CHAPTER 2: NUMERICAL & GRAPHICAL SUMMARIES OF QUANTITATIVE DATA
FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

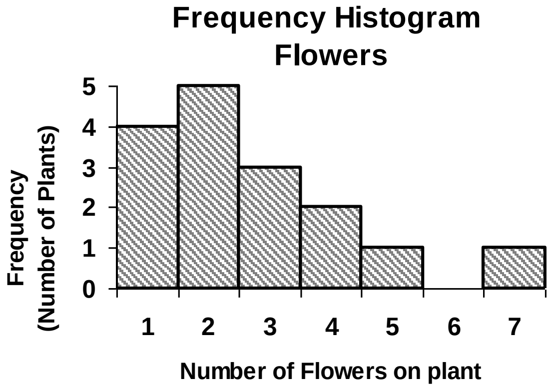
A HISTOGRAM is a bar graph displaying quantitative (numerical) data

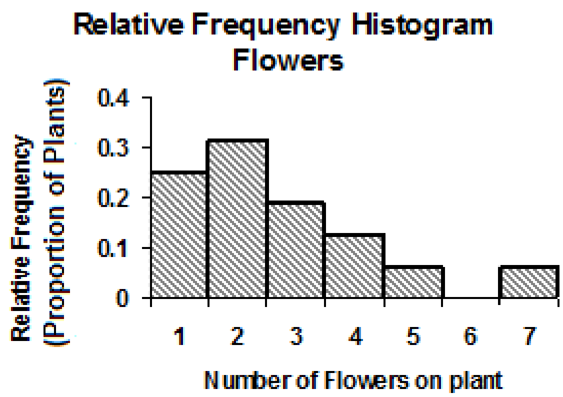
- Consecutive bars should be touching. There should not be a gap between consecutive bars.
- A "gap" should occur only if an interval does not have any data lying in it.
- Vertical axis can be frequency or can be relative frequency.

EXAMPLE 1:
Individual Data Values

Plants are being studied in a lab experiment.
The number of flowers on a plant, for a sample of 16 plants in this experiment are:
2,5,3,1,2,4,1,2,3,1,1,2,7,4,2,3

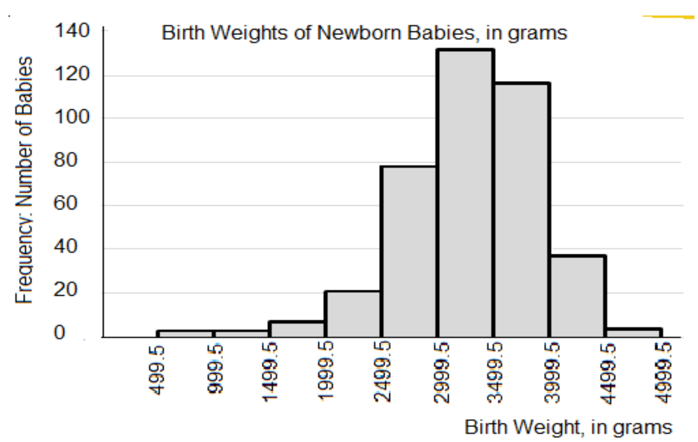
Number of Flowers	Frequency	Relative Frequency	Cumulative Relative Frequency
1	4	0.25	0.25
2	5	0.3125	0.5625
3	3	0.1875	0.75
4	2	0.125	0.875
5	1	0.0625	0.9375
7	1	0.0625	1.0





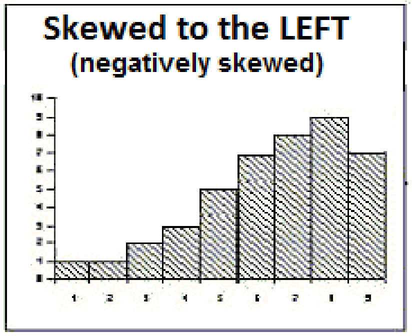
EXAMPLE 2:	Weight (grams) Interval Class Limits	Class Boundaries	Frequency	Relative Frequency	Cumulative Relative Frequency
Birthweights, in grams, for a sample of 400 newborn babies born at a hospital	500-999	499.5 – 999.5	3	0.0075	0.0075
	1000-1499	999.5-1499.5	3	0.0075	0.015
	1500-1999	1499.5-1999.5	7	0.0175	0.0325
	2000-2499	1999.5-2499.5	21	0.0525	0.085
Data is grouped into intervals	2500-2999	2499.5-2999.5	78	0.195	0.28
	3000-3499	2999.5-3499.5	131	0.3275	0.6075
	3500-3999	3499.5-3999.5	116	0.29	0.8975
	4000-4499	3999.5-4499.5	37	0.0925	0.99
	4500-4999	4499.5-4999.5	4	0.01	1

Describe the shape of the histogram, using proper terminology:



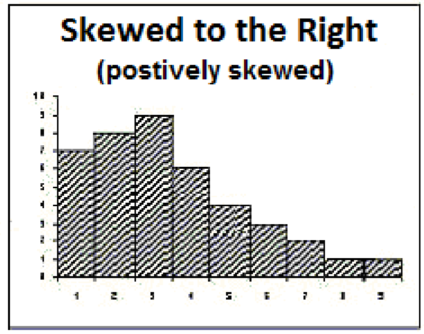
Note: In this class we will use intervals of equal width, as shown in the table and in the histogram; although unequal intervals can be used in some situations, the statistical work is easier if the intervals have equal width.

Shapes of Data Distributions



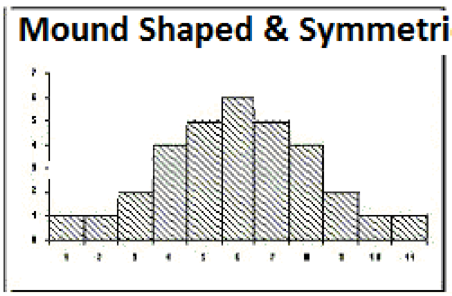
Skewed to the LEFT
(negatively skewed)

When data are skewed to the left
generally the mean is less than the median



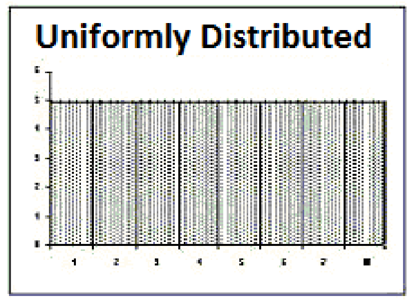
Skewed to the Right
(postively skewed)

When data are skewed to the right
generally the mean is greater than the median

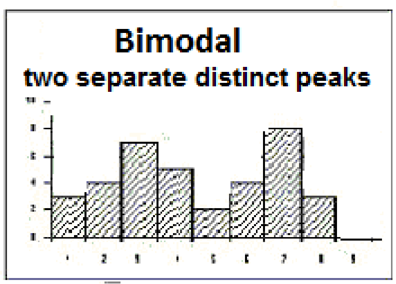


Mound Shaped & Symmetric

For Symmetric data
mean = median



Uniformly Distributed

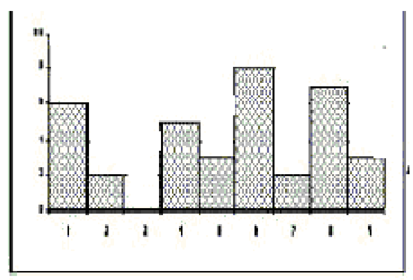


Bimodal
two separate distinct peaks

Distinct peaks appear as "hills"
separated by a "valley"

Peaks do not need to be exactly
the same height

IF data do not fit one of the descriptive terms for data, do not use a term that does not fit its shape.
Just describe what you see in the data if none of these descriptive terms apply



VOCABULARY

- **Class <u>Limits</u>: Lowest and highest possible data values in an interval.**
- **Class <u>Boundaries</u>: Numbers used to separate the classes, but without gaps.**
  Boundaries use one more decimal place than the actual data values and class limits. This prevents data values from falling on a boundary, so no ambiguity exists about where to place a particular data value
- **Class <u>Width</u>: Difference between two consecutive class boundaries**
  Can also calculate as difference between two consecutive <u>lower class limits</u>
- **Class <u>Midpoints</u>: Midpoint of a class = (lower limit + upper limit) / 2**

CHAPTER 2: CALCULATOR INSTRUCTIONS for TI-83 and TI-84 Calculators

Putting TI-84 calculator into Classic Mode with Stat Wizards "Off"

The TI-83 has only one way to display information on the screen and to do statistical functions.
Most newer TI-84 calculator have several ways to do this, but they can also be configured to match the TI-83.

In class the instructor will use a TI-84 in "classic" mode with "Stat Wizards" turned "off" to match how the TI-83 works. This will allow students using the TI-83 and those using the TI-84 to use the same keystrokes to match exactly what the instructor demonstrates.

Students using a TI-84 can use Classic Mode and turn off the Stat Wizards to match the instructor's calculator if they want to be able to do exactly what the instructor's calculator shows.
TI-84 only: Press MODE key. Arrow cursor to scroll down to next screen. Arrow cursor to CLASSIC and press ENTER. Arrow cursor down and right to highlight Stat Wizards OFF and press ENTER.

*Students using a TI-84 can choose to use Mathprint mode and/or turn on Stat Wizards if they prefer but the instructor will usually not demonstrate this in class.

Entering data into TI-83, 84 statistics list editor:

STAT "EDIT" Put data into list L1, press ENTER after each data value
If you have a frequencies for each value, enter frequencies into list L2, press ENTER after each value
2nd QUIT to exit stat list editor <u>after</u> you have entered data, checked it and corrected errors.

HISTOGRAM instructions for the TI-83, 84: Assuming your data has been entered in list L1

2nd STATPLOT 1
Highlight "**ON**" ; press **ENTER**
Type: Highlight histogram icon press **ENTER**
Xlist: **2nd L1 ENTER**
Freq: If there is no frequency list and all data is in one list type **1 ENTER**
OR If there is a frequency list, enter that list here **2nd L2 ENTER**

Set the appropriate window and scale for the histogram
WINDOW
XMin: lower boundary of first interval **XMax**: upper boundary of last interval **Xsc** =interval width
Example: For intervals 10 to <20, 20 to <30, . . . 60 to <70: Xmin = 9.5 Xmax=69.5 Xscl=10
YMin = 0 Estimate **YMax** to be large enough to display the tallest bar
Select an appropriate value of **YScl** for the tick marks on the y-axis
GRAPH **Calculator constructs the histogram**

TRACE You can use the left and right cursors (arrow keys) to move from bar to bar.
The screen indicates the frequency (count, height) for the bar that the cursor is positioned on.

Finding One Variable Summary Statistics on your TI-83,84 calculator

If <u>not</u> using a frequency list: Put data into list L1, press ENTER after each data value
2nd QUIT to exit stat list editor <u>after</u> you entered data, checked & corrected errors.
STAT "CALC" .1. *for 1 – Var Stats* 2nd L1 ENTER
If data is in a different list than L1, indicate the appropriate listname instead of L1

STATWIZARD
List: L1
FreqList:
Calculate

If using a frequency list: Put data into list L1, frequencies into list L2, press ENTER after each data value
2nd QUIT to exit stat list editor <u>after</u> you have entered data, checked it and corrected errors.
STAT "CALC" .1. *for 1 – Var Stats* 2nd L1 , 2nd L2 ENTER
order of lists should be data value list, frequency list

STATWIZARD
List: L1
FreqList: L2
Calculate

EXAMPLE 3: Elementary School Enrollment Cupertino Union School District

Number of Students at each elementary school in the district

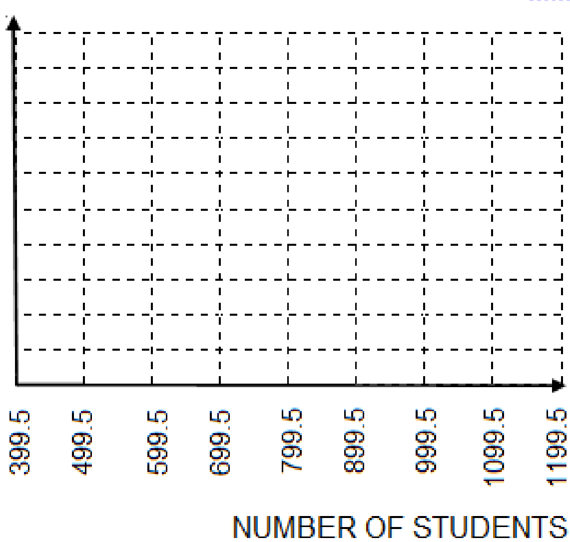
Christa McAuliffe Elementary School	443		William Faria Elementary School	632
Nelson S. Dilworth Elementary School	455		West Valley Elementary School	646
Blue Hills Elementary School	489		Stevens Creek Elementary School	658
John Muir Elementary School	493		Chester W. Nimitz Elementary School	704
Montclaire Elementary School	505		Abraham Lincoln Elementary School	735
Manuel De Vargas Elementary School	510		Dwight D. Eisenhower Elementary	742
D. J. Sedgwick Elementary School	522		R. I. Meyerholz Elementary School	742
William Regnart Elementary School	559		L. P. Collins Elementary School	746
Murdock-Portal Elementary School	562		Garden Gate Elementary School	766
C. B. Eaton Elementary School	611		Louis E. Stocklmeir Elementary	1172

Source: http://www.zillow.com/ca/districts/cupertino-union-428717/#/ca/districts/cupertino-union-428717/s=size_up&p=3

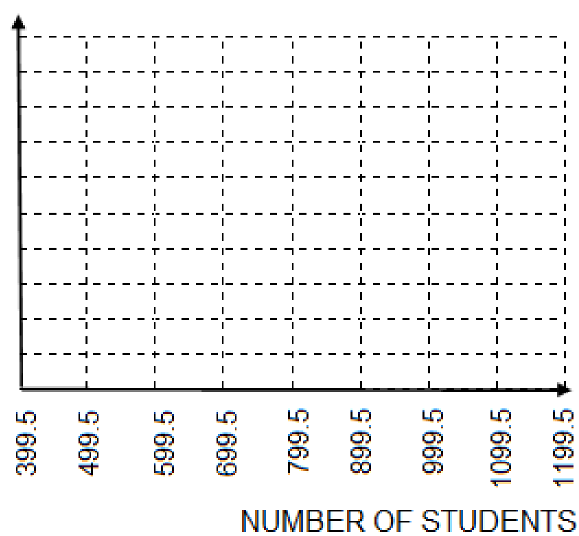
Interval (Class Limits)	Class Boundaries	Frequency	Relative Frequency
400-499			
500-599			
600-699			
700-799			
800-899			
900-999			
1000-1099			
1100-1199			

Create a histogram on your calculator using the lowest and highest class boundaries as the XMin and XMax; use the interval width as the Xscl.

Draw a **frequency histogram**.
Label and scale vertical axis using 0, 1, 2, 3, 4, ...



Draw a **relative frequency histogram**
Label and scale vertical axis using 0, 0.1, 0.2, . . .



CHAPTER 2: GRAPHICAL DISPLAYS OF QUANTITATIVE DATA: STEM AND LEAF PLOTS

Each data value is split into a stem and leaf using place value. Each stem shows only once but each data value gets is own leaf. A key indicating the place value representation by the stem and leaf should be shown.

EXAMPLE 4:

Suppose that a random sample of 18 mathematics classes at a community college showed the following data for the number of students enrolled per class:. Construct a stem and leaf plot.

Raw Data:	**37, 40, 38, 45, 28, 60, 42, 42, 32, 43, 36, 40, 82, 42, 39, 36, 60, 25**
Sorted Data:	**25, 28, 32, 36, 36, 37, 38, 39, 40, 40, 42, 42, 42, 43, 45, 60, 60, 82**

EXAMPLE 5

The table shows the number of baseball games won by each American League Major League Baseball Team in the 2010 regular season.

2010 Regular Season	Games Won
Tampa Bay Rays	96
New York Yankees	95
Boston Redsox	89
Toronto Blue Jays	85
Baltimore Orioles	66
Minnesota Twins	94
Chicago White Sox	88
Detroit Tigers	81
Cleveland Indians	69
Kansas City Royals	67
Texas Rangers	90
Oakland A's	81
LA Anaheim Angels	80
Seattle Mariners	61

Games Won (Sorted Data)
61
66
67
69
80
81
81
85
88
89
90
94
95
96

Construct a stem and leaf plot:

EXAMPLE 6: Read the data from this stem and leaf:

Weights of **18** randomly selected packages of meat in a supermarket, in pounds.

1	389999
2	00011268
3	27
4	
5	0
6	2

Leaf Unit = .1
Stem Unit = 1
1|9 = 1.9

What is the weight of the smallest package?______

What is the weight of the largest package? ______

How many packages weigh at least 2 but less than 4 pounds? _____

How many packages weigh at least 4 but less than 5 pounds? _____

How many packages weigh at least 5 pounds? ______

EXAMPLE 7: Read the data from this stem and leaf:

Number of students at each of **18** elementary schools in a city

1	389999
2	00011268
3	27
4	
5	0
6	2

Leaf Unit = 10
Stem Unit = 100
1|9 = 190

How many students in the smallest school? ______

How many students in the largest school? ______

Read back several data values from the stem and leaf plot.
Do you notice anything interesting about the data?
Do you think that these numbers could represent the actual raw data or might they have been altered in some way?

CHAPTER 2: PERCENTILES & QUARTILES (Measures of Relative Standing)

The P^{th} percentile is the value that divides the data between the lower P% and the upper (100 – P)% of the data:

P% of data values are less than (or equal to) the P^{th} percentile
(100-P)% of data values are greater than (or equal to) the P^{th} percentile

EXAMPLE 8: Interpreting Quartiles and Percentiles

A class of 20 students had a quiz in the sixth week of class. Their quiz grades were:

2 5 8 10 12 12 12 14 14 14 15 15 17 17 17 18 20 20 20 20

a. The 40^{th} percentile is a quiz grade of 14.

40% of students had quiz grades of 14 or less. ***60% of students had quiz grades of 14 or more***

2 5 8 10 12 12 12 14 14 14 15 15 17 17 17 18 20 20 20 20

$P_{40} = 14$

b. The 20^{th} percentile is a quiz grade of 11. Write a sentence that interprets (explains) what this means in the context of the quiz grade data.

"Special" Percentiles: First Quartile Q1 Median (Med) Third Quartile Q3

Your calculator can find these special percentiles using 1-variable statistics

c. The third quartile is 17.5. Write a sentence that interprets the third quartile in the context of this problem.

EXAMPLE 9: INTERQUARTILE RANGE (IQR) : difference between third and first quartiles. The IQR measures the spread of the middle 50% of the data : IQR = Q3 – Q1

Find the Interquartile Range Q1 = ______ Q3 = ______ IQR = ___________________

Interpretations:

The lowest 25% of data values for the quiz grades are less than or equal to (at most) ___________

The middle ________% of the data values for the quiz grades are located between _________and __________

The highest 25% of data values for the quiz grades are greater than or equal to (at least) ___________

CHAPTER 2: ESTIMATING PERCENTILES FROM CUMULATIVE RELATIVE FREQUENCY

(using the method from Collaborative Statistics, B. Illowsky & S. Dean, www.cnx.org)

EXAMPLE 10: Quiz Grades: 2 5 8 10 12 12 12 14 14 14 15 15 17 17 17 18 20 20 20 20

X =Quiz Grade	Frequency	Relative Frequency	Cumulative Relative Frequency
2	1	1/20 =0.05	0.05
5	1	0.05	0.10
8	1	0.05	0.15
10	1	0.05	0.20
12	3	3/20 = 0.15	0.35
14	3	0.15	0.50
15	2	2/20 =0.10	0.60
17	3	0.15	0.75
18	1	0.05	0.80
20	4	4/20 = .20	1.00

Sort data into ascending order and complete the cumulative relative frequency table.
Do NOT group the data into intervals. Each data value is on its own line in the table.

Procedure to estimate pth percentile using the cumulative relative frequency column.
Look down the cumulative relative frequency table to look for the decismal value of p.

- **IF YOU PASS BEYOND THE DECIMAL VALUE OF p:**
  then pth percentile is the data value (x) column at the first line in the table BEYOND the value of p
  Find the 40th percentile: Look down the cumulative relative frequency column for 0.40.
  You don't find 0.40, but pass it between 0.35 and 0.50
  The 40th percentile is the x value for the line at which you first pass 0.40.
  The 40th percentile is 14

- **IF YOU FIND THE EXACT DECIMAL VALUE OF p:**
  then pth percentile is the average of the data (x) value in that line and in the next line of the table
  Find the 20th percentile: Look down the cumulative relative frequency column for
  You find 0.20, on the line where x = 10.
  The 20th percentile is the average of the x values on that line (10) and on the line below it (12)
  The 20th percentile is (10+12)/2=11

Technical Note 1: Why do we do it this way?

This method finds the median correctly, for even or odd numbers of data values.
Then we use the same method for all other percentiles.
The median is 14.5 (If there are an even number of data values, the median is the average of the two middle values: 14 and 15.)
Using the table to find the 50th percentile, we see 0.50 exactly in the table; the procedure tells us to average the x value, 14, and the next x value, 15. This correctly gives 14.5 as the 50th percentile.
If you did not average, but used the x value for the line showing 0.50, you would incorrectly use 14 as the median which is not correct.

---

*Technical Note 2: **We'll use the method above to find percentiles in Math 10.***

There are other methods that are also sometimes used to find percentiles. Some books use a positional formula (p/100)(n+1).Different statistical software programs or calculators sometimes use slightly different methods and may obtain slightly different answers.

CHAPTER 2: PRACTICE WITH PERCENTILES

<u>You must learn to write the interpretation as shown below</u>

For the pth percentile that has value x, the interpretation is:

P% of the "data values" are less than or equal to x

(100-P)% of the "data values" are greater than or equal to x

In these sentences you must use the context of the story in the problem instead of saying the words "data values"

Read Section 2.3 and do practice problems in the textbook Introductory Statistics at OpenStax; see guidelines in textbook for how to write the interpretations of percentiles.

EXAMPLE 11:

12a. http://www.bls.gov/oes/current/oes353031.htm A survey about workers earnings showed that the 90th percentile of hourly earnings (including tips) for waiters and waitresses is $15.35 and the first quartile is $8.38.

Write the sentence that interprets the 90th percentile in the context of this problem.

Write the sentence that interprets the first quartile in the context of this problem.

12b. Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? ____________

Write the sentence that interprets the 85th percentile in the context of this problem.

12c. **PRACTICE** Here are wait times in minutes for a sample of 50 people waiting in line at the DMV. Find the 30th percentile and the 60th percentile; briefly explain how you found each.

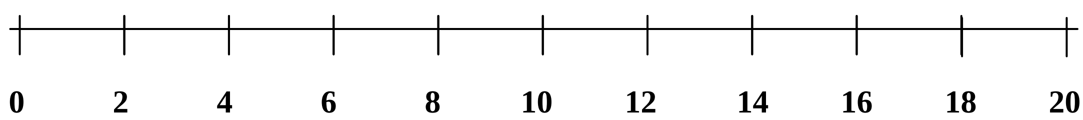
X = Wait Time at DMV	Frequency	Relative Frequency	CUMULATIVE Relative Frequency
12	4		
15	2		
18	6		
20	3		
24	5		
25	7		
27	6		
30	5		
32	6		
38	4		
45	2		

EXAMPLE 12 : Creating Box Plots using the “5 number summary” from 1–Var Stats

A class of 20 students had the following grades on a quiz during the 6th week of class

2 5 8 10 **12** **12** 12 14 14 **14** **15** 15 17 17 **17** **18** 20 20 20 20

Find the 5 number summary and draw a boxplot for the quiz grade data.
The box identifies the IQR. The lines (whiskers) extend to the minimum and maximum values.
Mark the median inside the box.



- The box shows where the middle 50% of the data values are located
- The IQR is represented by the length of the box.
  The left WHISKER shows where the lowest 25% of the data values are located
- The right WHISKER shows where the highest 25% of the data values are located

Boxplots are easy to do by hand once you have found the 5 number summary. If you want to learn how to create a boxplot on your calculator, refer to the technology section in the appendix of the textbook or to the online calculator handout instructions for your model of calculator.

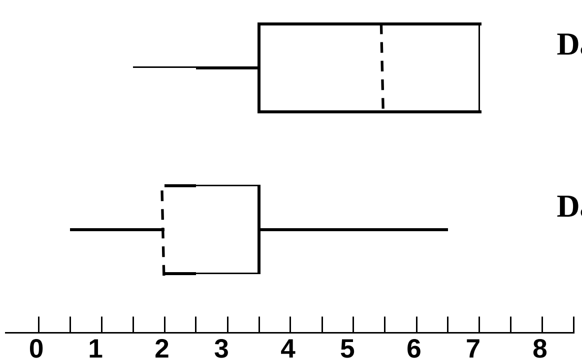
EXAMPLE 13: Find the 5 number summary and draw the boxplot

X	Frequency
3	40
5	25
6	11
7	3
10	2

EXAMPLE 14: Explain what is "strange" about each boxplot and what it means.

Data Set A

Data Set B



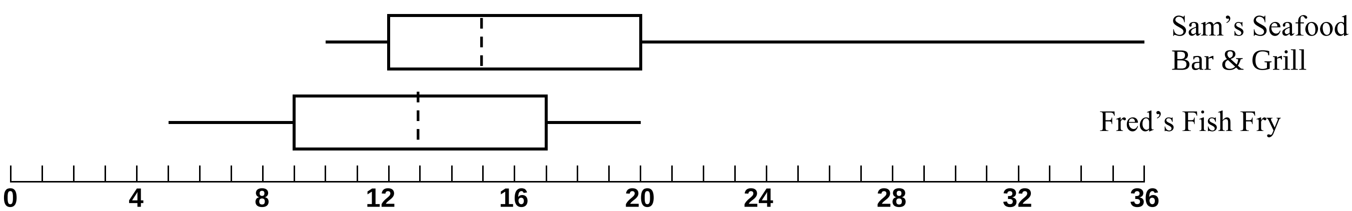
CHAPTER 2: INTERPRETING DATA BY USING BOXPLOTS

Using BOXPLOTS to compare two data sets

- We can compare which data set has higher or lower data values by comparing the location of the parts of the boxplot.
- We can compare spread by looking at the lengths of the whiskers compared to each other and as compared to the length of the box.

EXAMPLE 15: Interpreting Box Plots

The boxplots represent data for the amount a customer paid for his food and drink for random samples of customers in the last month at each of two restaurants



Find these values by reading the boxplot.

Sam's: Min _______ Q1 _______ Median _______ Q3 _______ Max _______ IQR________

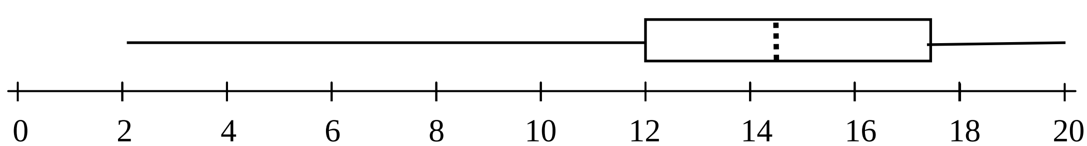
Fred's: Min _______ Q1 _______ Median _______ Q3 _______ Max _______ IQR________

Use the boxplots to compare the distributions of the data for the two restaurants. Look at the statistics for the center, quartiles, and extreme values, and the spread of the data. Discuss differences and/or similarities you see regarding the location of the data, the spread of the data, the shape of the data, and the existence of outliers.

EXAMPLE 16: Outliers and Boxplots: Graphical View; using quiz grade data from example 12.

2 5 8 10 **12** **12** 12 14 14 **14** **15** 15 17 17 **17** **18** 20 20 20 20

Outliers are data values that are unusually far away from the rest of the data.



The IQR is the length of the box; it measures the spread of the middle 50% of the data.
A data value is considered to be far enough away from the rest of the data to be an outlier if the distance between the data value and the closest end of the box is longer than 1½ times the length of the box

- The line from the box to the lowest data value is longer than 1½ times the length of the box. This indicates that there are data values at the low end of the data that are far away from the rest of the data. There are outliers at the low end of the data
- The line from the box to the highest data value is shorter than 1½ times the length of the box. This shows that there are not any outliers at the high end of the data.

CHAPTER 2: IDENTIFYING OUTLIERS USING QUARTILES & IQR

Outliers are data values that are unusually far away from the rest of the data.

We use values called "fences" as to decide if a data value is close to or far from the rest of the data. Any data values that are not between the fences (inclusive) are considered outliers.

Lower Fence: Q1 – 1.5*IQR **Upper Fence: Q3 + 1.5*IQR**

Outliers should be examined to determine if there is a problem (perhaps an error) in the data. Each situation involves individual judgment depending on the situation.

- If the outlier is due to an error that can not be corrected, or has properties that show it should not be part of the data set, it can be removed from the data.
- If the outlier is due to an error that can be corrected, the corrected data value should remain in the data.
- If the outlier is a valid data value for that data set, the outlier should be kept in the data set.

EXAMPLE 17: CALCULATING THE FENCES ; IDENTIFYING OUTLIERS

For a quiz, exam, or graded work, you must know be able to show your work doing the calculations to find the fences and explain your conclusion.

For the quiz grade data, find the lower and upper fences and identify any outliers.

2 5 8 10 **12** **12** 12 14 14 **14** **15** 15 17 17 **17** **18** 20 20 20 20

IQR =

Lower Fence: Q1 – 1.5(IQR) =

Upper Fence: Q3 + 1.5(IQR) =

Are there any outliers in the data? Justify your answer using the appropriate numerical test.

EXAMPLE 18: PRACTICE: CALCULATING THE FENCES ; IDENTIFYING OUTLIERS

The data show the lowest listed ticket prices in the San Jose Mercury News for 15 Bay Area concerts during one randomly selected week during a recent summer.

$33 $35 $35 $35 $35 $38 $40 $44 $45 $45 $45 $48 $54 $75 $89

Calculate the fences and identify all outliers. Clearly state your conclusion and show your work to justify it.

Technical Note: **In Math 10, we will find outliers by finding the fences using Q1, Q3 and IQR as above This method is usually considered appropriate for data sets of all shapes.**

There are many statistical methods of indentifying outliers or unusual values.
Different methods may be used in various situations and sometimes produce different results.
A statistics professor at UCLA wrote a 400+ page book about different methods of finding outliers!

CHAPTER 2: MEASURES OF CENTRAL TENDENCY (CENTER)

Mean = Average = $\frac{\text{sum of all data values}}{\text{number of data values}}$ Symbols: Sample Mean: $\overline{X}$ Population Mean μ

Median = Middle Value (if odd number of values)
OR Average of 2 middle values (if even number of values)

Mode = most frequent value

If data are not skew, the mean (average) is usually the most appropriate measure of center of the data.
If data are skew, the median is usually the most appropriate measure of center of the data.

EXAMPLE 19: The data show the lowest listed ticket prices in the San Jose Mercury News for 15 major Bay Area concerts during one randomly selected week during a recent summer. Consider this to be a sample of all concerts for that summer.

35 35 45 54 45 33 35 40 38 48 75 89 35 45 44

Ticket Price Data Sorted into Order

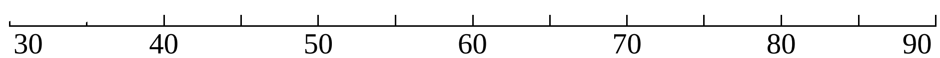
33 35 35 35 35 38 40 44 45 45 45 48 54 75 89

Find the mean

Find the median

Find the mode

Draw a dotplot of the data:



Which value should be used as the most appropriate measure of the center of this data?

The ___________ is the most appropriate measure of center because______________________________

EXAMPLE 20: Dawn's Diner has 10 employees who all worked on Friday last week.
The data show the number of hours that each employee at Dawn's Diner worked on Friday last week..

Data sorted into order **3 4.5 5 5 5 7 7 7.5 8 9 hours**

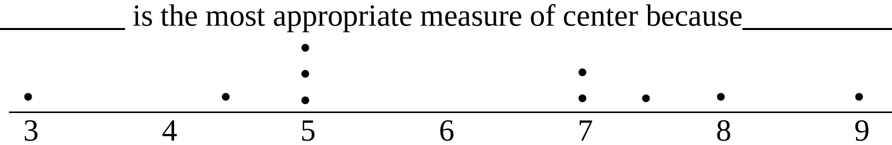
Find the mean

Find the median

Find the mode:____________________

Which value should be used as the most appropriate measure of the center of this data?

The ___________ is the most appropriate measure of center because______________________________



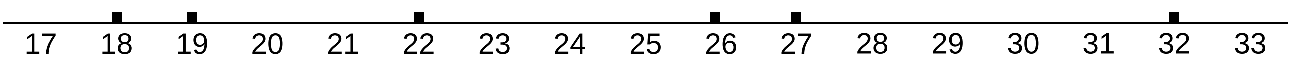
CHAPTER 2: MEASURES OF VARIATION (SPREAD)

EXAMPLE 21: Ages of students from two classes Random sample of 6 students from each class

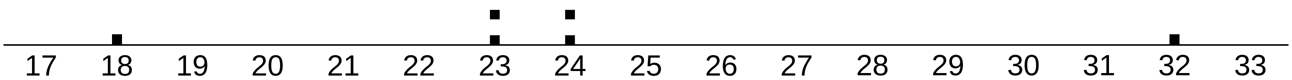
	Age Data						Mean	Range	Standard Deviation
Sample from Class 1	18	19	22	26	27	32	24	14	5.33
Sample from Class 2	18	23	23	24	24	32	24	14	4.52

Range = Maximum Value – Minimum Value = ______ –_____ =_____

DOTPLOT: Sample from Class 1



DOTPLOT: Sample from Class 2



Based on the dotplots, does one sample appear to have more variation than the other sample?___________

The **Standard Deviation** measures variation (spread) in the data by finding the distances (deviations) between each data value and the mean (average).

Sample from Class 1:			
x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
18	24		
19	24		
22	24		
26	24		
27	24		
32	24		
		$\sum_{all\ data}(x-\bar{x})^2 =$	
Sample Variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1} =$			
Sample Standard Deviation: $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} =$			

Sample from Class 2: **PRACTICE**			
x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
		$\sum_{all\ data}(x-\bar{x})^2 =$	
Sample Variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1} =$			
Sample Standard Deviation: $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} =$			

We will use the calculator or other technology to find the standard deviation.

If you need more practice to understand what the standard deviation represents, you can practice by finding the standard deviation for sample 2 at home.

CHAPTER 2: USING MEASURES OF VARIATION (SPREAD)

Use Standard Deviation as the most appropriate measure of variation	**SAMPLE STANDARD DEVIATION**	**POPULATION STANDARD DEVIATION**
	$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ n individuals in sample with mean $\bar{x}$ **If using sample data, use Sx from your calculator's 1VarStats**	$\sigma=\sqrt{\frac{\sum(x-\mu)^2}{N}}$ N individuals in population with mean μ **If using population data, use σx from your calculator's 1VarStats**

EXAMPLE 22: A class of 20 students has a quiz every week. All students in the class took the quizzes.

For the sixth week quiz, the grades are

2 5 8 10 12 12 12 14 14 14
15 15 17 17 17 18 20 20 20 20

x	Frequency
2	1
5	1
8	1
10	1
12	3
14	3
15	2
17	3
18	1
20	4

For the seventh week quiz, the grades are

1 8 8 12 13 13 13 14 14 14
14 14 15 15 17 17 18 18 18 20

x	Frequency
1	1
8	2
12	1
13	3
14	5
15	2
17	2
18	3
20	1

a. Use your calculator one variable statistics to find the mean, median and standard deviation for each quiz.
Which symbol is appropriate to use for the mean in this example: $\bar{x}$ or μ ? Why?
Which standard deviation is appropriate to use in this example: s or σ? Why?

6th week quiz: Mean ___ = _______ Standard Deviation ____ = ________ Variance ____ = _______

7th week quiz: Mean ___ = _______ Standard Deviation ____ = ________ Variance ____ = _______

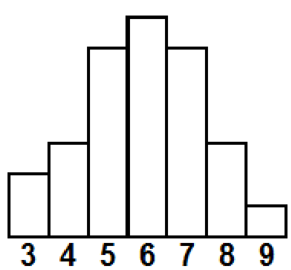
b. Which week's quiz exhibits more variation in the quiz grades? Justify your answer numerically.

c. Which week's quiz exhibits more consistency in the quiz grades? Justify your answer numerically

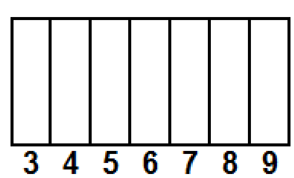
EXAMPLE 23:

Which graph represents data with the largest standard deviation?

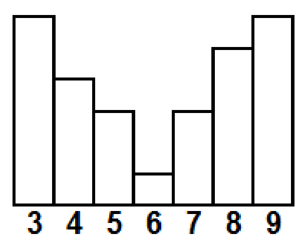
Which graph represents data with the smallest standard deviation?



Histogram A



Histogram B



Histogram C

CHAPTER 2: Z-SCORES (Measures of Relative Standing)

The "z-score" tells us how many standard deviations a data value is above or below the mean.
The "z-score" measures how far away a data value is from the mean, measured in "units" of standard deviations
It describes the location of a data value as "how many standard deviations above or below the mean"

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x-\mu}{\sigma} \quad or \quad \frac{x-\bar{x}}{s}$$

In our textbook this is sometimes noted as "**#of STDEVs**"

EXAMPLE 24: In the 6th week of class, the 20 students had the quiz grades below. Anya's quiz grade was 18.

2 5 8 10 12 12 **12** 14 14 14 15 15 17 17 17 **18** 20 20 20 20 **μ =14.1** **σ =** 4.89

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x-\mu}{\sigma} = \frac{18-14.1}{4.89} = \frac{3.9}{4.89} = 0.8$$

Anya's quiz grade was 3.9 *points* above average but it was 0.8 *standard deviations* above average.

Interpretation of Anya's z-score for the quiz:
Anya's quiz grade of 18 points is 0. 8 standard deviations above the average quiz grade of 14.1

EXAMPLE 25: In the 8th week of class, the 20 students had the exam grades below: Anya's exam grade was 90

44 52 56 59 **62** 65 70 71 72 74 74 75 77 79 84 85 **90** 91 94 100 μ = 73.7 σ = 14.25

Find and interpret Anya's z-score for the exam:

Did Anya perform better on the quiz or the exam when compared to the other students in her class?
Use the z-scores to explain and justify your answer.

EXAMPLE 26: In the same class as Anya, Beth's quiz grade was 12 points and her exam grade was 62 points.

Find and interpret Beth's z-score for the quiz.

Did Beth perform better on the quiz or the exam when compared to the other students in her class?
Use the z-scores to explain and justify your answer.

GUIDELINE: Writing a sentence interpreting a z-score in the context of the given data:

The (*description of variable*) of (*data value*) is |*z-score*| standard deviations (*above or below*) the average of (*value of the mean*)

Write absolute value of z (*drop the sign*)

Use
above if z score > 0
below if z score < 0

EXAMPLE 27: Z-scores for quiz grades on week 6 quiz for 4 students in the class:

Student	Anya	Beth	Carlos	Dan
Z-score			– 0.84	1.21

Based on the Z-scores, arrange the students quiz grades in order. Which is best? Which is worst?

_______________ _______________ _______________ _______________

EXAMPLE 28: Working Backwards from Z-score to Data Value

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x-\mu}{\sigma} \quad or \quad \frac{x-\bar{x}}{s}$$ can be solved for "x=":

A data value can be expressed as $\boxed{x = \text{mean} + (\text{z-score})(\text{standard deviation}) = \bar{x} + z\,s \quad or \quad \mu + z\,\sigma}$

For the week 6 quiz, $\mu = 14.1$ and $\sigma = 4.89$. Find the quiz scores for Carlos and Dan:

Carlos: z = – 0.84 x = ____

Dan: z = 1.21 x =____

Are high or low z-scores good or bad? It depends on the context of the problem.
Read the problem carefully. Think about the context and the meaning of the numbers for that problem.

Positive z-scores correspond to numbers that are larger than the average.
Higher than average is good for exam scores and salaries
Higher than average is bad for airline ticket costs or waiting time for a bus to arrive.
High z scores are good for race speeds (fast) but bad for race times (slow).
Negative z-scores correspond to numbers that are smaller than the average.
Lower than average is bad for exam scores and salaries.
Lower than average is good for airline ticket costs or waiting time for a bus to arrive.
Small z scores are bad for race speeds (slow) but good for race times (fast),
In some contexts, no value judgment applies; such as the number of children in a family

EXAMPLE 29: The air at an industrial site is tested for a sample of 30 days to measure the level of two pollutants: A and B. (A and B are measured in different units, have different "safe" levels, and different effects on public health, so are not directly comparable.)

Suppose that for today's pollution readings:

The level of pollutant A is 0.5 standard deviations below its average level: z = ______

The level of pollutant B is 0.8 standard deviations below its average level: z = ______

a. Compare today's pollution levels for A and B to the average readings for the 30 day sample at this site. Which of today's pollutant levels would be considered better for this site? Explain.

Today the level for pollutant _____ is better because

b *Practice: Working Backwards:* Suppose that the sample averages and standard deviations are
Pollutant A: $\bar{x}$ = 47 parts per billion, s = 4 Pollutant B: $\bar{x}$ = 10 micrograms per m^3, s = 1.5 ;
Find the actual levels for pollutants A and B.

(Note: Data underlying this example: h*ttp://www.epa.gov/air/criteria.html* The National Ambient Air Quality Standards , specify average "safe levels" that must be maintained in order to protect public health for various pollutants:
A: Nitrogen Dioxide NO_2 : 53 parts per billion ; B: Particulate Matter $PM_{2.5}$: 15 micrograms per m^3.)

CHAPTER 2: EMPIRICAL RULE for Mound Shaped Symmetric (Bell Shaped) Data

If the data are mound shaped and symmetric (bell shaped), then most of the data lie within two standard deviations away from the mean. Almost all the data lies within three standard deviations from the mean.

68% of the data is within ± 1 standard deviations of the mean
95% of the data is within ± 2 standard deviations of the mean
99% of the data is within ± 3 standard deviations of the mean

This provides another method for identifying unusual data values IF the data is known to be mound shaped and symmetric. Finding values further than 2 or 3 standard deviations from the mean is appropriate for data that is mound shaped and symmetric but may not be appropriate for skewed data
.
We will continue to use the outlier test we learned earlier using the fences because it is appropriate for data distributions of all shapes, including but not limited to skewed data.

EXAMPLE 30:

A food processing plant fills cereal into boxes that are labeled to contain 20 ounces of cereal.
The distribution of the amount of cereal per box is mound shaped and symmetric.

A machine fills boxes with an average of 20.6 ounces of cereal and a standard deviation is 0.2 ounces.

For quality assurance, the food processing plant manager needs to monitor how much cereal the boxes actually contain; each day a sample of randomly selected of boxes of cereal are weighed.

a. Approximately what percent of the boxes are filled with between 20.2 ounces and 21 ounces of cereal?

b. What value is 3 standard deviations below average? Why might the manager be concerned if there are boxes of cereal with weight less than 3 standard deviations below average?

c. What value is 3 standard deviations above average? Why might the manager be concerned if there are boxes of cereal weighing more than 3 standard deviations above average?