Table of Contents[1]

Descriptive Statistics

[1] Each lesson is ONE day, and ONE day is considered a 45-minute period.

Algebra I • Module 2

Descriptive Statistics

OVERVIEW

In this module, students reconnect with and deepen their understanding of statistics and probability concepts first introduced in Grades 6, 7, and 8. There is variability in data, and this variability often makes learning from data challenging. Students develop a set of tools for understanding and interpreting variability in data and begin to make more informed decisions from data. Students work with data distributions of various shapes, centers, and spreads. Measures of center and measures of spread are developed as ways of describing distributions. The choice of appropriate measures of center and spread is tied to distribution shape. Symmetric data distributions are summarized by the mean and mean absolute deviation, or standard deviation. The median and the interquartile range summarize data distributions that are skewed. Students calculate and interpret measures of center and spread and compare data distributions using numerical measures and visual representations.

Students build on their experience with bivariate quantitative data from Grade 8; they expand their understanding of linear relationships by connecting the data distribution to a model and informally assessing the selected model using residuals and residual plots. Students explore positive and negative linear relationships and use the correlation coefficient to describe the strength and direction of linear relationships. Students also analyze bivariate categorical data using two-way frequency tables and relative frequency tables. The possible association between two categorical variables is explored by using data summarized in a table to analyze differences in conditional relative frequencies.

This module sets the stage for more extensive work with sampling and inference in later grades.

Focus Standards

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

S-ID.A.1 Represent data with plots on the real number line (dot plots, histograms, and box plots).*

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.*

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).*

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

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.★

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.

Interpret linear models.

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.★

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

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

Foundational Standards

Develop understanding of statistical variability.

6.SP.A.1 Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers. *For example, "How old am I?" is not a statistical question, but "How old are the students in my school?" is a statistical question because one anticipates variability in students' ages.*

6.SP.A.2 Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and overall shape.

6.SP.A.3 Recognize that a measure of center for a numerical data set summarizes all of its values with a single number, while a measure of variation describes how its values vary with a single number.

Summarize and describe distributions.

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

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

a. Reporting the number of observations.

b. Describing the nature of the attribute under investigation, including how it was measured and its units of measurement.

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

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

Investigate patterns of association in bivariate data.

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

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

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

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

Focus Standards for Mathematical Practice

MP.1 **Make sense of problems and persevere in solving them.** Students choose an appropriate method of analysis based on problem context. They consider how the data were collected and how data can be summarized to answer statistical questions. Students select a graphical display appropriate to the problem context. They select numerical summaries appropriate to the shape of the data distribution. Students use multiple representations and numerical summaries and then determine the most appropriate representation and summary for a given data distribution.

MP.2 **Reason abstractly and quantitatively.** Students pose statistical questions and reason about how to collect and interpret data in order to answer these questions. Students form summaries of data using graphs, two-way tables, and other representations that are appropriate for a given context and the statistical question they are trying to answer. Students reason about whether two variables are associated by considering conditional relative frequencies.

MP.3 **Construct viable arguments and critique the reasoning of others.** Students examine the shape, center, and variability of a data distribution and use characteristics of the data distribution to communicate the answer to a statistical question in the form of a poster presentation. Students also have an opportunity to critique poster presentations made by other students.

MP.4 **Model with mathematics.** Students construct and interpret two-way tables to summarize bivariate categorical data. Students graph bivariate numerical data using a scatterplot and propose a linear, exponential, quadratic, or other model to describe the relationship between two numerical variables. Students use residuals and residual plots to assess if a linear model is an appropriate way to summarize the relationship between two numerical variables.

MP.5 **Use appropriate tools strategically.** Students visualize data distributions and relationships between numerical variables using graphing software. They select and analyze models that are fit using appropriate technology to determine whether or not the model is appropriate. Students use visual representations of data distributions from technology to answer statistical questions.

MP.6 **Attend to precision.** Students interpret and communicate conclusions in context based on graphical and numerical data summaries. Students use statistical terminology appropriately.

Terminology

New or Recently Introduced Terms

- **Skewed Data Distribution** (A data distribution is said to be *skewed* if the distribution is not symmetric with respect to its mean. Left-skewed or skewed to the left is indicated by the data spreading out longer (like a tail) on the left side. Right-skewed or skewed to the right is indicated by the data spreading out longer (like a tail) on the right side.)
- **Outlier** (An *outlier* of a finite numerical data set is a value that is greater than $Q3$ by a distance of $1.5 \cdot IQR$, or a value that is less than $Q1$ by a distance of $1.5 \cdot IQR$. Outliers are usually identified by an "*" or a "•" in a box plot.)
- **Sample Standard Deviation** (The *sample variance* for a numerical sample data set of n-values is the sum of the squared distances the values are from the mean divided by $(n-1)$. The *sample standard deviation* is the principle (positive) square root of the sample variance.)
- **Interquartile Range** (The *interquartile range* (or IQR) is the distance between the first quartile and the second quartile: $IQR = Q3 - Q1$. The IQR describes variability by identifying the length of the interval that contains the middle 50% of the data values.)

- **Association** (A *statistical association* is any relationship between measures of two types of quantities so that one is statistically dependent on the other.)
- **Conditional Relative Frequency** (A *conditional relative frequency* compares a frequency count to the marginal total that represents the condition of interest.)
- **Residual** (The *residual of the data point* (x_i, y_i) is the (actual y_i-value) $-$ (predicted y-value) for the given x_i.)
- **Residual Plot** (Given a bivariate data set and linear equation used to model the data set, a *residual plot* is the graph of all ordered pairs determined as follows: For each data point (x_i, y_i) in the data set, the first entry of the ordered pair is the x-value of the data point, and the second entry is the residual of the data point.)
- **Correlation Coefficient** (The *correlation coefficient*, often denoted by r, is a number between -1 and $+1$, inclusively, that measures the strength and direction of a linear relationship between the two types of quantities. If $r = 1$ or $r = -1$, then the graph of data points of the bivariate data set lie on a line of positive or negative slope.)

Familiar Terms and Symbols[2]

- Mean
- Median
- Data Distribution
- Variability
- Mean Absolute Deviation
- Box plot
- Quartile

Suggested Tools and Representations

- Graphing calculator
- Spreadsheet software
- Dot plot
- Box plot
- Histogram
- Residual plot

[2] These are terms and symbols students have seen previously.

Assessment Summary

Assessment Type	Administered	Format	Standards Addressed
Mid-Module Assessment Task	After Topic B	Constructed response with rubric	S-ID.A.1, S-ID.A.2, S-ID.A.3
End-of-Module Assessment Task	After Topic D	Constructed response with rubric	S-ID.A.2, S-ID.A.3, S-ID.B.5, S-ID.B.6, S-ID.C.7, S-ID.C.8, S-ID.C.9