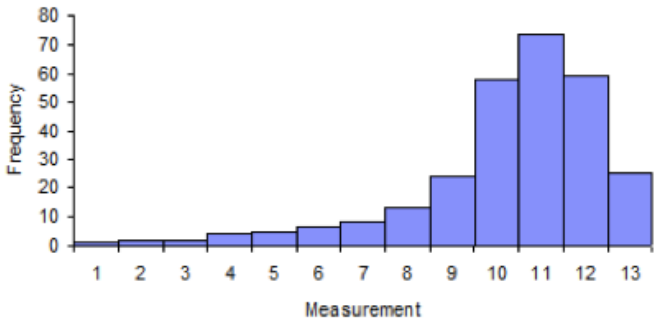




The Histograms below would be described as distributions that are **skewed left**.

- These distributions are stretched toward the large values
- When describing a distribution that is **skewed left** you should always give the **median as the center** of the distribution and the **Interquartile Range** as the **measure of spread**.





1. The following table gives nutritional information about some fast food sandwiches.

How Fast-Food Sandwiches Compare

Company	Sandwich	Total Calories
McDonald's	Cheeseburger	310
Wendy's	Jr. Cheeseburger	320
McDonald's	Quarter Pounder	420
McDonald's	Big Mac	560
Burger King	Whopper Jr.	390
Wendy's	Big Bacon Classic	580
Burger King	Whopper	700
Hardee's	1/3 lb Cheeseburger	680
Burger King	Double Whopper w/Cheese	1,060
Hardee's	Charbroiled Chicken Sandwich	590
Hardee's	Regular Roast Beef	330
Wendy's	Ultimate Chicken Grill	360
Wendy's	Homestyle Chicken Fillet	540
Burger King	Tendercrisp Chicken Sandwich	780
McDonald's	McChicken	370
Burger King	Original Chicken Sandwich	560
Subway	6" Chicken Parmesan	510
Subway	6" Oven Roasted Chicken Breast	330
Arby's	Regular Roast Beef	320
Arby's	Super Roast beef	440

Source: *McDonald's Nutrition Facts*, McDonald's Corporation, 2005; *U.S. Nutrition Information*, Wendy's International, Inc., 2005; *Nutrition Data* Food Systems, Inc., 2005; *Subway Nutrition Facts-US*, Subway, 2005; *Arby's Nutrition Information*, Arby's, Inc., 2005.

a. Use your calculator or data analysis software to make a histogram of the total calories for the sandwiches listed. Give the window from your calculator.

b. Describe the shape, center, and range of the distribution. Are there any outliers?

2. The table below gives the number of pennies stacked by 23 students with their dominant hand.

Dominant Hand							
27	35	41	36	34	6	42	20
47	41	51	48	49	32	29	21
50	51	49	35	36	53	54	

a. Use your calculator or data analysis software to make a histogram. Give the window from your calculator.

b. Describe the shape, center, and range of the distribution. Are there any outliers?

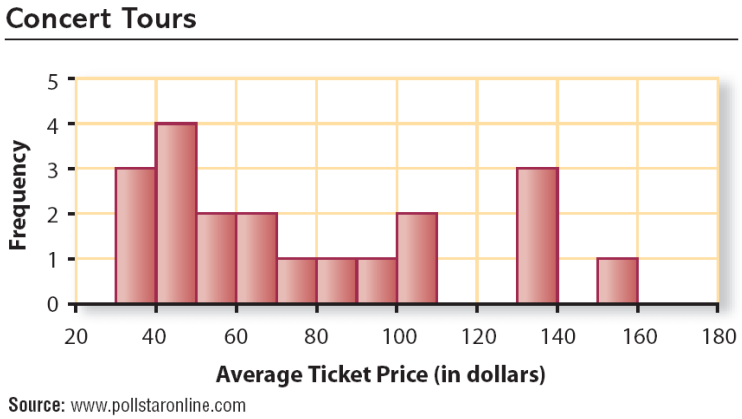
3. The data below represents the lengths of 32 black bears.

54 55 55 57 57 57 59 59 59 59 60 60 60 60 60 60 61 61 61 61 61 62 62 62 62 63 63 63 64 64 66 70

a. Use your calculator or data analysis software to make a histogram. Give the window from your calculator.

b. Describe the shape, center, and range of the distribution. Are there any outliers?

Pollstar estimates that revenue from all major North American concerts in 2005 was about $3.1 billion. The histogram below shows the average ticket price for the top 20 North American concert tours.



a. For how many of the concert tours was the average price $100 or more?

b. Barry Manilow had the highest average ticket price.
   i. In what interval does that price fall?

   ii. The 147,470 people who went to Barry Manilow concerts paid an average ticket price of $153.93. What was the total amount paid(gross) for all of the tickets?

c. The lowest average ticket price for Rascall Flatts.
   i. In what interval does that price fall?

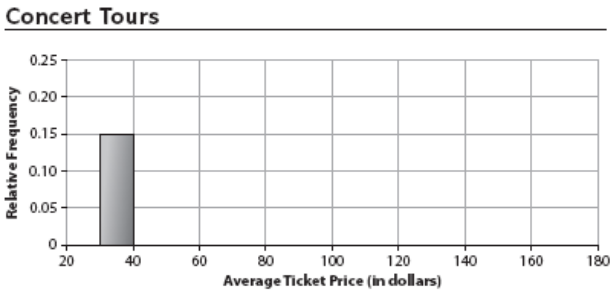
   ii. Their concert tour sold 807,560 tickets and had a gross of $28,199,995. What was the average price of a ticket to one of their concerts?

d. Describe the distribution of these average concert ticket prices (Shape/Center/Spread/Outliers)

5. Sometimes it is useful to display data showing the percentage or proportion of the data values that fall into each category. A **relative frequency histogram** has the proportion of percentage that fall into each bar of on the vertical axis rather than the frequency or count. Shown below is the start of a relative frequency histogram for the average concert ticket prices in problem 4.

a. Since prices between $30 and $40 happened 3 out of 20 times, the relative frequency for the first bar is 3/20 or .15. Complete a copy of the table and relative frequency histogram. Just as with the histogram an average price of $50 goes into the interval 50-60 in the table.

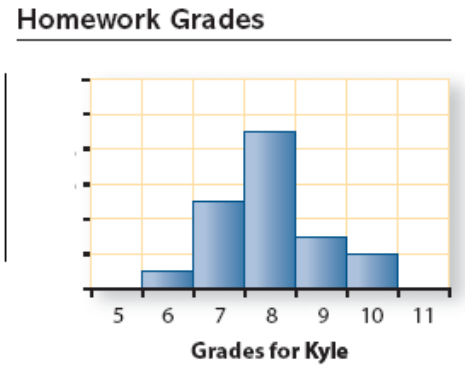
Average Price (in $)	Frequency	Relative Frequency
30–40	3	$\frac{3}{20} = 0.15$
40–50		
50–60		
60–70		
70–80		
80–90		
90–100		
100–110		
110–120		
120–130		
130–140		
140–150		
150–160		
Total		



b. When would it be better to use a relative frequency histogram for the average concert ticket prices rather than a histogram?

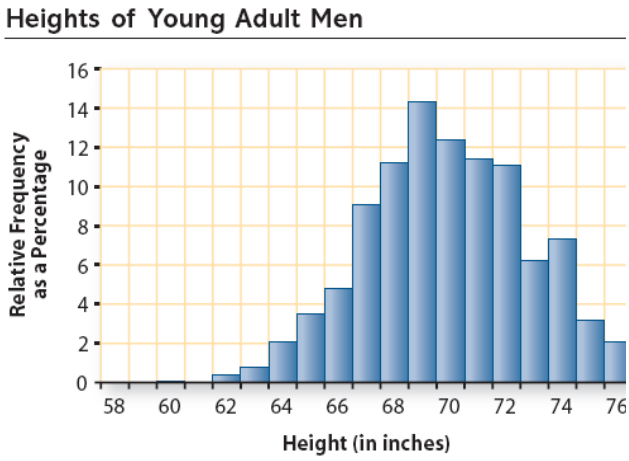
6. To study connections between a histogram and the corresponding relative frequency histogram, consider the histogram below showing Kyles's 20 homework grades for a semester. Notice that since each bar represents a single whole number (6, 7, 8, 9, or 10), those numbers are best placed in the middle of the bars on the horizontal axis. In this case, Kyle has one grade of 6 and five grades of 7.

a. Make a relative frequency histogram of these grades by copying the histogram but making a scale that shows proportion of all grades on the vertical axis rather than frequency.



b. Compare the shape, center, and spread of the two histograms.

The relative frequency histograms below show the heights of large samples of young adult men and women in the United States



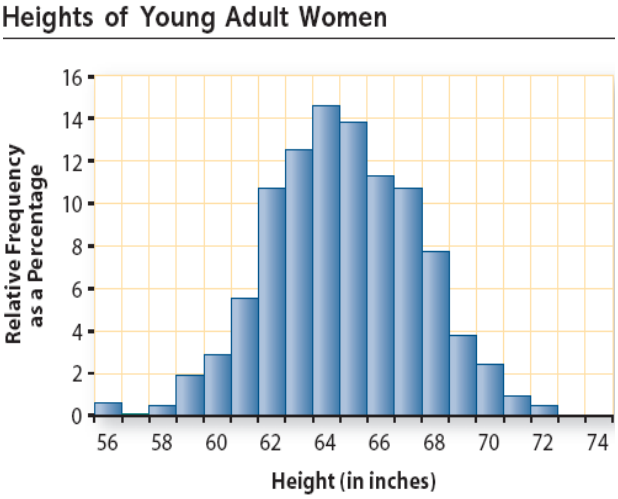
b. About what percentage of these young men are 6 feet tall.

c. About what percentage of these young men are at least 5 feet 9 inches tall?

d. If there are 5000 young men in this sample, how many are 6 feet tall.

e. If there are 5000 young men in this sample, how many are 5 feet 9 inches tall.

f. Walt Disney World recently advertised for singers to perform in Beauty and the Beast-Live on Stage.
   1) What percentage of men would meet the height requirement for Gaston (6'1" or taller)



g. About what percentage of these young women are 6 feet tall?

h. About what percentage these young women are 5 feet tall or less?

i. If there are 5000 young women in this sample, how many are 5 feet tall?

j. If there are 5000 young women in this sample, how many are 6 feet tall?

k. Walt Disney World recently advertised for singers to perform in Beauty and the Beast-Live on Stage.
   i. What percentage of women would meet the height requirement for Belle (5'5" -5'8:)

9.

Goals Scored	Number of Matches (frequency)
0	5
1	7
2	28
3	10
4	15

Goals Scored	Number of Matches (frequency)
5	8
6	5
7	1
8	1
9	1

a. What is the median number of goals scored per match?

b. What is the total number of goals scored in all matches?

c. What is the mean number of goals scored per match?

10. Suppose that, to estimate the mean number of children per household in a community, a survey was taken of 114 randomly selected households. The results are summarized in this frequency table.

Household Size

Number of Children	Number of Households
0	15
1	22
2	36
3	21
4	12
5	6
7	1
10	1

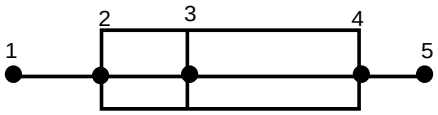
a. How many of the households had exactly 2 children?

b. Make a histogram of the distribution. Estimate the mean number of children per household from the histogram.

c. Calculate the mean number of children per household.

d. How will a frequency table of the number of children in the households of the students in your class be different from the one above? To check your answer, make a frequency table and describe how it differs from the one from the community survey. Would your class be a good sample to use to estimate the mean number of children per household in your community?

Box-and-Whisker Plot – is a data display that divides a set of data into four parts. The *median* or *second quartile* separates the set into two halves: the numbers that are below the median and the numbers that are above the median. The *first quartile* is the median of the lower half. The *third quartile* is the median of the upper half. The *lower extreme* is the least data value and the *upper extreme* is the greatest data value.



Five Number Summary

1 = minimum value, lower extreme

2 = 1st quartile, lower quartile (25th percentile)

3 = 2nd quartile, median (50th percentile)

4 = 3rd quartile, upper quartile (75th percentile)

5 = maximum value, upper extreme

Interquartile Range is the distance between the first quartile and third quartile. It accounts for the middle 50% of the data.

Outliers are values much lower or much higher than most of the data. In a box-and-whisker plot, outliers are data that fall more than 1.5 times the interquartile range from the quartiles. Do not extend whiskers to any outliers.

Lesson Summary:

- Non-symmetrical data distributions are referred to as skewed.
- Left-skewed or skewed to the left means the data spreads out longer (like a tail) on the left side.
- Right-skewed or skewed to the right means the data spreads out longer (like a tail) on the right side.
- The center of a skewed data distribution is described by the median.
- Variability of a skewed data distribution is described by the interquartile range (IQR).
- The IQR describes variability by specifying the length of the interval that contains the middle 50% of the data values.
- Outliers in a data set are defined as those values more than 1.5(IQR) from the nearest quartile. Outliers are usually identified by an "*" or a "•" in a box plot.

1. Thirty female users and twenty-five male users were selected at random from a database of people who play a video game regularly. Each of them agreed to be part of a research study and report their scores. A leadership score is based on a player's answers to leadership questions. A score of 1 to 40 is considered a beginning level leadership score, a score of 41 to 60 is considered a middle level leadership score, and a score of greater than 60 is considered an advanced level leadership score.

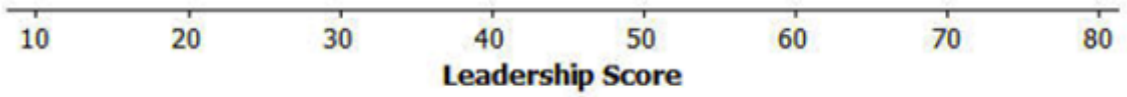
a. Use the following data to make a box plot of the female scores, and a box plot of the male scores on the line graph below. Give the 5 number summary for each set of data.

Female scores:

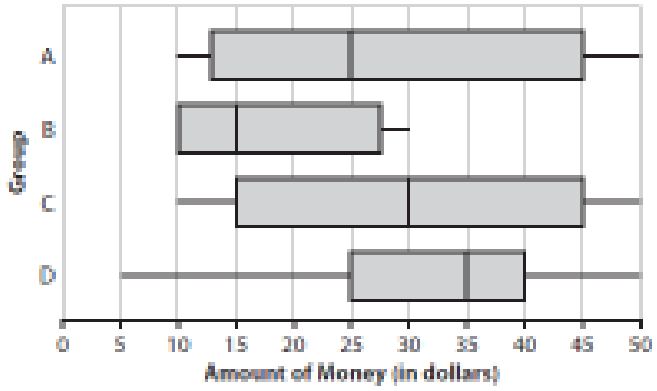
10	20	20	20	30	30	30	40	40	40
50	50	55	65	65	65	65	65	70	70
70	70	76	76	76	76	76	76	76	76

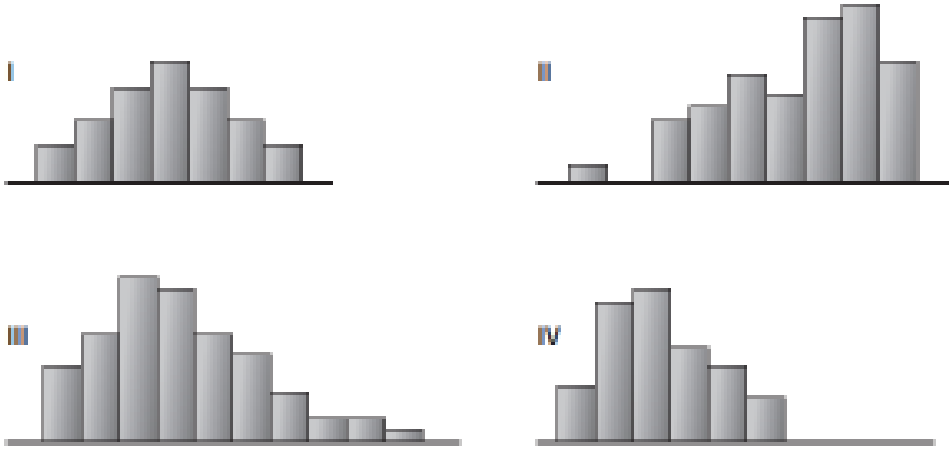
Male scores:

15	20	20	25	25	25	25	30	30	30
30	30	30	35	35	35	35	35	40	40
40	45	45	45	50					



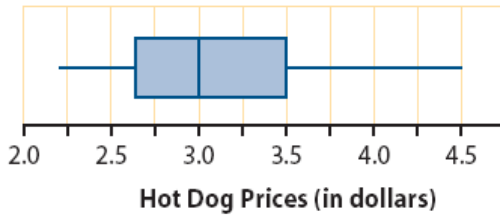
2. Match the histogram with the box plot and describe the shape of each.





The five-number summary can be displayed in a **box plot**. To make a box plot, first make a number line. Above this line draw a narrow box from the lower quartile to the upper quartile; then draw line segments connecting the ends of the box to each **extreme value** (the maximum and minimum). Draw a vertical line in the box to indicate the location of the median. The segments at either end are often called **whiskers**, and the plot is sometimes called a **box-and-whiskers plot**.

3. The following box plot shows the distribution of hot dog prices at Major League Baseball parks.

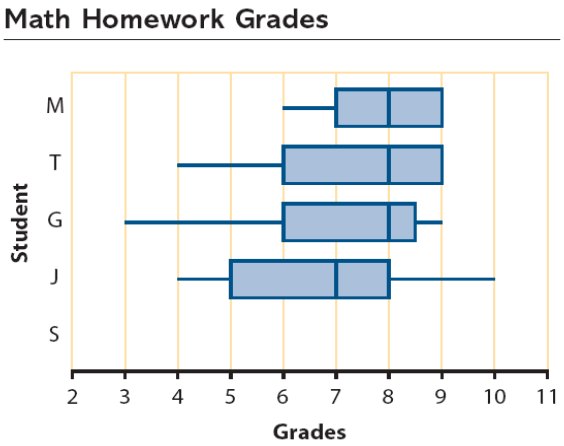


a. Is the distribution skewed to the left or to the right, or is it symmetric? Explain your reasoning.

b. Estimate the five-number summary. Explain what each value tells you about hot dog prices.

4. Use your calculator to make a box plot of Susan's grades. Add Susan box and whisker plot to the graph below

8, 8, 7, 9, 7, 8, 8, 6, 8, 7, 8, 8, 8, 7, 8, 8, 10, 9, 9, 9



Box plots are most useful when the distribution is skewed or has outliers or if you want to compare two or more distributions. The math homework grades for five ninth-grade students at Lakeview High School—Maria (M), Tran (T), Gia (G), Jack (J), and Susan (S)—are shown with corresponding box plots.

a. Why do the plots for Maria and Tran have no whisker at the upper end?

ci. Why is the lower whisker on Gia's box plot so long?

cii. Are there more grades for Gia in longer whisker than in the shorter whisker?

di. Which distribution is the most symmetric? **dii.** Which distributions are skewed to the left?

e. Which of the five students has the lowest median grade.

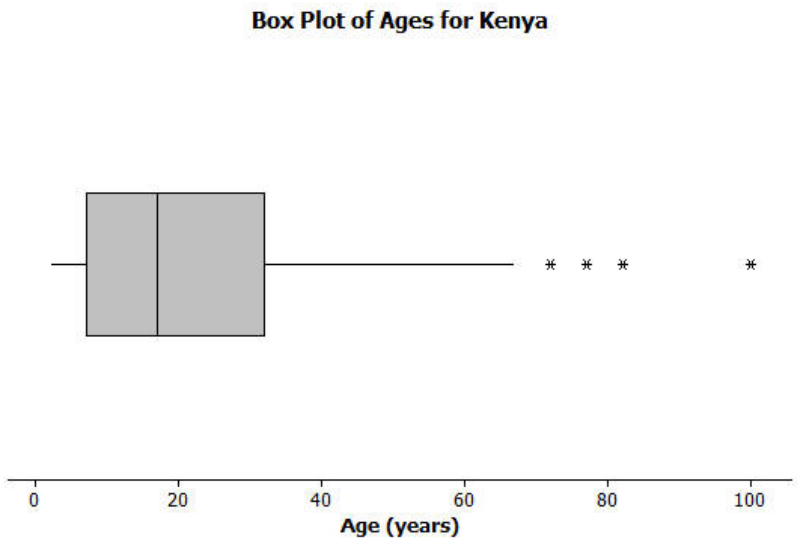
f. Which students have the smallest and largest **interquartile ranges**.

i. Does the student with the smallest interquartile range also have the smallest range?

ii. Does the student with the largest interquartile range also have the largest range?

g. Based on the box plots, which of the five students seems to have the best record?

Consider a box plot of the ages of 200 randomly selected people from Kenya:



A data distribution may contain extreme data. A box plot can be used to display extreme data values that are identified as outliers. The "*" in the box plot are the ages of 4 people from this sample. Based on the sample, these 4 ages were considered outliers. An outlier is defined to be any data value that is more than $\mathbf{1.5}\times(\boldsymbol{IQR})$ away from the nearest quartile.

a. Estimate the values of the 4 ages represented by an *.

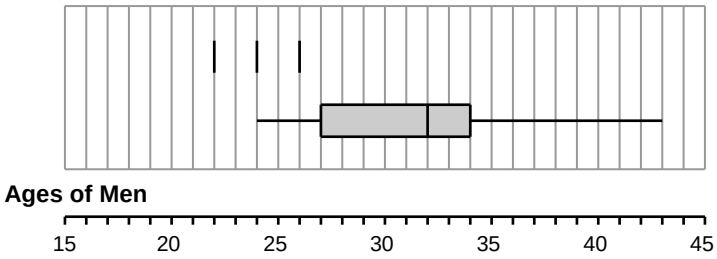
b. What is the median age of the sample of ages from Kenya? What are the approximate values of Q1 and Q3? What is the approximate interquartile range (IQR) of this sample?

c. Multiply the interquartile range (IQR) by 1.5. What value do you get?

d. Are there any age values that are greater than $\boldsymbol{Q}3+\mathbf{1.5}\times(\boldsymbol{IQR})$? If so, these ages would also be considered outliers.

e. Are there any age values that are less than $\boldsymbol{Q}\mathbf{1}-\mathbf{1.5}\times(\boldsymbol{IQR})$? If so, these ages would also be considered outliers.

8. The box plots below show the ages of the members of the 2006 U.S. Olympic Hockey team. Answer the questions that follow using the information in these plots.



a. Use the box plot to estimate the 5 number summary.

b. Are there any outliers? Show your work.

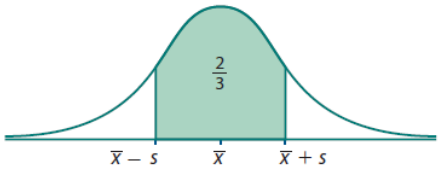
c. Describe the distribution.(Shape/Center/Spread/Outliers)

d. What percentile of the hockey players fall in the first quartile?

e. What percentile of the hockey players fall in the first 3 quartiles?

f. What percentage of the hockey players are 34 years or older? Explain how you know.

The IQR is very useful if the distribution **is skewed or has outliers**. For data that are **approximately normal—symmetric, mound-shaped, without outliers—**a different measure of spread called the ***standard deviation*** is typically used.

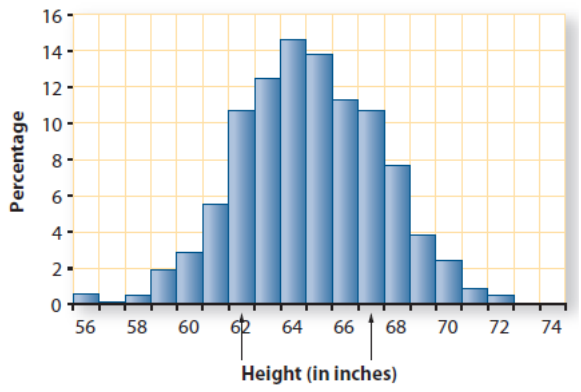


The **standard deviation *s*** is a distance that is used to describe the variability in a distribution. In the case of an approximately normal distribution, if you start at the mean and go the distance of one standard deviation to the left and one standard deviation to the right, you will enclose the middle 68% (about two-thirds) of the values. That is, in a distribution that is approximately normal, about two-thirds of the values lie between $\bar{x} - s$ and $\bar{x} + s$

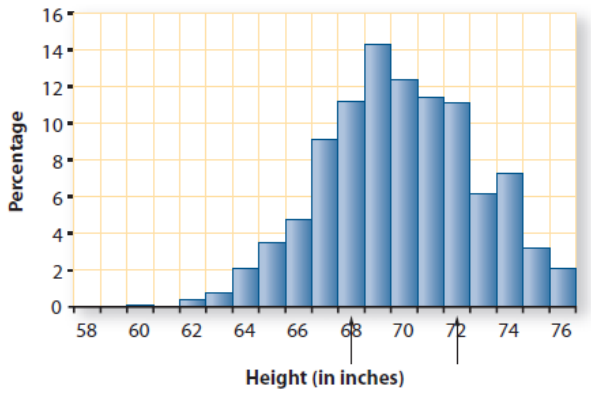
1. On each of the following distributions, the arrows enclose the middle two-thirds of the values. For each distribution:

i. Estimate the mean and the standard deviation(the distance from the mean to one of the two arrows)

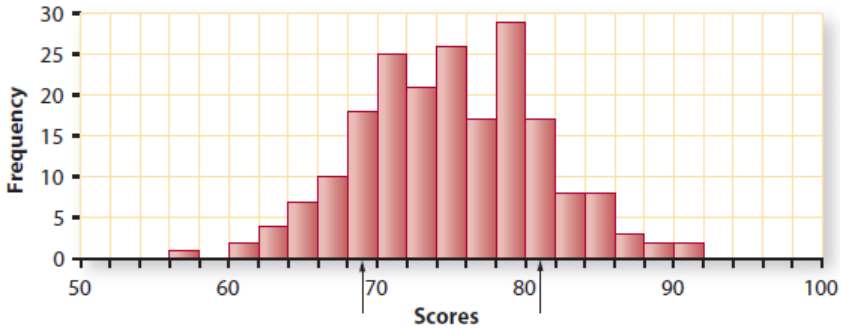
Heights of Young Adult Women



Heights of Young Adult Men



Achievement Test Scores



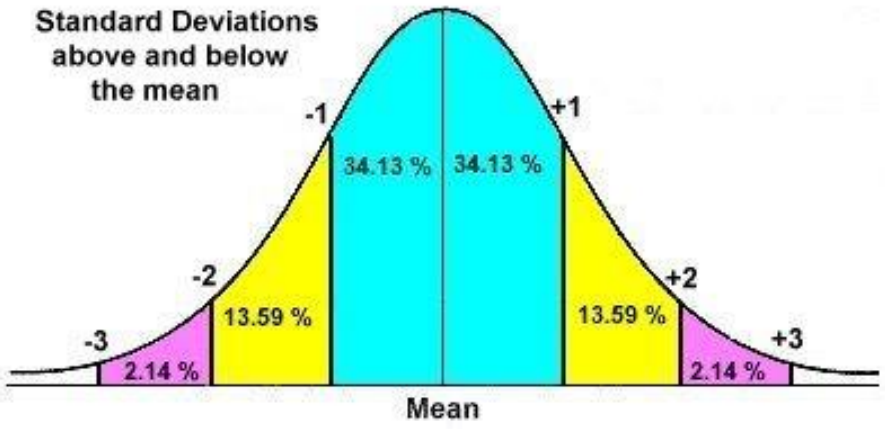
2. The sophomores who took the PSAT/NMSQT test in 2004 had a mean score of 44.2 on the mathematics section, with a standard deviation of 11.1. The distribution of scores was approximately normal. The highest possible score was 80 and the lowest was 20.

a. Sketch the shape of the histogram of the distribution of scores, including a scale on the *x*-axis.

b. A sophomore who scored 44 on this exam would be at about what percentile?

c. A sophomore who scored 33 on this exam would be at about what percentile?

d. A sophomore who scored 55 on this exam would be at about what percentile?



3. Using your calculator determine the mean and standard deviation for the data below. Then give the point values that are 1 standard deviation away from the mean.

Points Scored by LeBron James in His First Month

Date	Opponent	Total Points
Dec. 3	Cuyahoga Falls	15
Dec. 4	Cleveland Central Catholic	21
Dec. 7	Garfield	11
Dec. 17	Benedictine	27
Dec. 18	Detroit Redford	18
Dec. 28	Mansfield Temple Christian	20
Dec. 30	Mapleton	21

The data below represents the lengths of 32 black bears.

54 55 55 57 57 57 59 59 59 59 60 60 60 60 60 60 61 61 61 61 61 62 62 62 62 63 63 63 64 64 66 70

a. Calculate the mean and standard deviation.

b. Then give the point values that are 1 standard deviation away from the mean.

5. Using your calculator determine the mean and standard deviation for Susan and Jack

Susan's Homework Grades

8, 8, 7, 9, 7, 8, 8, 6, 8, 7,
8, 8, 8, 7, 8, 8, 10, 9, 9, 9

Jack's Homework Grades

10, 7, 7, 9, 5, 8, 7, 4, 7,
5, 8, 8, 8, 4, 5, 6, 5, 8, 7

a. Which student had the larger standard deviation? Explain why that makes sense?