NUMERICAL AND GRAPHICAL SUMMARIES OF QUANTITATIVE DATA: FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

Numerical data may be presented individually (ungrouped) or grouped into intervals
The frequency distribution table summarizes the data.

EXAMPLE 1: Individual Data Values *(ungrouped)*

Number of flowers on a plant, for a sample of 16 plants in a lab experiment: 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			
2			
3			
4			
5			
6			
7			

a. What percent of plants had 3 flowers?

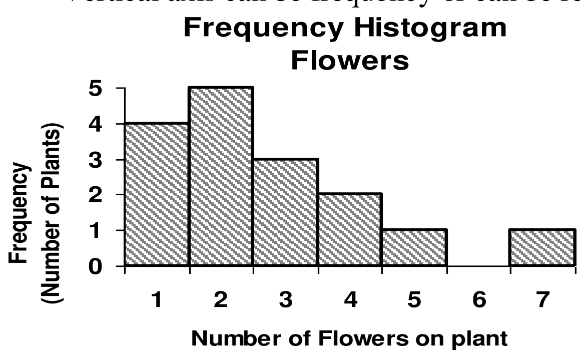
b. What percent of plants had <u>at most</u> 3 flowers?

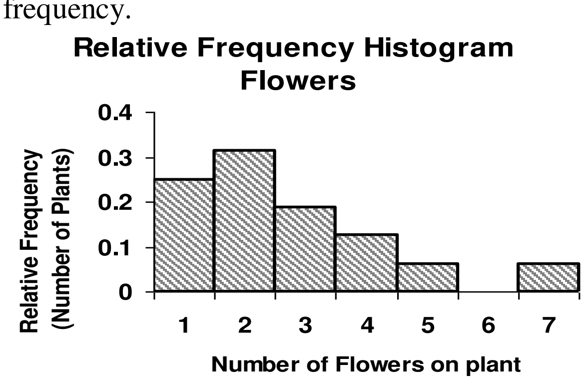
c. What percent of plants had <u>more than</u> 3 flowers?

d. What percent of plants had <u>at least</u> 5 flowers?

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 2: Life Expectancy at Birth In Years: 227 countries - Data is grouped into intervals

from U.S. Bureau of the Census 2005 International Data Base,	Interval Class Limits	Interval Class Boundaries	Frequency	Relative Frequency	Cumulative Relative Frequency
	30−39	29.5 to 39.5	6	6/227 = 0.026	0.026
	40−49	39.5 to 49.5	25	25/227 = 0.110	0.137
	50−59	49.5 to 59.5	19	19/227 = 0.084	0.220
	60−69	59.5 to 69.5	38	38/227 = 0.167	0.388
	70−-79	69.5 to 79.5	120	120/227 = 0.529	0.916
	80−-89	79.5 to 89.5	19	18/227 = 0.084	1.000

***Note: In Math 10*, we will use intervals of EQUAL WIDTH**

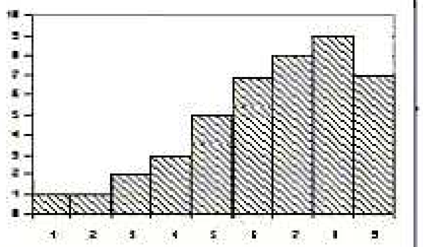
EXAMPLE 3: Ages of people:

Intervals varying in width (often used by U.S. Census Dep't.)
0-5, 6-14, 15-19, 20-24, 25-29, 30-39, 40-49, 50-59, 60-64, 65-79, 80+

Intervals of EQUAL WIDTH:
0-9, 10-19, 20-29, 30-39, 40-49, . . . , 80-89, 90-99

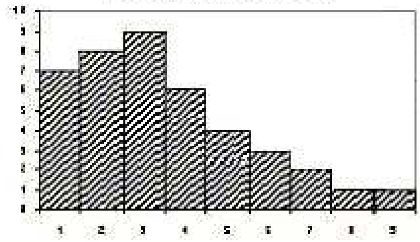
Shapes of Data Distributions

Skewed to the LEFT
(negatively skewed)



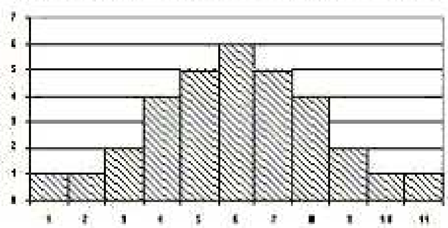
When data is skewed to the left, generally the mean is less than the median

Skewed to the RIGHT
(positively skewed)



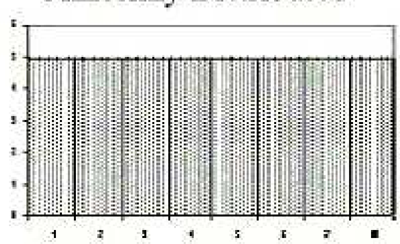
When data is skewed to the right, generally the mean is greater than the median

Mound Shaped & Symmetric

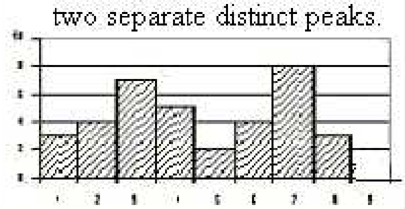


For symmetric data, mean = median

Uniformly Distributed

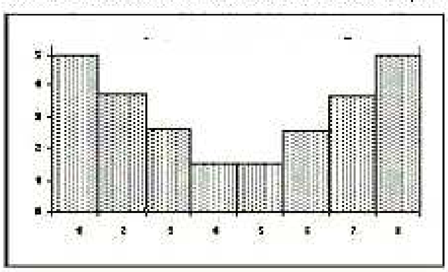


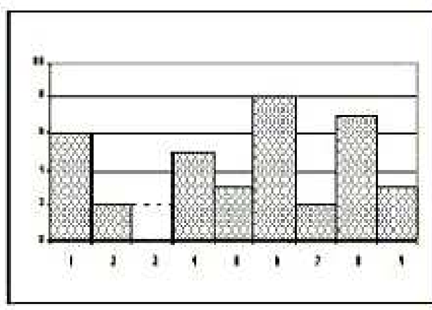
Bimodal
two separate distinct peaks.

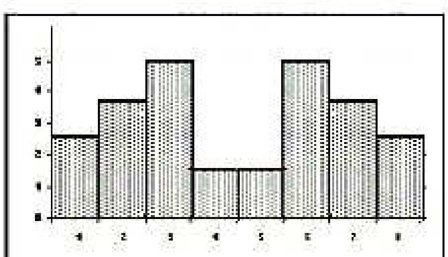


"hills" separated by a "valley"
Peaks do not need to be exactly the same height

Below are two additional examples of symmetric data.
Note that these are not moundshaped:







If data do not fit one of these descriptive terms, do not use a term that doesn't fit its shape.

Just describe what you see in the data if none of these descriptive terms apply.

Definitions and Calculator Instructions

- **Class Limits: Lowest and highest possible data values in an interval.**
- **Class Boundaries: 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 Width: Difference between two consecutive class boundaries**
  Can also calculate as difference between two consecutive lower class limits

EXAMPLE 3A: Age interval 30-39: 30 is the lower class limit 39 is the upper class limit
Class boundaries are 29.5 to 39.5

Age interval 40-49: 40 is the lower class limit 49 is the upper class limit
Class boundaries are 39.5 to 49.5

Class Width is 39.5 – 29.5 = 49.5 – 39.5 = 10

- **Class Midpoints: Midpoint of a class = (lower limit + upper limit) / 2**

EXAMPLE 3B: Age interval 30-39: class midpoint is (30 + 39)/2 = 34.5

- **Frequency = count = number of data values that lie in the interval**
  A **frequency distribution** counts the **number** of data items that fall into each interval.
- **Relative Frequency = proportion of data values that lie in the interval =** $\dfrac{\textbf{Frequency}}{\textbf{Number of Observations}}$
  A **relative frequency distribution** shows the **proportion** (fraction or percent) of data items in each interval.
- **Cumulative Relative Frequency**
  = sum of relative frequencies for all intervals up to and including current interval

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 after 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 **Xscl** = 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.

For TI-83, 84 Instructions for 1 variable statistics, see page 9 of notes.

EXAMPLE 4:
Student Total Headcount

Bay Area Community College Enrollment Fall 2014

27 Community Colleges comprising Regions III and IV of all CA community colleges (Bay and Interior Bay regions)

Note that the data has already been sorted into ascending numerical order.

http://datamart.cccco.edu/Students/Student_Term_Annual_Count.aspx

Community College Campus	Enrollment
Alameda	5461
Merritt	6085
Gavilan	6298
Berkeley City	6312
Canada	6315
Marin	6418
Contra Costa	6892
Las Positas	8364
Monterey	8464
Los Medanos	8689
Mission	8793
San Jose City	8906
San Mateo	8922
Evergreen Valley	8953
Hartnell	9624
Skyline	9690
West Valley	10174
Laney	10747
Ohlone	11065
Chabot Hayward	13177
Cabrillo	13444
Foothill	14924
Diablo Valley	19812
Deanza	22715
San Francisco Ctrs	23159
San Francisco	23575
Santa Rosa	26288

4a. When there are an odd number of data values, the median is the middle data value. The middle value of 27 values is the 14^{th} data value. Find the median enrollment:_______________

The average enrollment is (5461 + 6085 + 6298 + . . . +26288)/27 = 11602 students.
This is the “arithmetic average” and is also called the “mean”:

4b. Create a frequency/relative frequency/cumulative relative frequency table

Interval (Class Limits)	Class Boundaries	Frequency	Relative Frequency	Cumulative Relative Frequency
5000-9999				
10000-14999				
15000-19999				
20000-24999				
25000-29999				

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. See calculator instructions for histogram om page 3 if needed.

GRAPHING PRACTICE: DO THIS PAGE AT HOME FOR PRACTICE.

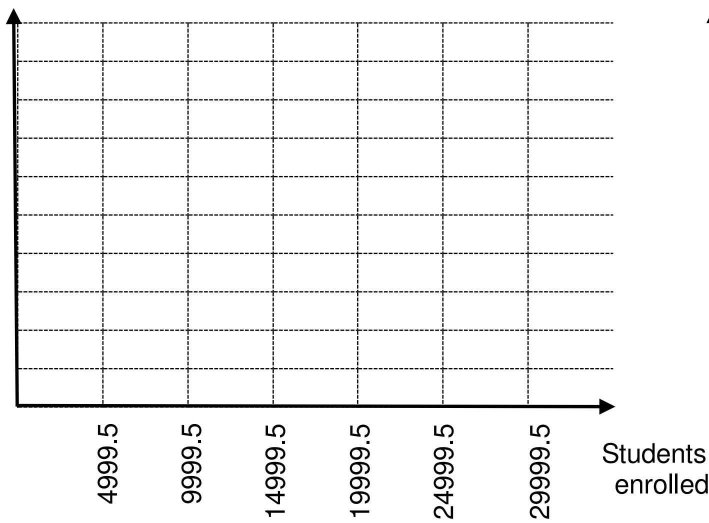
Draw the histograms by hand to be sure you understand how the calculator builds a histogram from the frequency table.

The frequency histogram should match the histogram we created in class on the calculator.

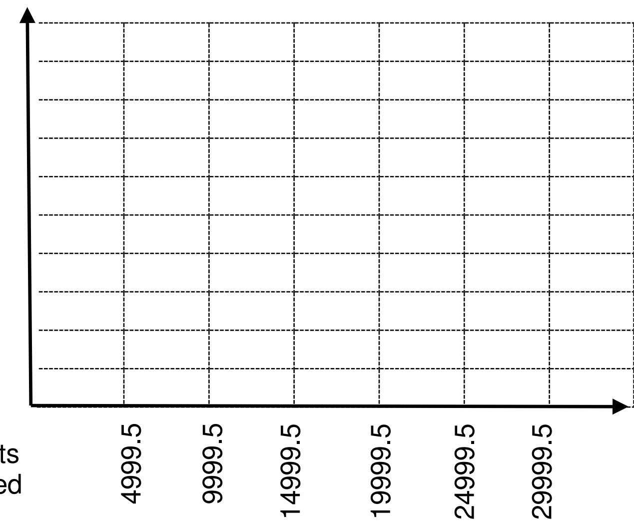
A HISTOGRAM is a bar graph displaying quantitative (numerical) data

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.

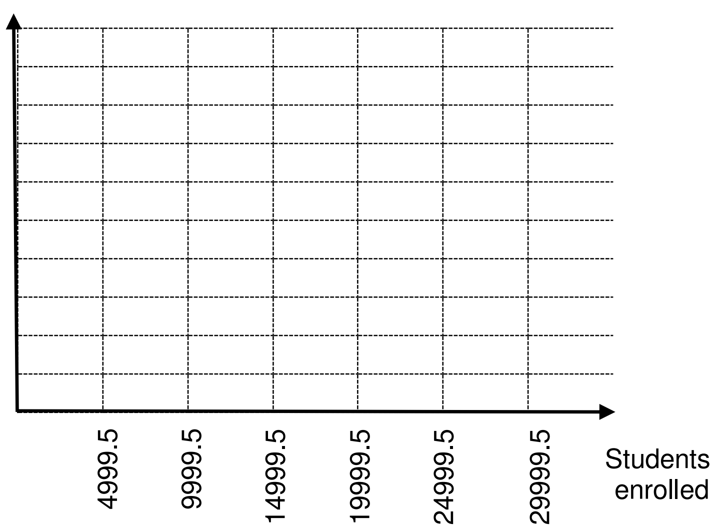
4c. Draw a **frequency histogram**.
Label and scale the vertical axis using 0, 2, 4, 6, 8, ...



4d. Draw a **relative frequency histogram**
Label and scale the vertical axis using 0, 0.1, 0.2, . . .



4e. ***OPTIONAL:*** The textbook also shows a graph called a **frequency polygon**.
You can draw one here if you want to see how it compares to the histogram.
At the midpoint of each interval draw a dot at the height of the frequency.
An interval with no data gets at dot at a height of 0 frequency at the midpoint of the interval.
Use a ruler to connect the dots. Label and scale the vertical axis using frequencies 0, 2, 4, 6, 8, ...



GRAPHICAL DISPLAYS OF QUANTITATIVE DATA: STEM AND LEAF PLOTS

Each data value is split into a stem and leaf using place value.
A key indicating the place value representation by the stem and leaf should be shown.

EXAMPLE 5:

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:

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**

PRACTICE Do at home if not done in class:
EXAMPLE 6

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 7: 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 8: 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?

DESCRIPTIVE STATISTICS:
MEASURES OF RELATIVE STANDING: PERCENTILES & QUARTILES

The **P^th percentile** is 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 9: ***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

$$\underbrace{2\quad 5\quad 8\quad 10\quad 12\quad 12\quad 12\quad 14}\quad \underbrace{14\quad 14\quad 15\quad 15\quad 17\quad 17\quad 17\quad 18\quad 20\quad 20\quad 20\quad 20}$$

$P_{40} = 14$

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 Third Quartile Q3

Your calculator can find these special percentiles using 1-variable statistics(Q1, Med, Q3).

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

c. Find the Interquartile Range Q1 = ______ Q3 = ______ IQR = __________________

<u>Finding summary statistics on your TI-83,84 calculator</u>

Enter data into the statistics list editor: [STAT] "EDIT" press enter

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 have entered data, checked it and corrected errors.

One Variable Summary Statistics: [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

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.

One Variable Summary Statistics: [STAT] "CALC" [1] *for 1 – Var Stats* [2nd] [L1] [,] [2nd] [L2] [ENTER]

order of lists should be data value list, frequency list

PRACTICE Chapter 2 Interpreting Percentiles, Quartiles and Median:

Read this subsection near the end of Section 2.3 in the textbook,*starting <u>after</u> Try-It problem 2.18)*

It provides practice understanding percentiles and guidelines to writing interpretations.

Read and do Examples and Try-It problems (2.19 through 2.22) to practice this important skill. You will be asked to write sentences interpreting percentiles, medians or quartiles on an exam, quiz or lab.

Estimating Percentiles From Cumulative Relative Frequency

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

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

x	Frequency	Relative Frequency	Cumulative Relative Frequency
2	1	1/20 =0.05	0.05
5	1	0.05	0.10
8	1	2/20 = 0.10	0.15
10	1	0.10	0.20
12	3	0.15	0.35
14	3	3/20 = 0.15	0.50
15	2	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 p^{th} 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 p^{th} percentile is the data value (x) column at the first line in the table BEYOND the value of p
  Find the 40^{th} 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 40^{th} percentile is the x value for the line at which you first pass 0.40.
  The 40^{th} percentile is 14
  TRY IT! Use the table to find the first quartiles.

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

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 (When 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 50^{th} 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 50^{th} percentile.
If you did not average, but used the x value for the line showing 0.50, you would use 14 as the median which is not correct.

NOTE: *We'll use the method above to find percentiles in Math 10.*
There are other methods that are also sometimes used to find percentiles.
Example 2.17 in the textbook chapter 2 shows how to use the positional formula (p/100)(n+1)
Different statistical software programs or calculators sometimes use slightly different methods and may obtain slightly different answers.

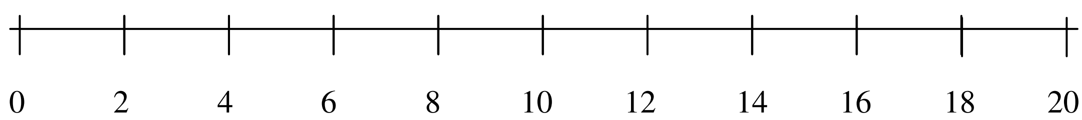
GRAPHICAL REPRESENTATION OF DATA: BOXPLOTS

EXAMPLE 11 : ***Creating Box Plots using the "5 number summary" from 1–Var Stats on your calculator***

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.

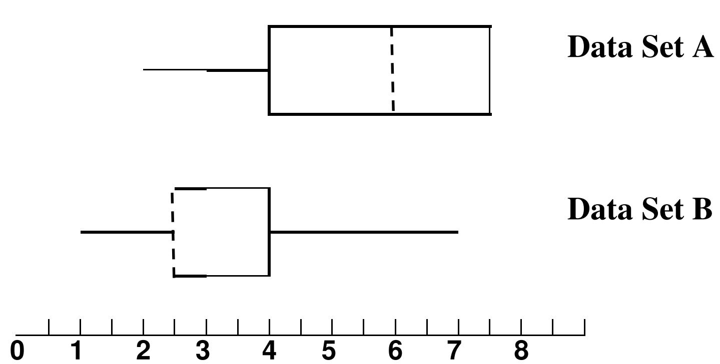


Boxplots are easy to do by hand once you have found the 5 number summary. If you want to learn how to create a boxplot on your calculator, refer to the technology section in the appendix of the textbook or to the online calculator handout instructions for your model of calculator.

EXAMPLE 12: Find the 5 number summary and draw the boxplot

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

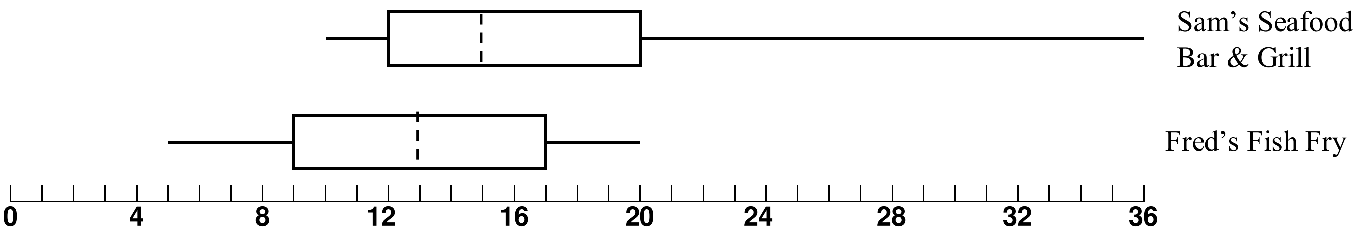
EXAMPLE 13: What is strange about these boxplots?
Explain what is "strange" and what it means in each boxplot



GRAPHICAL REPRESENTATION OF DATA: BOXPLOTS

EXAMPLE 14: ***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________

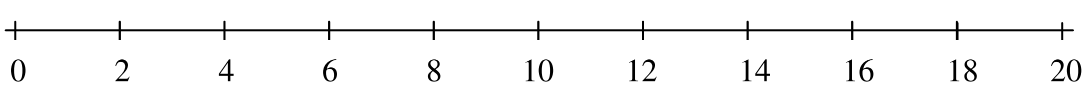
Fred's: Min _______ Q1 _______ Median _______ Q3 _______ Max _______ IQR________

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 15 ***(optional):***

Sometimes you may see a boxplot in which the whiskers (lines) are extended only until the lower and upper fences and any data that is more extreme than the fences are indicated by a dot • (or ∗, + or ∘).

It is more complicated to construct, but has the advantage that outliers are easily identified visually.



DESCRIPTIVE STATISTICS: Identifying Outliers Using Quartiles & IQR

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

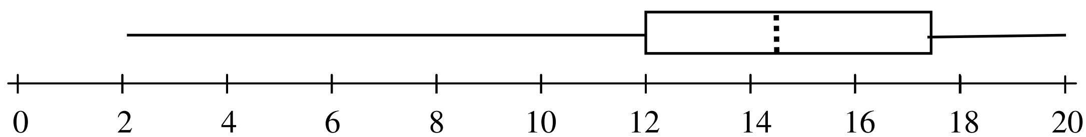
We use values called "fences" as to decide if a data value is close to or far from the rest of the data.
Any data values that are not between the fences (inclusive) are considered outliers.

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.

OUTLIER AND BOXPLOTS: Graphical View:



The IQR is the length of the box, the measures the spread of the middle 50% of the data.

- The line from the box to the lowest data value is longer than 1½ times the length of the box. This indicates that there <u>are</u> 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 <u>not</u> any outliers at the high end of the data.

OUTLIERS: Calculating the Fences and 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.

EXAMPLE 16: 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.

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

NOTE: ***There are many statistical methods of indentifying outliers or unusual values.***
The different methods sometimes produce different results.

For mound-shaped and symmetric data*, statisticians may flag outliers by finding values that are further than 2 (or further than 3) standard deviations away from the mean. This method is not generally appropriate for data distributions with other shapes. This method is based on the "Empirical Rule" and the "Normal Probability Distribution" that we will study later in this course.*

Chemistry students often learn another method called a "Q-test".

A statistics professor at UCLA wrote a 400+ page book about different methods of finding outliers!

DESCRIPTIVE STATISTICS: MEASURES OF CENTRAL TENDENCY (CENTER)

Mean = Average = $\dfrac{\text{sum of all data values}}{\text{number of data values}}$ 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

EXAMPLE 17: The table shows 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

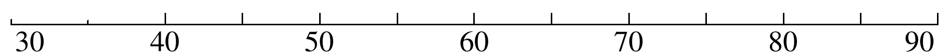
33	35	35	35	35	38	40	44	45	45	45	48	54	75	89

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 18:

CSU Enrollment for Fall 2009 : These data are for all 22 CSU "non-specialized" campuses.

Find the **mean** (average) number of students

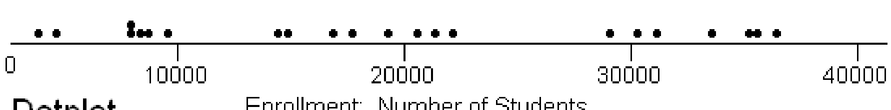
Find the **median** number of students

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____

CSU Campus	2009 Enrollment
Channel Islands	3,862
Monterey Bay	4,688
Humboldt	7,954
Bakersfield	8,003
Sonoma	8,546
Stanislaus	8,586
San Marcos	9,767
Dominguez Hills	14,477
East Bay	14,749
Chico	16,934
San Bernardino	17,852
San Luis Obispo	19,325
Los Angeles	20,619
Fresno	21,500
Pomona	22,273
Sacramento	29,241
San Francisco	30,469
San Jose	31,280
San Diego	33,790
Northridge	35,198
Long Beach	35,557
Fullerton	36,262



Dotplot

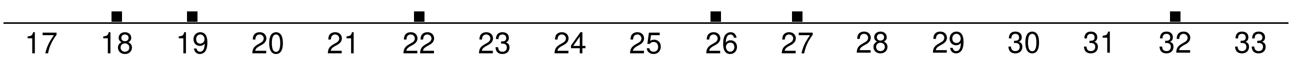
DESCRIPTIVE STATISTICS: MEASURES OF VARIATION (SPREAD)

EXAMPLE 19: Ages of students from two classes Random sample of 6 students from each class

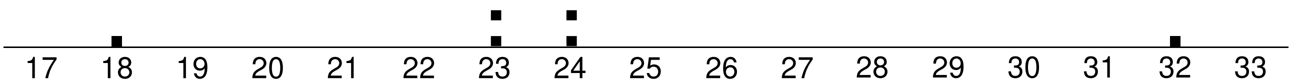
	Age Data						Mean	Range	Standard Deviation
Sample from Class 1	18	19	22	26	27	32	24	14	5.33
Sample from Class 2	18	23	23	24	24	32	24	14	4.52

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.

DESCRIPTIVE STATISTICS: MEASURES OF VARIATION (SPREAD)

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

EXAMPLE 20: 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: $\overline{x}$ or μ ? Why?
Which standard deviation is appropriate to use in this example: s or σ? Why?

6^{th} week quiz: Mean ___ = ______ Median = ______ Standard Deviation ____ = _______

7^{th} week quiz: Mean ___ = ______ Median = ______ Standard Deviation ____ = _______

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

d Find the variance for each week's quiz grades:

6^{th} week quiz: _________________________ 7^{th} week quiz: ___________________________

DESCRIPTIVE STATISTICS: Measures of Relative Standing: Z-SCORES

"z-score" tells us 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"

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

In our textbook this is sometimes noted as #of STDEVs

EXAMPLE 21: 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 22: 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 23: In the same class as Anya, Bob's quiz grade was 12 points and his exam grade was 62 points.

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

Did Bob perform better on the quiz or the exam when compared to the other students in his 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

Z-Scores Continued

EXAMPLE 24: Z-scores for quiz grades on week 6 quiz for 4 students in the class:

Student	Anya	Bob	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 25: 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 26: 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.)

The outlier test we learned earlier using the fences is appropriate for data distributions of all shapes, including but not limited to skewed data.

If data are mound shaped and symmetric, statisticians may use the values that are two or three standard deviations away from the mean as guidelines for data values that are "extreme".

Empirical Rule.

If the data are mound shaped and symmetric (bell shaped), then approximately

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

EXAMPLE 27:
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?

d. 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?