AP Statistics Unit 1 - Descriptive Statistics and Collecting Data

Content Area: **Math**
Course(s):
Time Period: **MP1**
Length: **45**
Status: **Published**

Unit Overview

Unit Summary	Unit Rationale
This foundational unit introduces students to the core concepts and methods used in exploring and analyzing data. Students will develop a statistical mindset by learning how to ask and answer questions using data, interpret data displays, and justify conclusions using appropriate methods. In part 1 of exploring one-variable data, students investigate distributions of a single variable to uncover patterns, trends, and outliers. They begin by identifying the individuals and variables in a dataset and classifying variables as categorical or quantitative. For categorical data, students construct and interpret frequency tables, relative frequency tables, and visual representations such as bar graphs and pie charts. For quantitative data, students use dot plots, histograms, and boxplots to display distributions and describe their shape, center, spread, and unusual features. Students calculate and interpret numerical summaries, including measures of center (mean, median), spread (range, IQR, standard deviation), and position (percentiles, z-scores), and learn when each is most appropriate. Finally, they compare distributions and assess how well a data set fits a normal distribution using standardized scores and the empirical rule. This part extends analysis to relationships between two variables. Students explore categorical-categorical relationships using two-way tables, calculating and comparing conditional frequencies. For quantitative-quantitative relationships, students construct and interpret scatterplots to describe form, direction, and strength of associations. They quantify linear associations using correlation and develop linear models using least-squares regression.	This unit lays the essential groundwork for the AP Statistics course by immersing students in the foundational skills of data exploration, representation, and collection. In a world increasingly driven by data, students must be able to ask meaningful questions, analyze data effectively, and make informed decisions based on evidence. These skills are not only central to the AP Statistics curriculum, but also vital for success in a wide range of academic fields and real-world applications. By focusing first on one-variable data, students learn how to summarize and describe individual distributions using both graphical and numerical techniques. This provides the basis for recognizing patterns, comparing groups, and interpreting variability in context. Building on that, the exploration of two-variable data introduces relationships between variables, helping students understand correlation, association, and the construction and evaluation of predictive models. These skills develop critical thinking and allow students to distinguish between causation and association—a core principle of statistical reasoning. Finally, by studying methods of data collection, students gain insight into how data is gathered and how design choices impact the validity of conclusions. Understanding sampling methods, experimental design, and potential biases empowers students to critically evaluate data presented to them in academic, professional, and media contexts. This unit emphasizes clarity in communication,

Through calculating predicted values, interpreting residual plots, and identifying influential points, students assess model fit and refine predictions. Transformations are introduced to linearize relationships when appropriate. Throughout, students interpret slope, intercept, and correlation in context and assess the suitability of a linear model for a given scenario. Students learn how data is generated through observational studies, surveys, and experiments, and the implications of each for drawing conclusions. They identify sampling methods (e.g., SRS, stratified, cluster), assess bias, and justify the appropriateness of methods for given scenarios. Students analyze the structure of well-designed experiments, recognizing the role of random assignment, control, replication, and comparison. They compare experimental designs (completely randomized, randomized block, matched pairs) and determine when causal conclusions are valid. Emphasis is placed on clearly defining the scope of inference, including when generalizations to populations or cause-and-effect conclusions are justified. Together, these three parts provide the statistical framework for understanding data in context. Students build skills in data analysis, graphical representation, and critical interpretation while learning how to draw meaningful and valid conclusions. This unit equips students with essential tools for thinking statistically and prepares them for deeper study in probability and inference.	justification of methods, and the use of technology to support analysis. It forms the backbone of the course and prepares students for more advanced statistical inference by ensuring they can work confidently with real-world data from the outset.

NJSLS

1.1.VAR-1.A.1	Numbers may convey meaningful information, when placed in context.
1.2.VAR-1.B.1	A variable is a characteristic that changes from one individual to another.
1.2.VAR-1.C.1	A categorical variable takes on values that are category names or group labels.
1.2.VAR-1.C.2	A quantitative variable is one that takes on numerical values for a measured or counted quantity.
1.3.UNC-1.A.1	A frequency table gives the number of cases falling into each category. A relative

	frequency table gives the proportion of cases falling into each category.
1.3.UNC-1.B.1	Percentages, relative frequencies, and rates all provide the same information as proportions.
1.3.UNC-1.B.2	Counts and relative frequencies of categorical data reveal information that can be used to justify claims about the data in context.
1.4.UNC-1.C.1	Bar charts (or bar graphs) are used to display frequencies (counts) or relative frequencies (proportions) for categorical data.
1.4.UNC-1.C.2	The height or length of each bar in a bar graph corresponds to either the number or proportion of observations falling within each category.
1.4.UNC-1.C.3	There are many additional ways to represent frequencies (counts) or relative frequencies (proportions) for categorical data.
1.4.UNC-1.D.1	Graphical representations of a categorical variable reveal information that can be used to justify claims about the data in context.
1.4.UNC-1.E.1	Frequency tables, bar graphs, or other representations can be used to compare two or more data sets in terms of the same categorical variable.
1.5.UNC-1.F.1	A discrete variable can take on a countable number of values. The number of values may be finite or countably infinite, as with the counting numbers.
1.5.UNC-1.F.2	A continuous variable can take on infinitely many values, but those values cannot be counted. No matter how small the interval between two values of a continuous variable, it is always possible to determine another value between them.
1.5.UNC-1.G.1	In a histogram, the height of each bar shows the number or proportion of observations that fall within the interval corresponding to that bar. Altering the interval widths can change the appearance of the histogram.
1.5.UNC-1.G.2	In a stem and leaf plot, each data value is split into a “stem” (the first digit or digits) and a “leaf” (usually the last digit).
1.5.UNC-1.G.3	A dotplot represents each observation by a dot, with the position on the horizontal axis corresponding to the data value of that observation, with nearly identical values stacked on top of each other.
1.5.UNC-1.G.4	A cumulative graph represents the number or proportion of a data set less than or equal to a given number.
1.5.UNC-1.G.5	There are many additional ways to graphically represent distributions of quantitative data.
1.6.UNC-1.H.1	Descriptions of the distribution of quantitative data include shape, center, and variability (spread), as well as any unusual features such as outliers, gaps, clusters, or multiple peaks.
1.6.UNC-1.H.2	Outliers for one-variable data are data points that are unusually small or large relative to the rest of the data.
1.6.UNC-1.H.3	A distribution is skewed to the right (positive skew) if the right tail is longer than the left. A distribution is skewed to the left (negative skew) if the left tail is longer than the right. A distribution is symmetric if the left half is the mirror image of the right half.
1.6.UNC-1.H.4	Univariate graphs with one main peak are known as unimodal. Graphs with two prominent peaks are bimodal. A graph where each bar height is approximately the same (no prominent peaks) is approximately uniform.
1.6.UNC-1.H.5	A gap is a region of a distribution between two data values where there are no observed data.
1.6.UNC-1.H.6	Clusters are concentrations of data usually separated by gaps.
1.6.UNC-1.H.7	Descriptive statistics does not attribute properties of a data set to a larger population, but may provide the basis for conjectures for subsequent testing.
1.7.UNC-1.I.1	A statistic is a numerical summary of sample data.

1.7.UNC-1.I.2	The mean is the sum of all the data values divided by the number of values. For a sample, the mean is denoted by x-bar: $\bar{x}$ = 1/n Σ [i = 1 to n] = x_i, where x_i represents the ith data point in the sample and n represents the number of data values in the sample.
1.7.UNC-1.I.3	The median of a data set is the middle value when data are ordered. When the number of data points is even, the median can take on any value between the two middle values. In AP Statistics, the most commonly used value for the median of a data set with an even number of values is the average of the two middle values.
1.7.UNC-1.I.4	The first quartile, Q1, is the median of the half of the ordered data set from the minimum to the position of the median. The third quartile, Q3, is the median of the half of the ordered data set from the position of the median to the maximum. Q1 and Q3 form the boundaries for the middle 50% of values in an ordered data set.
1.7.UNC-1.I.5	The p^{th} percentile is interpreted as the value that has p% of the data less than or equal to it.
1.7.UNC-1.J.1	Three commonly used measures of variability (or spread) in a distribution are the range, interquartile range, and standard deviation.
1.7.UNC-1.J.2	The range is defined as the difference between the maximum data value and the minimum data value. The interquartile range (IQR) is defined as the difference between the third and first quartiles: Q3 − Q1. Both the range and the interquartile range are possible ways of measuring variability of the distribution of a quantitative variable.
1.7.UNC-1.J.3	Standard deviation is a way to measure variability of the distribution of a quantitative variable. For a sample, the standard deviation is denoted by s: s_x = √[1/(n – 1) Σ(x_i – $\bar{x}$)²]. The square of the sample standard deviation, s^2, is called the sample variance.
1.7.UNC-1.J.4	Changing units of measurement affects the values of the calculated statistics.
1.7.UNC-1.K.1.i	An outlier is a value greater than 1.5 × IQR above the third quartile or more than 1.5 × IQR below the first quartile.
1.7.UNC-1.K.1.ii	An outlier is a value located 2 or more standard deviations above, or below, the mean.
1.7.UNC-1.K.2	The mean, standard deviation, and range are considered nonresistant (or non-robust) because they are influenced by outliers. The median and IQR are considered resistant (or robust), because outliers do not greatly (if at all) affect their value.
1.8.UNC-1.L.1	Taken together, the minimum data value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum data value make up the five-number summary.
1.8.UNC-1.L.2	A boxplot is a graphical representation of the five-number summary (minimum, first quartile, median, third quartile, maximum). The box represents the middle 50% of data, with a line at the median and the ends of the box corresponding to the quartiles. Lines (“whiskers”) extend from the quartiles to the most extreme point that is not an outlier, and outliers are indicated by their own symbol beyond this.
1.8.UNC-1.M.1	Summary statistics of quantitative data, or of sets of quantitative data, can be used to justify claims about the data in context.
1.8.UNC-1.M.2	If a distribution is relatively symmetric, then the mean and median are relatively close to one another. If a distribution is skewed right, then the mean is usually to the right of the median. If the distribution is skewed left, then the mean is usually to the left of the median.
1.9.UNC-1.N.1	Any of the graphical representations, e.g., histograms, side-by-side boxplots, etc., can be used to compare two or more independent samples on center, variability, clusters, gaps, outliers, and other features.
1.9.UNC-1.O.1	Any of the numerical summaries (e.g., mean, standard deviation, relative frequency, etc.) can be used to compare two or more independent samples.
1.10.VAR-2.A.1	A parameter is a numerical summary of a population.
1.10.VAR-2.A.2	Some sets of data may be described as approximately normally distributed. A normal

	curve is mound-shaped and symmetric. The parameters of a normal distribution are the population mean, μ, and the population standard deviation, σ.
1.10.VAR-2.A.3	For a normal distribution, approximately 68% of the observations are within 1 standard deviation of the mean, approximately 95% of observations are within 2 standard deviations of the mean, and approximately 99.7% of observations are within 3 standard deviations of the mean. This is called the empirical rule.
1.10.VAR-2.A.4	Many variables can be modeled by a normal distribution.
1.10.VAR-2.B.1	A standardized score for a particular data value is calculated as (data value – mean)/(standard deviation), and measures the number of standard deviations a data value falls above or below the mean.
1.10.VAR-2.B.2	One example of a standardized score is a z-score, which is calculated as z-score = ((x_i – μ)/σ). A z-score measures how many standard deviations a data value is from the mean.
1.10.VAR-2.B.3	Technology, such as a calculator, a standard normal table, or computer-generated output, can be used to find the proportion of data values located on a given interval of a normally distributed random variable.
1.10.VAR-2.B.4	Given the area of a region under the graph of the normal distribution curve, it is possible to use technology, such as a calculator, a standard normal table, or computer-generated output, to estimate parameters for some populations.
1.10.VAR-2.C.1	Percentiles and z-scores may be used to compare relative positions of points within a data set or between data sets.
2.1.VAR-1.D.1	Apparent patterns and associations in data may be random or not.
2.2.UNC-1.P.1	Side-by-side bar graphs, segmented bar graphs, and mosaic plots are examples of bar graphs for one categorical variable, broken down by categories of another categorical variable.
2.2.UNC-1.P.2	Graphical representations of two categorical variables can be used to compare distributions and/or determine if variables are associated.
2.2.UNC-1.P.3	A two-way table, also called a contingency table, is used to summarize two categorical variables. The entries in the cells can be frequency counts or relative frequencies.
2.2.UNC-1.P.4	A joint relative frequency is a cell frequency divided by the total for the entire table.
2.3.UNC-1.Q.1	The marginal relative frequencies are the row and column totals in a two-way table divided by the total for the entire table.
2.3.UNC-1.Q.2	A conditional relative frequency is a relative frequency for a specific part of the contingency table (e.g., cell frequencies in a row divided by the total for that row).
2.3.UNC-1.R.1	Summary statistics for two categorical variables can be used to compare distributions and/or determine if variables are associated.
2.4.UNC-1.S.1	A bivariate quantitative data set consists of observations of two different quantitative variables made on individuals in a sample or population.
2.4.UNC-1.S.2	A scatterplot shows two numeric values for each observation, one corresponding to the value on the x-axis and one corresponding to the value on the y-axis.
2.4.UNC-1.S.3	An explanatory variable is a variable whose values are used to explain or predict corresponding values for the response variable.
2.4.DAT-1.A.1	A description of a scatter plot includes form, direction, strength, and unusual features.
2.4.DAT-1.A.2	The direction of the association shown in a scatterplot, if any, can be described as positive or negative.
2.4.DAT-1.A.3	A positive association means that as values of one variable increase, the values of the other variable tend to increase. A negative association means that as values of one variable increase, values of the other variable tend to decrease.
2.4.DAT-1.A.4	The form of the association shown in a scatterplot, if any, can be described as linear or

	non-linear to varying degrees.
2.4.DAT-1.A.5	The strength of the association is how closely the individual points follow a specific pattern, e.g., linear, and can be shown in a scatterplot. Strength can be described as strong, moderate, or weak.
2.4.DAT-1.A.6	Unusual features of a scatter plot include clusters of points or points with relatively large discrepancies between the value of the response variable and a predicted value for the response variable.
2.5.DAT-1.B.1	The correlation, r, gives the direction and quantifies the strength of the linear association between two quantitative variables.
2.5.DAT-1.B.2	The correlation coefficient can be calculated by: r = 1/(n − 1)Σ((x_i −¯x)/s_x)((y_i − ¯y)/s[subscript y]). However, the most common way to determine r is by using technology.
2.5.DAT-1.B.3	A correlation coefficient close to 1 or −1 does not necessarily mean that a linear model is appropriate.
2.5.DAT-1.C.1	The correlation, r, is unit-free, and always between −1 and 1, inclusive. A value of r = 0 indicates that there is no linear association. A value of r = 1 or r = −1 indicates that there is a perfect linear association.
2.5.DAT-1.C.2	A perceived or real relationship between two variables does not mean that changes in one variable cause changes in the other. That is, correlation does not necessarily imply causation.
2.6.DAT-1.D.1	A simple linear regression model is an equation that uses an explanatory variable, x, to predict the response variable, y.
2.6.DAT-1.D.2	The predicted response value, denoted by $y\hat{}$, is calculated as $y\hat{}$ = a + bx, where a is the y-intercept and b is the slope of the regression line, and x is the value of the explanatory variable.
2.6.DAT-1.D.3	Extrapolation is predicting a response value using a value for the explanatory variable that is beyond the interval of x-values used to determine the regression line. The predicted value is less reliable as an estimate the further we extrapolate.
2.7.DAT-1.E.1	The residual is the difference between the actual value and the predicted value: residual = $y - y\hat{}$.
2.7.DAT-1.E.2	A residual plot is a plot of residuals versus explanatory variable values or predicted response values.
2.7.DAT-1.F.1	Apparent randomness in a residual plot for a linear model is evidence of a linear form to the association between the variables.
2.7.DAT-1.F.2	Residual plots can be used to investigate the appropriateness of a selected model.
2.8.DAT-1.G.1	The least-squares regression model minimizes the sum of the squares of the residuals and contains the point ¯x, ¯y).
2.8.DAT-1.G.2	The slope, b, of the regression line can be calculated as b = r(s[subscript y]/s_x) where r is the correlation between x and y, s[subscript y] is the sample standard deviation of the response variable, y, and s_x is the sample standard deviation of the explanatory variable, x.
2.8.DAT-1.G.3	Sometimes, the y-intercept of the line does not have a logical interpretation in context.
2.8.DAT-1.G.4	In simple linear regression, r^2 is the square of the correlation, r. It is also called the coefficient of determination. r^2 is the proportion of variation in the response variable that is explained by the explanatory variable in the model.
2.8.DAT-1.H.1	The coefficients of the least-squares regression model are the estimated slope and y-intercept.
2.8.DAT-1.H.2	The slope is the amount that the predicted y-value changes for every unit increase in x.
2.8.DAT-1.H.3	The y-intercept value is the predicted value of the response variable when the explanatory

	variable is equal to 0. The formula for the y-intercept, a, is $a = \bar{y} - b\bar{x}$.
2.9.DAT-1.I.1	An outlier in regression is a point that does not follow the general trend shown in the rest of the data and has a large residual when the Least Squares Regression Line (LSRL) is calculated.
2.9.DAT-1.I.2	A high-leverage point in regression has a substantially larger or smaller x-value than the other observations have.
2.9.DAT-1.I.3	An influential point in regression is any point that, if removed, changes the relationship substantially. Examples include much different slope, y-intercept, and/or correlation. Outliers and high leverage points are often influential.
2.9.DAT-1.J.1	Transformations of variables, such as evaluating the natural logarithm of each value of the response variable or squaring each value of the explanatory variable, can be used to create transformed data sets, which may be more linear in form than the untransformed data.
2.9.DAT-1.J.2	Increased randomness in residual plots after transformation of data and/or movement of r^2 to a value closer to 1 offers evidence that the least-squares regression line for the transformed data is a more appropriate model to use to predict responses to the explanatory variable than the regression line for the untransformed data.
3.1.VAR-1.E.1	Methods for data collection that do not rely on chance result in untrustworthy conclusions.
3.2.DAT-2.A.1	A population consists of all items or subjects of interest.
3.2.DAT-2.A.2	A sample selected for study is a subset of the population.
3.2.DAT-2.A.3	In an observational study, treatments are not imposed. Investigators examine data for a sample of individuals (retrospective) or follow a sample of individuals into the future collecting data (prospective) in order to investigate a topic of interest about the population. A sample survey is a type of observational study that collects data from a sample in an attempt to learn about the population from which the sample was taken.
3.2.DAT-2.A.4	In an experiment, different conditions (treatments) are assigned to experimental units (participants or subjects).
3.2.DAT-2.B.1	It is only appropriate to make generalizations about a population based on samples that are randomly selected or otherwise representative of that population.
3.2.DAT-2.B.2	A sample is only generalizable to the population from which the sample was selected.
3.2.DAT-2.B.3	It is not possible to determine causal relationships between variables using data collected in an observational study.
3.3.DAT-2.C.1	When an item from a population can be selected only once, this is called sampling without replacement. When an item from the population can be selected more than once, this is called sampling with replacement.
3.3.DAT-2.C.2	A simple random sample (SRS) is a sample in which every group of a given size has an equal chance of being chosen. This method is the basis for many types of sampling mechanisms. A few examples of mechanisms used to obtain SRSs include numbering individuals and using a random number generator to select which ones to include in the sample, ignoring repeats, using a table of random numbers, or drawing a card from a deck without replacement.
3.3.DAT-2.C.3	A stratified random sample involves the division of a population into separate groups, called strata, based on shared attributes or characteristics (homogeneous grouping). Within each stratum a simple random sample is selected, and the selected units are combined to form the sample.
3.3.DAT-2.C.4	A cluster sample involves the division of a population into smaller groups, called clusters. Ideally, there is heterogeneity within each cluster, and clusters are similar to one another in their composition. A simple random sample of clusters is selected from the population to form the sample of clusters. Data are collected from all observations in the selected

	clusters.
3.3.DAT-2.C.5	A systematic random sample is a method in which sample members from a population are selected according to a random starting point and a fixed, periodic interval.
3.3.DAT-2.C.6	A census selects all items/subjects in a population.
3.3.DAT-2.D.1	There are advantages and disadvantages for each sampling method depending upon the question that is to be answered and the population from which the sample will be drawn.
3.4.DAT-2.E.1	Bias occurs when certain responses are systematically favored over others.
3.4.DAT-2.E.2	When a sample is comprised entirely of volunteers or people who choose to participate, the sample will typically not be representative of the population (voluntary response bias).
3.4.DAT-2.E.3	When part of the population has a reduced chance of being included in the sample, the sample will typically not be representative of the population (undercoverage bias).
3.4.DAT-2.E.4	Individuals chosen for the sample for whom data cannot be obtained (or who refuse to respond) may differ from those for whom data can be obtained (nonresponse bias).
3.4.DAT-2.E.5	Problems in the data gathering instrument or process result in response bias. Examples include questions that are confusing or leading (question wording bias) and self-reported responses.
3.4.DAT-2.E.6	Non-random sampling methods (for example, samples chosen by convenience or voluntary response) introduce potential for bias because they do not use chance to select the individuals.
3.5.VAR-3.A.1	The experimental units are the individuals (which may be people or other objects of study) that are assigned treatments. When experimental units consist of people, they are sometimes referred to as participants or subjects.
3.5.VAR-3.A.2	An explanatory variable (or factor) in an experiment is a variable whose levels are manipulated intentionally. The levels or combination of levels of the explanatory variable(s) are called treatments.
3.5.VAR-3.A.3	A response variable in an experiment is an outcome from the experimental units that is measured after the treatments have been administered.
3.5.VAR-3.A.4	A confounding variable in an experiment is a variable that is related to the explanatory variable and influences the response variable and may create a false perception of association between the two.
3.5.VAR-3.B.1.a	Comparisons of at least two treatment groups, one of which could be a control group.
3.5.VAR-3.B.1.b	Random assignment/allocation of treatments to experimental units.
3.5.VAR-3.B.1.c	Replication (more than one experimental unit in each treatment group).
3.5.VAR-3.B.1.d	Control of potential confounding variables where appropriate.
3.5.VAR-3.C.1	In a completely randomized design, treatments are assigned to experimental units completely at random. Random assignment tends to balance the effects of uncontrolled (confounding) variables so that differences in responses can be attributed to the treatments.
3.5.VAR-3.C.2	Methods for randomly assigning treatments to experimental units in a completely randomized design include using a random number generator, a table of random values, drawing chips without replacement, etc.
3.5.VAR-3.C.3	In a single-blind experiment, subjects do not know which treatment they are receiving, but members of the research team do, or vice versa.
3.5.VAR-3.C.4	In a double-blind experiment neither the subjects nor the members of the research team who interact with them know which treatment a subject is receiving.
3.5.VAR-3.C.5	A control group is a collection of experimental units either not given a treatment of interest or given a treatment with an inactive substance (placebo) in order to determine if the treatment of interest has an effect.

3.5.VAR-3.C.6	The placebo effect occurs when experimental units have a response to a placebo.
3.5.VAR-3.C.7	For randomized complete block designs, treatments are assigned completely at random within each block.
3.5.VAR-3.C.8	Blocking ensures that at the beginning of the experiment the units within each block are similar to each other with respect to at least one blocking variable. A randomized block design helps to separate natural variability from differences due to the blocking variable.
3.5.VAR-3.C.9	A matched pairs design is a special case of a randomized block design. Using a blocking variable, subjects (whether they are people or not) are arranged in pairs matched on relevant factors. Matched pairs may be formed naturally or by the experimenter. Every pair receives both treatments by randomly assigning one treatment to one member of the pair and subsequently assigning the remaining treatment to the second member of the pair. Alternately, each subject may get both treatments.
3.6.VAR-3.D.1	There are advantages and disadvantages for each experimental design depending on the question of interest, the resources available, and the nature of the experimental units.
3.7.VAR-3.E.1	Statistical inference attributes conclusions based on data to the distribution from which the data were collected.
3.7.VAR-3.E.2	Random assignment of treatments to experimental units allows researchers to conclude that some observed changes are so large as to be unlikely to have occurred by chance. Such changes are said to be statistically significant.
3.7.VAR-3.E.3	Statistically significant differences between or among experimental treatment groups are evidence that the treatments caused the effect.
3.7.VAR-3.E.4	If the experimental units used in an experiment are representative of some larger group of units, the results of an experiment can be generalized to the larger group. Random selection of experimental units gives a better chance that the units will be representative.

Standards for Mathematical Practice

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

Unit Focus

Enduring Understandings	Essential Questions
Part 1: Exploring One-Variable Data • Given that variation may be random or not, conclusions are uncertain. • Graphical representations and statistics allow us to identify and represent key features of data.	*Part 1: Exploring One-Variable Data* • Is my cat old, compared to other cats? • How certain are we that what seems to be a pattern is not just a coincidence?

• The normal distribution can be used to represent some population distributions. *Part 2: Exploring Two-Variable Data* • Given that variation may be random or not, conclusions are uncertain. • Graphical representations and statistics allow us to identify and represent key features of data. • Regression models may allow us to predict responses to changes in an explanatory variable. *Part 3: Collecting Data* • Given that variation may be random or not, conclusions are uncertain. • The way we collect data influences what we can and cannot say about a population. • Well-designed experiments can establish evidence of causal relationships.	*Part 2: Exploring Two-Variable Data* • Does the fact that the number of shark attacks increases with ice cream sales necessarily mean that ice cream sales cause shark attacks? • How might you represent incomes of individuals with and without a college degree to help describe similarities and/or differences between the two groups? • How can you determine the effectiveness of a linear model that uses the number of cricket chirps per minute to predict temperature? *Part 3: Collecting Data* • What do our data tell us? • Why might the data we collected not be valid for drawing conclusions about an entire population?

Instructional Focus

Learning Targets
Part 1: Exploring One-Variable Data: • Identify questions to be answered, based on variation in one-variable data. • Identify variables in a set of data. • Classify types of variables. • Represent categorical data using frequency or relative frequency tables. • Describe categorical data represented in frequency or relative tables. • Represent categorical data graphically. • Describe categorical data represented graphically. • Compare multiple sets of categorical data. • Classify types of quantitative variables. • Represent quantitative data graphically. • Describe the characteristics of quantitative data distributions. • Calculate measures of center and position for quantitative data. • Calculate measures of variability for quantitative data. • Explain the selection of a particular measure of center and/or variability for describing a set of quantitative data.

- Represent summary statistics for quantitative data graphically.
- Describe summary statistics of quantitative data represented graphically.
- Compare graphical representations for multiple sets of quantitative data.
- Compare summary statistics for multiple sets of quantitative data.
- Compare a data distribution to the normal distribution model.
- Determine proportions and percentiles from a normal distribution.
- Compare measures of relative position in data sets.

Part 2: Exploring Two-Variable Data

- Identify questions to be answered about possible relationships in data.
- Compare numerical and graphical representations for two categorical variables.
- Calculate statistics for two categorical variables.
- Compare statistics for two categorical variables.
- Represent bivariate quantitative data using scatterplots.
- Describe the characteristics of a scatter plot.
- Determine the correlation for a linear relationship.
- Interpret the correlation for a linear relationship.
- Calculate a predicted response value using a linear regression model.
- Represent differences between measured and predicted responses using residual plots.
- Describe the form of association of bivariate data using residual plots.
- Estimate parameters for the least-squares regression line model.
- Interpret coefficients for the least-squares regression line model.
- Identify influential points in regression.
- Calculate a predicted response using a leastsquares regression line for a transformed data set.

Part 3: Collecting Data

- Identify questions to be answered about data collection methods.
- Identify the type of a study.
- Identify appropriate generalizations and determinations based on observational studies.
- Identify a sampling method, given a description of a study.
- Explain why a particular sampling method is or is not appropriate for a given situation.
- Identify potential sources of bias in sampling methods.
- Identify the components of an experiment.
- Describe elements of a well-designed experiment.
- Compare experimental designs and methods.
- Explain why a particular experimental design is appropriate.
- Interpret the results of a well-designed experiment.

Prerequisite Skills

- Ability to read and interpret word problems and academic texts.
- Comfort with explaining thinking in writing and orally.
- Understanding of cause and effect, comparisons, and trends in everyday contexts.
- Basic understanding of drawing conclusions from evidence.

- Identify individuals and variables in a data set (e.g., from prior science or math experiences).
- Differentiate between types of variables (categorical vs. numerical).
- Create and interpret bar graphs, dot plots, and basic histograms.
- Calculate and interpret mean, median, mode, and range.
- Understand order and ranking of values.
- Calculate percentages and proportions.
- Understand and apply concepts of variability (e.g., differences between numbers).
- Work with and compare units (such as inches vs. centimeters).
- Use a calculator or technology to input data and compute summary statistics.
- Recognize and interpret basic normal distribution shapes (bell curve familiarity is helpful).
- Recognize ordered pairs and coordinate graphs from algebra (x-y relationships).
- Plot and interpret points on the Cartesian plane.
- Understand linear relationships and slope from Algebra I or II.
- Solve and interpret linear equations and expressions.
- Recognize trends and associations visually from scatterplots.
- Calculate and interpret proportions and percentages in two-way tables.
- Basic understanding of correlation and residuals from prior algebra work (helpful, but not always required).
- Identify the purpose of a study or investigation.
- Recognize differences between observation and experimentation (often covered in science classes).
- Understand random sampling and its role in fairness and generalization (basic probability concepts).
- Distinguish between population and sample.
- Recognize bias in surveys or studies (e.g., from real-life examples or media).
- Follow a procedure or steps in a scientific method (often from middle or high school science).
- Communicate logical reasoning and justify choices or interpretations.

Common Misconceptions

- Confusing categorical and quantitative variables

    - Students may classify numerical-looking variables (e.g., zip codes, phone numbers) as quantitative when they are actually categorical.

- Assuming the mean is always the best measure of center
    - Students often default to the mean without considering the impact of skewed data or outliers.
- Thinking variability is only about the range
    - Students may not understand standard deviation or IQR and use the range as the sole indicator of spread.
- Misinterpreting histograms and boxplots
    - Students may think the height of a histogram bar represents individual data values or misunderstand what quartiles in a boxplot represent.
- Using frequency and relative frequency interchangeably
    - Students may confuse counts with proportions or percentages, especially when comparing distributions.
- Overgeneralizing the normal distribution
    - Students may assume all data follows a bell-shaped curve or apply the Empirical Rule inappropriately.
- Assuming correlation implies causation
    - Students may incorrectly conclude that a strong relationship between two variables proves one causes the other.
- Misinterpreting correlation coefficients
    - Students may believe correlation and slope are the same, or misjudge the strength and direction of association based on correlation alone.
- Ignoring nonlinear associations
    - Students may think that if the data isn't linear, there's no relationship—even when a curved trend is evident.
- Drawing regression lines visually without justification
    - Students may estimate lines of best fit by eye and assume they are accurate without calculating or using technology.
- Overlooking influential points or outliers in scatterplots
    - Students may not realize how a single point can drastically affect correlation or the regression line.

- Believing all relationships can be made linear through transformation
    - Students may expect a transformation to always yield a valid linear model, even when it's inappropriate.
- Confusing observational studies with experiments
    - Students may not recognize that only experiments involve treatments and random assignment.
- Assuming all surveys or samples are unbiased
    - Students may not consider how poor wording, voluntary response, or undercoverage introduce bias.
- Thinking large samples always ensure accuracy
    - Students may believe sample size alone determines validity, ignoring sampling method and design.
- Confusing random sampling with random assignment
    - Students may not understand that random sampling allows generalization to a population, while random assignment supports causal inference.
- Misinterpreting experimental diagrams or design
    - Students may incorrectly draw conclusions about causality or generalizability based on poor understanding of the design.
- Assuming blocking is always necessary
    - Students may want to include blocking for every variable, even when it doesn't address a source of variability in the response.

Spiraling For Mastery

Current Unit Content/Skills	Spiral Focus	Activity
• **Topic 1 Intro:** ○ Identify the question to be answered or problem to be solved. • **1A The Language of Variation:** ○ Individuals and variables	• **Describing and Displaying Data** ○ ***Categorical vs. Quantitative*** Introduced in 1A and revisited when choosing display types (bar graphs, dot plots, boxplots,	• **"Same Variable, New Context" Warm-Ups** ○ Revisit prior skills in new scenarios. ▪ Provide a new dot plot and ask for shape, center, and spread. ▪ Give a two-way table

- Classify variables and categorical or quantitative
- Make and Interpret a frequency table or a relative frequency table for a distribution of data.

- **1B - Displaying and Describing Categorical Data**
  - Make and Interpret bar graphs of categorical data.
  - Compare distributions of categorical data.
  - Identify what makes some graphs of categorical data misleading.
- **1C - Displaying Quantitative Data with Graphs**
    - Make and interpret dot plots of quantitative data.
    - Describe the distribution of a quantitative variable.
    - Compare distributions of quantitative data.
    - Make and interpret stemplots of quantitative data.
    - Make and interpret histograms of quantitative data.
- **1D - Describing Quantitative Data with Numbers**
    - Find the median of a distribution of quantitative data.
    - Calculate the mean of a distribution of quantitative data.

etc.), calculating statistics, and interpreting results in 2A/2B.

- ***Distributions and Comparative Language***
  First appears in 1B–1D and spirals through 2A (comparing two-way tables) and 2B (comparing scatterplot data). Revisits again in experimental results (3C).
- ***Graphical Representation***
  Graphing skills build from 1B–1F (bar graphs, dot plots, histograms, box plots, etc.) and spiral into 2B (scatterplots), 2C (residual plots), and 3B (survey representation).
- ***Center and Variability***
  Measures like mean, median, IQR, and standard deviation from 1D spiral into data comparisons in 1F (normal distributions) and 2C (standard deviation of residuals).
- ***Z-scores and Percentiles***
  First introduced in 1E and revisited when comparing across distributions (e.g., transformed models in 2D or estimating population responses in 3C).

- **Analyzing Relationships Between Variables**
  - ***Direction, Form,***

and ask students to identify marginal/conditional distributions.
    - Use a boxplot from a different context (e.g., SAT scores vs. race times) for comparison practice.
- **Data Matching Tasks**
  - Encourage students to connect multiple representations.
    - Give students cards with: Histograms, Boxplots, Summary statistics, Descriptions in words. Students match the visual, numerical, and verbal representations of the same data set.
- **"Design an Investigation" Projects**
  - Integrate data collection, display, and analysis.
    - Students pose a question, identify variables (categorical vs. quantitative), decide on sampling/experimental design, and plan analysis.
    - Can be structured over multiple lessons.
- **Graph Gallery Walk**
  - Interpret and critique graphs and visual displays.
    - Post different graphs (some with errors/misleading elements). Students rotate, interpret, and leave feedback or correct errors.

- **Statistical Scramble (Spiral Review Stations)**
  - Rotating practice that combines multiple concepts.
    - Set up stations with mixed tasks:
      - Station 1: Describe a distribution

- Find the range of a distribution of quantitative data.
- Calculate and interpret the standard deviation of a distribution of quantitative data.
- Find the interquartile range (IQR) of a distribution of quantitative data.
- Choose appropriate measures of center and variability to summarize a distribution of quantitative data.
- Identify outliers in a distribution of quantitative data.
- Make and interpret box plots of quantitative data.
- Use box plots and summary statistics to compare distribution of quantitative data.

- **1E - Describing Position and Transforming Data**
  - Calculate and interpret a percentile in a distribution of quantitative data.
  - Calculate and interpret a standardized score (z-score) in a distribution of quantitative data.
  - Use percentiles or standardizes scores (z-scores) to compare the relative positions of individual values in

and Strength
Initial concepts in 2B continue into regression analysis in 2C and evaluation of linearity in 2D.

- ***Residuals and Model Appropriateness***
  Connects visual interpretation skills (1C–1D) with model evaluation in 2C and 2D. Interpreting residual plots relies on understanding spread and center.
- ***Correlation ≠ Causation***
  Taught in 2B and spiraled into 3A–3C during discussion of observational vs. experimental design and appropriate inferences.
- ***Influence of Outliers***
  First addressed in 1D and revisited in 2D (how outliers affect regression, r^2, and model fit).

- **Designing and Collecting Data**
  - ***Study Design Vocabulary (Population, Sample, Bias, etc.)***
    Vocabulary and classification skills from 1A are reused when evaluating sampling methods and identifying sources of bias.
  - ***Randomness and Representativeness***
    Connects back to ideas of variability from 1D and fairness in data comparison (2B)

(dot plot, histogram)

- Station 2: Calculate mean, IQR, SD
- Station 3: Make a boxplot and identify outliers
- Station 4: Interpret scatterplots and correlation
- Station 5: Identify bias in a study design

- **"Defend Your Choice" Discussions**
  - Reinforce conceptual decision-making.
    - "Should we use the mean or median here? Why?"
    - "Is this sampling method appropriate?"
    - "Does the graph show a misleading representation?"

- **Quick Writes with Prompts**
  - Strengthen statistical communication and reasoning.
    - "Describe the shape, center, and spread of this distribution."
    - "Explain whether the study supports a cause-and-effect conclusion."
    - "Compare two data sets using both visual and numerical

distributions of quantitative data. • Use a cumulative relative frequency graph to estimate percentiles and individual values in a distribution of quantitative data. • Analyze the effect of adding, subtraction, multiplying by, or dividing by a constant on the shape, center, and variability of a distribution of quantitative data. • **1F - Normal Distributions** • Draw a normal curve to model the distribution of a quantitative variable. • Use the empirical rule to estimate the proportion of values in a specified interval in a normal distribution. • Determine whether a distribution of quantitative data is approximately normal using graphical and numerical evidence. • Find the proportion of values in a specified interval in a normal distribution. • Find the value that corresponds to a given percentile in a normal distribution. • Calculate the mean or standard deviation of a	and is central to drawing valid inferences from studies (3B–3C). ○ ***Interpreting Experimental Results*** Inference skills developed from graphical and numerical comparisons (1D–1F, 2C–2D) are used when interpreting findings from experiments (3C). ○ ***Cause and Effect vs. Generalization*** Links back to distinguishing correlation from causation (2B) and extends to justify valid conclusions based on study design (3A–3C).	summaries." • **Tech-Enhanced Exploration with Data Tools** ○ Reinforce skills using real data and tools like Desmos, CODAP, or StatKey. ▪ Create and interpret graphs. ▪ Fit regression models and assess residuals. ▪ Simulate sampling or random assignment. • **Exit Tickets with Cumulative Prompts** ○ Daily spiral and quick formative assessment. ▪ Example prompt: "Today we talked about sampling bias. On your ticket, also write down one way we used dot plots earlier this week." • **Weekly Spiral Quizzes** ○ Interleave past and present learning to build fluency. ○ Structure each short quiz to include: ▪ 1–2 questions from current content ▪ 1 question from last week ▪ 1 question from a prior unit • **AP Exam Free Response Spiral Practice (FRAPPYs)** ○ Practice real AP problems aligned to both current and past units. ▪ Choose problems that span multiple skills (e.g., describing distributions,

normal distribution given the value of a percentile.
 	
• **2A - Relationships Between Two Categorical Values:**
  • Identify the explanatory and response variables in a given setting.
  • Calculate statistics or two categorical variables.
  • Display the relationship between two categorical variables.
  • Describe the relationship between two categorical variables.
• **2B - Relationships Between Two Quantitative Variables**
  • Make a scatter plot to display the relationship between two quantitative variables.
  • Describe the direction, form, and strength of a relationship displayed in a scatter plot and identify unusual features.
  • Interpret the correlation for a linear relationship between two quantitative variables.
  • Distinguish correlation from causation.
		interpreting simulation results, calculating probabilities).
		▪ Use a gradual release: Individual → Group → Class Discussion.
		▪ Include student reflection: “Which part felt familiar? From which earlier unit?”

• Calculate the correlation of a linear relationship between two quantitative variables. • **2C - Linear Regression Models** • Make predictions using a regressions line, keeping in mind the dangers of extrapolation. • Calculate and interpret a residual. • Interpret the slope and y intercept of a regression line. • Determine the equation of a least-squares regression line using formulas. • Determine the equation of a least-squares regression line using technology. • Construct and interpret residual plots to assess whether a regression model is appropriate. • Interpret the coefficient of determination r^2 and the standard deviation of the residuals s. • **2D - Analyzing Departures from Linearity** • Describe how unusual points influence the least-squares regression line, the correlation, r^2, and the standard deviation of the residuals. • Calculate the		

predicted values from linear models using variables that have been transformed with logarithms or powers. • Determine which of several models does a better job of describing the relationship between two quantitative variables. • **3A - Introduction to Data Collection** • Identify the population and sample in a statistical study. • Distinguish between an observational study and an experiment. • Determine what inferences are appropriate from an observational study. • **3B - Sampling and Surveys** • Describe how to select a simple random sample. • Describe other random sampling methods: stratified, cluster, systematic; explain the advantages and disadvantages of each method. • Identify voluntary response sampling and convenience sampling, and explain how these sampling methods can lead to bias. • Explain how undercoverage, nonresponse,		

questions wording, and other aspects of a sample survey can lead to bias.
 	
• **3C - Experiments**
 	
  • Explain the concept of confounding and how it limits the ability to make cause-and-effect conclusions.
 	
  • Identify the experimental units and treatments in an experiment.
 	
  • Explain the purpose of a control group in an experiment.
 	
  • Describe the placebo effect and explain the purpose of blinding in an experiment.
 	
  • Describe how to randomly assign treatments in an experiment and explain the purpose of random assignment.
 	
  • Explain the purpose of controlling other variables in an experiment.
 	
  • Describe a randomized block design for an experiment and explain the benefits of blocking in an experiment.
 	
  • Describe a matched pairs design for an experiment.
 	
  • Explain the meaning of statistically significant in the context of an		

experiment. • Identify when it is appropriate to make an inference about a population and when it is appropriate to make an inference about cause and effect.		

Assessment

Formative Assessment	Summative Assessment
• Homework • Lesson Checks • Quizzes • Exit Tickets • Lesson Reflections • Performance Tasks • AP Classroom Progress Checks	• Topic Tests • Unit 1 Benchmark (Link-It)

Resources

Key Resources	Supplemental Resources
AP Classroom AP Statistics CED Pacing Guide	iXL Delta Math Desmos Khan Academy Math Medic Skew the Script Teacher Made worksheets Textbook - Starnes, D., & Tabor, J. (2024). *The Practice of Statistics for the AP Classroom* (7th ed.).

	Beford, Freeman & Worth.

Career Readiness, Life Literacies, and Key Skills

9.4.12.CI.1	Demonstrate the ability to reflect, analyze, and use creative skills and ideas (e.g., 1.1.12prof.CR3a).
9.2.12.CAP.4	Evaluate different careers and develop various plans (e.g., costs of public, private, training schools) and timetables for achieving them, including educational/training requirements, costs, loans, and debt repayment.
9.4.12.CT.2	Explain the potential benefits of collaborating to enhance critical thinking and problem solving (e.g., 1.3E.12profCR3.a).
9.4.12.IML.3	Analyze data using tools and models to make valid and reliable claims, or to determine optimal design solutions (e.g., S-ID.B.6a., 8.1.12.DA.5, 7.1.IH.IPRET.8).
9.4.12.IML.7	Develop an argument to support a claim regarding a current workplace or societal/ethical issue such as climate change (e.g., NJSLSA.W1, 7.1.AL.PRSNT.4).
9.4.12.TL.4	Collaborate in online learning communities or social networks or virtual worlds to analyze and propose a resolution to a real-world problem (e.g., 7.1.AL.IPERS.6).

Interdisciplinary Connections

RI.CR.11–12.1	Accurately cite a range of thorough textual evidence and make relevant connections to strongly support a comprehensive analysis of multiple aspects of what an informational text says explicitly and inferentially, as well as interpretations of the text.
8.1.12.DA.5	Create data visualizations from large data sets to summarize, communicate, and support different interpretations of real-world phenomena.
8.1.12.DA.6	Create and refine computational models to better represent the relationships among different elements of data collected from a phenomenon or process.
W.AW.11–12.1.B	Develop claim(s) and counterclaims avoiding common logical fallacies and using sound reasoning and thoroughly, supplying the most relevant evidence for each while pointing out the strengths and limitations of both in a manner that anticipates the audience's knowledge level, concerns, values, and possible biases.
SL.PE.11–12.1.A	Come to discussions prepared, having read and researched material under study; explicitly draw on that preparation by referring to evidence from texts and other research on the topic or issue to stimulate a thoughtful, well-reasoned exchange of ideas.
	Global economic activities involve decisions based on national interests, the exchange of different units of exchange, decisions of public and private institutions, and the ability to distribute goods and services safely.
SCI.HS-LS2-6	Evaluate the claims, evidence, and reasoning that the complex interactions in ecosystems maintain relatively consistent numbers and types of organisms in stable conditions, but changing conditions may result in a new ecosystem.