STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

New Jersey Student Learning Standards
Mathematics

Algebra 1

Instructional Support Tool for Teachers: **Statistics & Probability**

This instructional support tool is designed to assist educators in interpreting and implementing the New Jersey Student Learning Standards for Mathematics. It contains explanations or examples of each Algebra 1 course standard to answer questions about the standard's meaning and how it applies to student knowledge and performance. To ensure that descriptions are helpful and meaningful to teachers, this document also identifies grades 6 to 8 prerequisite standards upon which each Algebra 1 standard builds. It includes the following: sample items aligned to each identified prerequisite; identification of course level connections; instructional considerations and common misconceptions; sample academic vocabulary and associated standards for mathematical practice.
Examples are samples only and should not be considered an exhaustive list.

This instructional support tool is considered a living document. The New Jersey Department of Education believe that teachers and other educators will find ways to improve the document as they use it. Please send feedback to mathematics@doe.state.nj.us so that the Department may use your input when updating this guide.

Please consult the New Jersey Student Learning Standards for Mathematics for more information.

Statistics and Probability Standards Overview

Decisions or predictions are often based on data—numbers in context. These decisions or predictions would be easy if the data always sent a clear message, but the message is often obscured by variability. Statistics provides tools for describing variability in data and for making informed decisions that take it into account. Technology plays an important role in statistics and probability by making it possible to generate plots, regression functions, and correlation coefficients, and to simulate many possible outcomes in a short amount of time.

Interpreting Categorical and Quantitative Data (S-ID)

Summarize, represent, and interpret data on a single count or measurement variable

S-ID.A.1 S-ID.A.2 S-ID.A.3

Summarize, represent, and interpret data on two categorical and quantitative variables

S-ID.B.5 S-ID.B.6

Interpret linear models

S-ID.C.7 S-ID.C.8 S-ID.C.9

Mathematical Practices

1. Make sense of problems and persevere in solving them
2. Reason abstractly and quantitatively
3. Construct viable arguments and critique the reasoning of others
4. Model with mathematics
5. Use appropriate tools strategically
6. Attend to precision
7. Look for and make use of structure
8. Look for and express regularity in repeated reasoning

Standard S-ID.A.1.
Represent data with plots on the **REAL** number line (dot plots, histograms, and box plots). *(Content Emphases: Additional)*

Explanation and/or Examples

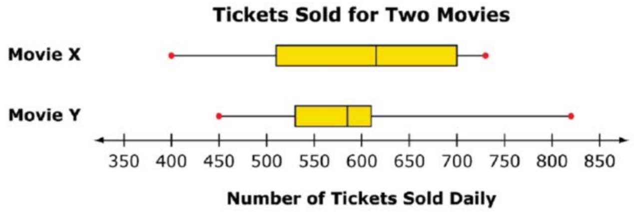
Example 1:
A movie theater recorded the number of tickets sold for two movies each day during one week and generated the following lists:

Movie X: 400, 402, 432, 501, 510, 520, 583, 591, 620, 625, 636, 690, 700, 710, 710, 734, 740

Movie Y: 500, 450, 650, 820, 600 530, 500, 750, 610, 580, 560

Create box plots of the data on the real number line.

Ex. 1 Solution**:**



Building Blocks (Grades 6 to 8 prerequisites):

Grade 7

7.SP.B.3
Informally assess the degree of **visual overlap of two numerical data distributions** with similar variabilities, **measuring the difference between the centers** by expressing it as a multiple of a measure of variability.

7.SP.B.4
Use **measures of center and measures of variability** for numerical data from random samples to **draw** informal comparative **inferences about two populations**.

Grade 6

6.SP.B.4
Display numerical data in plots on a number line, including **dot plots, histograms, and box plots**

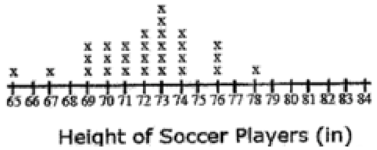
Prerequisite Examples:

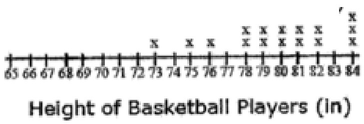
Grade 7

7.SP.B.3
Jason wanted to compare the mean height of the players on his favorite basketball and soccer teams. He thinks the mean height of the players on the basketball team will be greater but doesn't know how much greater. He also wonders if the variability of heights of the athletes is related to the sport they play. He thinks that there will be a greater variability in the heights of soccer players as compared to basketball players. He used the rosters and player statistics from the team websites to generate the following lists.

- Height of Basketball Team Players in inches for 2010 Season: 75, 73, 76, 78, 79, 78, 79, 81, 80, 82, 81, 84, 82, 84, 80, 84
- Height of Soccer Team Players in inches for 2010: 73, 73, 73, 72, 69, 76, 72, 73, 74, 70, 65, 71, 74, 76, 70, 72, 71, 74, 71, 74, 73, 67, 70, 72, 69, 78, 73, 76, 69

To compare the data sets, Jason creates a two dot plots on the same scale.





Describe any overlap between the dot plots and determine the difference between their measures of center.

Standard S-ID.A.1.
Represent data with plots on the **REAL** number line (dot plots, histograms, and box plots). *(Content Emphases: Additional)*

Prerequisite Examples:

Grade 7

7.SP.B.4

The two data sets below depict random samples of the management salaries in two companies. Based on the salaries below which measure of center will provide the most accurate estimation of the salaries for each company?

- Company A: 1.2 million, 242,000, 265,500, 140,000, 281,000, 265,000, 211,000
- Company B: 5 million, 154,000, 250,000, 250,000, 200,000, 160,000, 190,000

Grade 6

6.SP.B.4

Grade 6 students were collecting data for a math class project. They decided they would survey the other two grade 6 classes to determine how many DVDs each student owns. A total of 48 students were surveyed. The data are shown in the table below in no specific order. Create a data display.

- 11, 21, 5, 12, 10, 31, 19, 13, 23, 33, 10, 11, 25, 14, 34, 15, 14, 29, 8, 5, 22, 26, 23, 12, 27, 4, 25, 15, 7, 2, 19, 12, 39, 17, 16, 15, 28, 16

Grade/Course Level Connections:

S-ID.A.2 and S-ID.A.3

Instructional Considerations:

- A histogram has units of **measurement** of a **numerical** variable on the horizontal axis (e.g., ages with intervals of equal length).
- A bar graph is appropriate when the horizontal axis has **categories** and the vertical axis is labeled by either frequency (e.g., book titles on the horizontal and number of students who like the respective books on the vertical) or measurement of some numerical variable (e.g., days of the week on the horizontal and median length of root growth of radish seeds on the vertical).

Common Misconceptions:

- Students may believe that a bar graph and a histogram are the same.
- Students may believe that the lengths of the intervals of a boxplot - (min,Q1), (Q1,Q2), (Q2,Q3) and (Q3,max) – are related to the number of subjects in each interval. Students should understand that each interval theoretically contains one-fourth of the total number of subjects.

Academic Vocabulary:

dot plots, box plots, histograms, quartiles, median, ranges, bins, intervals

Standards for Mathematical Practice:

(4) Model with mathematics
(5) Use appropriate tools strategically

Standard S-ID.A.2
Use statistics **appropriate to the shape** of the data distribution to **compare** center (median, mean) and spread (interquartile range, **standard deviation**) of **two or more** different data sets. *(Content Emphases: Additional)*

Explanation and/or Examples

Given two sets of data or two graphs, students:

- Use the appropriate measures of center and spread to describe symmetric and skewed distributions
- Compare two or more data sets using measures of center and spread of the data sets.

Example 1:
A statistically-minded state trooper wondered if the speed distributions are similar for cars traveling northbound and for cars traveling southbound on an isolated stretch of interstate highway. He uses a radar gun to measure the speed of all northbound cars and all southbound cars passing a particular location during a fifteen minute period. Here are his results:

- Northbound Cars 60, 62, 62, 63, 63, 63, 64, 64, 64, 64, 65, 65, 65, 65, 66, 66, 67, 68, 70, 83
- Southbound Cars 55, 56, 57, 57, 58, 60, 61, 61, 62, 63, 64, 65, 65, 67, 67, 68, 68, 68, 68, 71

Draw box plots of these two data sets, and then use the plots and *appropriate numerical summaries* of the data to write a few sentences comparing the speeds of northbound cars and southbound cars at this location during the fifteen minute period of time.

Ex. 1 Solution:
Task 1027 by IM by CC BY-NC-SA 4.0

Building Blocks (Grades 6 to 8 prerequisites):

Grade 7

7.SP.B.4
Use measures of center and measures of variability for numerical data from random samples to **draw informal comparative inferences** about two populations.

Grade 6

6.SP.B.5c:
Summarize numerical data sets **in relation to their context**, such as by:

Giving quantitative measures of center (**median and/or mean**) and variability (**interquartile range and/or mean absolute deviation**), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered.

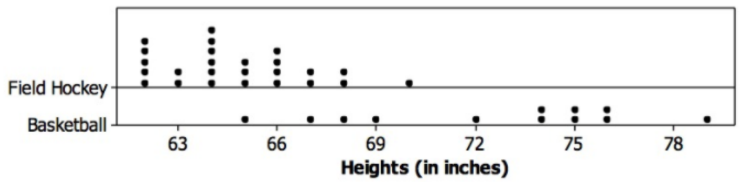
6.SP.B.5d:
Summarize numerical data sets **in relation to their context**, such as by:

Relating the choice of measures of center and variability **to the shape of the data distribution** and the context in which the data were gathered.

Prerequisite Examples:

Grade 7

7.SP.B.4:
The heights of the players on the University of Maryland women's basketball team for the 2012-2013 season and the heights of the players on the women's field hockey team for the 2012 season are represented in the dot plots below.



a. Based on visual inspection of the dot plots, which group appears to have the larger average height? Which group appears to have the greater variability in the heights?
b. Compute the mean and mean absolute deviation (MAD) for each group. Do these values support your answers in part (a)?

Task 1340 by IM by CC BY-NC-SA 4.0

Algebra 1 Instructional Support Tool – Statistics & Probability

Standard S-ID.A.2.
Use statistics **appropriate to the shape** of the data distribution to **compare** center (median, mean) and spread (interquartile range, **standard deviation**) of **two or more** different data sets. *(Content Emphases: Additional)*

Prerequisite Examples:

Grade 6

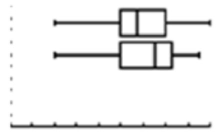
6.SP.B.5c:
You are planning to take on a part time job as a server at a local restaurant. During your interview, the boss told you that their best server, Jenni, typically made $70 a night in tips last week. However, when you asked Jenni about this, she said that she typically made only $50 per night last week. She provides you with a copy of her nightly tip amounts from last week. Determine the mean and median tip amount.

a. Which measure of center is Jenni's boss using to describe the typical tip? Why might he choose this value?
b. Which measure of center did Jenni use? Why might she choose this value?
c. Which measure of center best describes the typical amount of tips per night? Justify your answer.

Jenni's Tips

Day	Tip Amount
Sunday	$50
Monday	$45
Wednesday	$48
Friday	$125
Saturday	$85

6.SP.B.5d:
The boxplots show the distribution of scores on a district writing test of two classes at a school



a. Describe the shape of each data distribution. Is the median an appropriate measure of center for the distributions? Justify your answer.

Grade/Course Level Connections:

S-ID.A.1 and S-ID.A.3

Instructional Considerations:

- Introduce the formula of standard deviation by reviewing previously learned mean absolute deviation.
- The mean and standard deviation are most commonly used to describe sets of data. However, if the distribution is extremely skewed and/or has outliers, it may be best to use the median and the interquartile range to describe the distribution since these measures are not sensitive to outliers.

Common Misconceptions:

Students may believe that the mean and median are interchangeable and that either is appropriate for any data set.

Academic Vocabulary:

shape, skewed, symmetric, center, spread, mean, median, interquartile range, standard deviation, outliers, extreme data points

Standards for Mathematical Practice:

(4) Model with mathematics
(5) Use appropriate tools strategically

Standard S-ID.A.3.
Interpret differences in shape, center, and spread **in the context of the data** sets, accounting for possible effects of extreme data points (outliers). *(Content Emphases: Additional)*

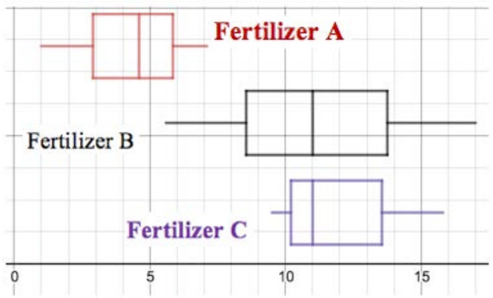
Explanation and/or Examples

Example 1:
Delia wanted to find the best type of fertilizer for her tomato plants. She purchased three types of fertilizer and used each on a set of seedlings. After 10 days, she measured the heights (cm) of each set of seedlings. The data she collected is shown to the right in tabular form and as boxplots.

Write a brief description comparing the three types of fertilizer. Use appropriate statistics in your description.

Seedling Heights

Fertilizer A	Fertilizer B	Fertilizer C
7.1	11.0	10.5
6.3	9.2	11.8
1.0	5.6	15.5
5.0	8.4	14.7
4.5	7.2	11.0
5.2	12.1	10.8
3.2	10.5	13.9
4.6	14.0	12.7
2.4	15.3	9.9
5.5	6.3	10.3
3.8	8.7	10.1
1.5	11.3	15.8
6.2	17.0	9.5
6.9	13.5	13.2
2.6	14.2	9.7



Ex. 1 Sample Solution*:*
The median height for plants receiving fertilizer A is 4.6 cm. *Approximately* 75% of the plants receiving fertilizer A are shorter than the smallest plant that received fertilizer B or fertilizer C. Fertilizer A generally resulted in shorter tomato plants as compared to plants receiving fertilizer B or C. The distribution for fertilizer A is slightly skewed left while the distribution for fertilizer B is approximately normal and the distribution for fertilizer C is strongly skewed right. The median plant height for fertilizer B and fertilizer C are the same (11cm). However, fertilizer C resulted in less variation in plant heights (IQR=3.35) as compared to plants receiving fertilizer B (IQR=5.2). Using the 1.5IQR rule, there are no outliers.

Building Blocks (Grades 6 to 8 prerequisites):

Grade 6

6.SP.B.5c:
Summarize numerical data sets **in relation to their context**, such as by:

Giving quantitative measures of center (**median and/or mean**) and variability (**interquartile range and/or mean absolute deviation**), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered.

Grade 7

7.SP.B.4
Use measures of center and measures of variability for numerical data from random samples to **draw informal comparative inferences** about two populations.

Standard S-ID.A.3
Interpret differences in shape, center, and spread **in the context of the data** sets, accounting for possible effects of extreme data points (outliers). *(Content Emphases: Additional)*

Prerequisite Examples:

Grade 6

6.SP.B.5c:

In Mrs. Sanchez's math classroom, more people sit on the right-hand side of the room than the left. The students on the right-hand side of the classroom received the following scores on an exam worth 100 points: 85, 90, 100, 95, 0, 0, 90, 70, 100, 95, 80, 95.The students on the left received these test scores: 65, 80, 90, 65, 80, 60, 95, 85

a. Make two dot plots of the students' scores, one for each side of the room.
b. Describe any outliers (striking deviations) for each of the dot plots.

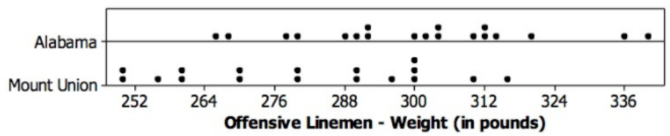
Grade 7

7.SP.B.4:

College football teams are grouped with similar teams into "divisions" (and in some cases, "subdivisions") based on many factors such as game attendance, level of competition, athletic department resources, and so on. Schools from the Football Bowl Subdivision (FBS, formerly known as Division 1-A) are typically much larger schools than schools of any other division in terms of enrollment and revenue. "Division III" is a division of schools with typically smaller enrollment and resources.

One particular position on a football team is called "offensive lineman," and it is generally believed that the offensive linemen of FBS schools are heavier on average than the offensive linemen of Division III schools.

For the 2012 season, the University of Mount Union Purple Raiders football team won the Division III National Football Championship while the University of Alabama Crimson Tide football team won the FBS National Championship. Below are the weights of the offensive linemen for both teams from that season represented with dot plots.



a. Based on visual inspection of the dot plots, which group appears to have the larger average weight? Does one group seem to have greater variability in its weights than the other, or do the two groups look similar?
b. Compute the mean and mean absolute deviation (MAD) for each group. Do your measures support your answers in part (a)?

Grade/Course Level Connections:

S-ID.A.1 and S-ID.A.2

Standard S-ID.A.3.
Interpret differences in shape, center, and spread **in the context of the data** sets, accounting for possible effects of extreme data points (outliers). *(Content Emphases: Additional)*

Instructional Considerations:

- 1.5 x IQR rule: One definition of outlier is any data point more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile.
- The mean and standard deviation are most commonly used to describe sets of data. However, if the distribution is extremely skewed and/or has outliers, it may be best to use the median and the interquartile range to describe the distribution since these measures are not sensitive to outliers.

Common Misconceptions:

- Students may believe that the lengths of the intervals of a boxplot - (min,Q1), (Q1,Q2), (Q2,Q3) and (Q3,max) – are related to the number of subjects in each interval. Students should understand that each interval theoretically contains one-fourth of the total number of subjects.
- Students may believe that all bell-shaped curves are normal distributions. For a bell-shaped curve to be Normal, 68% of the values must lie within one standard deviation of the mean, 95% within two, and 99.7% within three.

Academic Vocabulary:

outliers, extreme data points, skewed, symmetric, normal, bell-shaped, mean, median, interquartile range, standard deviation, Normal distribution

Standards for Mathematical Practice:

(2) Reason abstractly and quantitatively
(6) Attend to precision

Standard S-ID.B.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. *(Content Emphases: Supporting)*

Explanation and/or Examples

The table entries are the *joint frequencies* of how many subjects displayed the respective cross-classified values (e.g. 38 students in 7^{th} grade report English as their favorite subject)

Favorite Subject by Grade					
Grade	**English**	**History**	**Math/Science**	**Other**	**Totals**
7^{th} Grade	38	36	28	14	116
8^{th} Grade	47	45	72	18	182
Totals	85	81	100	32	298

Row totals and column totals constitute the *marginal frequencies*.

- row totals - 116 (7^{th} grade students); 182 (8^{th} grade students);
- column totals - 85 (English); 51 (History); 100 (Math/Science); 32 (Other);
- total number of subjects – 298

Dividing joint and marginal frequencies by the total number of subjects define *relative frequencies*.

Favorite Subject by Grade					
Grade	**English**	**History**	**Math/Science**	**Other**	**Totals**
7^{th} Grade	$\frac{38}{298} = .13$	$\frac{36}{298} = .12$	$\frac{28}{298} = .09$	$\frac{14}{298} = .05$	$\frac{116}{298} = .39$
8^{th} Grade	$\frac{47}{298} = .16$	$\frac{45}{298} = .15$	$\frac{72}{298} = .24$	$\frac{18}{298} = .06$	$\frac{182}{298} = .61$
Totals	$\frac{85}{298} = .29$	$\frac{81}{298} = .27$	$\frac{100}{298} = .33$	$\frac{32}{298} = .11$	$\frac{298}{298} = 1$

Conditional relative frequencies are determined by focusing on a specific row or column of the table and are particularly useful in determining any associations between the two variables.

Conditional relative frequency for **Columns**

Favorite Subject by Grade					
Grade	**English**	**History**	**Math/Science**	**Other**	**Totals**
7^{th} Grade	$\frac{38}{85} = .45$	$\frac{36}{81} = .44$	$\frac{28}{100} = .28$	$\frac{14}{32} = .44$	$\frac{116}{298} = .39$
8^{th} Grade	$\frac{47}{85} = .55$	$\frac{45}{81} = .56$	$\frac{72}{100} = .72$	$\frac{18}{32} = .56$	$\frac{182}{298} = .61$
Totals	$\frac{85}{85} = 1$	$\frac{81}{81} = 1$	$\frac{100}{100} = 1$	$\frac{32}{32} = 1$	$\frac{298}{298} = 1$

Conditional relative frequency for **Rows**

Favorite Subject by Grade					
Grade	**English**	**History**	**Math/Science**	**Other**	**Totals**
7^{th} Grade	$\frac{38}{116} = .33$	$\frac{36}{116} = .31$	$\frac{28}{116} = .24$	$\frac{14}{116} = .12$	$\frac{116}{116} = 1$
8^{th} Grade	$\frac{47}{182} = .26$	$\frac{45}{182} = .25$	$\frac{72}{182} = .39$	$\frac{18}{182} = .1$	$\frac{182}{182} = 1$
Totals	$\frac{85}{298} = .29$	$\frac{81}{298} = .27$	$\frac{100}{298} = .33$	$\frac{32}{298} = .11$	$\frac{298}{298} = 1$

Standard S-ID.B.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. *(Content Emphases: Supporting)*

Explanation and/or Examples

Example 1:
At the NC Zoo, 23 interns were asked their preference of where they would like to work. There were three choices: African Region, Aviary, or North American Region. There were 13 who preferred the African Region, 5 of them were male. There were 6 who preferred the Aviary, 2 males and 4 females. A total of 4 preferred the North American Region and only 1 of them was female.

a. Use the data on the NC Zoo internship to create a two-way frequency table.
b. Using the two-way frequency table from the NC Zoo Internship, calculate the proportion of males who prefer the African Region. Calculate the proportion of females who prefer the African Region
c. How does the proportion of males who prefer the African Region compare to the proportion of females who prefer the African Region?
d. 15% of the paid employees are male and work in the Aviary. How does that compare to the interns who are male and prefer to work in the Aviary? Explain how you made your comparison.

Ex. 1 Solution*:*

a.

Location Preference				
Gender	**African**	**Aviary**	**NA**	**Totals**
Male	5	2	3	10
Female	8	4	1	13
Totals	13	6	4	23

b. The proportion of males who prefer the African Region is 5/10 and the proportion of females who prefer the African Region is 8/13
c. A greater proportion of females (.62) prefer the African Region
d. 15% of the paid employees are male and work in the Aviary. 2 out of 23 interns (approximately 8.7%) are male and prefer to work in the Aviary. Therefore a greater percentage of paid male employees prefer working in the Aviary.

Example 2:
The 54 students in one of several middle school classrooms were asked two questions about musical preferences: “Do you like rock?” “Do you like rap?” The responses are summarized in the table below.

	Likes Rap Music		
Likes Rock Music	**Yes**	**No**	**Totals**
Yes	27	6	33
No	4	17	21
Totals	31	23	54

Is there evidence in this sample of a positive association in this class between liking rock and liking rap? Justify your answer by pointing out a feature of the table that supports it.

Ex. 2 Solution*:*

	Likes Rap Music		
Likes Rock Music	**Yes**	**No**	**Totals**
Yes	.87	.26	.61
No	.13	.74	.39
Totals	1	1	1

Yes, there is evidence of a positive association. Of those who like Rap, 87% like Rock, too. 74% of those who do not like Rap also do not like Rock.

Standard S-ID.B.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. *(Content Emphases: Supporting)*

Building Blocks (Grades 6 to 8 prerequisites):

Grade 8

8.SP.A.4
Understand that **patterns of association** can also be seen in bivariate categorical data by **displaying frequencies and relative frequencies in a two-way table**. **Construct and interpret a two-way table** summarizing data on two categorical variables collected from the same subjects. **Use relative frequencies calculated for rows or columns to describe possible association** between the two variables. *For example, collect data from students in your class on whether or not they have a curfew on school nights and whether or not they have assigned chores at home. Is there evidence that those who have a curfew also tend to have chores?*

Prerequisite Examples:

Grade 8

8.SP.A.4
The table above illustrates the results when 100 students were asked the survey questions: "Do you have a curfew?" and "Do you have assigned chores?" Is there evidence that those who have a curfew also tend to have chores?

	Curfew	
Chores	**Yes**	**No**
Yes	40	10
No	10	40

Sample Solution a: Calculating relative frequencies for rows, 80% of students that have chores have curfew and 80% of those without chores do not have curfew

	Curfew		
Chores	**Yes**	**No**	**Totals**
Yes	.8	.2	1
No	.2	.8	1
Totals	.5	.5	1

Of the students who answered that they have chores, 80% have a curfew. Of the students who answered they did not have chores, 80% did not have curfew. From this sample, there appears to be evidence of a positive association between having a curfew and having chores.

Sample Solution b: Calculating relative frequencies for columns, 80% of those with a curfew have chores and 80% of those without a curfew do not have chores

	Curfew		
Chores	**Yes**	**No**	**Totals**
Yes	.8	.2	.5
No	.2	.8	.5
Totals	1	1	1

Of the students who answered that they had a curfew, 80% had chores. Of the students who answered they did not have a curfew, 80% did not have chores. From this sample, there appears to be evidence of a positive association between having a curfew and having chores.

Standard S-ID.B.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. *(Content Emphases: Supporting)*

Grade/Course Level Connections:

n/a

Instructional Considerations:

- The focus is that two categorical variables are being measured on the same subject.
- Begin with two categories for each variable and represent them in a two-way table with the two values of one variable defining the rows and the two values of the other variable defining the columns. (Extending the number of rows and columns is easily done once students become comfortable with the 2x2 case.)
- When students are proficient with analyzing two-way frequency tables, build upon their understanding to develop the vocabulary.
- Students may use spreadsheets, graphing calculators, and statistical software to create frequency tables and determine associations or trends in the data.

Common Misconceptions:

- Students may believe that row totals and column totals (marginal frequencies) should be the same
- Students may believe that only proportions relative to the total number of subjects can be determined

Academic Vocabulary:

frequency, joint frequencies, marginal frequencies, row totals, column totals, relative frequencies, conditional relative frequency, categorical, proportion

Standards for Mathematical Practice:

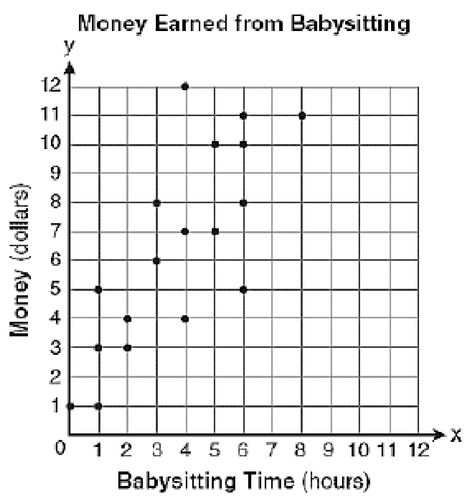
(2) Reason abstractly and quantitatively

(6) Attend to precision

Standard S-ID.B.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** (including the use of technology); 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**, including with the use of technology.
c. **Fit a linear function** for a scatter plot that suggests a linear association *(Content Emphases: Supporting)*

Explanation and/or Examples

Students choose a function suggested by the context to model data.
Example 1:



Which of the following best models the data?

a. $y = x$

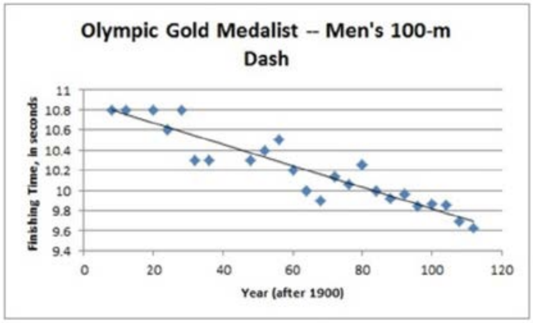
b. $y = \frac{6}{5}x + 2$

c. $y = \frac{3}{2}x + 4$

d. $y = \frac{1}{4}x + 4$

Ex. 1 Solution: **b**

Example 2:
The scatterplot below shows the finishing times for the Olympic gold medalist in the men's 100-meter dash for many previous Olympic games. The line of best fit is also shown.



Is a linear model a good fit for the data? Explain, commenting on the strength and direction of the association.

<u>Ex. 2 Solution</u>:

Standard S-ID.B.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** (including the use of technology); 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**, including with the use of technology.
c. **Fit a linear function** for a scatter plot that suggests a linear association *(Content Emphases: Supporting)*

Explanation and/or Examples

Fit a quadratic function for a scatter plot that suggests a quadratic association.

Example 3:
A study was done to compare the speed x (in miles per hour) with the mileage y (in miles per gallon) of an automobile. The results are shown in the table.

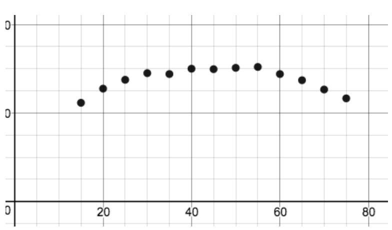
Auto Mileage by Speed

Speed, x (mph)	Mileage, y (mpg)
15	22.3
20	25.5
25	27.5
30	29.0
35	28.8
40	30.0
45	29.9
50	30.2
55	30.4
60	28.8
65	27.4
70	25.3
75	23.3

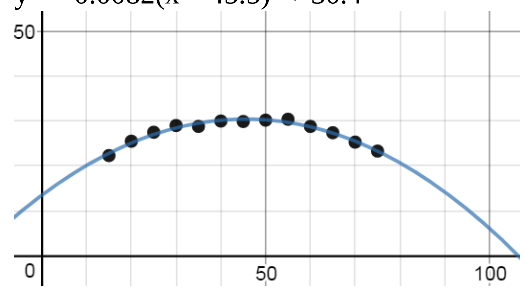
a. Use your calculator or desmos.com to make a scatter plot of the data.
b. Use the regression feature to find a model that best fits the data.
c. Approximate the speed at which the mileage is the greatest.

Ex. 3 Solution*:*

a.



b. $y = -0.0082(x - 45.5)^2 + 30.4$



c. The approximate speed at which the mileage is greatest is 45.5 mph

Note:
A ***residual*** is the difference between the actual y-value and the predicted y-value ($y - \hat{y}$), which is a measure of the error in prediction. (Note: $\hat{y}$ is the symbol for the predicted y-value for a given x-value.) A residual is represented on the graph of the data by the *vertical* distance between a data point (x,y) and the point on the graph of the function (x, $\hat{y}$).

A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Standard S-ID.B.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** (including the use of technology); 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**, including with the use of technology.
c. **Fit a linear function** for a scatter plot that suggests a linear association *(Content Emphases: Supporting)*

Explanation and/or Examples

Students informally assess fit by plotting and analyzing residuals.

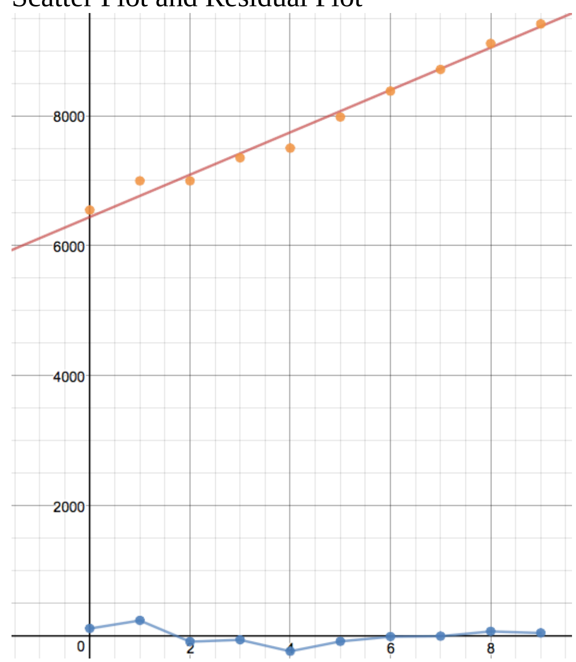
Example 4:

The table below displays the annual tuition rates of a state college in the U.S. between 1990 and 2000, inclusively. The linear function $(t) = 326x + 6440$ has been suggested as a good fit for the data. Use a residual plot to determine the goodness of fit of the function for the data provided in the table.

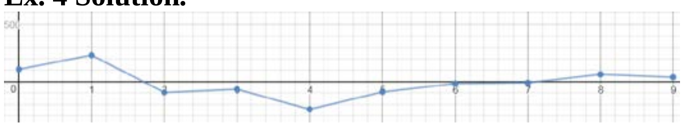
Tuition Rate over time

Year (0=1990)	Tuition Rate
0	6546
1	6996
2	6996
3	7350
4	7500
5	7978
6	8377
7	8710
8	9110
9	9411

Scatter Plot and Residual Plot



Ex. 4 Solution:



Building Blocks (Grades 6 to 8 prerequisites)

Grade 8

8.SP.A.1
Construct and interpret scatter plots for bivariate measurement data to **investigate patterns of association** between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association

8.SP.A.2
Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, **informally fit a straight line, and informally assess the model fit** (e.g. line of best fit) **by judging the closeness of the data points to the line.**

Standard S-ID.B.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** (including the use of technology); 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**, including with the use of technology.

c. **Fit a linear function** for a scatter plot that suggests a linear association *(Content Emphases: Supporting)*

Building Blocks (Grades 6 to 8 prerequisites)

8.SP.A.3

Use the equation of a linear model to **solve problems in the context of bivariate measurement data.** *For example, in a linear model for a biology experiment, interpret a slope of 1.5 cm/hr as meaning that an additional hour of sunlight each day is associated with an additional 1.5 cm in mature plant height*

Prerequisite Examples:

Grade 8

8.SP.A.1

Data from a local fast food restaurant is provided showing the number of staff members and the average time for filling an order are provided in the table shown. Describe the association between the number of staff and the average time for filling an order.

Staff Efficiency

Number of staff	Average time to fill order (seconds)
3	180
4	138
5	120
6	108
7	96
8	84

8.SP.A.2

The capacity of the fuel tank in a car is 13.5 gallons. The table at right shows the number of miles traveled and how many gallons of gas have been used. Describe the relationship between the variables. If the data is linear, determine a line of best fit. Do you think the line represents a good fit for the data set? Why or why not?

Fuel Efficiency

Miles Traveled	Gallons Used
0	0
75	2.3
120	4.5
160	5.7
250	9.7
300	10.7

8.SP.A.3

Given the data from students' math scores and absences, a linear model that can be used to predict a student's math score for a given number of absences is constructed. The linear model is $s = -\frac{25}{3}a + 95$ where a represents the number of absences and s represents the predicted score.. Determine the math score that a student with 4 absences can expect.

Standard S-ID.B.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** (including the use of technology); 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**, including with the use of technology.
c. **Fit a linear function** for a scatter plot that suggests a linear association *(Content Emphases: Supporting)*

Grade/Course Level Connections:

F-LE.A.2, S-ID.C.7 and S-ID.C.8

Instructional Considerations:

- A **residual** is the difference between the actual y-value and the predicted y-value ($y - \hat{y}$), which is a measure of the error in prediction. (Note: $\hat{y}$ is the symbol for the predicted y-value for a given *x*-value.) A residual is represented on the graph of the data by the *vertical* distance between points (x,y) and (x, $\hat{y}$).
- A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate

Common Misconceptions:

- Students may believe that residual plots should show a pattern of some sort. Just the opposite is the case.
- Students may believe that predicted values are precise values.
- Students may believe that there is only one correct possible function representation.

Academic Vocabulary:

predicted value, residual, residual plot, scatter plot, quantitative variables, linear association

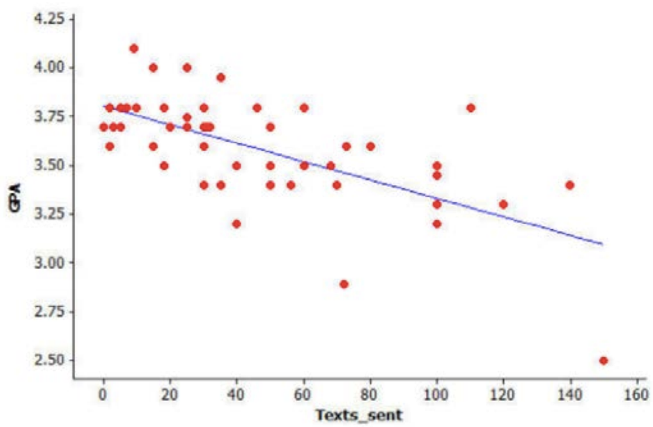
Standards for Mathematical Practice:

(4) Model with mathematics
(6) Attend to precision

Standard S-ID.C.7
Interpret the slope (rate of change) and the intercept (constant term) of a linear model **in the context** of the data. *(Content Emphases: Major)*

Explanations and/or Examples

Example 1:
Medhavi suspects that there is a relationship between the number of text messages high school students send and their academic achievement. To explore this, she asks a random sample of 52 students at her school how many text messages they sent yesterday and what their grade point average (GPA) was during the most recent marking period. Her data are summarized in the scatter plot at right. The line of best fit is also shown.



a. Use the graph to find the equation for the line of best fit
b. What do the slope and y-intercept of the line of best fit represent in the context of the problem

Ex. 1 Solution:

Building Blocks (Grades 6 to 8 prerequisites):

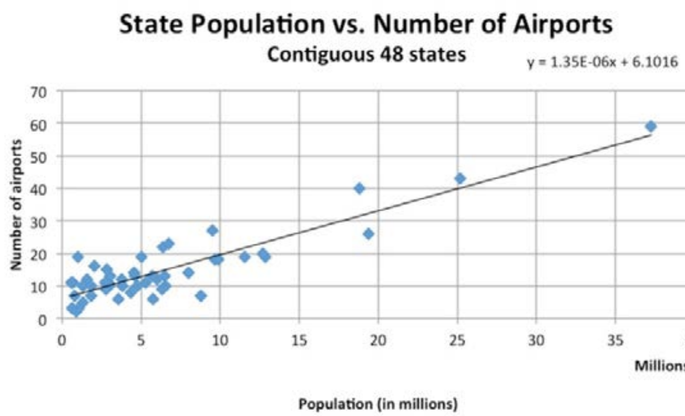
Grade 8

8.SP.A.3
Use the equation of a linear model to solve problems in the context of bivariate measurement data, **interpreting the slope and intercept.** *For example, in a linear model for a biology experiment, interpret a slope of 1.5 cm/hr as meaning that an additional hour of sunlight each day is associated with an additional 1.5 cm in mature plant height*

Prerequisite Examples

Grade 8

8.SP.A.3
The scatter plot below shows the relationship between the number of airports in a state and the population of that state according to the 2010 census. Each dot represents a single state.



LaToya uses the function $y = (1.35\times10^{-6})x + 6.1$ to model the relationship between the number of airports, y and the population in a state, x.

c. What does the number 6.1 that appears in LaToya's function mean in the context of airports vs. populations?
d. What does the number 6.1 that appears in LaToya's function mean in the context of airports vs. populations?

Standard S-ID.C.7
Interpret the slope (rate of change) and the intercept (constant term) of a linear model **in the context** of the data. *(Content Emphases: Major)*

Grade/Course Level Connections

S-ID.B.6c and F-LE.B.5

Instructional Considerations:

- Note that unlike a two-dimensional graph in mathematics, the scales of a scatterplot need not be the same, and even if they are similar (such as SAT Math and SAT Verbal), they still need not have the same spacing. So, visual rendering of slope makes no sense in most scatterplots, i.e., a 45 degree line on a scatterplot need not mean a slope of 1.
- Often the interpretation of the intercept (constant term) is not meaningful in the context of the data. For example, this is the case when the zero point on the horizontal is of considerable distance from the values of the horizontal variable, or in some case has no meaning such as for SAT variables.

Common Misconceptions:

- Students may believe that the slope of a linear function is a number used only to sketch the graph of the line. The idea of slope as a rate of change is especially important, and extends to understanding concepts in calculus and geometry.
- Students may believe that one variable *causes* the other to occur, especially in a linear relationship.

Academic Vocabulary:

rate of change**,** slope, linear models

Standards for Mathematical Practice:

(2) Reason abstractly and quantitatively
(6) Attend to precision

Standard S-ID.C.8
Compute (using technology) and **interpret** the correlation coefficient of a linear fit. *(Content Emphases: Major)*

Explanation and/or Examples

Example 1:
A study compared the number of years of education a person received and that person's average yearly salary. It was determined that the relationship between these two quantities was linear and the correlation coefficient was 0.91. Which conclusion can be made based on the findings of this study?

a. There was a weak relationship.
b. There was a strong relationship.
c. There was no relationship.
d. There was an unpredictable relationship.

***Ex. 1 Solution*: b**

Example 2:
The correlation coefficient of a given data set is 0.97. List three specific things this tells you about the data.

Ex 2. Sample Solution:
1. Since this number is nonzero, there is a correlation.
2. Sine this number is positive, the two variables move in the same direction, i.e. as one increases, so does the other; as one decreases, so does the other.
3. Since this number is close to 1, the positive correlation is strong

Example 3:
As shown in the table, a person's target heart rate during exercise changes as the person gets older.

Target Heart Rate by Age

Age (years)	**Target Heart Rate** (beats per minute)
20	135
25	132
30	129
35	125
40	122
45	119
50	115

Which value represents the linear correlation coefficient, rounded to the nearest thousandth, between a person's age, in years, and that person's target heart rate, in beats per minute?
a. -0.999
b. -0.664
c. 0.998
d. 1.503

***Ex. 3 Solution*: a**

Building Blocks (Grades 6 to 8 prerequisites):

Grade 8
8.SP.A.1
Construct and interpret scatter plots for bivariate measurement data to **investigate patterns of association** between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.

Standard S-ID.C.8.
Compute (using technology) and **interpret** the correlation coefficient of a linear fit. *(Content Emphases: Major)*

Prerequisite Examples
Grade 8

8.SP.A.1
Data for 10 students' math and science scores are provided in the chart.
Describe the association between the math and science scores.

Math and Science Scores

Student	Math	Science
1	64	68
2	50	70
3	85	83
4	34	33
5	56	60
6	24	27
7	72	74
8	63	63
9	42	40
10	93	96

Grade/Course Level Connections:

S-ID.B.6c

Instructional Considerations:
- The correlation coefficient, typically denoted by r, must be between -1 and +1, inclusively ($-1 \leq r \leq 1$)
- The correlation coefficient may indicate a weak positive, strong positive, weak negative, strong negative, or no correlation.
- Students should assess the impact of outliers on the correlation coefficient in order to understand that an outlier in a data set can make the correlation look weaker or stronger than it actually is.

Common Misconceptions:
- Students may believe that outliers have no effect on the correlation coefficient and the line of best fit.
- Students may believe that a negative correlation coefficient is problematic.
- Students may believe that the values of the slope and the correlation coefficient are interchangeable

Academic Vocabulary:

correlation, correlation coefficient, linear fit, linear models

Standards for Mathematical Practice:

(2) Reason abstractly and quantitatively
(6) Attend to precision

Standard S-ID.C.9.
Distinguish between correlation and causation. *(Content Emphases: Major)*

Explanation and/or Examples

Example 1:
Diane did a study for a health class about the effects of a student's semester math exam score on height. Based on a graph of her data, she found that there was a direct relationship between students' math scores and height. She concluded that "doing well on your semester math exam makes you tall." Is this conclusion justified? Explain any flaws in Diane's reasoning.

Ex 1. Sample Solution:
Though math exam scores and height may be related, this does not mean that one causes the other. So, the flaw in Diane's reasoning centers on confusion between correlation and causation. A possible explanation of the relationship may be that the tall students in the sample are all college athletes and because of their demanding practice and competition schedules, are required to keep study hours. During these study hours, a graduate tutor is on hand to answer questions. Therefore, this could explain the higher math test scores associated with these tall students, but it does not mean that "If you do well on your math exam, then you will be tall."

Building Blocks (Grades 6 to 8 prerequisites):
n/a

Prerequisite Examples
n/a

Grade/Course Level Connections:
S-ID.C.8

Instructional Considerations:
- Noting that a correlated relationship between two quantitative variables is not causal (unless the variables are in an experiment) is a very important topic and a substantial amount of time should be spent on it.
- A causal relationship among correlated observances cannot be concluded when causality was not even considered by the study itself.
- Determining causation requires a controlled randomized experiment.

Common Misconceptions:
- Students may believe that when two quantitative variables are related, i.e., correlated, that one causes the other to occur. Causation is not necessarily the case. For example, at a theme park, the daily temperature and number of bottles of water sold are demonstrably correlated, but an increase in the number of bottles of water sold does not cause the day's temperature to rise or fall.
- Students may believe that a positive correlation coefficient is evidence of a causal relationship.
- Students may believe that strong correlation means strong causation

Academic Vocabulary:
causation

Standards for Mathematical Practice:
(2) Reason abstractly and quantitatively
(6) Attend to precision