Part 6

Statistics and Probability

Statistics and Probability

Conceptual Category Overview

Standards in this conceptual category relate to either statistics, probability, or a connection between the two. Earlier domains specifically address statistical topics, such as interpreting categorical and quantitative data (S.ID) and making inferences and justifying conclusions (S.IC) from both numerical and graphical approaches. Later domains emphasize mathematical concepts of probability. These include the domains of Conditional Probability and the Rules of Probability (S.CP) and Using Probability to Make Decisions (S.MD). Contrary to traditional statistics courses and standards, the Common Core Standards for Mathematics place less emphasis on the normal distribution but place more emphasis on simulation and its relationship to probability and inference. Statistics and probability rely heavily on the relationship between chance outcomes and inference. Students use simulation to gain perspective on making and justifying inferences while other simulations are more discrete in nature, pushing students to find expectations of different events using probability as justification.

An overarching theme within these standards is the need for and use of technology to meet standards. Teachers and students can use technology to create displays of data and calculate summary statistics. Given this technology, the standards emphasize less the actual calculation of summary values or statistics but rather focus on their comparisons, interpretation, and proper use. Terminology in the standards requires students to interpret best-fit lines, conduct hypothesis testing through simulation, and experiment with data to make inferences.

Direct Connections to Statistics and Probability in Middle Grades

Statistics and probability at the high school level build on initial middle grades standards. In Grade 6, students describe a univariate data set using different displays, find ways to describe the center of a data set, and begin to understand variability in univariate distributions. Building on these experiences, students in high school will use measures of center and variation to create margins of error and test for statistical differences.

In Grade 7, students expand on their experiences to make comparisons between two different univariate distributions. Students at the high school level will use understandings from this level as a foundation for understanding and making statistical inferences using simulations. Teachers introduce students to probabilistic concepts in Grade 7, which continue to be developed in the high school standards. Students in Grade 7 investigate chance processes and develop, use, and evaluate probabilistic models. They learn to use tree diagrams, organized lists, tables, and simulation to determine the probability of compound events. The high school standards expand the Grade 7 standards to include understanding of independent variables and of conditional probability. They will also use ideas developed in Grade 7, such as the fundamental counting principle, to make sense of permutations and combinations.

In Grade 8, students investigate patterns of association in bivariate data both in graphs and in tables. Students may describe patterns of association with lines and will explain the slopes and intercepts of the lines in context of the data. Students at the high school level will extend these standards by using the least squares regression line, analyzing residuals in best-fit models, interpreting the correlation coefficient, and expanding their concept of function of best fit to include nonlinear functions.

SUGGESTED MATERIALS

Technology is a significant part of statistics and probability because it can generate plots, regression functions, and correlation coefficients rapidly. Technology is also pivotal because the standards rely on simulation of many possible outcomes to make inference.

S.ID	S.IC	S.CP	S.MD	
✓	✓			Statistical software for displaying data sets and calculating summary statistics.
✓				Calculators, spreadsheets, and/or tables to estimate areas under the normal curve.
✓	✓			Multiple data sets to engage students in making and testing conjectures.
	✓			Technology to perform statistical simulations, simulation of differences, or randomizations.
		✓	✓	Manipulatives to simulate random events, such as spinners, dice, etc.
		✓		Resources and data sets from various media outlets.
			✓	Technology to investigate effects of decisions related to probability over multiple trials.
			✓	Applets to investigate long-run frequency and its relationship to expected value.

STATISTICS AND PROBABILITY—OVERARCHING KEY VOCABULARY

S.ID	S.IC	S.CP	S.MD	
	✓		✓	**Center** – Often described as the mean or median of a data set or distribution but not limited to these procedures. It can be considered a balance or middle point of the distribution.
✓	✓	✓	✓	**Correlation** – A statistical relationship between two random variables or two sets of data. Correlation *typically* refers to relationships in quantitative variables while association refers to relationships between categorical variables or between a categorical variable and a quantitative variable. related words: *association*, *dependence*
✓	✓		✓	**Dispersion** – Denotes how stretched or squeezed a distribution is graphically. Common examples of measures of statistical dispersion are the variance, standard deviation, mean absolute deviation, and interquartile range. These measures, excluding the interquartile range, are often interpreted as the average distance data points are from a given center. related words: *variability*, *spread*
✓	✓		✓	**Distribution** – A representation of a variable that tells us what values the variable takes and how often it takes these values.
	✓	✓	✓	**Expected value** – A measure of the center of the distribution. It is often described as the mean value or average of a data set. In a probability distribution, the expected value is the weighted average of all possible values. To find the expected value of a discrete random variable, multiply each possible value by its probability, then add all of the products over the entire sample space or find $\sum x_i$ and $P(x_i)$.
✓	✓	✓	✓	**Probability distribution** – A distribution that assigns a probability to each measurable subset of the possible outcomes of a random experiment, survey, or procedure of statistical inference.

STATISTICS AND PROBABILITY—OVERARCHING KEY VOCABULARY

S.ID	S.IC	S.CP	S.MD	
✓	✓		✓	**Shape** – Distributional shape is described by its number of peaks and by its possession of symmetry, its tendency to skew, or its uniformity. A unimodal or bimodal data set is a data set that has one or two peaks, relatively.
	✓	✓	✓	**Simulation** – A process used to model random events, such that simulated outcomes closely match random experimental outcomes. By observing simulated outcomes, students should gain insight into random phenomena. Commonly induced phenomena include repeated samples from a population or differences between two treatment groups to obtain a margin of error.

Interpreting Categorical and Quantitative Data (S.ID)

Domain Overview

Standards in S.ID focus on the understanding of univariate and bivariate data both numerically and graphically. Univariate data can be either categorical (qualitative) or quantitative. Teachers and students represent categorical data with bar graphs and quantitative data with stem-and-leaf plots, dot plots, and histograms. Bivariate data consists of observational matches between either two quantitative, two categorical, or a quantitative and categorical measurement. Teachers and students use scatterplots and their associated statistics to interpret relationships between two quantitative measures, and use two-way frequency tables to represent the relationship between two categorical variables. Teachers and students use side-by-side bar graphs and box plots for comparison between two quantitative variables by their categorical features.

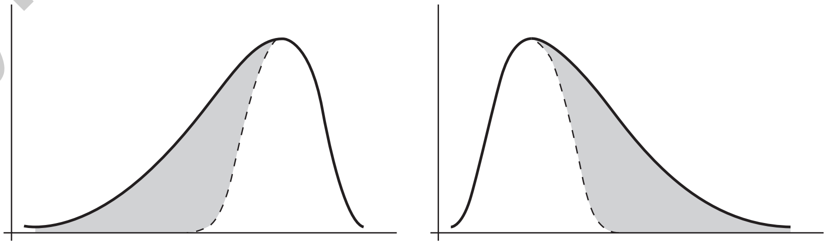
S.ID—KEY VOCABULARY

S.ID.A	S.ID.B	S.ID.C	
✓			**Bin width** – A value that divides a data set into a series of consecutive, non-overlapping intervals. The bin width is the width of a rectangle in a histogram.
	✓	✓	**Bivariate data** – A data set in which two variables are matched between one observational unit such that correlation can be analyzed between the two variables. Example: The height and arm span of students in which the data is matched by student.
✓	✓		**Boxplot** – A standardized way of displaying the distribution of data based on the five-number summary: minimum, first quartile, median, third quartile, and maximum. This data display is particularly useful for observing spread by observing the size of each quartile.
✓	✓	✓	**Center** – Often described as the mean or median but not limited to these procedures. It is a balance or middle point of the distribution.
		✓	**Central limit theorem** – A theorem stating that given a distribution with a mean μ and variance σ^2 or standard deviation σ, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance $\frac{\sigma^2}{n}$ or standard error $\frac{\sigma}{\sqrt{n}}$ as n, a sample size, increases.
	✓	✓	**Correlation** – A statistical relationship between two random variables or two sets of data. Correlation *typically* refers to relationships in quantitative variables while association refers to relationships between categorical variables. related words: *association, dependence*
	✓	✓	**Correlation coefficient** – A number that illustrates a quantitative measure of some type of correlation and dependence. The value *r*, Pearson's correlation coefficient, is most commonly used to represent this relationship between two quantitative variables.

S.ID—KEY VOCABULARY

<table>
<tr><th>S.ID.A</th><th>S.ID.B</th><th>S.ID.C</th><th></th></tr>
<tr><td>✓</td><td></td><td></td><td><b>Dispersion</b> – Denotes how stretched or squeezed a distribution is graphically. Common examples of measures of statistical dispersion are the variance, standard deviation, mean absolute deviation, and interquartile range. Dispersion is often interpreted as the average distance data points are from a given center. related words: <i>variability, spread</i></td></tr>
<tr><td>✓</td><td>✓</td><td>✓</td><td><b>Distribution</b> – A representation of a variable that tells us what values the variable takes and how often it takes these values.</td></tr>
<tr><td>✓</td><td></td><td></td><td><b>Dotplot</b> – A graph that assesses the distribution of quantitative data. Each dot above the number line represents an observation from the data. (It sometimes is also called a line plot.)</td></tr>
<tr><td>✓</td><td></td><td></td><td><b>Empirical rule</b> – A rule that describes the distribution of data values in relationship to the normal curve. Data from a normal distribution will have 99.7% of its values between 3 standard deviations above the mean and 3 standard deviations below the mean; 95% of its values between 2 standard deviations above the mean and 2 standard deviations below the mean; and 68% of its values between 1 standard deviation above the mean and 1 standard deviation below the mean.</td></tr>
<tr><td></td><td>✓</td><td></td><td><b>Extrapolation</b> – The process of estimating the value of a variable on the basis of its relationship with another variable beyond the original observation range.</td></tr>
<tr><td></td><td>✓</td><td></td><td><b>Frequency</b> – Generally, a count of the number of occurrences of a random variable. This may consist of counts on one variable or multiple variables. To help illustrate the difference between different types of frequencies in a two-way (or contingency) table, the following table represents two categorical variables illustrating the distribution of red and white cars and trucks in a parking lot.
<table>
<tr><th></th><th>Red</th><th>White</th><th>Total</th></tr>
<tr><td><b>Car</b></td><td>18</td><td>12</td><td>30</td></tr>
<tr><td><b>Truck</b></td><td>22</td><td>10</td><td>33</td></tr>
<tr><td><b>Total</b></td><td>40</td><td>22</td><td>63</td></tr>
</table>
Each set of numbers in a two-way table has a specific name. The numbers in the middle cells are the <b>joint frequency</b>. These are the frequency of both events happening such as there were 18 cars that were red. The numbers in the total row and column are called the <b>marginal frequencies</b>. These numbers describe the totals for one particular categorical variable.<b>Conditional relative frequency</b> numbers are the ratio of a joint relative frequency and its related marginal relative frequency. For example, the conditional relative frequency a vehicle is white given that it is a car is $\frac{12}{30}$.</td></tr>
<tr><td>✓</td><td>✓</td><td></td><td><b>Histogram</b> – A graphical representation of a distribution consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to a given interval. It can be used to show an estimate of the probability distribution of a continuous quantitative variable.</td></tr>
<tr><td>✓</td><td>✓</td><td></td><td><b>Interquartile range (IQR)</b> – A measure of spread, based on dividing a data set into quartiles, which divide a data set arranged in order into four equal parts. The values that divide each part are called the first, second, and third quartiles, denoted by Q1, Q2, and Q3, respectively. The IQR is Q3–Q1.</td></tr>
<tr><td></td><td>✓</td><td></td><td><b>Joint frequency</b> – See frequency.</td></tr>
<tr><td></td><td>✓</td><td></td><td><b>Marginal frequency</b> – See frequency.</td></tr>
<tr><td>✓</td><td>✓</td><td></td><td><b>Mean</b> – The arithmetic <b>mean</b>, or average, is the sum of a collection of numbers divided by the number of numbers in the collection.</td></tr>
</table>

S.ID—KEY VOCABULARY

S.ID.A	S.ID.B	S.ID.C	
✓	✓		**Median (Q2)** – A number that represents the middle of the data set. The median is found by arranging the data set into numerical order. The median is the number that represents the halfway point of the data set.
✓			**Normal distribution** (also known as normal curve) – A naturally occurring distribution that is symmetric about the mean, bell shaped, and dispersed systematically. This distribution relates to sampling distributions and sampling variation. related words: *bell curve*
✓		✓	**Outliers** – Extreme values in a univariate data set identified by using the interquartile range (IQR). One method of identifying outliers is to consider any data larger than Q3 + 1.5 · IQR and less than Q1 – 1.5 · IQR as an outlier. Visually, this is data that appear 1.5 times farther than the ends of the box on the box plot. Outliers may also be present in bivariate data sets; however, identification using mathematical procedures is not necessary in the standards.In addition to outliers, *unusual data* points can influence data set descriptions and summaries. These data points are inconsistent with other data points and may be identified as possibly influential.
✓	✓		**Probability distribution** – A distribution that assigns a probability to each measurable subset of the possible outcomes of a random experiment, survey, or procedure of statistical inference.
✓	✓	✓	**Categorical variables** – Data that takes on values that are names or labels and not numerical. Examples include eye color, hair color, gender, etc. related words: *qualitative variable*
✓	✓	✓	**Quantitative variables** – Data that is numerical and represents a measurable quantity. Examples include height of students, a state's population, or fuel efficiency of a car.
	✓		**Residual** – The difference between the observed dependent variable and the estimated (predicted) value for the independent variable. If the predicted model is called *f(x), x* is the observed independent variable, and *y* is the observed dependent variable, the residual is represented by *y* – *f(x)* = residual.
✓	✓	✓	**Shape** – Distributional shape is described by its number of peaks and by its possession of symmetry, its tendency to skew, or its uniformity. A unimodal or bimodal data set is a data set that has one or two peaks, relatively.
✓			**Skewness** – A measure of the asymmetry of the probability distribution of a variable about its mean. Negative skew is often referred to as left skew and positive skew is often referred to as right skew.  Negative skew Positive skew
✓			**Standard deviation (σ)** – A measurement of dispersion that measures how far each number in the data set is from the mean. It is the square root of the variance.
✓	✓		**Univariate data** – A data set that is described by "one variable" (one type of data). Example: the height of students in class.
✓			**Variance (σ^2)** – A measure of dispersion that can be described as a holistic measure of how far each number in the data set is from the mean. Variance is calculated by squaring the difference between each number in the set and the mean, then dividing the sum of the squares by the number of values in the set.

Statistics and Probability | Interpreting Categorical and Quantitative Data

S.ID.A

Cluster A

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

STANDARD 1	**S.ID.A.1:** Represent data with plots on the real number line (dot plots, histograms, and box plots).
STANDARD 2	**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.
STANDARD 3	**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).
STANDARD 4	**S.ID.A.4:** Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

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

Students find and interpret summary information representing a univariate data set, including measures of central tendency and spread. Students develop values of spread, building on the conceptual foundations of mean absolute deviation from middle grades content and calculated using technology such as calculators, computer software designed for student learning, and/or tables. Students may find relationships between the mean and median of a data set and may also describe when each is appropriate for use based on characteristics of the distribution. In addition, students justify the placement of means and median based on spread measurements, distributional shape, and extreme values in the data set. Teachers should emphasize the connection of the mean and standard deviation for symmetric distributions and the median and inter-quartile range for non-symmetric distributions or distributions with outliers. After students have related and developed measures of center and spread, students relate these measures to the number of observations and percentages under a normal curve. Though the empirical rule is not addressed in the standard, this antecedent can serve as a conceptual framework to investigate the relationship between population percentages under the normal curve using technology.

Standards for Mathematical Practice

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

Students begin this cluster by using different display models to represent quantitative data. They attend to precision in their choice of display by justifying their reasoning for use of each model. Teachers should have students use software such as NCTM's Core Math Tools, Tinkerplots, and other statistical software to look for and express repeated reasoning for the effects of skewed data sets and extreme values on their measures of center. Students may also justify and express relationships between different measures by the structure for finding the measures of center and dispersion using quantitative reasoning. Teachers need to give students opportunities to construct viable arguments for the use of certain measures of center and analyze the reasoning of others during this process. Students should use calculators or other display tools such as Core Math Tools to build understanding of the relationship between measures of spread and center to their quantitative measures.

Students use quantitative reasoning to relate the area under a distribution to a probability of occurrence, particularly the normal distribution in Standard 4. Students should use the symmetric structure of the normal distribution to find the probability less than, greater than, or between given values. Students should use tools such as graphing calculators, spreadsheets, and statistical software strategically to build their understanding and calculation of the relationship between data displays and probability of occurrence under the normal curve. Students do this by counting the number of occurrences between given events and dividing by the total. Students need to attend to precision in using and describing the probability of occurrence for what events should and should not be included. For example, $P(-2 > x > 2)$, where x is a normally distributed random variable with mean 0 and standard deviation 1, would not include values of the variable that are between −2 and 2. Thus, using the empirical rule, students could calculate $P(-2 > x > 2)$ as $1 - .95 = .05$.

Related Content Standards

S.ID.B S.IC.A S.MD.A 6.SP.A 6.SP.B

Notes

STANDARD 1 (S.ID.A.1)

Represent data with plots on the real number line (dot plots, histograms, and box plots).

Students use dot plots, histograms, and boxplots to represent quantitative data collected from their world or through purposely given data sets. Though these graphs are separate, graphing them simultaneously on one number line builds a strong foundation and understanding of each.

Students describe features that each data representation possesses for the description of a particular distribution. Histograms are useful to represent distributional shape but limit the ability to determine exact measures of center and spread. The shape of histograms may change based on the chosen bin widths. Dot plots allow for the calculation of summary statistics but are tedious to draw and, if represented by numbers that are not integers, may be difficult to summarize. Boxplots can emphasize the spread of a distribution by comparing the size of quartiles and can help with the understanding of skewness of a data set while being plotted simultaneously with histograms and dot plots.

What the TEACHER does:

- Provides opportunities for students to make sense of different data sets themselves and relate them to their context.
- Provides opportunities to relate different displays of the same data set to determine the usefulness and relationship of each.
- Chooses purposeful data sets that can highlight certain relationships to other standards within the domain.

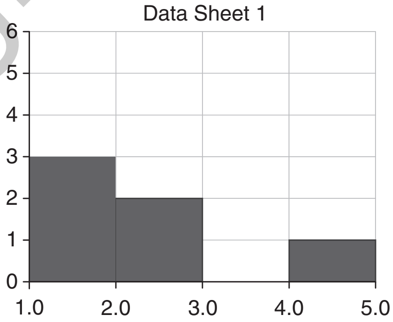
What the STUDENTS do:

- Produce dot plots, histograms, and box plots, either by hand or with technology.
- Explore different data sets and determine their usefulness in prediction and description.
- Explain what the display is telling the viewer in the context of the situation.
- Understand how extreme cases are represented based on the creation of each display.

Addressing Student Misconceptions and Common Errors

When students collect data from their class and represent it in dot plots, they typically like to remember or recall their own input to the data collection. When this happens, students are less likely to view data sets as a whole to generalize about a population and describe data features holistically. Generally, don't allow different colors or initials when plotting points as a class to help in generalizing data summaries.

When creating histograms, students have difficulty in interval and scale creation. Teachers should emphasize the difference between histograms as plots of continuous data and bar graphs as categorical data displays. They should clarify whether inclusion of observations at the interval are included in the adjacent left or right bars of the histogram. Inclusion of observations at interval values is different and can be altered in many statistical software packages but typically is inclusive on the right adjacent bar. For example, the data set {1, 1, 1, 2, 2, 4} is represented below as a histogram.

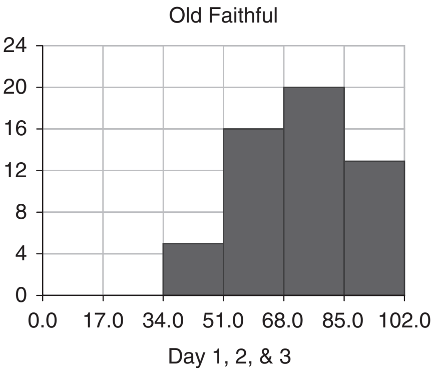


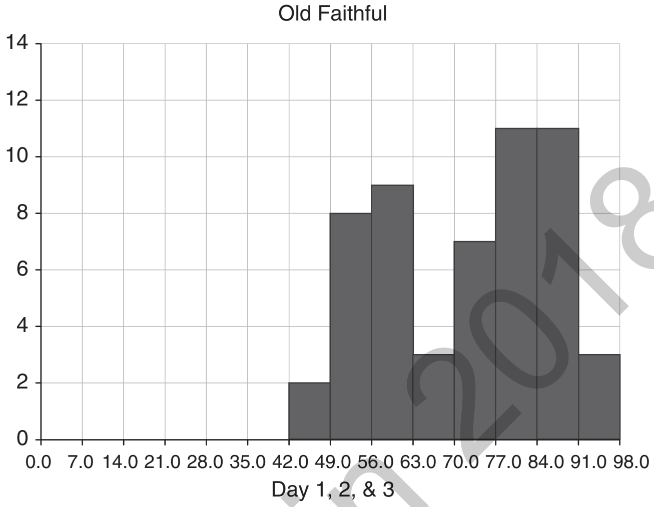
Students often use means instead of medians when plotting or creating boxplots. They often find difficulty in dividing a data set into quartiles when quartile measurements are not observations from the data set. It is important for teachers to emphasize students' conceptual understanding of order-based statistics, such as the boxplot, IQR, and median, in contrast to the mean and standard deviation, using multiple graphical representations of data sets.

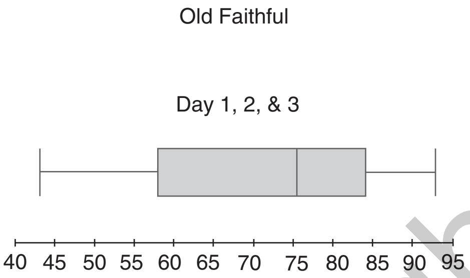
Connections to Modeling

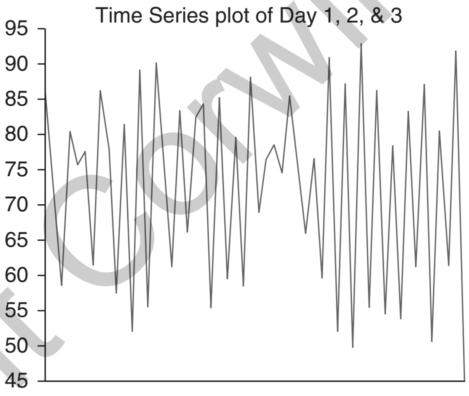
Students collect real-world data, use simulations to create data about the real world, or use purposeful data sets to highlight specific statistical ideas. They model these distributions with a range of different graphic displays, finding usefulness for each in different situations. The data that follow are the wait times between eruptions of the geyser “Old Faithful.”

Example: **Describe the distribution of the “Old Faithful” data set.**









Using a histogram, students may see one (unimodal) or two peaks (bimodal). Students with larger bin widths may fail to notice the unlikely occurrence of the mean. Based on the bimodal and symmetric histogram, the center of the entire distribution is less likely than the first and third quartiles. Students using a boxplot to represent the distribution may be more likely to notice a slight left skew in the data set. Though a time series plot is not required in the standard, it provides detail into the relationship between high wait times and low wait times in the “Old Faithful” data set that other plots fail to represent.

Related Content Standards

S.ID.A.2 S.ID.A.3 S.ID.A.4 S.ID.B.5 6.SP.A 6.SP.B

Notes

STANDARD 2 (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.

The overarching purpose of statistics is to use data to summarize, compare, and predict. This particular standard focuses on the issue of comparison between different data sets. Students are required to understand the differences between the centers and spreads of two distributions. In some cases, summary information may be similar or different based on the context of the data set; thus, students are required to use and justify appropriate measures of center and spread.

What the TEACHER does:

- Provides multiple data sets for comparison that have equal centers but different measures of dispersion.
- Provides multiple data sets for comparison that have equal measures of variance and different measures of center.
- Provides data sets in which appropriate measures of center are medians or means and measures of dispersion are IQR and standard deviation respectively.
- Ensures students justify parameters based on shapes of distributions.

What the STUDENTS do:

- Describe distributions in terms of shape, center, dispersion, and unusual features that will connect previous standards of absolute mean deviation from middle grades to standard deviation and standard deviation of distributions.
- Justify appropriateness of measures of center based on distributional shape and unusual features.
- Relate the appropriate measures of dispersion by the best method for measure of center because measures of dispersion are described by the measure of center.

Addressing Student Misconceptions and Common Errors

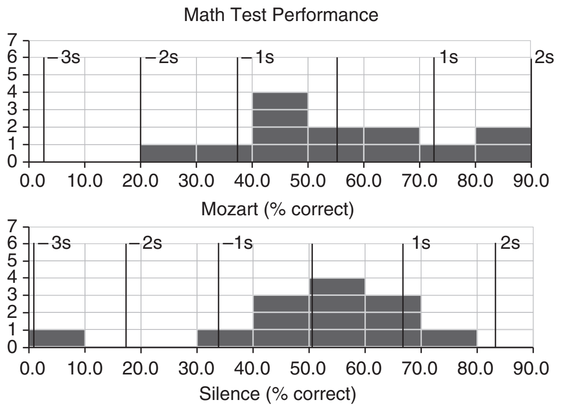
Students, if allowed, will recite summary statistics of center and dispersion without making inference of their relationship during comparison. Make sure that students are comparing distribution centers while comparing their measures of dispersion and skew. Ensure students are interpreting these measures relative to their context.

Students can become confused with when using multiple measures of center and dispersion. Ensuring students have made relational connections among these different measures and have time to investigate their properties with different shaped distributions can reduce this confusion.

Connections to Modeling

Statistics is useful for objective comparisons between different treatments and population parameters from real-world experiments and sample surveys. Students should properly analyze comparisons between distributions with appropriate center and dispersion summaries. A common measure of wealth for a population is the median household income. Statisticians use this measure as an alternative to the mean household income because of its strength to measure center in the presence of outliers and skewness.

The example that follows highlights the difference between performance on a mathematics test taken in silence and a test taken while listening to Mozart (from a data set archived in NCTM's Core Math Tools). The vertical segments labeled 1s, 2s, and so forth highlight the standard deviations while the unlabeled middle lines represent the mean. Because of the extreme value in the data for students taking the test in silence, a comparison of means shows a larger difference than a comparison of medians. A comparison of the size of the standard deviations also implies a larger spread for the class that took their test while listening to Mozart. Based on the occurrence of an extreme value in the one data set, the IQR is a better tool for comparison of the two distributions' spread.



Related Content Standards

S.ID.A.1 S.ID.A.3 S.ID.A.4 6.SP.A 6.SP.B

STANDARD 3 (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).

We often describe distributions by their center. We may identify one or multiple peaks or modes of the data set, which may or may not relate to the distribution's measure of center. Generally in the CCSS-M, the mean is best used for the measure of center if there are no extreme values or there is no skewness in the distribution. When distributions are skewed to the right or left, the mean tends to be more toward the skew than the center of the distribution. In cases of skewed data sets or cases with extreme values, it is more appropriate to use the median as a center because it is considered robust, or less subjective to extreme values or skewness. The uniform distribution or uniform probability model may also be useful for students because it is a good model for introduction of probability models for continuous data.

Students should use summaries of quantitative spread that use the relationship of dispersion and center. The interquartile range (IQR), which represents the width of the box plot, or Quartile 3 minus Quartile 1, does not use a measure of center in its calculation. For this reason, students should use the IQR as a preferred description measure when in the presence of skewness or outliers (that is, when the median is the more appropriate measure of center). Students use the standard deviation when the mean of a data set is appropriate, such as when the distribution is symmetric or when a sample is intended to measure the population mean.

What the TEACHER does:

- Provides opportunities for students to investigate different-shaped distributions.
- Provides tools to calculate different measures of center and spread.
- Focuses student conception on the relationships of summary statistics to their distributional shape.

What the STUDENTS do:

- Recognize and name different shapes and their characteristics of center, shape, and spread.
- Recognize the tendency of the mean to be toward a skew or extreme value.
- Understand and be able to identify which measures of center and spread are appropriate when given certain distributional characteristics.

Addressing Student Misconceptions and Common Errors

Students often use the word *outlier* inaccurately, failing to implement appropriate mathematical models to check for appropriate characterization of a data point as an *outlier*. Using the terms *unusual feature* or *data point* reduces this issue and sets the word *outlier* aside for the use of an assigned mathematical procedure. Also using the word *unusual* can describe data shapes and points that may be different from the majority of values in a data set but which do not fit the mathematical procedure to be characterized as an outlier. Students using the word *unusual* provides for a separation between the mathematical procedure to determine outliers and interesting facets about a distribution that seem unique.

Depending on the software, applet, or calculator used to construct box plots, a modified box plot may be used. Using modified box plots can reduce the time needed to investigate and describe all of the characteristics of a particular data set. A modified boxplot is identical to a boxplot with the exception that the tails extend to the most extreme values that are not outliers. Outliers of the data set are represented by dots to the left or right of and in line with the boxplot. Using a modified boxplot, students can trace the width of the box and use this to justify the distance away from Q1 and Q3 of outliers being more than 1.5 box lengths.

Connections to Modeling

Unusual data points or characteristics are common in real-world experiments and surveys. Students learn to deal with these using proper measures of center and dispersion or removal of erroneous data entries. Students apply reasoning for the use of measures of center and dispersion by describing the influence the unusual or extreme data point may have on the summary statistics. Statisticians report household annual income statistics in terms of median income because of the effects from persons making extreme amounts of money.

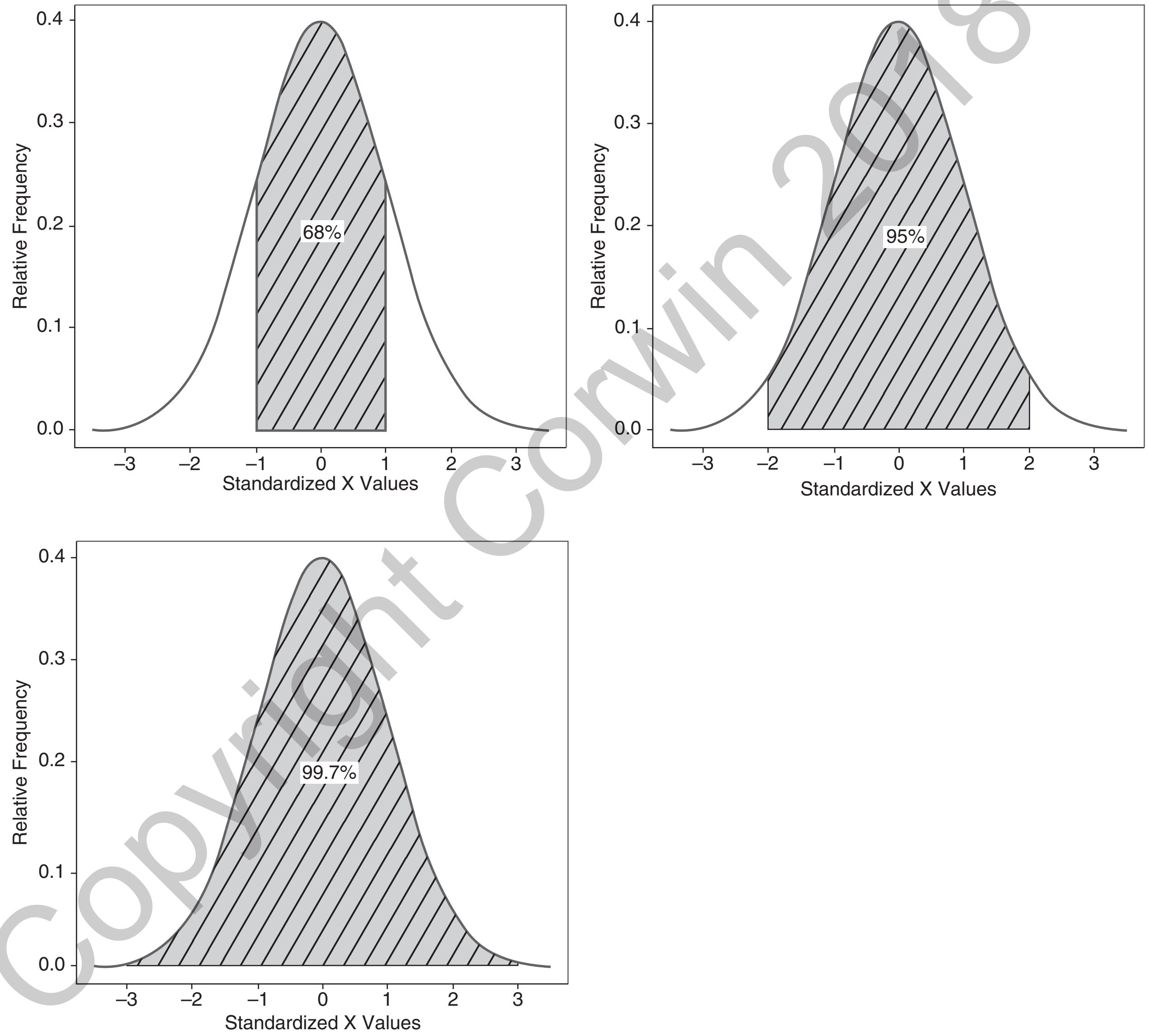
Related Content Standards

S.ID.A.1 S.ID.A.2 6.SP.A 6.SP.B

STANDARD 4 (S.ID.A.4)

Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

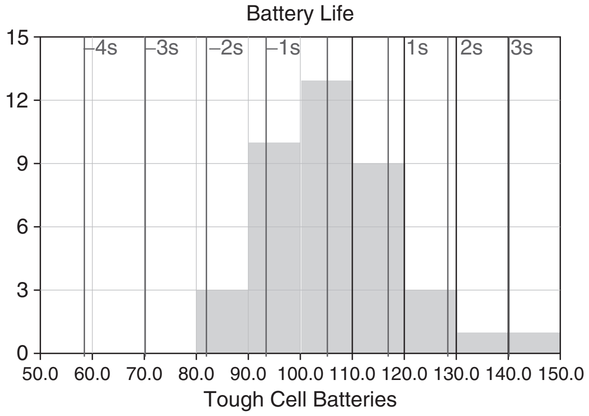
This particular standard emphasizes the use of the normal distribution to obtain the probability or likelihood of certain events. Though the empirical rule is not explicitly stated, it can form a basis for the relational understanding of area under the distribution and likelihood of occurrence. The empirical rule states that 68% of the data in a normal distribution is between 1 standard deviation below the mean and 1 standard deviation above the mean; 95% of data is between 2 standard deviations below the mean and 2 standard deviations above the mean; 99.7% of the data from a normal distribution is between 3 standard deviations below the mean and 3 standard deviations above the mean, shown below.

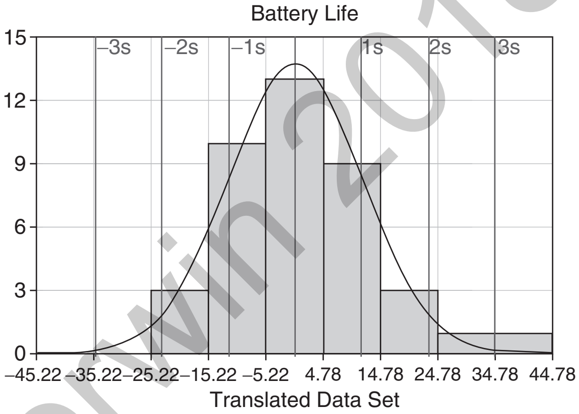


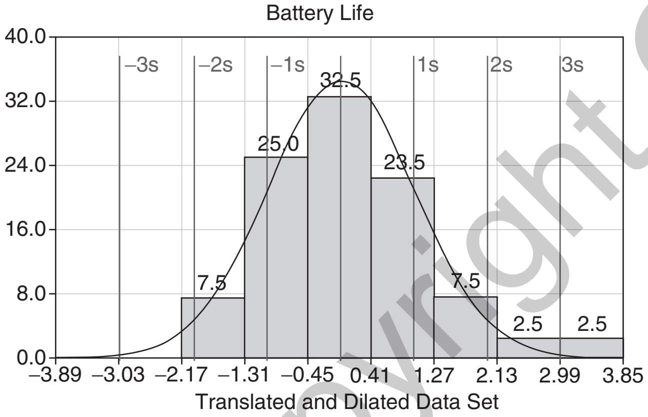
The standard emphasizes the use of calculators, spreadsheets, and tables to estimate the area under the curve. Students can use functions within advanced calculators to identify boundaries of interest. For example, a problem may require finding the probability of an event being between –1 standard deviation and 1.5 standard deviations of the mean. This is an interval contained within two bounds. Intervals may be upward-bound restricted, lower-bound restricted, or both lower- and upward-bound restricted. When using tables or spreadsheets to identify areas or probability of interest, it is important to identify the table or spreadsheet design. Some tables may provide areas below or above a value of interest.

If students use tables, they will need to learn how to standardize data using data transformations. Teachers should devote attention and time to ensuring students understand the impact of operations on data sets. Addition and subtraction of a data set by the same number relates to translation. This keeps the data set shape and spread equal; however, the center shifts. Multiplication and division of the data set relates to dilation. The center and spread are numerically changed; however, the shape is equivalent. To students, the shape of the graph of the data set may appear to change during experimentation, but the shapes of the original and the transformed data are equivalent. Students can verify this when changing the scale of the transformed graph or the bin width of the histogram by the scale of the multiplier. Standardization uses these properties to transform a data set by shifting it to a center of zero and making its standard deviation 1.

For example, a histogram of "Tough Cell Batteries" has been created using Core Math Tools using a data set embedded in the program from *Navigating Through Data Analysis in Grades 6–8*. The next histogram is the same data set after subtracting the mean of 105.22 from each data point. Notice the new center of 0 that has shifted or translated left. The last histogram represents the data set after dividing by the standard deviation of 11.63. Notice that the shape of the distribution is identical as long as the appropriate bin width is created. After subtracting the mean and dividing by the standard deviation, the final histogram has a mean of zero and standard deviation of 1.







What the TEACHER does:

- Begins by having students relate concepts from irregular area in geometry to the normal distribution.
- Plans on experiences that allow students to measure a random event over multiple times that will create a normal distribution.
- Uses the relative frequency between certain intervals to discover the relationship between the distribution and certain cumulative percentiles.
- Provides software, calculators, and tables for calculating exact cumulative percentiles under the normal curve.

What the STUDENTS do:

- Use geometric concepts of irregular area to estimate and calculate area under the curve.
- Learn to see visually the relationship between the area under a curve in terms and probabilistic inequalities.
- Use standardization to transform data for calculation of area under a curve using tables.
- Represent normal distributions by their mean and standard deviation and relate to contextual probability distributions.
- Use the frequencies of real-world data to make probability distributions.

Addressing Student Misconceptions and Common Errors

Students commonly believe that any data that is collected should follow a normal distribution. In actuality, many populations are not normally distributed. The misconception arises from lack of experiences in graphing distributions of a sample or a misunderstanding of a sampling distribution of means. As an extension to help students understand the usefulness of the normal distribution, students can plot multiple means from multiple sample sizes to verify this distribution shape would approach normality as the sample size of each sample increases with no regard to the population distribution. The population distribution will not matter; thus, choosing a population that has skew or other unusual features can help illustrate this idea. A sampling of pennies from a finite population works as a useful illustration.

Connections to Modeling

Measurement errors in physical experiments are often modeled by the normal distribution. Having students collect measurements of a particular phenomenon will create data entries that are normally distributed. An example may consist of having students repeatedly measure the time of swinging a pendulum of a known length 10 times. Having students find these values and describe relative frequencies related to standard deviations from the mean can help develop the empirical rule. Students can use this understanding to perform more precise percentile analyses using the normal curve.

Related Content Standards

S.ID.A.1 S.ID.A.2 6.SP.B

Notes

Statistics and Probability | Interpreting Categorical and Quantitative Data
S.ID.B

Cluster B

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

STANDARD 5

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.

STANDARD 6

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; 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.

Cluster B: Summarize, represent, and interpret data on two categorical and quantitative variables.

This particular cluster emphasizes the association of two variables that may be either quantitative or categorical. Teachers and students use two-way frequency tables to summarize and analyze categorical data. Similarly, students and teachers use scatter plots and function fitting to describe and represent correlations between two quantitative variables. Problem-solving situations in this domain should include different forms of variability. In particular, students should be looking at variability within and between groups, co-variability, variability in model fitting, sampling variability, chance variability from sampling, and chance variability resulting from assignment to groups in experiments. To learn more about each of these types of variability, read the publicly accessible document called the *Guidelines for Assessment and Instruction in Statistics Education*. Different types of variability will possibly lead to different interpretations and explanations of association from students.

Standards for Mathematical Practice

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

Students answer qualitative and quantitative statistical questions. Teachers should incorporate the four roles of variability (Franklin et al., 2007, pp. 11–12) during problem-solving tasks in this cluster by formulating statistical questions that anticipate variability, collecting data that is designed for acknowledging and reducing variability, analyzing data in ways that account for variability, and interpreting results in ways that allow for variability from both categorical and quantitative data sets to reinforce this cluster. Attention to variability in the problem-solving process allows for students to gain statistical habits of mind, enjoy statistics, and complete the standards for mathematical practice. Students use software, such as Core Math Tools, graphing calculators, and applets to use different functions to model data sets. This technology should provide both graphical and numerical displays that develop students' understanding of the standards. Students strategically use these tools to make residuals and model fittings. Students use the structure of a best-fit line and its related residual plot to determine characteristics that represent appropriate models and describe these in context. Students should compare different fits and models with justification related to residuals, model fits, least squares, and more. Teachers and students should attend to precision of the language they use to ensure that students do not confuse correlation with wording that implies causation.

Related Content Standards

S.ID.C S.ID.A F.LE.A F.LE.B 7.SP.C 8.SP.A

STANDARD 5 (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.

Teachers and students use categorical data to record counts or numbers of occurrence in experiments. Students begin working on associated standards when creating tally marks of certain events in elementary and middle grades. This standard requires the recording of occurrences of a bi-conditional event in a table, as follows. A bi-conditional event will require students to record two characteristics for each observational unit. The table was generated from a survey by the U.S. Centers for Disease Control and Prevention in which one observational unit was recorded for gender and use of helmets.

Seniors Who Rarely, Never, or Always Wear a Helmet While Biking

	Grade 12 Boys	Grade 12 Girls	Total
Rarely or Never	1,067	758	1,825
Yes	93	109	202
Total	1,160	867	2,027

Entries in the “Total” row and “Total” column are called marginal frequencies or the marginal distribution. Entries in the body of the table are called joint frequencies because they describe the joint events, such as Grade 12 boys who rarely or never wear helmets while biking or Grade 12 girls who wear helmets while biking. Students look for association in this table by analyzing the differences in “risky behavior” of not wearing helmets by boys and girls. Making possible associations or trends should relate to conditional probabilities and comparisons.

What the TEACHER does:

- Provides experiences for students to collect survey data and input this data into two-way frequency tables.
- Ensures students interpret marginal distributions in terms of the context of the situation and using the correct conditionality statement.
- Allows students to make multiple comparisons in two-way tables and develop relationships to other standards of conditionality.
- Ensures students are interpreting relationships in tables as associations rather than causality, especially if the design of the table is not given.

What the STUDENTS do:

- Use relative frequencies to make inferences about associations or trends in the data.
- Interpret relative frequencies in terms of a subset of a conditioned event.
- Use probabilistic independence as a way to show and justify no association. Using exact independence may be considered at entry points. Independence is described in more detail in S.CP.
- Begin to allow for variation in inferences of independence as students progress.Though a two-way table may not support mathematical independence, the data may be so close in estimation that it can be assumed. This reasoning lays the foundation for the use of a chi-squared statistic in further statistical courses.
- Use standards CP.A.2, CP.A.3, and CP.A.4 as possible methods to show and justify independence.

Addressing Student Misconceptions and Common Errors

Students often make arguments for independence or other contextual comparisons based solely on number of counts on one variable without adjustment for marginal distributions. The need to use the bi-conditional structure and proportionality within the table is imperative. Focus student attention on the differences in total amounts to move them toward conditional arguments for comparison. Questions related to proportional comparisons in the table are useful in prompting student reasoning.

Students can look at marginal distributions either vertically or horizontally. This provides challenges when students are making arguments of comparison and association from the data. Whether students justify with the vertical or horizontal marginal distributions when describing association is mathematically irrelevant, the argument for association can seem conflicting based on students’ pre-conceived view of dependency and causation. Ensure students have correlated the correct marginal distribution with their conditional argument.

Students often have difficulty separating causation and association in contextual data sets. Using data sets that are obviously associated but do not imply causal relationships can help reduce this misconception and provide entry points for discussion among peers.

Connections to Modeling

The middle grades and other high school standards place much emphasis on correlation between quantitative variables; however, the association of categorical variables is also pertinent in many real-world contexts. Students collect data on two categorical variables that are dichotomous and represent their frequencies in a two-way frequency table. Students model association by testing for independence of the two variables using multiple strategies.

Related Content Standards

S.ID.A.1 S.ID.B.6.a S.ID.B.6.c 8.SP.A

Notes

STANDARD 6 (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; 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.

This standard requires students to plot paired data sets on the coordinate plane. This subpart of standard six is an extension of the Grade 8 statistics and probability standards. Of particular note is the second part of the stem of the standard that requires a description of the association in context of the data. Students can use words such as strongly related, moderately related, weakly related, not related, or similar words. Teachers should focus student attention to the description of the relationship of the points in terms of the context, including the parameters and *y*-intercepts of the model. For a negative slope, students describe how the increase in the *x* variable (contextual name such as speed of a car) is associated with a decrease in the *y* variable (contextual name such as fuel efficiency). For a positive slope, students relate the relationship of the increase in the *x* variable with an increase in the *y* variable, again ensuring contextual description.

Students begin to see patterns emerge from the association of the quantitative bivariate data after plotting on the coordinate plane. These associations may appear linear, quadratic, or exponential but should be limited from prediction outside the range of provided data. When students attempt to predict outside the range of a given data set for a particular model, this is called extrapolation. An example would be a quadratic fit with negative concavity that appears to only increase in the scatter plot but will decrease at other values of the independent variable as it increases. Eventually the quadratic will decrease, limiting the model's prediction ability outside the fit of the data. This often happens when students interpret the *y*-intercept in a model because the independent variable being zero may not make contextual sense. See the hamburger nutrition in subpart c of this standard for an example.

Generally, students should use functional shapes they have learned in F.LE to help in choosing appropriate models. Scatter plots that appear to increase on the *y*-axis at rates much faster than linear functions may be fit well with quadratic, power, or exponential models. Scatter plots that have varying peaks may be fit with appropriate polynomials of appropriate degree. Scatter plots that appear to increase quickly then level out may be fit with logarithmic models. Once students have chosen a specific model, say a quadratic, they may use the vertex form of the quadratic to transform the model to fit the data in the scatter plot by shifting the vertex and dilating its stretch. Even if students have little exposure or prerequisite knowledge of F.LE standards and transformations of functions from other courses, students can explore through trial and error using technology to build their knowledge of the general shape of these functions.

Typically, teachers and students pick models and describe reasoning for these models based on their closeness of the model to the points on the scatter plot. In addition to the general shape of a data set, students should also analyze the residual plots of their model fit to determine its appropriateness. Refer to subpart b of this standard for more detail on residual plots.

As an extension, students may linearly transform a data set after determining an appropriate model. They may do this by performing the inverse operation to the dependent variable, which will linearize the relationship. For example, if a scatter plot appears to be quadratic, students may square root each value in the data set to reduce the quadratic form to a linear form. Students may then interpret the graph as linear function using the independent variable in relationship to the transformed square root, in this case, dependent variable. Using the transformation of data sets rather than complex functions helps students and others understand the relationship between the variables in a contextual situation rather than using complex functions.

Teachers should use simulations or real-world contexts in which students can test their models. This can help build students' understanding of model fitting and the appropriateness of their model. A non-linear example or task designed to teach this standard follows in the Sample Planning Page.

What the TEACHER does:

- Asks students to describe patterns in the scatter plot as associations between the increase or decrease in the two variables.
- Provides multiple data sets that can be represented by linear, quadratic, and exponential models.
- Uses data programs that allow students to explore and analyze their own conjectures. Technology that allows for student interaction of graphed data points can help extend this standard to others in this cluster.
- Provides opportunities for students to sustain statistical arguments for different models of best fit.

What the STUDENTS do:

- Describe how variables are related within context of the situation.
- Use technology and other tools to represent scatter plots.
- Describe associations in terms of strength and direction.
- Discuss extrapolation and its limitations for different contextual and mathematical models.

Addressing Student Misconceptions and Common Errors

Students may describe the relationship of the data in different manners. Some may argue data are moderately associated while others say weakly associated. Some may describe that a decrease in the *x* variable is associated with a decrease in the *y* variable. These are neither misconceptions nor errors but can require others to think about the situation from a different perspective. Allowing students time to grapple with, discuss, and even collaborate on objective terminology can lead to further understanding and consistency from student to student.

Students often have difficulty separating causation and association with contextual data sets. Using data sets that are obviously associated but not causal can help reduce this misconception and provide entry points for discussion among peers.

Connections to Modeling

Scatter plots are essentially models used to describe the relationship between real-world variables that are quantitative in nature. Students purposefully select quantitative variables that are being tested for correlation, formulate a model to describe this relationship using an algebraic function and scatter plot, analyze the relationships of the graph to draw inference, interpret the results of the relationship in terms of the original situation, validate the conclusions by comparing them with the contextual situation, and report on the conclusions and the reasoning behind them. These essential steps represent the cycle of modeling for scatter plots.

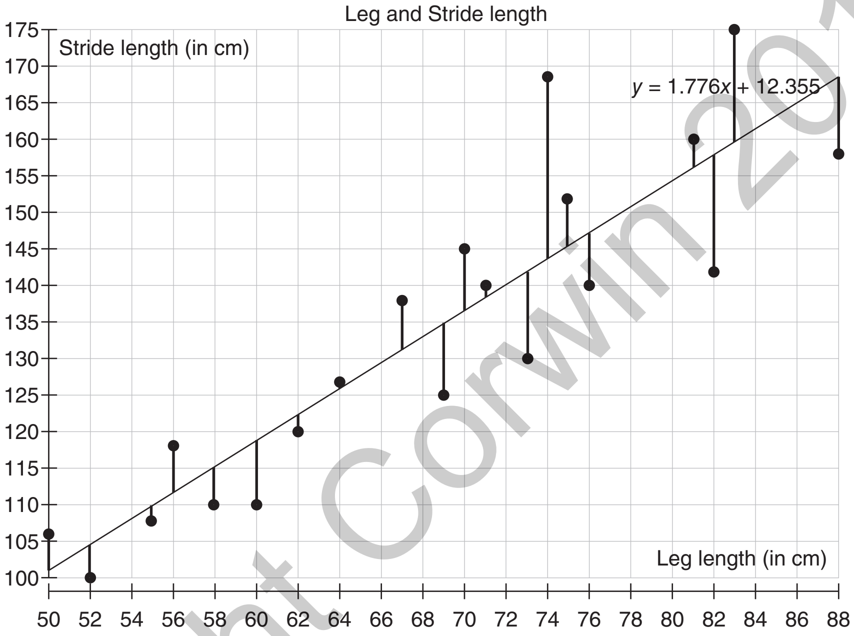
Notes

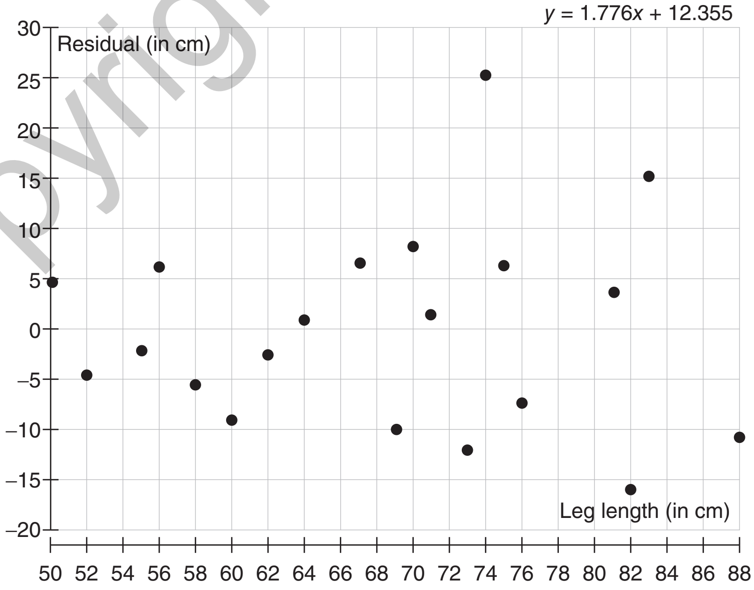
b. Informally assess the fit of a function by plotting and analyzing residuals.

The residuals of a model are defined by the difference of actual (observed) value and the predicted value for the same independent variable. The residual symbolizes how different the predicted value is from the actual. A residual plot is created using the independent variable as the *x* component of an ordered pair and residuals as the *y* value of the ordered pair. A residual plot that has no pattern about the horizontal line, $y = 0$, suggests a potentially appropriate fit while patterns in the residual plot suggest an inappropriate model fit.

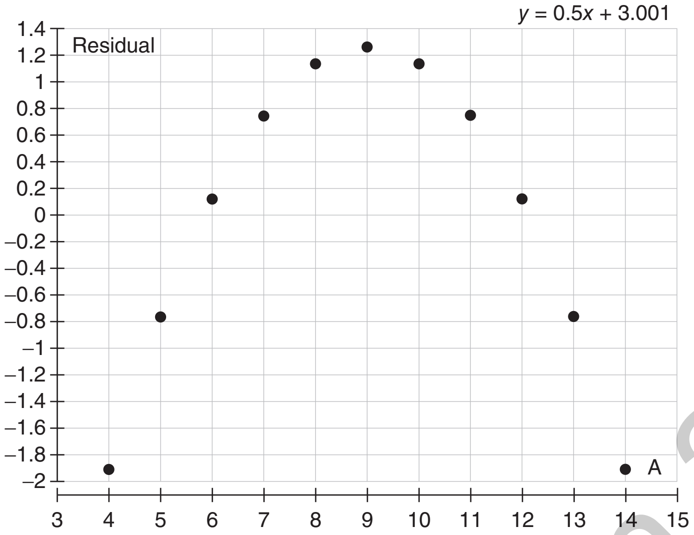
The creation of a residual plot can be very tedious for large data sets and reduce time to build understanding from multiple data sets. Using technology to create these plots allows students to analyze the relationship of the placement of points, the model, and their respective residual. Teachers should use different statistical software and web-based applets to help students understand the analysis of residuals by moving points on the scatter plot to see the effects of the model and residual.

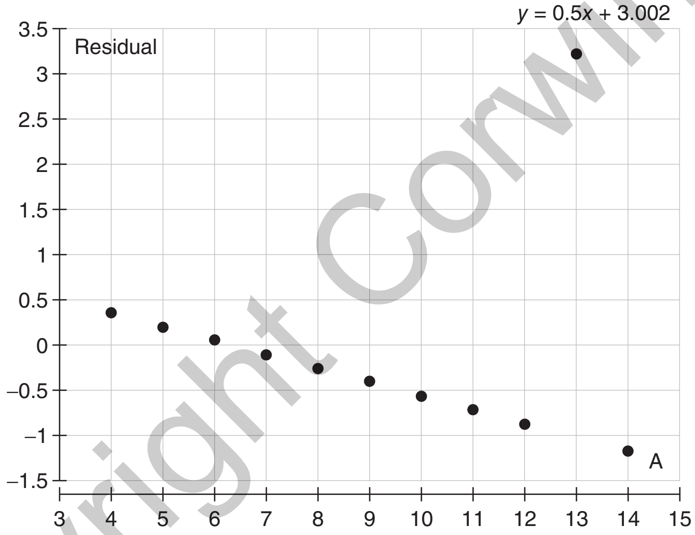
To help illustrate and teach this, Core Math Tools provides a tab and plot in which points and fits can be moved and interactively changed to illustrate their relationship:

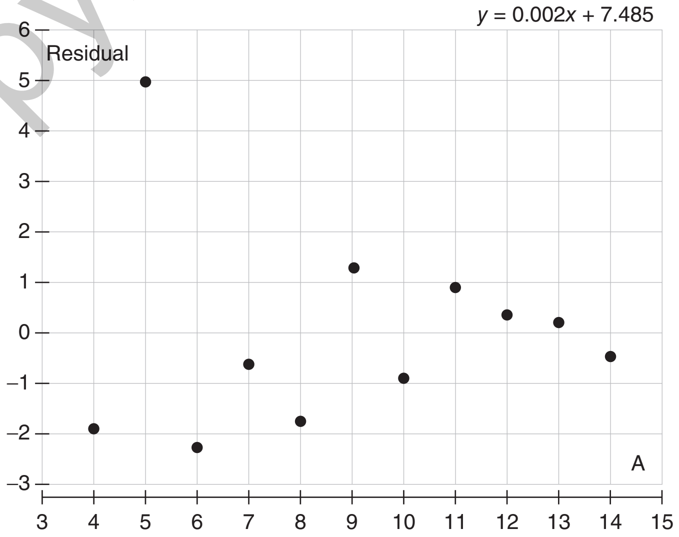




Students should look for patterns in the residual plot to justify appropriate model fits as described in subpart a of this standard. The previous residual plot in subpart a of this standard represents a random pattern in a residual plot, indicating a good fit for the linear model. In these later cases, the residual plots have a range of different patterns. Conceptually, students should be able to argue that the predictions have a systematic flaw that can still be adjusted for by a different model or the removal of outliers.







What the TEACHER does:

- Provides students with a range of data sets to analyze different residual plots.
- Connects the predicted model representation with the residual plot and the horizontal line at zero.
- Allows for connections in this cluster by having multiple contextual situations, best-fit model types, and residual plots.
- Interprets a residual as the difference between the observed value and the predicted value.
- Helps students begin abstraction of bivariate relationships by providing experiences with technology that allow them to explore the relationships between model predictions and residual plots.

What the STUDENTS do:

- Move lines of best fit and/or data points to deduce their interdependence.
- Analyze different residual plots for different predictive models.
- Recognize that points above the horizontal line on a residual plot are underestimates and points below the horizontal line at zero are overestimates.
- Recognize the relationship of patterns in prediction with inappropriate model fitting.
- Deduce that points randomly distributed above and below the horizontal line at zero on a residual plot imply appropriate fits for a model.

Addressing Student Misconceptions and Common Errors

Students often have difficulty connecting observed values from a scatter plot to their respective residual plot. Similarly, students have difficulty visually conceptualizing a residual plot and defining points from this plot contextually. Providing technology that allows for these two plots to move interactively can reduce this confusion. Questions to students that focus on interpreting different points on a residual plot and their relationships to the model and scatter plot may also reduce this confusion.

Connections to Modeling

Students need a fundamental understanding of best-fit models and how to ensure their best fits. Using a residual plot graphically depicts the appropriateness of the best fit by plotting the model's error values in respect to its independent values. Students should verify that the residuals in a residual plot are randomly distributed on the plot, with zero as the center. Having students fit an appropriate model and a residual plot while simultaneously moving points can increase students' conceptual understanding of their relationship.

The free online applet called "Regression" from http://www.shodor.org/interactivate/activities/Regression allows students to create points on a scatter plot, fit a line, remove points, move points, and move the best-fit line while simultaneously seeing the effects on the residual plot. Many statistical software programs for teaching statistics also include this feature, including NCTM's Core Math Tools.

Notes

c. Fit a linear function for a scatter plot that suggests a linear association.

This substandard directly places attention to the fitting of a linear function to a scatter plot. Though this standard could be extended to the calculation of different statistics to determine a best-fit line using least squares or mean/sum absolute differences, the wordage of "*a* linear function" makes explicit that using a procedure to determine a line of best fit is not the point of the standard. Students should grapple with ideas about what makes a best-fit line. Must there be the same number of points above the best-fit line as below? Must the best-fit line go through two points, one point, or no points?.

Technology can help in creating a line of best fit. Using technology that incorporates a specific procedure for calculating a best-fit line allows students to contextually describe the same parameters for a model, which, in turn, reduces confusion in the class. In addition, using technology to determine a best-fit line saves time in creating equations by not requiring them to use two points on the coordinate plane to create a function. As students fit an appropriate linear model, it is important to focus on the contextual description for the parameters of slope and *y*-intercept.

What the TEACHER does:

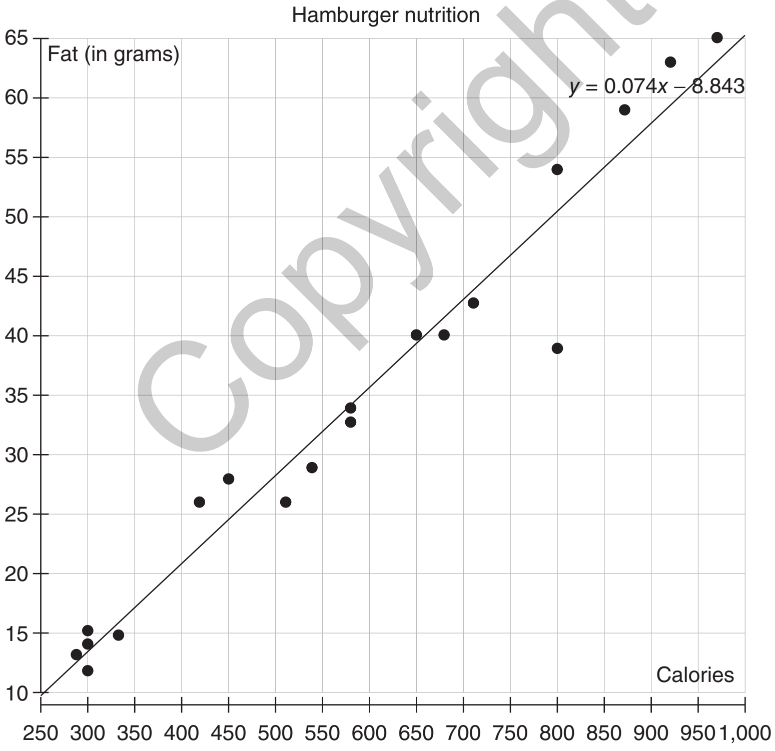
- Provides tools that can be used to fit and build students' understanding of linear functions.
- Uses purposeful data sets that do not suggest linear association but rather other models.
- Uses purposeful data sets that require students to grapple with extreme data points.
- Always uses purposeful terminology to imply association rather than causation. Makes clear through discussions the objectivity of the association by mathematical reasoning.

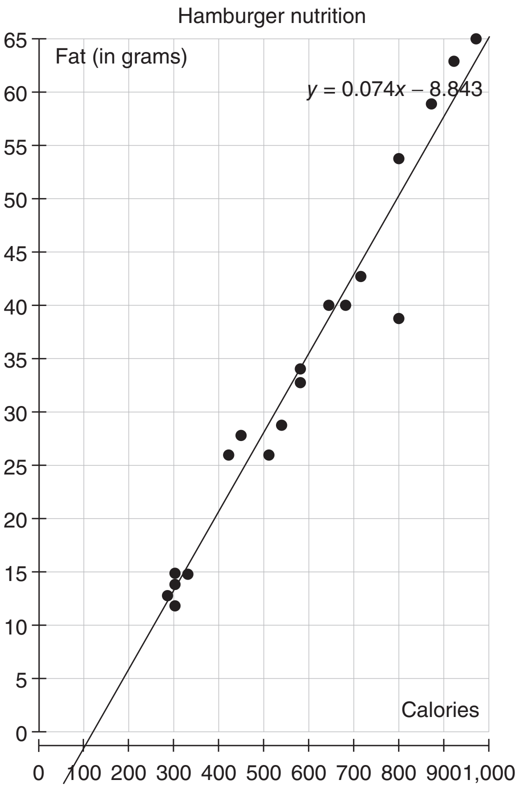
What the STUDENTS do:

- Use technology to make inferences about associations in data sets.
- Use technology to test conjectures about and explore the correlation between two quantitative variables.
- Create their own best-fit lines and discuss reasons for choosing the associated lines.
- Use purposeful terminology to imply association rather than causation.

Addressing Student Misconceptions and Common Errors

When students use technology to graph best-fit lines, breaks in the *x*- and/or *y*-axis can cause issues in relating the parameters of the linear model to the graphical display. Breaks in the *x*- and *y*-axis often encourage students to wrongly interpret or create a *y*-intercept for a model because the *x*-axis is not zero. It is important to ensure breaks are not used or that students have dealt with this potential misconception. To help with this misconception, place two graphs of the same data where one includes the point (0, 0) or includes both the *x*- and *y*-axis, and another has breaks in both the *x*- and *y*-axis and, in turn, doesn't include the *x*- and *y*-axis side by side. Have students discuss the benefits and drawbacks of both graphs. In the illustration, the Hamburger Nutrition data set in Core Math Tools is provided.





Interpreting the parameters, especially the *y*-intercept, of a best-fit line may not be appropriate if the *x* value cannot be zero. Connecting this idea to extrapolation and prediction can help students handle this misconception. The previous illustration of the data set Hamburger Nutrition is one such case. A hamburger with zero calories is non-existent; thus, the *y*-intercept is used for model fitting rather than contextual description.

Teachers should pay careful attention to student descriptions of lines of best fit. Students often say things like, "I want the same number of points on both sides of the line." This is a lack of precision by students in vocabulary and understanding. Another student attempting to create the same model may come up with a completely different solution that does not necessarily make a best-fit line. Other ideas, such as, "I needed the line to go through two values," are also erroneous. Have students relate the best-fit line to the reduction of the residuals. Also, look for students who may think that every line should go through the point (0, 0). Teachers can help clear up this misconception through additional investigations with negative slopes or by interpreting the *y*-intercept in a model that does not go through the origin; however, technology often uses an intersection of the axes that is not (0, 0), making this misconception harder to investigate in student understanding.

Connections to Modeling

A scatter plot represented by a linear function simply describes a particular phenomenon or summarizes it in a compact form. There are many scenarios in which teachers can use descriptive modeling to represent bivariate data. One example may be a graph of car weight versus miles per gallon of gasoline used, displayed in a scatter plot; another example might be ACT scores for states and percentage of states' student population taking the test. Descriptive modeling seeks for students to understand and explain the relationship of the data specifically in a description. Analytical modeling in this standard seeks to explain this bivariate data based on linear function fitting by using parameters that come from data collection. Linear functions and their associated attributes of slope and intercept are important tools for analyzing such problems and how they influence contextual situations. In addition to descriptive and analytical modeling, graphing utilities, spreadsheets, calculators, statistical software, and more are powerful tools that can be used to model linear functions and their associated scatter plots.

Related Content Standards

S.ID.B.5 S.ID.C.7 S.ID.C.8 S.ID.C.9 A.CED.A.2 F.LE.A F.LE.B 8.SP.A

Notes

Statistics and Probability | Interpreting Categorical and Quantitative Data
S.ID.C

Cluster C

Interpret linear models.

STANDARD 7 **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.

STANDARD 8 **S.ID.C.8:** Compute (using technology) and interpret the correlation coefficient of a linear fit.

STANDARD 9 **S.ID.C.9:** Distinguish between correlation and causation.

Cluster C: Interpret linear models.

Proper interpretation of linear models is essential for understanding, critiquing, and applying research results. Teachers should emphasize the link between the context of a problem and statistical models of the situation. Students should describe slope and intercept in words relative to the variables of interest in the study. Students should interpret the correlation coefficient as the amount of directional association rather than a measure of causation between the two variables in context. The quadrant count ratio described later in this cluster and in the Guidelines for Assessment in Statistics Education report (GAISE report; Franklin et al., 2007) can help build the foundation for interpreting the correlation coefficient.

Standards for Mathematical Practice

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

The focus of cluster ID.C is the creation and interpretation of linear models. Allowing students opportunities to investigate a range of data sets and statistical questions can allow for students to make use of structure, justify results, and critique the reasoning of others. Students should draw conclusions from these different investigations and compare how well their interpretations hold up across different statistical problems with particular attention to correlation versus causation. Teachers can provide opportunities for students to model in this cluster by allowing students to create their own statistical questions that expect association and analyze these results using content standards. Students create linear models and compare these models' correlation coefficients across multiple problem situations to develop reasoning and understanding of its value. Teachers should ensure students are attending to precision when interpreting linear models by ensuring the context is effectively described and does not imply causation inappropriately. Students should have opportunities to relate different values of correlation coefficients to different levels of fit by using repeated reasoning.

Related Content Standards

S.ID.B F.IF.C F.LE.A F.LE.B 8.SP.A

Notes

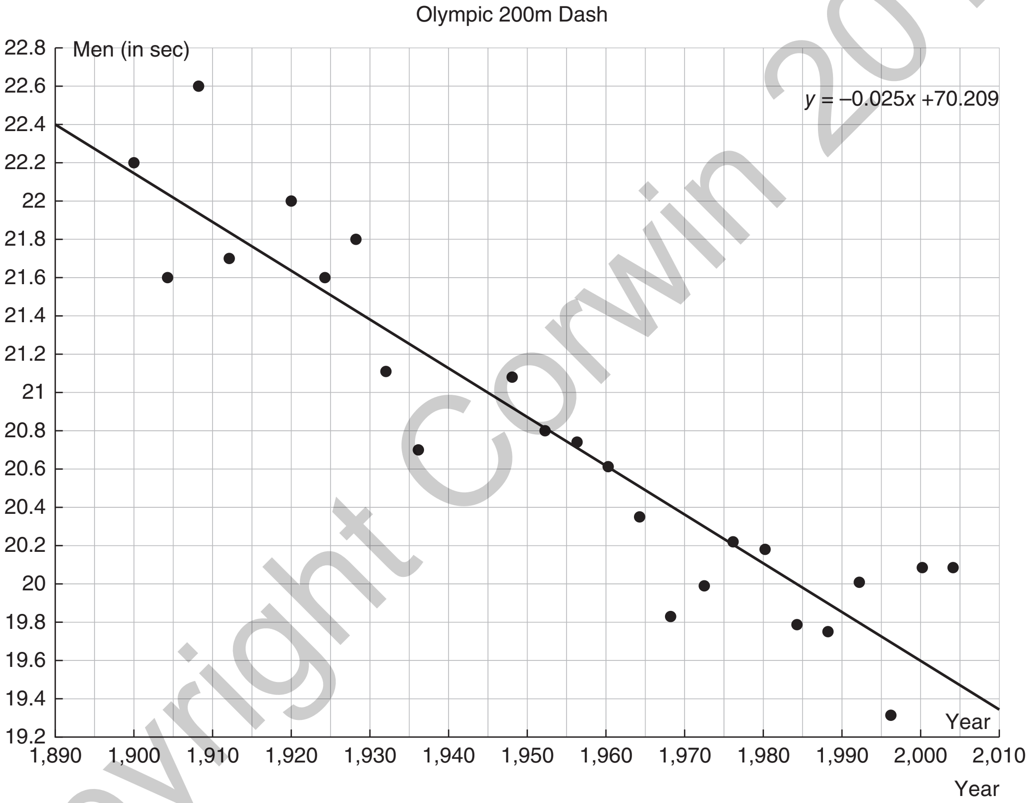
STANDARD 7 (S.ID.C.7)

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

Students interpret slope from a given linear model in terms of units of increase of the *x* variable and the associated increase or decrease on the *y* variable. For a positive slope, students describe how a specific increase in the *x* variable contributes to a certain increase in the *y* variable (in context). For a negative slope, students relate the increase in the *x* variable (in context of the situation) with a decrease in the *y* variable. For every 1 unit of increase in the *x* variable, the *y* variable will increase or decrease the amount and direction of the slope.

The intercept of a linear model cannot always be interpreted within a context. A correct interpretation discusses the *y* variable in context when the *x* variable in context is zero. If the *x* variable cannot contextually be zero or close to zero, students should address the inappropriateness of interpretation. In addition, the *y* intercept may be far from the points in the data set and be useless for prediction (and is considered a case of unwarranted extrapolation).

Consider the 200 m dash times data for men from *Focus in High School Mathematics Reasoning and Sense Making Statistics and Probability* (Shaughnessy, Chance, & Kranendonk, 2009). The scatter plot and best-fit line for this data is provided below.



The best-fit line, $y = -.025x + 70.209$, can be described in context. A student may interpret the slope as follows: the completion time for the 200 m dash decreases, on average, by .025 seconds each year. A student may interpret the *y*-intercept, 70.2 seconds, as the prediction of completion time of the 200 m dash in the year 0. This prediction is actually not feasible, and students should describe it as such because of impracticality and extrapolation. Similarly, students may attempt to predict too far into the future and find prediction times that are less than 0.

What the TEACHER does:

- Ensures that students are interpreting slope within context. Pays careful attention to students' wordage to ensure they understand the rate in terms of both the independent and dependent variable.
- Scaffolding during student discovery should move students from abstract concepts of the linear model to contextual description.

What the STUDENTS do:

- Explain the slope and intercept estimates in terms of problem context.
- Extend connections from linear functions to making connections to contextual situations.
- Explain reasoning for slope and *y*-intercept parameters in terms of predicted values.

Addressing Student Misconceptions and Common Errors

Students are often told to describe slope as “the increase in the *y* variable for every one *x* variable” where *y* variable and *x* variable are replaced with the context and units of the problem situation because most statistical programs output slope as a unit rate. This is often difficult for students to understand because data provided may have differences in the independent variable that is larger than 1 or in which a difference in 1 is unmeasurable. In addition, students rarely see slope in mathematics written in decimal notation. Having students focus on the interpretation of slope as a unit rate can help reduce this confusion and provide a conceptual foundation for interpreting the slope within the context of the situation. Another alternative to following a predetermined contextual description is to use focusing questions to develop slope as a rate of change. This uses the more general definition of slope, which can represent change over any interval. Students may convert a unit rate provided by technology to a mixed number, then interpret in context of the independent and dependent variables. The benefit of such an approach may be useful for connecting to other mathematical standards.

Interpretations of linear models may not make sense contextually for students. This should be justified through student discourse and writing. In the previous example, the *y*-intercept does not have contextual meaning since a prediction for the year 0 would imply that it took approximately 70 seconds, the *y*-intercept, to complete the 200 m dash. This is unrealistic; however, students may not immediately see the ramifications.

A graph, such as the one in the 200 m dash example, may cause confusion because it appears that the *y*-intercept may be 22.4 seconds. Students’ attention needs to focus on the scales that the graphing utility created for the scatterplot and graph to recognize that the origin is not on many scatterplot graphs created via technology.

Connections to Modeling

S.ID.C.7 requires the use of context. If data suggests a linear relationship, students can describe the relationship with a best-fit line and describe the line’s strength and direction through a correlation coefficient. Then, the students link their linear model to its context to see if the model makes sense. Testing the model is part of the modeling cycle. Students’ interpretations of slope and *y*-intercept should tie back to the context in which they were derived or being used for description.

Related Content Standards

S.ID.B.6 F.IF.B.6

Notes

STANDARD 8 (S.ID.C.8)

Compute (using technology) and interpret the correlation coefficient of a linear fit.

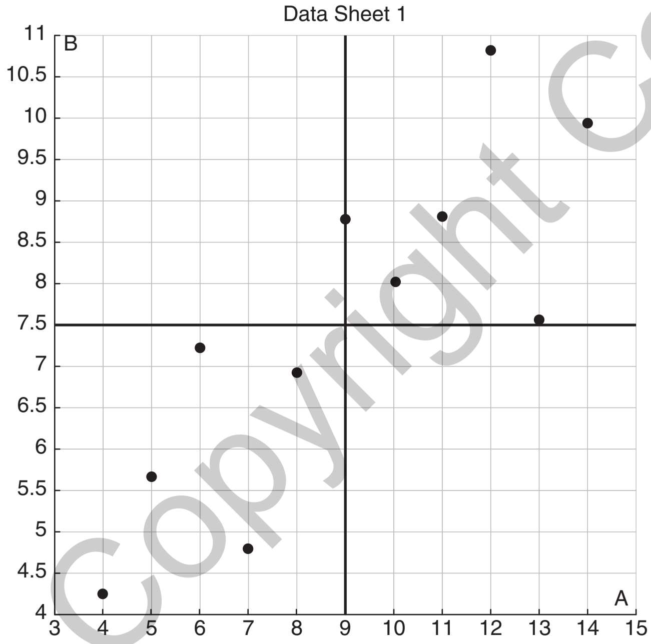
Students use technology to compute the correlation coefficient of a linear fit. This is a measure of the strength and direction of the linear relationship between two variables, or the correlation coefficient (r). A model is said to be perfectly positively related when the r value is 1 and perfectly negatively related when the r value is –1. No linear association would be represented with an r value of 0. Some common interpretations for specific r values are in the table below but are not a gold standard. Tables such as this should not be used alone but in conjunction with the scatter and residual plots to determine appropriate fit.

|*r* value|	Description
.0–.19	Very weak
.2–.39	Weak
.4–.59	Moderate
.6–.79	Strong
.80–1.0	Very strong

Students can model the correlation coefficient by visually displaying points' relationships to both their independent and dependent variables' mean using vertical and horizontal lines. Though not a part of this standard, the calculation of the correlation coefficient is the slope of the line when the x and y have been standardized or shifted to an intercept of (0, 0) with both x and y having standard deviation 1.

The GAISE report discusses the use of Quadrant Count Ratio (QCR) to help in conceptually understanding or statistically reasoning about the correlation coefficient. The QCR is useful in that it is interpreted in the same way the correlation coefficient is, yet students can easily compute its value by hand. In the later graph, the QCR calculation is calculated and described along with the correlation coefficient outputted for comparison.

Data Set 1

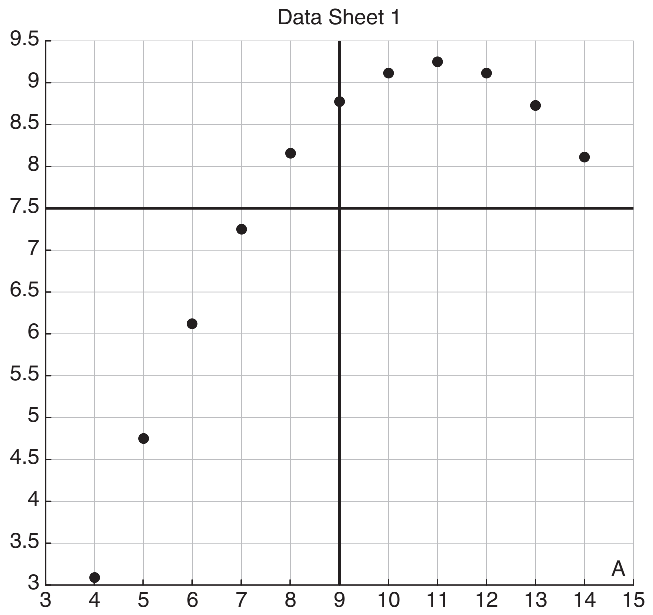


$$r = .82$$

$$QCR = \frac{Quadrant\,1\,and\,3 - Quadrant\,2\,and\,4}{Total\,Points}$$

$$QCR = \frac{10-0}{11} = \frac{10}{11} = .909$$

Data Set 2



$$r = .82$$

$$QCR = \frac{Quadrant\,1\,and\,3 - Quadrant\,2\,and\,4}{Total\,Points}$$

$$QCR = \frac{9-1}{11} = \frac{8}{11} = .727$$

Like all statistical summaries, the QCR and correlation coefficient have their shortcomings; however, the QCR can help students interpret and create a foundation for understanding the correlation coefficient (r). Though not required for this standard, the similar structure of calculation by dividing the data set into quadrants brings insight to why the formula for the correlation coefficient behaves in the way it does. It should be emphasized that though both of these former illustrations have "strong" associations according to the table, only the first is an appropriate model based on the scatter plot. The r value in the second data set is not interpretable because it is not a linear fit.

What the TEACHER does:

- Provides technology to calculate correlation coefficients.
- Provides opportunities for students to connect different forms of strength to different model fits and scatter plots.
- Makes connections for interpretation of the correlation coefficient that may come from the quadrant count ratio and relationship to mean and standardization.
- Emphasizes the connection between the mean of both the independent and dependent variables with the correlation coefficient.

What the STUDENTS do:

- Use technology to compare different models and their respective correlation coefficients.
- Reason and make sense of different correlation coefficients and their relationship to different situations.
- Interpret points in relationship to the means of the x and y values by graphing a vertical and horizontal line to represent the mean of both x and y variables.

Addressing Student Misconceptions and Common Errors

The r value should not be used alone to determine the appropriateness of fit of the best-fit line. Data sets are available that suggest the same linear model and r value, but the residual plot reveals patterns that suggest inappropriateness of prediction. The r value should be used as an additional tool to determine and describe the line of best fit.

Many students often interpret high correlation coefficients as "good fits" to the data; however, appropriate linear fits cannot be determined solely by the correlation coefficient. The following two data sets have approximately the same correlation coefficient; however, observing the residual plot and linear fit implies only one data set should be appropriately considered as having a linear association.

Data Set 1

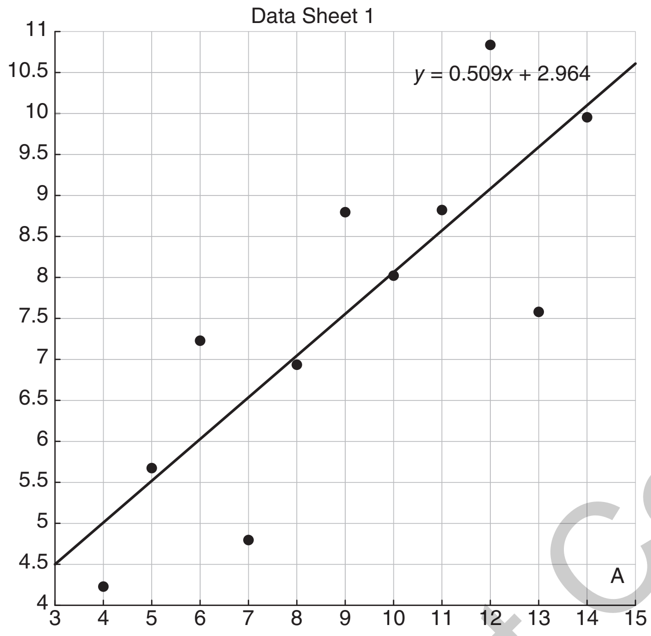
x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.83	9.96	7.24	4.26	10.84	4.82	5.68

Data Set 2

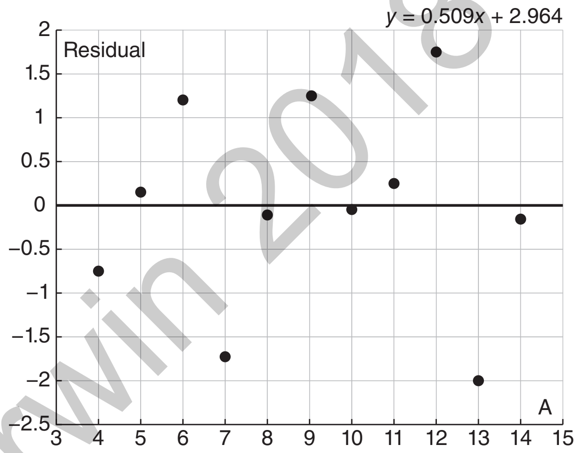
x	10	8	13	9	11	14	6	4	12	7	5
y	9.13	8.15	8.73	8.78	9.25	8.1	6.13	3.1	9.13	7.26	4.74

Data Set 1

Linear Fit to Scatter Plot

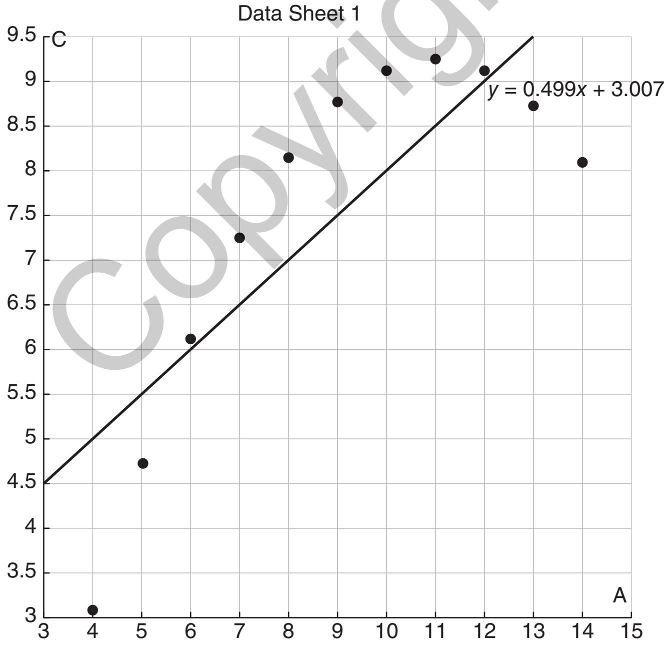


Residual Plot of Linear Fit

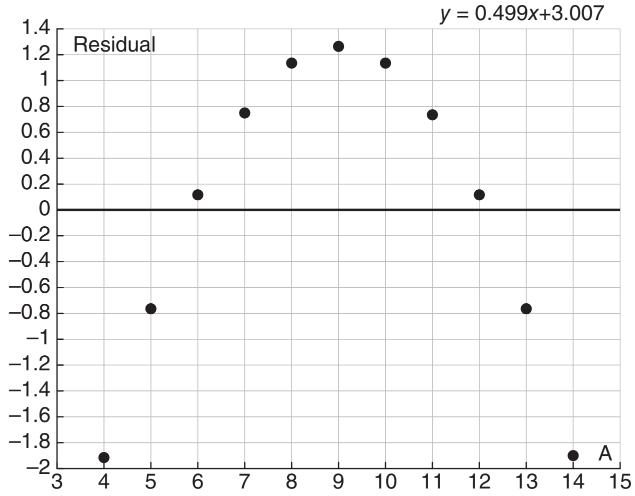


Data Set 2

Linear Fit to Scatter Plot



Residual Plot of Linear Fit



Connections to Modeling

As with all of this domain, students should use real-world data or create their own data to explore this standard. Particular to this standard, students should explore situations in which similar *r* values exist but are inappropriate based on the model fit and the relationship of the correlation coefficient to different contextual situations. As students look at the correlation coefficient from different contexts with both weak and strong correlation, they become more familiar with general guidelines for interpreting different values of the correlation coefficient.

Related Content Standards

S.ID.B.6   8.SP.A

Notes

STANDARD 9 (S.ID.C.9)

Distinguish between correlation and causation.

Data alone cannot be used to determine causation; thus, experimentation and science is involved to help deduce causation. Students and the teacher should pay careful attention to terminology used when describing bivariate relationships (both quantitative versus quantitative, quantitative versus categorical, and categorical versus categorical), even when causal effects seem evident. In order to determine causation to a population or from a treatment, at the very least, researchers must complete a random sampling from the population and a random assignment to the treatment group. In many situations, this is impossible and unethical.

Teachers can exemplify this standard with the use of two variables that are associated but obviously not causal. An example may entail having students plot their arm length and height. Though the two variables will be associated, having a large arm length does not cause you to be tall. And using a stretching regimen designed to increase your arm length will likely have no impact on your height.

In addition, teachers should directly tie this standard to S.IC.B.3 in which students understand the differences between experiments, surveys, and observational studies.

What the TEACHER does:

- Always models correct terminology when analyzing reports using the associated terminology of association, correlation, or causation.
- Listens for correct use of terminology while students express their interpretations.
- Provides experiences for students that have association, but the variables are not causal.

What the STUDENTS do:

- Relate standard IC.B.3 to this standard by distinguishing when experiments may imply causality.
- Use precision in the wording of relationships between two variables using words such as *correlated* or *associated* for two variables that are linearly related.
- Justify, explain, and provide reasons for the difference between correlation and causation.

Addressing Student Misconceptions and Common Errors

This standard directly addresses misconceptions in this cluster. The teacher's use of correct terminology and attention to wordage in students' descriptions can help engrain this standard throughout instruction.

Connections to Modeling

Relationships between different random variables are present all around us. Students model different situations that are correlated but not necessarily causal. Examples of associations in class may be students' height versus arm span or weight of a car versus price. Multiple data sets exist on the Internet or in software programs to reduce the time for exploring this standard as well. NCTM's Core Math Tools has a large range of bivariate data sets that exemplify this standard. One excellent example from NCTM's Core Math Tools demonstrating this standard is the time to complete the 200 m dash for different years of the Olympics for men and/or women, discussed in Standard 7 (S.ID.C.7).

Related Content Standards

S.ID.B.5 S.ID.B.6 8.SP.A

Notes

Sample PLANNING PAGE

Statistics and Probability

Domain: Interpreting Categorical and Quantitative Data

Cluster B: Summarize, represent, and interpret data on two categorical and quantitative variables.

Standards:

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

S.ID.B.6.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

Standards for Mathematical Practice:

SFMP 1. Make sense of problems and persevere in solving them.

Students may already have conjectures about the relationship between body weight and brain weight. They will need to make sense of problems that potential outliers have on the prediction of brain weights of animals.

SFMP 2. Use quantitative reasoning.

Students use quantitative reasoning through different mathematical models to predict an animal's brain weight.

SFMP 3. Construct viable arguments and critique the reasoning of others.

Students explain their models and predictions in context. Justification and reasoning is presented to the inclusion or exclusion of certain animals for adequate prediction.

SFMP 4. Model with mathematics.

Students model a contextual situation of body versus brain weight using scatter plots. Mathematical models are compared and justified on the basis of better predictions.

SFMP 5. Use appropriate tools strategically.

Students use Core Math Tools to explore different mathematical models and their predictive validity. The tool will allow for the inclusion or exclusion of certain animals based on student observations.

SFMP 6. Attend to precision.

Students attend to precision by comparing the predicted values to their actual values. Students also attend to precision by using their proposed model to predict for an animal not included in the data set.

SFMP 7. Look for and make use of structure.

Students use the structure of the data set to find unreasonable or potential outliers in the data set. Students may acknowledge non-linearity or the difference in the type of animal for justification.

Goal:

Students experiment with fitting different types of models to a set of data on a scatter plot and making predictions. Students justify appropriate models and use these models to solve problems.

Planning:

Materials: The "Animal Brain and Body Weight" data set from NCTM Core Math Tools, NCTM's Core Math Tools or other statistical software.

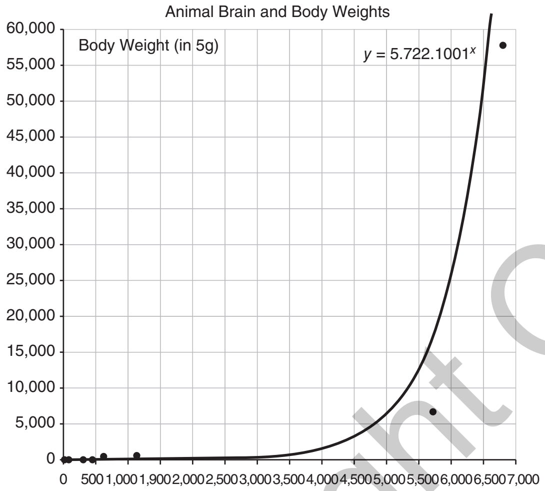
Sample Activity: The rhinoceros weighs approximately 3600 kg. Use the data provided to determine the best estimate possible for its brain weight. Justify the model you choose by analyzing the model's ability to predict other animals' brain weights.

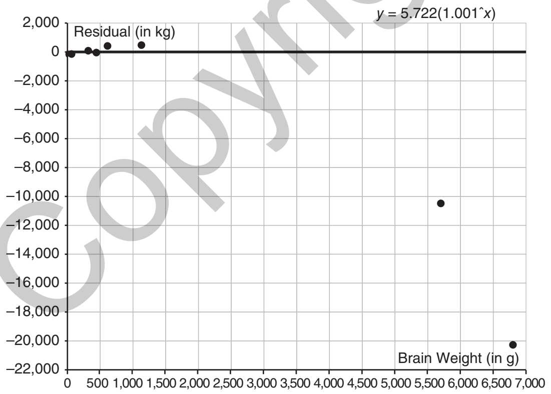
Sample PLANNING PAGE

Questions/Prompts:

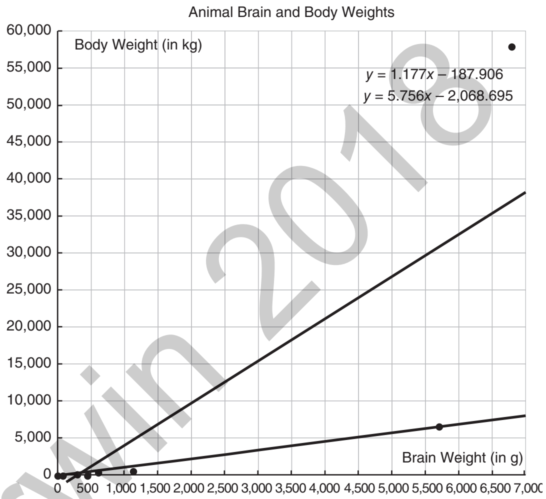
The “Animal Brain and Body Weight” data set provides interesting characteristics that may be fit using multiple models and specific observations. The data set includes the blue whale and walrus, which students may consider eliminating one or both from the data set because they are measurements of non-land animals. Most models have issues in residual plots without the elimination of the blue whale observation. Using Core Math Tools, students can fit a range of different models and use the residual plot button to analyze their ability to predict.

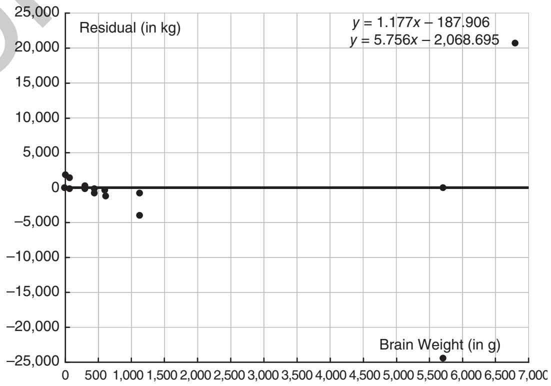
An obvious choice by students is to first try an exponential model. The model for this best-fit model and residual plot follow:





Having students remove just one or two extreme observations and fitting a linear model appears to be another useful model, as follows. Color coding in the program allows for students to compare best-fit models with and without the inclusion of certain points.





Questions to consider asking students during the exploration may consist of the following:

Do you think the model you fit would accurately predict a rhino’s brain weight? Why?

When you are observing your residuals, where do you think the error would be for a rhino? Why? How does the pattern of incorrect prediction help know this?

Sample PLANNING PAGE (Continued)

What does the pattern in errors imply about the model? Why?

What do you think is causing the model to have such poor predictions?

How do you think body weight and brain weight relate to one another? How does the scatter plot show this relationship? Do you see any deviations from what you are describing? Where?

The brain weight of a rhinoceros is approximately 500 grams. Do not share this information with students until after they have created and defended their models with others. Teachers can foster discussion afterward revealing the actual amount about residuals from student models.

Differentiating Instruction:

Struggling Students: Allow time for students to learn to use the statistical software you have for your class. Familiarity with a program will allow students more time to explore free conjectures and different model fits.

Students need to use their final model's function as a tool to predict the rhinoceros brain weight. They need to connect this function to their graphical model. Core Math Tools does this by relating the function description and graph with the same color. Once students have determined this function, have them use the function for prediction on paper.

Extensions: Students are often encouraged to predict outside the range of their observed values, or extrapolate. Having students explore different animals' weights could potentially exemplify this issue. Teachers may pre-plan for the lesson by removing the blue whale entry from students' data sets and then prompting students about this animal after exploring the rhinoceros. The fitting of models should always include an analysis of its residuals, so extensions should focus on development of the understanding of residual plots. Students may also consider using models that are not quadratic, exponential, and linear simply by clicking different model fit buttons. This will naturally occur as students explore the problem and should not be restricted for only advanced or honors students.

Notes

Reflection Questions: Interpreting Categorical and Quantitative Data

1. How can you ensure students use appropriate technological tools to test conjectures, explore data, analyze data, and develop statistical reasoning?

2. Explain how natural sampling, measurement, and variability are a part of different standards in this domain.

3. How can you select student presentations to prompt classroom discourse and foster students' conceptual understanding of key statistical ideas in this domain? Chose one cluster as an example to exemplify your reasoning.

4. How will you incorporate technology into your formative and summative assessments?

   a. How will you use this to monitor the development of student understanding?

   b. How will this affect your instruction?

   c. How will you use technology in your assessments to evaluate student progress?

   d. How can you ensure that students are using appropriate tools strategically?