



SOUTH CAROLINA
STATE DEPARTMENT
OF EDUCATION

STATISTICS UNIT

for Algebra I

By: Wendy Tumolo

Adapted from a lesson by Lisa Donnahoo

July 2012

COMMON CORE STATE STANDARDS

- S.ID.1** - Represent data with plots on the real number line (dot plots, histograms, and box plots).
- S.ID.2** - Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread "*variability*" (interquartile range, standard deviation) of two or more different data sets.
- S.ID.3** - Interpret differences in shape (uniform, skewed, symmetrical), center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

VOCABULARY

population - the entire collection of individuals or objects about which information is desired

sample - a subset (*small group*) of the population, must accurately represent the population

When to choose a population/sample for your study?

- approval rating of the president
- national average for price of gas
- salaries of workers at a specific company

MEASURES OF CENTER

mean - most used measure of center, "equal portions"

- sample average " \bar{x} " *x-bar*

- population average " μ " *mu*

**affected by outliers - "not resistant to outliers"*

median - middle number of ordered data, for an even number of data - the mean of the 2 middle numbers, "ascending order - $1/2$ the data above and $1/2$ below"

**not affected by outliers - "resistant to outliers"*

When to use the mean/median for your study?

- income

- home values

- test scores

MEASURES OF CENTER

mean

Find the mean for the following data:

Ages (in years) of Little League soccer players

10, 9, 10, 11, 8, 15, 9, 7, 8, 6, 12, 10

By hand

Calculator

Answer in the context of the problem:

The average _____ of _____ is
approximately _____.

MEASURES OF CENTER

mean

Find the mean for the following data:

Ages (in years) of Little League soccer players

10, 9, 10, 11, 8, 15, 9, 7, 8, 6, 12, 10

By hand $\frac{10 + 9 + 10 + 11 + 8 + 15 + 9 + 7 + 8 + 6 + 12 + 10}{12} \approx 9.5833$

Calculator 2ND MEM ClrAllLists ENTER
STAT Edit *enter data in list* 2ND QUIT
STAT CALC 1-Var Stats 2ND L1 ENTER
 $\bar{x} = 9.5833$

Answer in the context of the problem:

The average age of the Little League soccer players is approximately 9.6 years old.

MEASURES OF CENTER

median - *odd number of data*

Find the median for the following data:

Racecars top speeds (in mph) 180, 201, 220, 191, 219, 209, 186

By hand

Calculator

Answer:

**One-half of the _____ were _____ than
_____ and one-half were _____.**

MEASURES OF CENTER

median - *odd number of data*

Find the median for the following data:

Racecars top speeds (in mph) 180, 201, 220, 191, 219, 209, 186

By hand Put data in order 180 186 191 **201** 209 219 220
Find the middle piece of data *there should be an equal number of data above and below the median

Calculator 2ND MEM ClrAllLists ENTER
STAT Edit enter data in list 2ND QUIT
STAT CALC 1-Var Stats 2ND L1 ENTER
Med = 201

Answer:

One-half of the racecars top speeds were faster than 201 mph and one-half were slower.

MEASURES OF CENTER

median - *even number of data*

Find the median for the following data:

684, 764, 656, 702, 855, 1133, 1132, 1303 **have students give context*

By hand

Calculator

Answer:

**One-half of the _____ were _____ than
_____ and one-half were _____.**

MEASURES OF CENTER

median - *even number of data*

Find the median for the following data:

684, 764, 656, 702, 855, 1133, 1132, 1303 **have students give context*

By hand

Put data in order 656 684 702 **764** | **855** 1132 1133 1303

Find the mean of the two middle pieces of data

$$(764 + 855) \div 2 = 809.5$$

Calculator

2ND MEM ClrAllLists ENTER

STAT Edit *enter data in list* **2ND QUIT**

STAT CALC 1-Var Stats 2ND L1 ENTER

Med = 809.5

Answer:

One-half of the _____ **were** _____ **than**
_____ **and one-half were** _____.

MEASURES OF VARIABILITY

variance - average of the squares of the distances each value is from the mean

" σ^2 " population *sigma squared*

" s^2 " sample *s squared*

standard deviation - square root of the variance

" σ " population " s " population

*used most often - returns the variance to the original units of measure of the data set

*large measures of variability (spread) data is spread out from the mean, small measures of variability (spread) data is close to the mean

MEASURES OF VARIABILITY

variance

Find the variance for the following data:

Weights of dogs at a dog park 10, 60, 50, 30, 40, 20 mean =

By hand

x	x - mean	(x - mean) ²

$\Sigma =$ _____

For a population: $\sigma^2 = \Sigma / \text{number of data pieces}$

**note: for a sample: $s^2 = \Sigma / (\text{number of data pieces} - 1)$*

Calculator

MEASURES OF VARIABILITY

variance

Find the variance for the following data:

Weights of dogs at a dog park 10, 60, 50, 30, 40, 20 mean = 35

By hand

x	x - mean	(x - mean) ²
10	10 - 35 = -25	(-25) ² = 625
60	60 - 35 = 25	(25) ² = 625
50	15	225
30	-5	25
40	5	25
20	-15	225

***adds up to 0 if
you didn't round*

$$\Sigma = \underline{1750}$$

For a population: $\sigma^2 = \Sigma / \text{number of data pieces} = 1750 / 6 = 291.667$

*note: for a sample: $s^2 = \Sigma / (\text{number of data pieces} - 1) = 1750 / 5 = 350$

Calculator **STAT** **Edit** *enter data in list* **2ND** **QUIT**
STAT **CALC** **1-Var Stats** **2ND** **L1** **ENTER**
 $\sigma x^2 = 17.078^2$ (population) $sx^2 = 18.708^2$ (sample)

MEASURES OF VARIABILITY

standard deviation

Find the standard deviation for the following data:

Weights of dogs at a dog park 10, 60, 50, 30, 40, 20

By hand

For a population: $\sigma = \sqrt{\sigma^2}$

**note: for a sample: $s = \sqrt{s^2}$*

Calculator

Answer: **add and subtract the standard deviation from the mean to get a 'typical' range*

The typical _____ of _____ is
between _____ and _____.

MEASURES OF VARIABILITY

standard deviation

Find the standard deviation for the following data:

Weights of dogs at a dog park 10, 60, 50, 30, 40, 20

By hand

For a population: $\sigma = \sqrt{\sigma^2} = \sqrt{291.667} = 17.078$

*note: for a sample: $s = \sqrt{s^2} = \sqrt{350} = 18.708$

Calculator **STAT** **Edit** *enter data in list* **2ND** **QUIT**
STAT **CALC** **1-Var Stats** **2ND** **L1** **ENTER**
 $\sigma x = 17.078$ (*population*) **$s x = 18.708$** (*sample*)

Answer: **add and subtract the standard deviation from the mean to get a 'typical' range*

The typical weight of a dog at the dog park is between 18 and 52 pounds.

MEASURES OF VARIABILITY

quartiles - divides the data into 4 equal sized groups

Q₁ lower quartile - median of the 1st half of the data

Q₃ upper quartile - median of the 2nd half of the data

Find Q1 and Q3 for the following data (*even number of data*):

18, 15, 12, 6, 8, 2, 3, 5, 20, 10 **have students give context*

By hand *Put data in ascending order* ◊

*Split the data into 2 equal portions - **Median***

COVER UP THE MEDIAN

*Find the median of the lower $\frac{1}{2}$ of the data - **Q₁***

*Find the median of the upper $\frac{1}{2}$ of the data - **Q₃***

Calculator

MEASURES OF VARIABILITY

quartiles - divides the data into 4 equal sized groups

Q_1 lower quartile - median of the 1st half of the data

Q_3 upper quartile - median of the 2nd half of the data

Find Q_1 and Q_3 for the following data (*even number of data*):

18, 15, 12, 6, 8, 2, 3, 5, 20, 10 **have students give context*

By hand *Put data in ascending order* 2 3 5 6 8 | 10 12 15 18 20

Split the data into 2 equal portions - Median 9

COVER UP THE MEDIAN

Find the median of the lower $1/2$ of the data - Q_1 5

Find the median of the upper $1/2$ of the data - Q_3 15

Calculator **STAT** **Edit** **enter data in list** **2ND** **QUIT**

STAT **CALC** **1-Var Stats** **2ND** **L1** **ENTER**

$Q_1 = 5$ **$Q_3 = 15$**

MEASURES OF VARIABILITY

quartiles - divides the data into 4 equal sized groups

Q₁ lower quartile - median of the 1st half of the data

Q₃ upper quartile - median of the 2nd half of the data

Find Q1 and Q3 for the following data (*odd number of data*):

18, 15, 14, 20, 25, 28, 22, 14, 8 **have students give context*

By hand *Put data in ascending order*

*Split the data into 2 equal portions - **Median***

COVER UP THE MEDIAN

*Find the median of the lower $\frac{1}{2}$ of the data - **Q₁***

*Find the median of the upper $\frac{1}{2}$ of the data - **Q₃***

Calculator

MEASURES OF VARIABILITY

quartiles - divides the data into 4 equal sized groups

Q_1 lower quartile - median of the 1st half of the data

Q_3 upper quartile - median of the 2nd half of the data

Find Q_1 and Q_3 for the following data (*odd number of data*):

18, 15, 14, 20, 25, 28, 22, 14, 8 **have students give context*

By hand *Put data in ascending order* **8 14 14 15 18 20 22 25 28**

Split the data into 2 equal portions - Median **18**

COVER UP THE MEDIAN

Find the median of the lower $1/2$ of the data - Q_1 **14**

Find the median of the upper $1/2$ of the data - Q_3 **23.5** **don't round or it won't be in the 'middle'*

Calculator **STAT Edit enter data in list 2ND QUIT**
STAT CALC 1-Var Stats 2ND L1 ENTER
Q1 = 14 Q3 = 23.5

MEASURES OF VARIABILITY

interquartile range (IQR) - distance from Q_1 to Q_3

***holds the middle half (50%) of the data**

Formula $IQR = Q_3 - Q_1$

Find the IQR for the following data (*already found Q_1 and Q_3*):

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

The middle half of the _____ varies by at most _____.

Find the IQR for the following data (*already found Q_1 and Q_3*):

18, 15, 14, 20, 25, 28, 22, 14, 8

The middle half of the _____ varies by at most _____.

MEASURES OF VARIABILITY

interquartile range (IQR) - distance from Q_1 to Q_3

***holds the middle half (50%) of the data**

Formula $IQR = Q_3 - Q_1$

Find the IQR for the following data (*already found Q_1 and Q_3*):

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

$$IQR = Q_3 - Q_1$$

$$IQR = 15 - 5 = 10$$

The middle half of the _____ varies by at most _____.

Find the IQR for the following data (*already found Q_1 and Q_3*):

18, 15, 14, 20, 25, 28, 22, 14, 8

$$IQR = Q_3 - Q_1$$

$$IQR = 23.5 - 14 = 9.5$$

The middle half of the _____ varies by at most _____.

1-VARIABLE STATISTICS

Calculator

- 1.) clear lists 2ND MEM ClrAllLists ENTER
- 2.) enter data in list STAT EDIT Edit ENTER
- 3.) calculate 1-variable statistics STAT CALC 1-Var Stats ENTER

\bar{x} = mean

Σx = sum of the data

Σx^2 = sum of the data squared

S_x = sample standard deviation

σ_x = population standard deviation

n = number of data pieces

$\min X$ = minimum data value

Q_1 = median of lower half of data

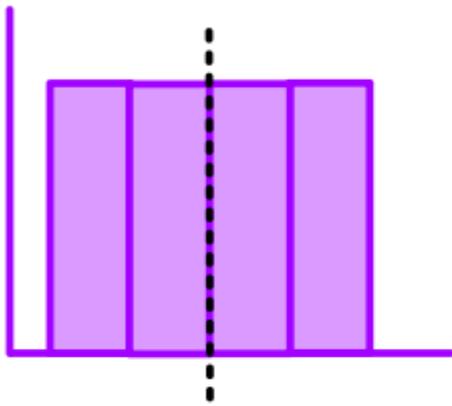
Med = median

Q_3 = median of upper half of data

$\max X$ = maximum data value

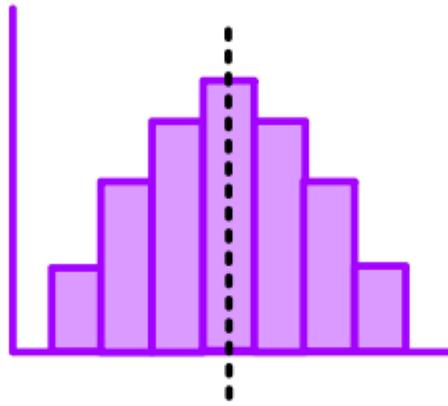
SHAPES OF GRAPHS

uniform



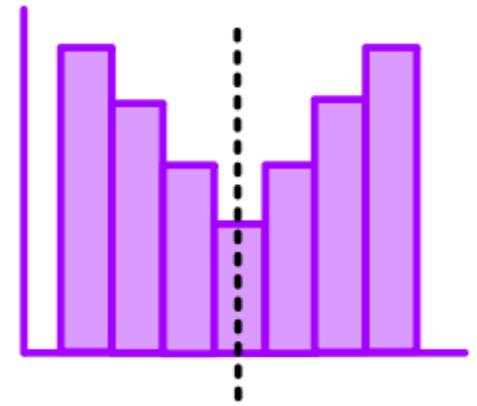
symmetric

- normal



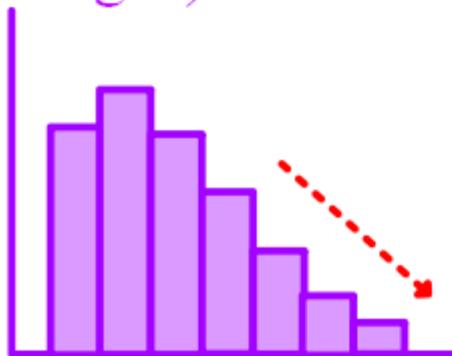
symmetric

- other than normal



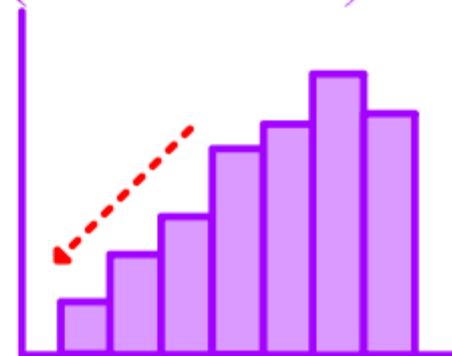
positively skewed

(skewed right)



negatively skewed

(skewed left)



*named
by the
direction
of the tail*

DOTPLOTS

- 1.) Draw a horizontal number line with appropriate scale
- 2.) Place dots above the number line to represent the data
- 3.) Stack dots vertically for data values that repeat
- 4.) Label the axis and make a title

Construct a dotplot for the following data:

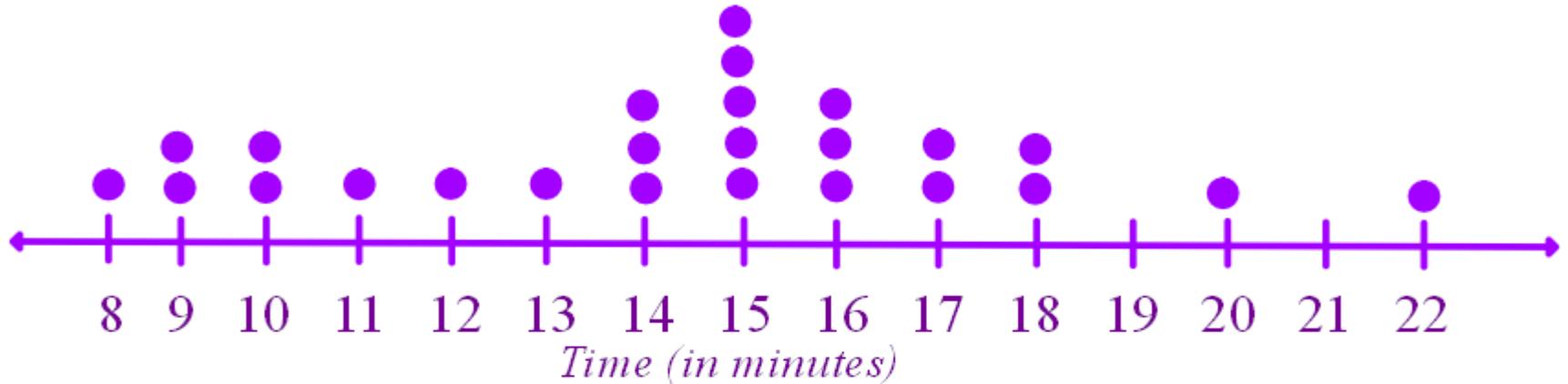
Time in minutes that it takes students in the same math class to get to school

15	15	18	20	15	12	16	18	11	8
17	9	9	10	14	22	15	16	15	10
13	14	16	17	14					



DOTPLOTS

Math Students Travel Time to School



■ DESCRIBE THE GRAPH

Shape -

Mean -

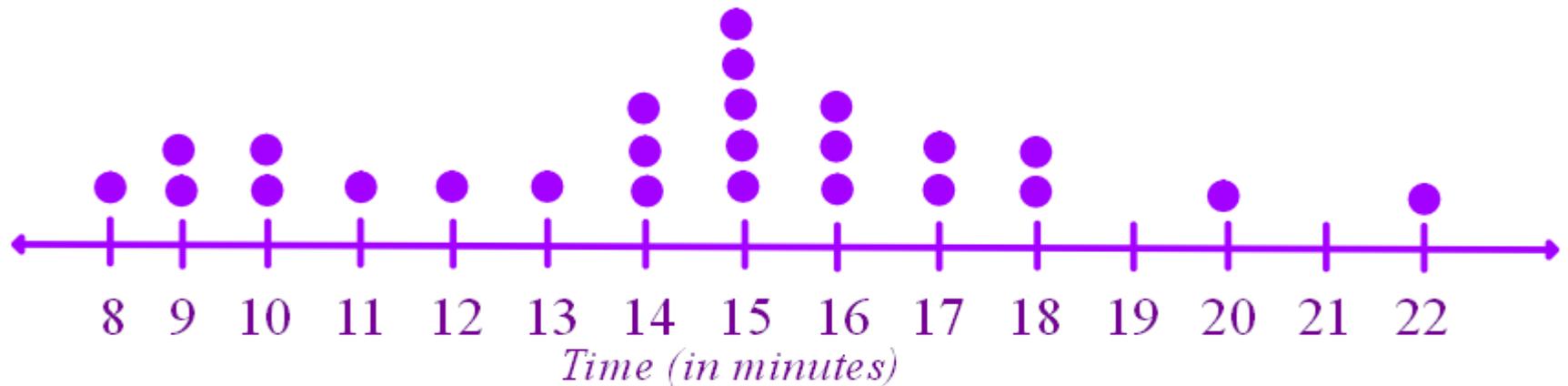
Median -

Standard deviation -

Interquartile range -

DOTPLOTS

Math Students Travel Time to School



■ DESCRIBE THE GRAPH

Shape - **symmetric/normal** **use mean and standard deviation*

Mean - **14.36 min.** *average travel time to school for a math student*

Median - **15 min.** *$\frac{1}{2}$ of students travel less than 15 min. and $\frac{1}{2}$ travel longer*

Standard deviation - **population 3.44** *typical travel time 10.9 min - 17.8 min*

Interquartile range - **16.5 - 11.5 = 5** *the travel time for the middle half of the students varies by at most 5 minutes.*

FREQUENCY DISTRIBUTION

Ungrouped

Use for discrete numerical data with a range of less than 15

**Do not omit a number even if it is not in the data set*

Relative frequency (percent) - $\frac{\text{frequency}}{\text{total frequency}}$

Construct a frequency distribution for the following data:

Minutes for third-graders to complete a dexterity test:

4 8 8 9 8 5 9
9 10 11 7 7 8 7
7 8 4 8 7 5 7
6 5 10 8 9

CLASS	FREQUENCY	RELATIVE FREQUENCY
-------	-----------	--------------------

$\Sigma =$ _____

$\Sigma =$ _____

FREQUENCY DISTRIBUTION

Ungrouped

Use for discrete numerical data with a range of less than 15

**Do not omit a number even if it is not in the data set*

Relative frequency (percent) - $\frac{\text{frequency}}{\text{total frequency}}$

Construct a frequency distribution for the following data:

Minutes for third-graders to complete a dexterity test:

4 8 8 9 8 5 9
 9 10 11 7 7 8 7
 7 8 4 8 7 5 7
 6 5 10 8 9

CLASS	FREQUENCY	RELATIVE FREQUENCY
4	2	2/26 = 7.7%
5	3	3/26 = 11.5%
6	1	1/26 = 3.8%
7	6	23.1%
8	7	26.9%
9	4	15.4%
10	2	7.7%
11	1	3.8%
	$\Sigma =$ <u>26</u>	$\Sigma =$ <u>99.9%</u>

HISTOGRAM (Ungrouped)

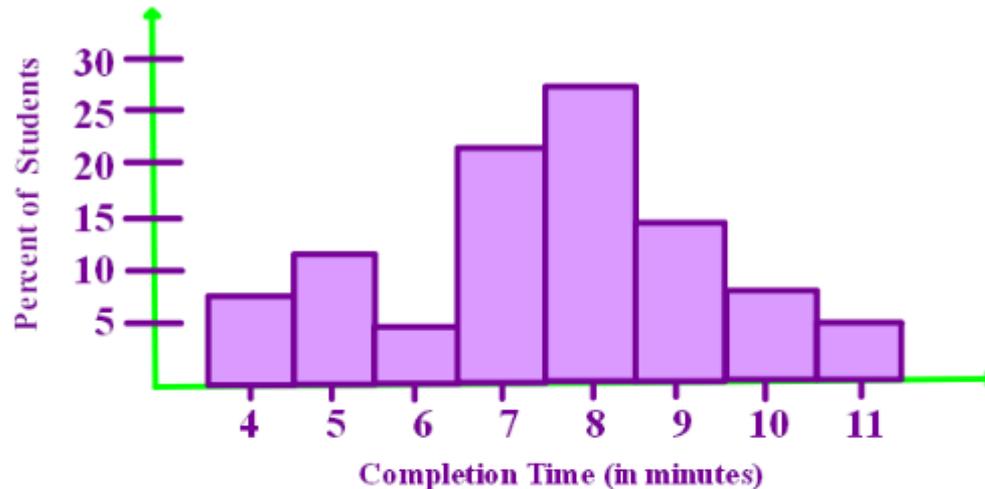
- 1.) Class values are placed on the x-axis
- 2.) Frequency/relative frequency is placed on the y-axis
- 3.) Bars are centered over their data values
- 4.) Bars touch (*unless there is a class with 0 frequency*)

Construct a frequency distribution for the data in the table.



HISTOGRAM (Ungrouped)

THIRD GRADERS DEXTERITY TEST



■ DESCRIBE THE GRAPH

Shape -

Mean -

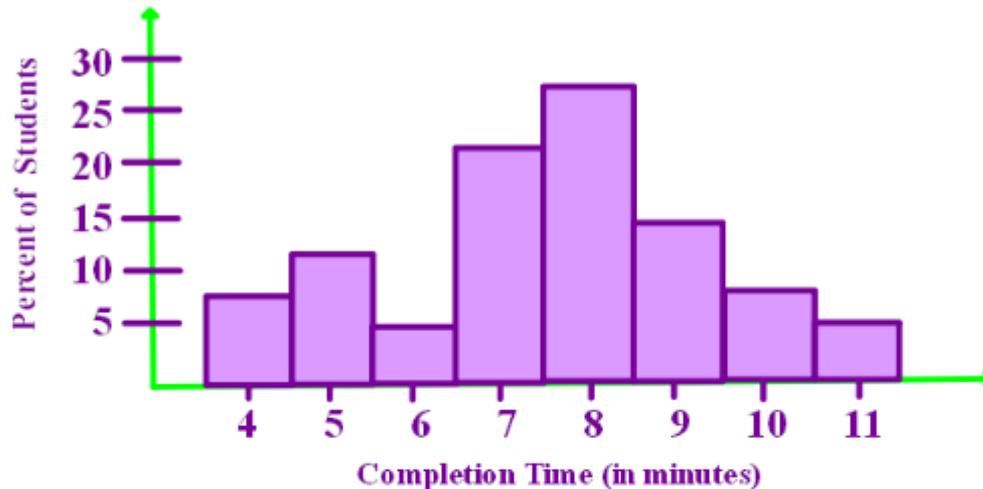
Median -

Standard deviation -

Interquartile range -

HISTOGRAM (Ungrouped)

THIRD GRADERS DEXTERITY TEST



■ DESCRIBE THE GRAPH

Shape - symmetric/normal **use mean and standard deviation*

Mean - **7.46 min.** *average completion time on 3rd-graders dexterity test*

Median - **8 min.** *$\frac{1}{2}$ of students finished in less than 8 min. and $\frac{1}{2}$ took longer*

Standard deviation - **population 1.78** *typical completion time 5.68 - 9.24 min*

Interquartile range - **$9 - 7 = 2$** *the completion time for the middle half of the students varies by at most 2 minutes.*

FREQUENCY DISTRIBUTION

Grouped

Use for discrete numerical data with a range of 15 or larger

Sort the data into intervals (classes)

Each class must be the same width - $(\text{max \#} - \text{min \#}) / \# \text{ of classes}$

Construct a frequency distribution for the following data:

Percentage of college students in public college for each state in the US

95 90 75 80 62 86 87 82 95 84 76 52 72 81 77 87 55
 85 80 91 87 88 76 84 89 63 96 55 79 70 81 84 89 82
 82 92 81 84 81 73 80 74 85 40 83 56 56 73 89 75

CLASS INTERVAL

TALLY

FREQUENCY

**RELATIVE
FREQUENCY**

40 - 49

50 - 59

60 - 69

70 - 79

80 - 89

90 - 99

$\Sigma =$ _____

$\Sigma =$ _____

FREQUENCY DISTRIBUTION

Grouped

Use for discrete numerical data with a range of 15 or larger

Sort the data into intervals (classes)

Each class must be the same width - $(\text{max \#} - \text{min \#}) / \# \text{ of classes}$

Construct a frequency distribution for the following data:

Percentage of college students in public college for each state in the US

95 90 75 80 62 86 87 82 95 84 76 52 72 81 77 87 55
 85 80 91 87 88 76 84 89 63 96 55 79 70 81 84 89 82
 82 92 81 84 81 73 80 74 85 40 83 56 56 73 89 75

CLASS INTERVAL	TALLY	FREQUENCY	RELATIVE FREQUENCY
40 - 49		1	$1/50 = 2\%$
50 - 59		5	$5/50 = 10\%$
60 - 69		2	4%
70 - 79		11	22%
80 - 89		25	50%
90 - 99		6	12%
		$\Sigma = 50$	$\Sigma = 100\%$

FREQUENCY DISTRIBUTION

Grouped

Class width = $(\text{max \#} - \text{min \#}) / \# \text{ of classes}$

CLASS INTERVAL	CLASS BOUNDARIES
----------------	------------------

40 - 49	39.5 - 49.5
---------	-------------

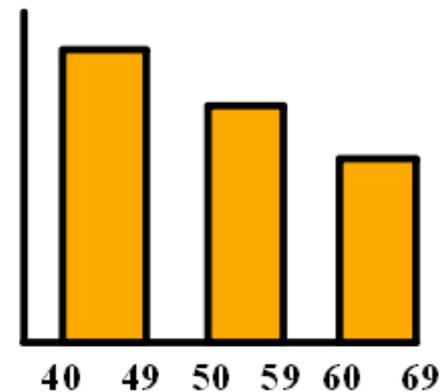
50 - 59	49.5 - 59.5
---------	-------------

60 - 69	59.5 - 69.5
---------	-------------

70 - 79	69.5 - 79.5
---------	-------------

80 - 89	79.5 - 89.5
---------	-------------

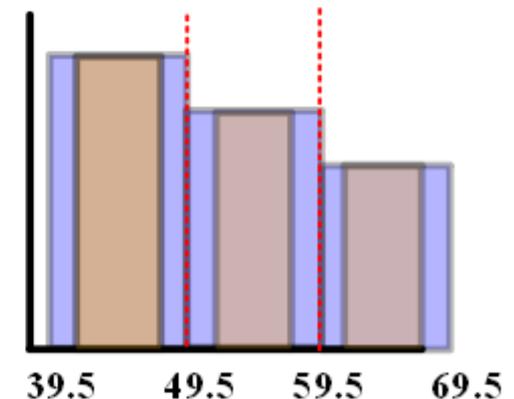
90 - 99	89.5 - 99.5
---------	-------------



HISTOGRAM - BARS MUST TOUCH

Subtract .5 from lower class limit to stretch bar to the left

Add .5 to upper class limit to stretch bar to the right



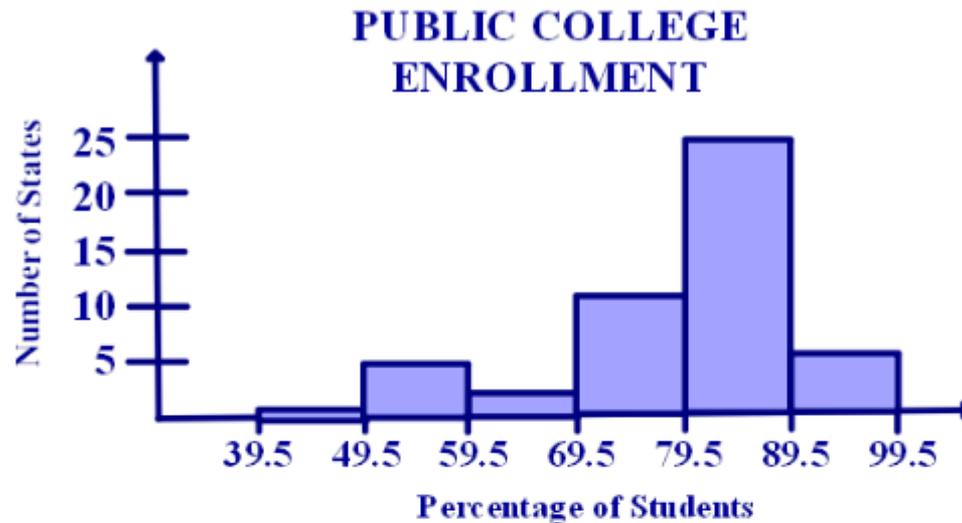
HISTOGRAM (Grouped)

- 1.) Place lower class boundaries (+ one extra) on the x-axis
- 2.) Frequency/relative frequency is placed on the y-axis
- 3.) Bars begin/end at the lower class boundaries
- 4.) Bars touch (*unless there is a class with 0 frequency*)

Construct a frequency distribution for the data in the table.



HISTOGRAM (Grouped)



■ DESCRIBE THE GRAPH

Shape -

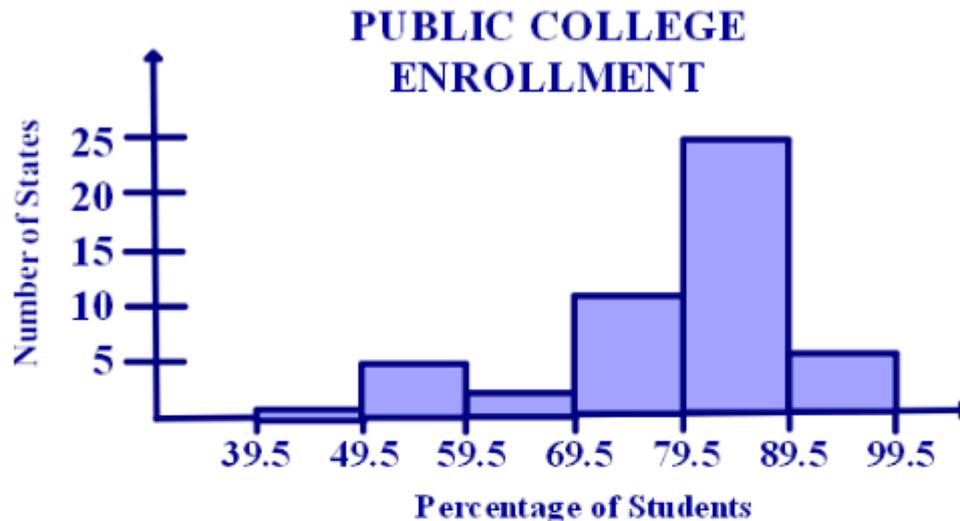
Mean -

Median -

Standard deviation -

Interquartile range -

HISTOGRAM (Grouped)



■ DESCRIBE THE GRAPH

Shape - negatively skewed **use median and interquartile range*

Mean - **78.38%** *average % of students enrolled in public college*

Median - **81%** *$1/2$ of states had enrollment less than 81% and $1/2$ had higher*

Standard deviation - **population 12.13** *typical enrollment 66.25% - 90.51%*

Interquartile range - **$87 - 74 = 13$** *public college enrollment for the middle half of the states varies by at most 13%.*

FIVE NUMBER SUMMARY

**PUT DATA IN ASCENDING ORDER*

minimum - smallest data number

Q₁ lower quartile - median of lower half of the data

median - middle number

Q₃ upper quartile - median of upper half of the data

maximum - largest data number

- **outlier(s)** - any piece of data that is
lower than $Q_1 - 1.5 \cdot \text{IQR}$
or
higher than $Q_3 + 1.5 \cdot \text{IQR}$

FIVE NUMBER SUMMARY

Find the five number summary and identify any outliers for the following data: **have students give context for the data*

13 23 26 16 33 65 28 39 14 8

By hand

Put data in ascending order

8 13 14 16 23 26 28 33 39 65

Find the smallest data value

Find the median

Find Q_1

Find Q_3

Find the greatest data value

Check for outliers

Calculator

FIVE NUMBER SUMMARY

Find the five number summary and identify any outliers for the following data: **have students give context for the data*

13 23 26 16 33 65 28 39 14 8

By hand

Put data in ascending order 8 13 14 16 23 | 26 28 33 39 65

Find the smallest data value 8

Find the median $(23 + 26) / 2 = 24.5$

Find Q_1 14

Find Q_3 33

Find the greatest data value 65

Check for outliers IQR $33 - 14 = 19$ $14 - 1.5(19) = -14.5$
 $33 + 1.5(19) = 61.5$ *65 is an outlier

Calculator

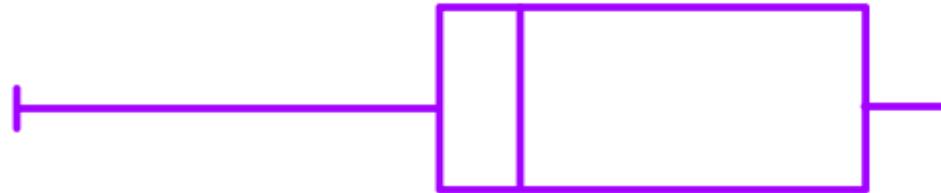
STAT 1: Edit enter data in list 2ND QUIT

STAT CALC 1: 1-Var Stats 2ND L1 ENTER

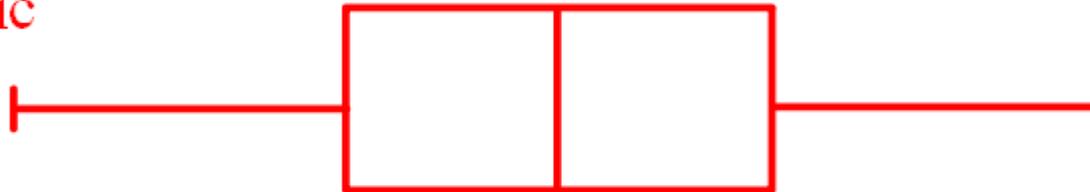
minX = 8 $Q_1 = 14$ Med = 24.5 $Q_3 = 33$ maxX = 65

BOXPLOT SHAPES

Negative Skew



Symmetric



Positive Skew



BOXPLOT

- 1.) Calculate the 5-number summary and check for outliers
- 2.) Draw a number line with the appropriate scale
- 3.) Make dots above the line representing the 5-number summary
- 4.) Draw a whisker between the min and Q_1 and Q_3 and the max
(or lowest/highest number that is not an outlier), place star(s) at the outlier(s)
- 5.) Draw a box from Q_1 to Q_3 with a vertical line at the median

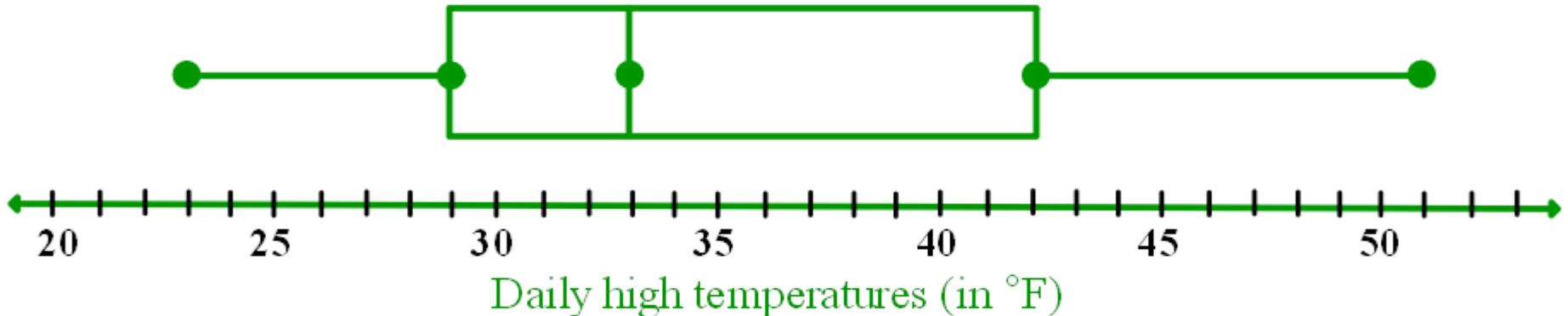
Construct a boxplot for the following data:

High temperatures (°F) in Clover, SC for 11 days in January 2012

23 27 29 30 31 33 38 40 42 43 51

BOXPLOT

Clover, SC - January 2012



■ DESCRIBE THE GRAPH

Shape -

Mean -

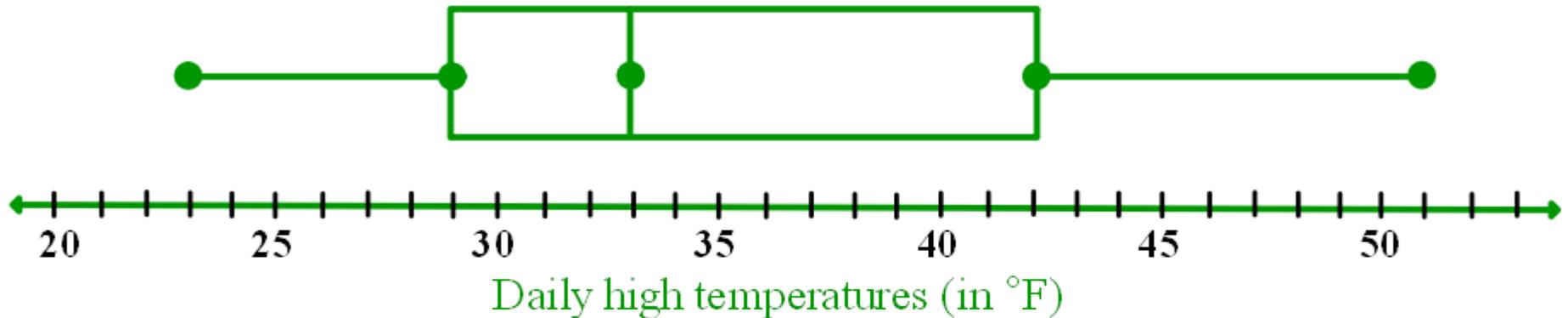
Median -

Standard deviation -

Interquartile range -

BOXPLOT

Clover, SC - January 2012



■ DESCRIBE THE GRAPH

Shape - **positively skewed** **use median and interquartile range*

Mean - **35.18°** *average daily high temperature in Clover during January 2012*

Median - **33°** *1/2 of the days it was colder than 33° and 1/2 of the days were hotter*

Standard deviation - **population 7.93** *typical high temp. 27.3° - 43.1°*

Interquartile range - **42 - 29 = 13** *the middle 50% of the daily high temperatures varied by at most 13°*

BOXPLOT

Construct a boxplot for the following data:

8 13 14 16 23 26 28 33 39 65

What might the data represent? Helps the students to give meaning to what they are doing.

Calculator

STAT **Edit** *enter data in list* **2ND** **QUIT**

2ND **STATPLOT** **Plot1** **ENTER**

Set the following:

ON **ENTER**

Type: 1st graph, 2nd row *(to show outliers)*

Xlist: **2ND** **L1** *(where you put your data)*

Freq: **1** *(or an appropriate # - sets the x-axis)*

Mark: select how you want outliers marked

ZOOM **ZoomStat** *(sets viewing window for the data)*

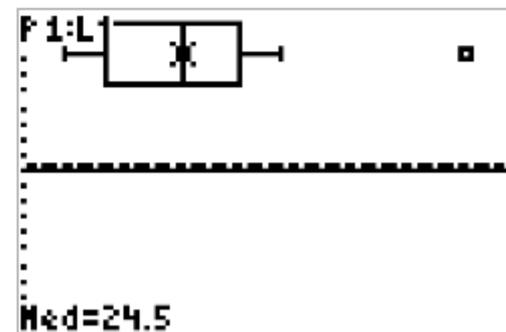
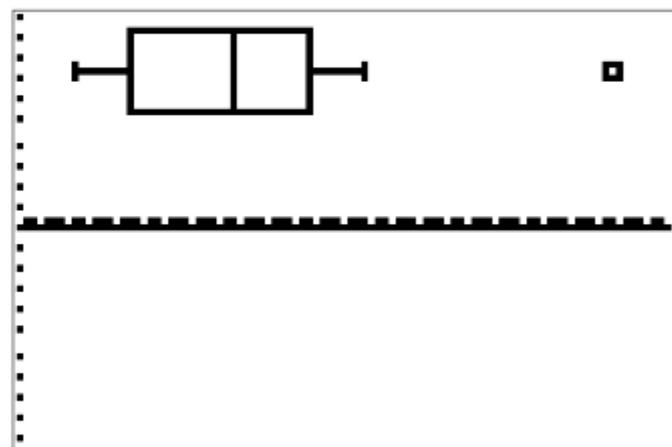
TRACE **Use the arrow keys to find the**

5-number summary:

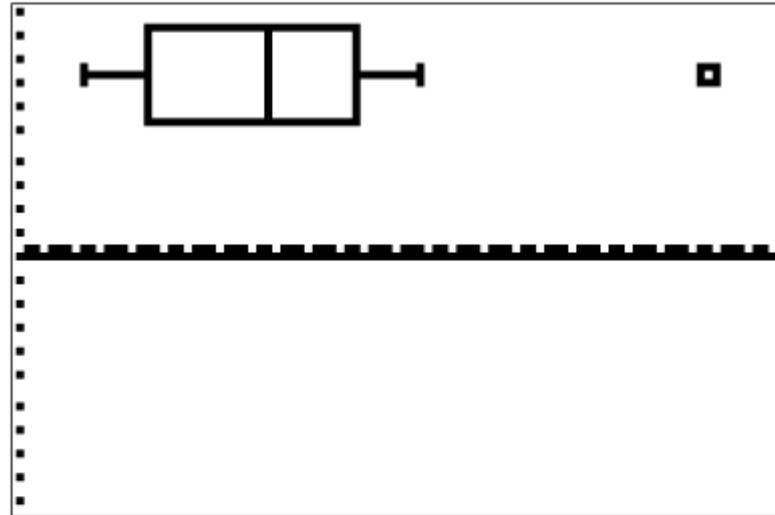
minX = 8 **Q₁ = 14** **Med = 24.5**

Q₃ = 33 **maxX = 65**

**note - the outlier is still the "maximum" number*



BOXPLOT



■ DESCRIBE THE GRAPH

Shape -

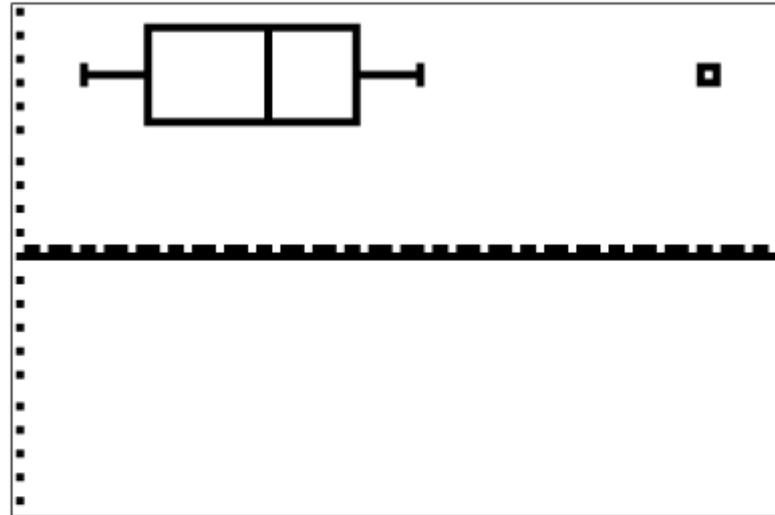
Mean -

Median -

Standard deviation -

Interquartile range -

BOXPLOT



■ DESCRIBE THE GRAPH

Shape - **positively skewed with an outlier at 65** **use median and IQR*

Mean - **26.5**

Median - **24.5**

Standard deviation - **population 15.77**

Interquartile range - **$33 - 14 = 19$**

Activity

Glued to the Tube or Hooked to the Books?

Authors: Jill Holowaty and Dave Muse

15-20 minutes

SCATTERPLOTS
and
LINEAR REGRESSION

COMMON CORE STATE STANDARDS

S.ID.5 - Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

S.ID.6 - Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

- a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. *Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.*
- b. Informally assess the fit of a function by plotting and analyzing residuals.
- c. Fit a linear function for a scatter plot that suggests a linear association.

COMMON CORE STATE STANDARDS

S.ID.7 - Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

S.ID.8 - Compute (using technology) and interpret the correlation coefficient of a linear fit.

S.ID.9 - Distinguish between correlation and causation.

VOCABULARY

bivariate data - two variables of interest recorded for each individual in a group

two-way table - a rectangular table that consists of rows for the first variable (x) and columns for the second variable (y)

- cell - the intersection of a row and column
- cell count - numeric value in the cell
- marginals - the sum of the cell counts for each row/column
- table total - the sum of the marginal row (or column) counts; the sum of the cell counts

PERCENTS/PROBABILITIES

- marginal relative frequency - ratio of marginal total to table total
- joint relative frequency - ratio of cell count to table total
- conditional relative frequency - ratio of cell count to marginal total

Ex: 1 Students' After School Activities

	9th	10th	11th	12th
Athletics	150	160	140	150
Fine Arts	100	90	120	125
Other	125	140	150	150

This a 3 x 4 (rows x columns) two-way table for a student's grade level and the type of after school activity that the student participates in.

Calculate: a.) marginal totals b.) table total

c.) marginal relative frequency for 9th graders

d.) marginal relative frequency for athletics

Ex: 1 Students' After School Activities

	9th	10th	11th	12th	
Athletics	150	160	140	150	600
Fine Arts	100	90	120	125	435
Other	125	140	150	150	565
	375	390	410	425	1600

This a 3 x 4 (rows x columns) two-way table for a student's grade level and the type of after school activity that the student participates in.

Calculate: a.) marginal totals b.) table total

c.) marginal relative frequency for 9th graders

$$375/1600 = 15/64 = .2344 = 23.4\% \mid 23.4\% \text{ of the students are 9th graders.}$$

d.) marginal relative frequency for athletics

$$600/1600 = 3/8 = .375 = 37.5\% \mid 37.5\% \text{ of the students participate in athletics.}$$

Ex: 1 Students' After School Activities cont.

	9th	10th	11th	12th	
Athletics	150	160	140	150	600
Fine Arts	100	90	120	125	435
Other	125	140	150	150	565
	375	390	410	425	1600

e.) joint relative frequency for 10th graders and students in fine arts

f.) joint relative frequency for 12th graders and students in other activities

g.) conditional relative frequency for students in athletics who are 12th graders

h.) conditional relative frequency for 11th graders who are in fine arts

Ex: 1 Students' After School Activities cont.

	9th	10th	11th	12th	
Athletics	150	160	140	150	600
Fine Arts	100	90	120	125	435
Other	125	140	150	150	565
	375	390	410	425	1600

- e.) joint relative frequency for 10th graders and students in fine arts
 $90/1600 = 9/160 = .0563 = 5.6\%$ | 5.6% of students are 10th graders who participate in fine arts.
- f.) joint relative frequency for 12th graders and students in other activities
 $150/1600 = 3/32 = .0938 = 9.4\%$ | 9.4% of students are 12th graders who participate in other activities.
- g.) conditional relative frequency for students in athletics who are 12th graders
 $150/425 = 6/17 = .3529 = 35.3\%$ | 35.3% of 12th graders participate in athletics.
- h.) conditional relative frequency for 11th graders who are in fine arts
 $120/435 = 8/29 = .2759 = 27.6\%$ | 27.6% of students who participate in fine arts are 11th graders.

Ex: 1 Students' After School Activities cont.

	9th	10th	11th	12th	
Athletics	150	160	140	150	600
Fine Arts	100	90	120	125	435
Other	125	140	150	150	565
	375	390	410	425	1600

i.) the ratio of students in athletics for each grade level

9th

11th

10th

12th

j.) the ratio of 9th graders for each activity

athletics

fine arts

other

Ex: 1 Students' After School Activities cont.

	9th	10th	11th	12th	
Athletics	150	160	140	150	600
Fine Arts	100	90	120	125	435
Other	125	140	150	150	565
	375	390	410	425	1600

i.) the ratio of students in athletics for each grade level

9th $150/375 = 2/5 = .4 = 40\%$

11th $140/410 = 14/41 = .3415 = 34.1\%$

10th $160/390 = 16/39 = .4103 = 41\%$

12th $150/425 = 6/17 = .3529 = 35.3\%$

_____ % of _____-graders participate in athletics.

j.) the ratio of 9th graders for each activity

athletics $150/600 = 1/4 = .25 = 25\%$

fine arts $100/435 = 20/87 = .2299 = 23\%$

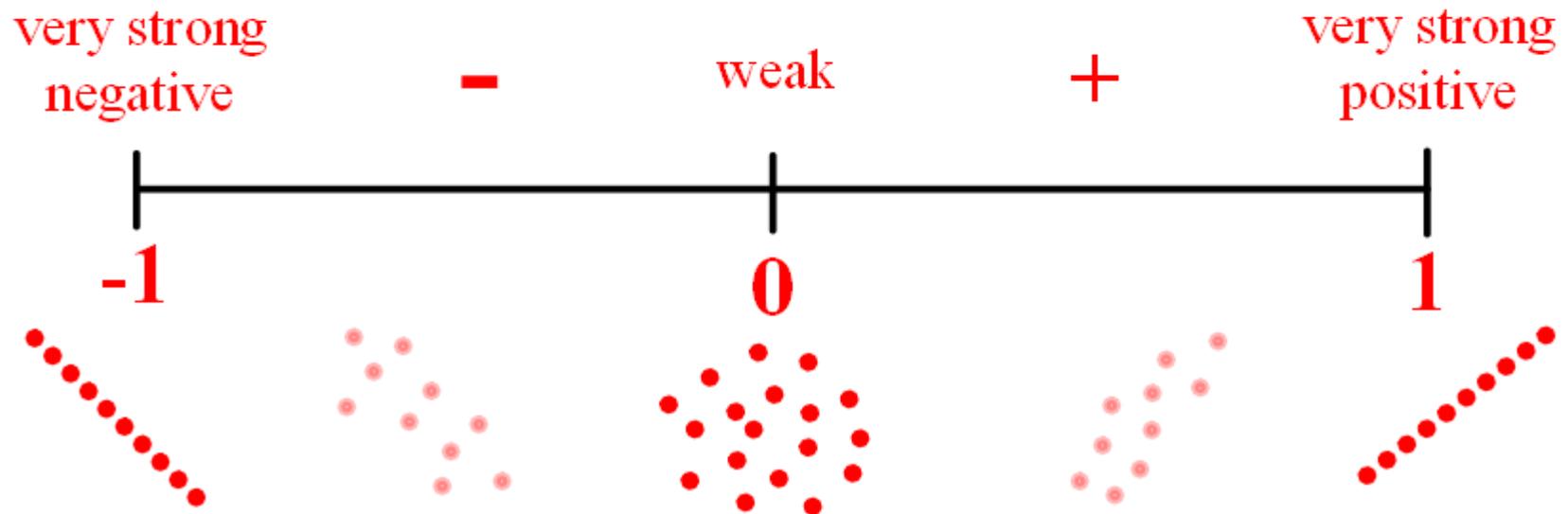
other $125/565 = 25/113 = .2212 = 22\%$

_____ % of students who
participate in _____
are 9th graders.

VOCABULARY

linear regression - linear relationship between two variables

correlation coefficient - "r", strength of the relationship

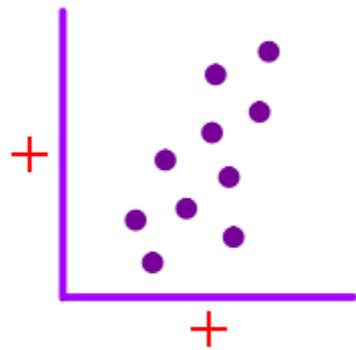


CORRELATION (Relationship)

Linear

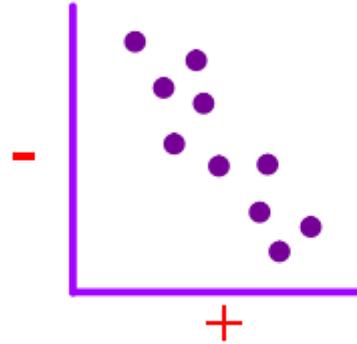
positive

Ex: $y = ax + b$



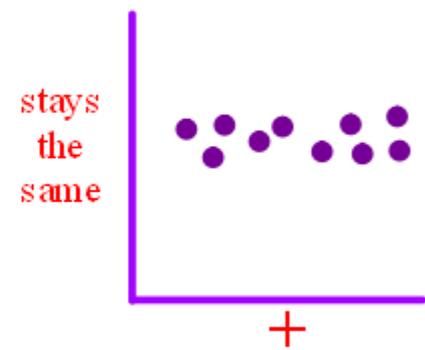
negative

Ex: $y = -ax + b$



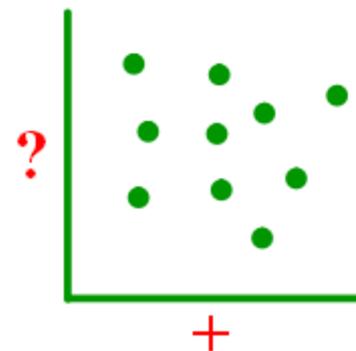
constant

Ex: $y = b$



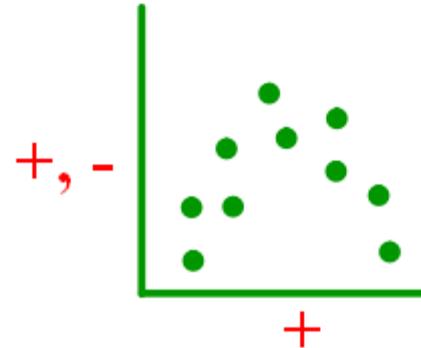
Non-linear

no correlation



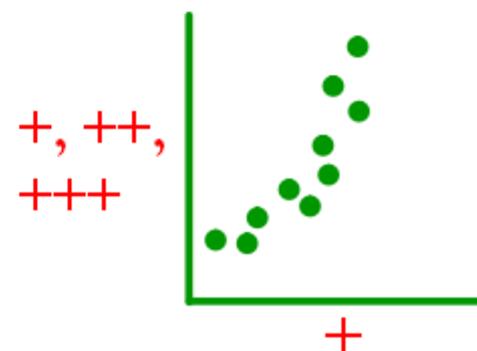
quadratic

Ex: $y = -ax^2 + bx + c$



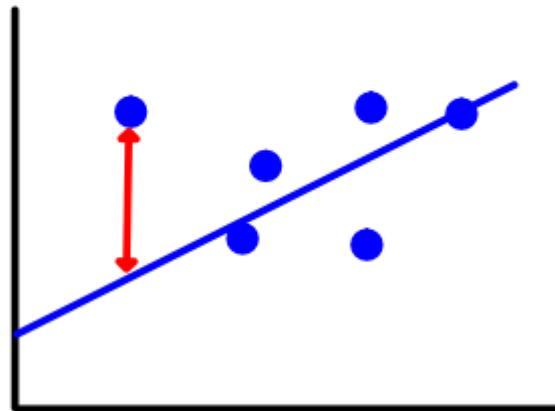
exponential

Ex: $y = ab^x$



VOCABULARY

residuals - vertical distances between data points and the linear regression line; smaller residuals - the better the linear equation fits the data



causation - the idea that one variable causes another to happen

CORRELATION \neq CAUSATION

<http://mathbits.com/mathbits/tisection/statistics2/correlationcausation.htm>

Ex: 2 Amount of Money Raised During a Community Service Project

Time spent in hours	1.5	2.3	2.4	2.8	1.8	2	3.2	4	4.5
Money raised in dollars	50	75	70	80	65	75	90	95	100

Construct a scatterplot with time as the x-variable and money as the y-variable

By hand and with a graphing calculator

Describe the scatterplot:

association - positive or negative, how strong

shape - linear, quadratic, or exponential

outliers - unusual values away from the rest of the data

Ex: 2 Amount of Money Raised During a Community Service Project

Time spent in hours	1.5	2.3	2.4	2.8	1.8	2	3.2	4	4.5
Money raised in dollars	50	75	70	80	65	75	90	95	100

Construct a scatterplot with time as the x-variable and money as the y-variable

Calculator

STAT **Edit** *enter data in lists* **2ND** **QUIT**

2ND **STATPLOT** **Plot1** **ENTER**

Set the following:

ON **ENTER**

Type: 1st graph, 1st row

Xlist: **2ND L1** (*where you put your data*)

Ylist: **2ND L2** (*where you put your data*)

Mark: select how you ordered pairs marked

ZOOM **ZoomStat** (*sets viewing window for the data*)

TRACE Use the arrow keys to see the coordinates for each point

Describe the scatterplot: **association** - strong positive

shape - linear

outliers - none



LINEAR REGRESSION EQUATION

$$y = mx + b$$

$$\hat{y} = ax + b$$


Use your calculator to find "a" and "b"

STAT **Calc** **LinReg(ax + b)** **2nd** **L1** , **2nd** **L2** **ENTER**

*if "r" does not show in list turn 'Diagnostics On'

2nd **CATALOG** **Diagnostics On** **ENTER** **ENTER**

LINEAR REGRESSION EQUATION

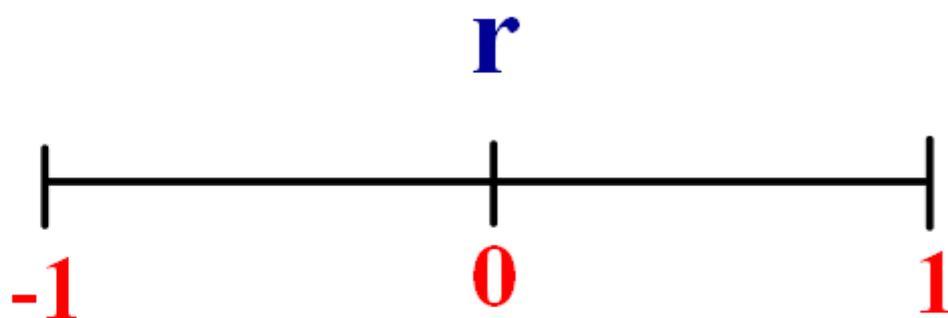
a = *slope*

$$y = \underline{\mathbf{a}} x + \underline{\mathbf{b}}$$

b = *y-intercept*

r² = **don't need*

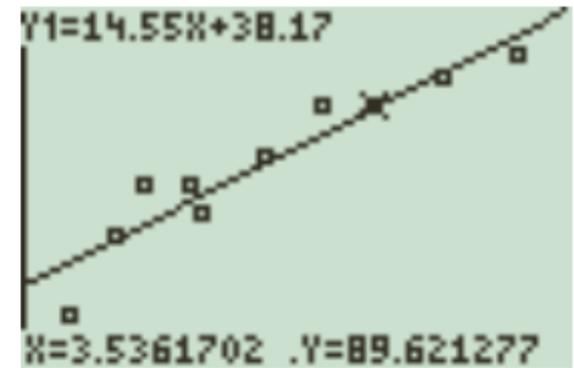
r = *correlation coefficient*



```
LinReg
y=ax+b
a=14.54879043
b=38.17273716
r2=.8849136453
r=.940698488
```

Ex: 2 Amount of Money Raised During a Community Service Project

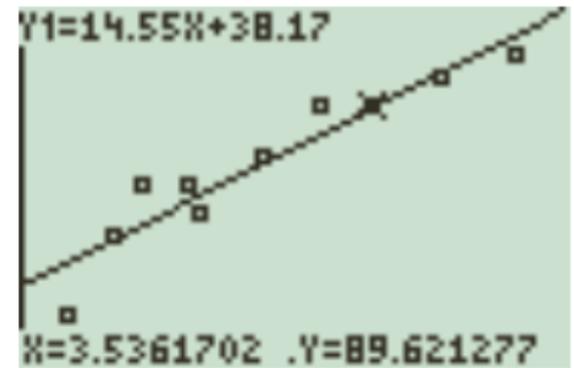
Graph the linear regression equation on the scatterplot and interpret the results.



- What is the slope of your equation? Interpret the slope in the context of the problem.
- What is the y-intercept of your equation? Interpret the y-intercept in the context of the problem. Does your interpretation make sense, is it possible?
- What is the correlation coefficient for your data. What does it imply about the linear relationship between the two variables?
- How well does the line fit the data? Look at the residuals - are the points far away or close to the line?

Ex: 2 Amount of Money Raised During a Community Service Project

Graph the linear regression equation on the scatterplot and interpret the results.



a.) What is the slope of your equation? Interpret the slope in the context of the problem.

14.55 - for every hour that a student works, they raise \$14.55

b.) What is the y-intercept of your equation? Interpret the y-intercept in the context of the problem. Does your interpretation make sense, is it possible?

38.17 - if a student does not work at all (0 hours), they will raise \$38.17, this does not make sense

c.) What is the correlation coefficient for your data. What does it imply about the linear relationship between the two variables?

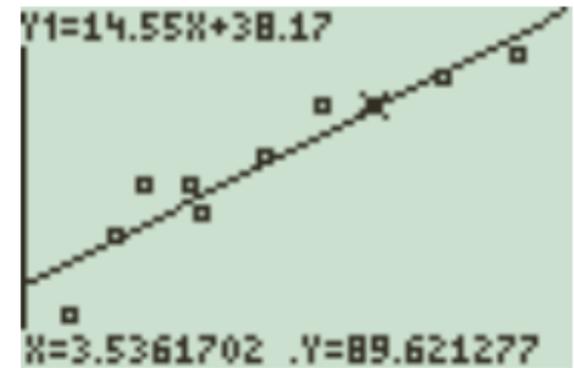
r = .94 - there is a strong positive linear relationship between hours worked and money raised, the more hours worked - the more money raised

d.) How well does the line fit the data? Look at the residuals - are the points far away or close to the line?

close to the line

Ex: 2 Amount of Money Raised During a Community Service Project

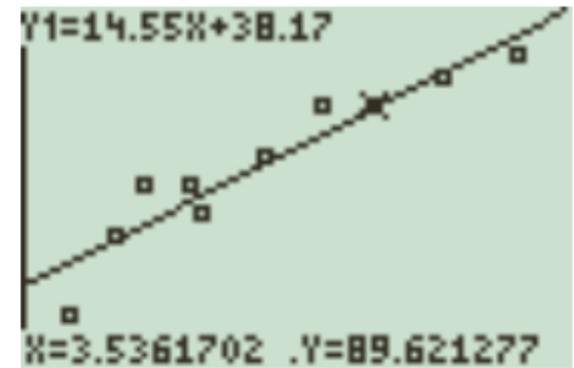
Graph the linear regression equation on the scatterplot and interpret the results.



- e.) There is a strong positive linear relationship between the number of hours worked and the amount of money raised. Do you think the number of hours a student works causes them to raise more money? What might be some other factors involved besides the number of hours worked that could be causing this positive relationship between the two variables?
- f.) Based on this relationship how much money would you expect this student to raise if they worked three and one-half hours?

Ex: 2 Amount of Money Raised During a Community Service Project

Graph the linear regression equation on the scatterplot and interpret the results.



e.) There is a strong positive linear relationship between the number of hours worked and the amount of money raised. Do you think the number of hours a student works causes them to raise more money? What might be some other factors involved besides the number of hours worked that could be causing this positive relationship between the two variables?

Possible answer: Just because they are working does not mean that they will automatically raise money. Other factors that may influence the amount of money they raise may be the reason they are raising money, what type of service they are performing, the location where they are working, and the time of year are some possibilities.

f.) Based on this relationship how much money would you expect this student to raise if they worked three and one-half hours?

$y = 14.55(3.5) + 38.17 = 89.095$ | I would expect this student to raise approximately \$89.10 when working for 3.5 hours.

Ex: 3 Time Between Flowering and Harvesting of Grain

Time in days (x)	16	18	20	22	24	26	28	30	32
	34	36	38	40	42	44	46		
Yield of grain (y)	2508	2518	3306	3423	3057	3190	3500	3883	3823
	3646	3708	3333	3517	3214	3103	2776		

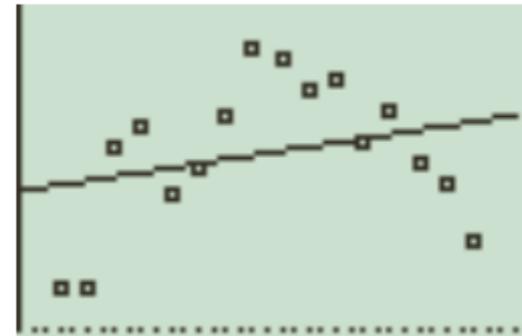
- Use your calculator to sketch the graph with the linear regression line.
- What is the value of the correlation coefficient?
- Does the line fit the data?
- What type of equation might fit the data better?

Ex: 3 Time Between Flowering and Harvesting of Grain

Time in days (x)	16	18	20	22	24	26	28	30	32
	34	36	38	40	42	44	46		
Yield of grain (y)	2508	2518	3306	3423	3057	3190	3500	3883	3823
	3646	3708	3333	3517	3214	3103	2776		

a.) Use your calculator to sketch the graph with the linear regression line.

```
LinReg  
y=ax+b  
a=12.02867647  
b=2908.673529  
r2=.0748708719  
r=.2736254227
```



b.) What is the value of the correlation coefficient?

r = .27 Closer to 0 means the relationship is weak

c.) Does the line fit the data?

No The data is curved, not linear

d.) What type of equation might fit the data better?

Quadratic A parabola

Ex: 3 Time Between Flowering and Harvesting of Grain

Time in days (x)	16	18	20	22	24	26	28	30	32
	34	36	38	40	42	44	46		
Yield of grain (y)	2508	2518	3306	3423	3057	3190	3500	3883	3823
	3646	3708	3333	3517	3214	3103	2776		

e.) Using your calculator fit a quadratic equation to the data

$$y = ax^2 + bx + c$$

f.) Graph the quadratic equation on your scatterplot.
Which equation fits the data best?

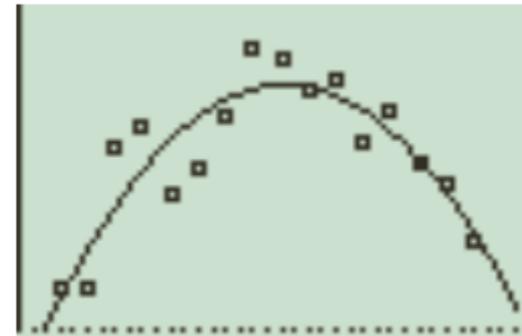
Ex: 3 Time Between Flowering and Harvesting of Grain

Time in days (x)	16	18	20	22	24	26	28	30	32
	34	36	38	40	42	44	46		
Yield of grain (y)	2508	2518	3306	3423	3057	3190	3500	3883	3823
	3646	3708	3333	3517	3214	3103	2776		

e.) Using your calculator fit a quadratic equation to the data

$$y = ax^2 + bx + c$$

```
QuadReg
y=ax2+bx+c
a=-4.54564951
b=293.8589461
c=-1073.315441
R2=.7933900578
```



STAT Calc QuadReg 2nd L1 , 2nd L2 ENTER

$$y = -4.55x^2 + 293.86x - 1073.32$$

f.) Graph the quadratic equation on your scatterplot.

Which equation fits the data best?

Quadratic It has the same shape as the data.

Ex: 4 Cooling Temperatures of a Freshly Brewed Cup of Coffee
after it is poured from the brewing pot into a serving cup. The brewing pot temperature is approximately 180°F.

Time in min (x)	0	5	8	11	15	18	22	25
	30	34	38	42	45	50		
Temp in °F (y)	179.5	168.7	158.1	149.2	141.7	134.6	125.4	123.5
	116.3	113.2	109.1	105.7	102.2	100.5		

- a.) Use your calculator to sketch the graph of the data. Describe the shape of the graph.
- b.) Calculate the linear regression equation.
What is the value of the correlation coefficient?
- c.) Draw the linear equation on the scatterplot.
Does the line fit the data?

Ex: 4 Cooling Temperatures of a Freshly Brewed Cup of Coffee
 after it is poured from the brewing pot into a serving cup. The brewing pot temperature is approximately 180°F.

Time in min (x)	0	5	8	11	15	18	22	25
	30	34	38	42	45	50		
Temp in °F (y)	179.5	168.7	158.1	149.2	141.7	134.6	125.4	123.5
	116.3	113.2	109.1	105.7	102.2	100.5		

a.) Use your calculator to sketch the graph of the data. Describe the shape of the graph.

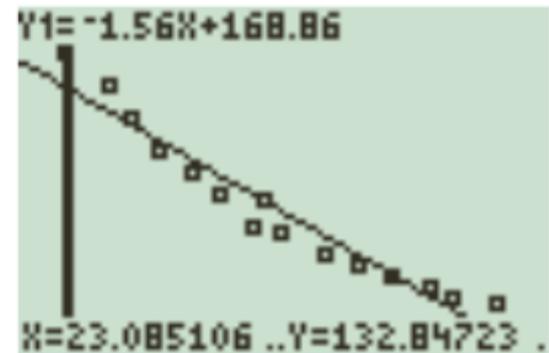
Linear/curved Negative slope with no outliers

b.) Calculate the linear regression equation. What is the value of the correlation coefficient?

$$y = -1.56x + 168.86 \quad r = -.97$$

c.) Draw the linear equation on the scatterplot. Does the line fit the data?

The data points are close to the line, but the line appears to have a dip (curve) in the middle of it.



```
LinReg
y=ax+b
a=-1.563715998
b=168.861042
r^2=.9399494817
r=-.9695099183
```

Ex: 4 Cooling Temperatures of a Freshly Brewed Cup of Coffee
after it is poured from the brewing pot into a serving cup. The brewing pot temperature is approximately 180°F.

Time in min (x)	0	5	8	11	15	18	22	25
	30	34	38	42	45	50		
Temp in °F (y)	179.5	168.7	158.1	149.2	141.7	134.6	125.4	123.5
	116.3	113.2	109.1	105.7	102.2	100.5		

d.) What type of equation might fit the data better?

e.) Using your calculator fit an exponential equation

to the data. $y = ab^x$

f.) What is the correlation coefficient for the exponential?

g.) Graph the exponential equation on your scatterplot.
Which equation fits the data best?

Ex: 4 Cooling Temperatures of a Freshly Brewed Cup of Coffee
 after it is poured from the brewing pot into a serving cup. The brewing pot temperature is approximately 180°F.

Time in min (x)	0	5	8	11	15	18	22	25
	30	34	38	42	45	50		
Temp in °F (y)	179.5	168.7	158.1	149.2	141.7	134.6	125.4	123.5
	116.3	113.2	109.1	105.7	102.2	100.5		

d.) What type of equation might fit the data better?

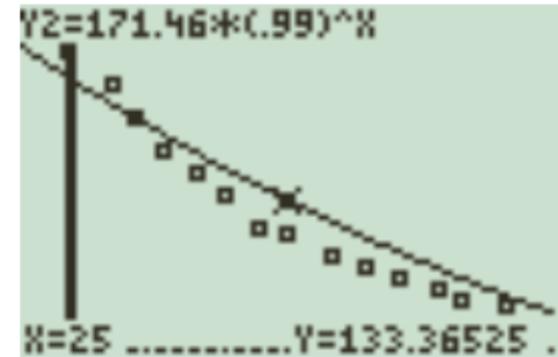
Exponential

e.) Using your calculator fit an exponential equation

to the data. $y = ab^x$

STAT Calc ExpReg 2nd L1 , 2nd L2 ENTER

$y = 171.46(.99)^x$



f.) What is the correlation coefficient for the exponential?

$r = -.98$

```
ExpReg
y=a*b^x
a=171.4617283
b=.9882469577
r^2=.9701377262
r=-.9849556976
```

g.) Graph the exponential equation on your scatterplot.

Which equation fits the data best?

Exponential

Linear Regression

NUMB3RS ACTIVITY

"How Tall Is The Criminal?"

15-20 minutes