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
Unit 1 AP Statistics gives a comprehensive introduction to the statistical problem-solving cycle by synthesizing three foundational pillars: data production, graphical representation, and statistical modeling. Rather than presenting numbers in a vacuum, this 45-day curriculum uses the Experience First, Formalize Later (EFFL) model to ground each mathematical concept in a real-world context. By introducing data collection at the beginning of the year, students develop a key understanding that the integrity of any subsequent visual display, calculated statistic, or regression model is bound by how safely and representatively the underlying data was gathered. In the first phase of the unit, students move from managing raw categorical and quantitative variables to constructing and interpreting data displays such as stem-and-leaf plots, histograms, box plots, and mosaic plots. When evaluating quantitative data, students must move past isolated descriptions and address Shape, Outliers, Center, and Variability (CUSS) while weaving the problem's context into their descriptions. This phase also establishes methods for measuring relative standing through percentiles and z-scores, culminating in density curves and the Normal distribution. Students master the Empirical Rule (68-95-99.7) and use graphing calculators to execute standard proportion calculations and inverse boundary lookups. The second phase shifts the focus to the mechanics of sound research design, teaching students how to collect high-quality data via surveys and regulated experiments. Students learn to implement and identify sampling methods, including Simple Random Sampling (SRS), stratified sampling, cluster sampling, and systematic sampling, while weighing the operational advantages of each. Crucially, students learn to detect major sources of bias, such as undercoverage, nonresponse, and response bias, and	Starting the course with this unit gives students a firm foundation for AP Statistics right from the beginning. In most math classes, data is often just a set of numbers for students to work with. This unit takes a different approach by showing statistics as a full scientific process. Students begin with real-world questions, collect data systematically, and then use that data to describe and predict outcomes. By combining the College Board's Unit 1 (One-Variable Descriptive Statistics and Collecting Data) and Unit 5 (Two-Variable Relationships) with Math Medic's Unit 10 (Two-Variable Relationships), this unit highlights a key rule: the way data is collected determines what conclusions can be made. Students quickly learn that even the most elaborate calculations or graphs are useless if the data comes from poor sampling or biased research. Organizing the course in this way helps students understand more during each 47-minute class. Starting with data collection lets students see how statistics work in real situations before they deal with complex notation or formulas. When they later create displays such as mosaic plots, histograms, and boxplots, or calculate z-scores, they already comprehend the context and the variables involved. Introducing bivariate linear regression early also helps students apply the same principles—such as describing patterns, evaluating models, calculating residuals, and considering outliers—to both single- and two-variable data before moving on to probability and inference. This unit is also important for teaching students how to communicate clearly, which is important for the AP Statistics Exam. The College Board penalizes students for using certain language or for listing numbers without context. This unit addresses those issues by teaching clear frameworks for interpretation, such as CUSS & BS for distributions, DUFS & BS for scatterplots, and templates for regression slope, r^2, and s. Spreading this unit over

explicitly predict the direction of the bias as a systematic over- or underestimation of the true population parameter. By studying the four core principles of experimental design—Comparison, Randomization, Replication, and Control—across completely randomized, randomized block, and matched pairs layouts, students learn to determine a study's true scope of inference. They discover that random selection allows for generalization to a larger population, whereas random assignment is the unique mechanism required to establish a causal relationship. Finally, the unit concludes by examining bivariate relationships between two quantitative variables. Students begin by constructing scatterplots to analyze patterns observed from Direction, Form, Strength, and Outliers (DUFS) before building Least-Squares Regression Lines (LSRL) in the statistical format $y(hat) = a + bx$. To satisfy rigorous grading standards, students must master precise template scripts to interpret the real-world meaning of the slope (b), y-intercept (a), standard deviation of the residuals (s), and the coefficient of determination (r^2). Students are taught to avoid deterministic language, using probabilistic phrasing such as "on average" or "predicted value" to reflect the uncertainty inherent in statistical models. To ensure the highest level of mastery before their benchmark assessment, students learn to validate their linear models by auditing residual plots for random scatter and evaluating how high-leverage outliers or influential points change the trajectory and predictive strength of their regression lines.	45 days allows time for the Experience First, Formalize Later (EFFL) model, so students can discover concepts through collaboration before learning the formal concepts. This approach gives students a firm foundation in data collection, visual understanding, and explicit communication, helping them avoid common mistakes and preparing them to make strong statistical arguments throughout the year.

NJSLS

1.1.A.1	A statistical study is a study in which data are collected from a sample to answer an investigative question about a larger population.
1.1.A.2	Statistical studies are necessary when the population is too large or it is too difficult to collect data from every item or individual in the population.
1.1.A.3	A datum (singular form of data) is a piece of information about an item or individual. A collection of data is called a data set.
1.1.A.4	A population consists of all items or individuals of interest. The population size is represented by the symbol □.

1.1.A.5	A sample selected for study is a subset of the population from which data are obtained. The number of items in the sample, called the sample size, is represented by the symbol □.
1.1.A.6	Each component of a statistical study and the resulting calculations can be related to an aspect of the corresponding real-world context from which the components were derived. This identification of a statistical result with the corresponding contextual component is what is meant by “in context.”
1.1.B.1	An investigative question for a specific study should have a defined purpose and should not be changed based on the data analysis or results.
1.1.B.2	An investigative question should be posed so that the required data can be collected and analyzed.
1.2.A.1	An observational unit is an item or individual from which a datum is collected.
1.2.A.2	A variable is a characteristic that may change from one observational unit to another.
1.2.A.3	Data collected on numerical and categorical variables measured on observational units, including photographs, sounds, videos, and text, can convey meaningful information.
1.2.A.4	A parameter is a numerical attribute or summary of the variable of interest for a population.
1.2.A.5	A statistic is a numerical attribute or summary of the variable of interest for a sample. The value of a statistic from a certain sample is often not equal to the unknown value of the population parameter but may provide the basis for making inferences about the population parameter.
1.2.B.1	A categorical variable, also called a qualitative variable, takes on values that are category names or group labels.
1.2.B.2	A quantitative variable, also called a numerical variable, takes on numerical values for a measured or counted quantity and generally has units of measure.
1.2.C.1	A discrete quantitative variable can take on a countable number of values. The number of values may be finite or countably infinite, as with the whole numbers.
1.2.C.2	A continuous quantitative variable can take on an infinite number of possible values within a given interval. The number of values the variable can take on is measurable but not countable. This variable can take on all possible values between any pair of values.
1.3.A.1	A frequency table shows the number of observational units in each category of a categorical variable.
1.3.A.2	A relative frequency table shows the proportion of observational units in each category of a categorical variable.
1.3.B.1	Percentages, relative frequencies, and ratios all provide the same information as proportions.
1.3.B.2	Counts and relative frequencies of categorical variables reveal information that can be used to justify claims about the variables in context.
1.4.A.1	Bar charts, also called bar graphs, display frequencies (counts) or relative frequencies (proportions) for the categories of a single categorical variable. Each bar on a bar chart represents a category of the categorical variable of interest. The height or length of each bar corresponds to the frequency or relative frequency of the observational units in each category.
1.4.A.2	Pie charts are used to display frequencies (counts) or relative frequencies (proportions) for categorical data. Each slice on a pie chart represents a category of the categorical variable of interest. The area of each slice, as a fraction of the total area, corresponds to the relative frequency of observational units falling within each category. The sum of the slices’ areas together will equal 1, or 100% of the total area.
1.4.B.1	Graphical representations of a categorical variable reveal information that can be used to justify claims about the variable in context.

1.4.C.1	Frequency and relative frequency tables, bar charts, and pie charts can be used to compare two or more data sets in terms of the same categorical variable.
1.5.A.1	Histograms, stem-and-leaf plots, and dotplots provide a visual representation of the distribution of the values of a quantitative variable. These graphs show the frequency or relative frequency of the quantitative variable values or intervals of values and maintain the natural ordering, smallest to largest, of the quantitative variable.
1.5.A.2	A histogram places the observed values of the quantitative variable into ordered intervals, or bins, along the horizontal axis. Each bar represents an interval or bin, and the height of each bar shows the frequency or relative frequency of the observations within that interval. Altering the interval widths, or bin widths, can change the appearance of the histogram. Alternatively, a histogram can be constructed with bins on the vertical axis with bars appearing horizontally.
1.5.A.3	A stem-and-leaf plot splits each value of the quantitative variable into two parts: a “stem” (the first digit or digits) and a “leaf” (usually the single digit after the stem digit or digits). Both stems and leaves are ordered from smallest to largest.
1.5.A.4	A dotplot represents each value of the quantitative variable by a dot. Each dot is placed above the horizontal or beside the vertical axis corresponding to the value of that observation, with nearly identical values stacked on top of each other.
1.6.A.1	Descriptions of the distribution of one quantitative variable include shape, center, and variability (spread) as well as any unusual features such as outliers, gaps, or clusters in context.
1.6.A.2	The shape of the distribution of one quantitative variable is skewed to the right (positively skewed) if the right tail (toward larger values) is longer than the left. The shape of the distribution is skewed to the left (negatively skewed) if the left tail (toward smaller values) is longer than the right. The shape of the distribution is approximately symmetric if the left half is approximately the mirror image of the right half.
1.6.A.3	Distributions of one quantitative variable with one main peak are called unimodal. Distributions with two prominent peaks are called bimodal. A distribution in which each frequency or each relative frequency is approximately the same with no prominent peaks is approximately uniform.
1.6.A.4	Outliers for one quantitative variable are data points that are unusually small or large relative to the rest of the data.
1.6.A.5	A gap is a region in a distribution between two values in which there are no observed data.
1.6.A.6	Clusters are concentrations of values usually separated by gaps.
1.6.B.1	Graphical representations of a quantitative variable may reveal information that can be used to justify claims about the variable in context.
1.7.A.1	Two commonly used measures of center in the distribution of a quantitative variable are the mean and median.
1.7.A.2	The mean is the sum of all the values divided by the number of values and can be found with and without using technology. For a sample, the mean is denoted by x-bar: $\bar{x} = 1/n \Sigma [i = 1 \text{ to } n] x_i$, where x_i represents the ith data point in the sample and n represents the number of data values in the sample.
1.7.A.3	The median is the middle value when the data set is ordered from smallest to largest and can be found with and without using technology. One common method for determining the median of a data set with an even number of values is to use the mean of the two middle values. A common method for determining the median of a data set with an odd number of values is to use the value in the middle of all the values.
1.7.A.4	In an ordered data set, the smallest value is the minimum value, and the largest value is the maximum value.

1.7.A.5	The first quartile, denoted by Q1, is the median value of the lower half of the ordered data set from the minimum value to the position of the median. Approximately 25% of the values in the data set are less than or equal to Q1. The third quartile, denoted by Q3, is the median value of the upper half of the ordered data set from the position of the median to the maximum value. Approximately 75% of the values in the data set are less than or equal to Q3. The second quartile, Q2, is also the median of the data set. Q1 and Q3 form the boundaries for the middle 50% of values in an ordered data set.
1.7.A.6	The □th percentile is the value that has □% of the data less than or equal to it when the data set is ordered from smallest to largest. The first and third quartiles are the 25th and 75th percentiles, respectively.
1.7.B.1	Three commonly used measures of variability (or spread) in the distribution of a quantitative variable are the range, interquartile range, and standard deviation.
1.7.B.2	The range is the difference between the maximum data value and the minimum data value.
1.7.B.3	The interquartile range (IQR) is the difference between the third and first quartiles: Q3 −Q1.
1.7.B.4	The standard deviation is a typical deviation of the data values from their mean and can be found with and without using technology. The sample standard deviation is denoted by □ and calculated by □ = √[1/(□ − 1) Σ(□$_i$ − $\bar{□}$)²], where □ is the data value, $\bar{□}$ is the mean, and □ is the number of data values in the sample. The square of the sample standard deviation, □² is called the sample variance.
1.7.C.1	Changing units of measurement affects the values of the calculated statistics.
1.7.D.1.i	An outlier is a value located more than 1.5 I × QR above the third quartile or more than 1.5 I × QR below the first quartile.
1.7.D.1.ii	An outlier is a value located more than 2 standard deviations above, or below, the mean.
1.7.E.1	Summary statistics can be used to compare features of two or more independent samples, including center, variability, shape, and outliers.
1.7.F.1	The median and IQR are considered a resistant (or robust) measure of center and measure of variability, respectively, because outliers do not greatly (if at all) affect their values. Because outliers can affect their values greatly, the mean is considered a nonresistant (or non-robust) measure of center, and the range and standard deviation are considered nonresistant (or non-robust) measures of variability.
1.7.F.2	Summary statistics of a quantitative variable may reveal information that can be used to justify claims about the variable in context.
1.8.A.1	A five-number summary is made up of the minimum data value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum data value.
1.8.A.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”) that represent 25% of the data extend from the first quartile to the minimum and from the third quartile to the maximum. If there are outliers in the data, the whiskers extend to the most extreme data values that are not outliers, and outliers are usually denoted with an asterisk or other symbol.
1.8.B.1	If a distribution is relatively symmetric, then the values of the mean and median are relatively close to each other. If a distribution is skewed right, then the value of the mean is usually larger than the median. If the distribution is skewed left, then the value of the mean is usually smaller than the median.
1.9.A.1	Graphical representations of a quantitative variable can be used to compare important features between two or more distributions of the same quantitative variable. Histograms, back-to-back stem-and-leaf plots, and dotplots may be used to compare center, variability, shape, outliers, clusters, or gaps in two or more distributions. Boxplots may be used to

	compare center, variability, outliers, and skewness (or symmetry).
1.9.B.1	A comparison of graphical representations for two or more distributions can include any of the numerical summaries (e.g., mean, standard deviation, etc.).
1.9.C.1	Multiple quantitative one-variable graphical representations may reveal information that can be used to justify claims about the variable in context.
1.9.D.1	A standardized score measures the number of standard deviations a data value falls above or below the mean.
1.9.D.2	A z-score is calculated as $(x_i - \mu)/\sigma$, where x_i is the data value, μ is the population mean, and σ is the population standard deviation. A z-score measures how many standard deviations a data value is above (positive z-score) or below (negative z-score) the mean. When the population mean and standard deviation are unknown, the sample mean and standard deviation may be used to determine a z-score.
1.9.E.1	z-scores may be used to compare relative positions of individual values within a distribution or between distributions.
1.10.A.1	The first component of an investigative question should guide the data collection process and should be phrased in terms of the variable(s) of interest in the study.
1.10.A.2	The second component of an investigative question should guide the data analysis choice.
1.10.A.2.i	In the case of a hypothesis test, the investigative question should make clear the parameter and the direction of the alternative hypothesis (i.e., not equal, greater than, less than, association, not independent).
1.10.A.2.ii	In the case of a confidence interval, the investigative question should make clear the parameter and the goal of estimation of that parameter within a range of potential values.
1.10.A.3	The third component of an investigative question should indicate the type(s) of conclusion(s) applicable from the study. The investigative question should provide the population to which the conclusions will be applicable and, in the case of an experiment that uses random assignment, a cause-and-effect conclusion.
1.10.B.1	A census consists of recording information from all items or individuals in a population.
1.10.C.1	An experiment is a study in which a researcher assigns conditions, or treatments, to experimental units to explore an investigative question of interest about the population.
1.10.C.2	The experimental unit is the observational unit to which the treatment is assigned. When experimental units consist of people, they are sometimes referred to as subjects or participants.
1.10.C.3	An explanatory variable, or factor, is a variable whose different categories, or levels, are imposed on the experimental units. The different categories, or levels, of the explanatory variable are called treatments. When there is more than one explanatory variable, the combinations of the categories, or levels, of the explanatory variables are called treatments.
1.10.C.4	A response variable is an outcome measured on each experimental unit after the treatment has been administered.
1.10.D.1	An observational study is a study where treatments are not imposed. The researcher records the values of the variables of interest in order to explore an investigative question of interest.
1.10.D.2	A prospective study is one in which the observational units of study are selected at a point in time, and data are gathered both at that time and into the future.
1.10.D.3	A retrospective study is one in which the observational units of study are selected at a point in time and data from the past are gathered.
1.10.D.4	A survey is an observational study in which the data are collected from humans using a standard set of questions.
1.10.D.5	A confounding variable in an observational study provides an alternative explanation for

	the observed relationship between the explanatory and response variables determined in the study, thereby reducing the possibility of concluding a causal relationship between the explanatory and response variables of interest. To be a confounding variable, a variable must be associated with both the explanatory variable and the response variable.
1.10.E.1	A sample is considered random when all observational units in the sample are selected from the population using some type of random mechanism, such as a random number generator.
1.10.E.2	When observational units, or experimental units, in a sample are randomly selected from a population, it is appropriate to make generalizations about the entire population of individuals from which the sample was selected.
1.10.E.3	A sample is not randomly selected when observational units are deliberately chosen or volunteer themselves to be in the sample.
1.10.E.4	When observational units, or experimental units, in a sample are not randomly selected from a population, it is appropriate to make generalizations only about a population of individuals that are similar to those used in the study.
1.11.A.1	Sampling without replacement is a sampling strategy in which an observational unit from a population can be selected only once. The observational unit is not returned to the population before subsequent selections of observational units are made, so there is no chance that the observational unit can be selected again.
1.11.A.2	Sampling with replacement is a sampling strategy in which an observational unit from the population can be selected more than once. The observational unit is returned to the population before subsequent selections of observational units are made, so it is possible that the observational unit could be selected again.
1.11.A.3	In a simple random sample (SRS) of size n, every sample of the size n has the same chance of being selected. This method is the basis for many types of sampling mechanisms. There are several procedures to obtain a simple random sample; for example, using a random number generator or randomly selecting numbered slips of paper.
1.11.A.4	A stratified random sample involves the division of all individuals in a population into non-overlapping groups, called strata, based on one or more shared attributes or characteristics (homogeneous grouping). Within each stratum a simple random sample is selected, and the selected individuals are combined to form one sample.
1.11.A.5	A cluster random sample involves the division of a population into smaller groups, called clusters. Ideally, each cluster mirrors the heterogeneity of the population, with clusters similar to one another. A simple random sample of clusters is selected from the population to form the sample of clusters. Data are collected from all observational units in each of the selected clusters.
1.11.A.6	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 between successive sampling units.
1.11.B.1	Each random sampling method has different characteristics that make it more appropriate for sampling populations depending on the question being investigated.
1.12.A.1	Bias in a sampling method is a systematic error in the sampling procedure that results in a statistic being consistently larger or consistently smaller than the parameter the statistic is used to estimate.
1.12.A.2	Voluntary response bias is a bias that may occur when a sample consists entirely of volunteers.
1.12.A.3	Undercoverage bias may occur when the sampling method fails to include part of the population or a part of the population is less likely to be selected based on the sampling method.
1.12.A.4	Nonresponse bias may occur because of a failure to obtain responses from some individuals chosen to be sampled. The respondents and nonrespondents could differ

	significantly in ways that are important for the study.
1.12.A.5	Response bias may occur when responses to a survey or measurements of observational units tend to differ from the “true” value in one direction. Examples include questions that are confusing or leading (question wording bias) or self-reported responses.
1.12.A.6	Nonrandom sampling methods (e.g., samples chosen by convenience or voluntary response) introduce potential bias because they do not use random chance to select the individuals.
1.13.A.1.i	Comparisons of at least two treatment groups, one of which could be a control group
1.13.A.1.ii	Random assignment of treatments to experimental units
1.13.A.1.iii	Replication
1.13.A.1.iv	Direct control of potential extraneous sources of variation in the response
1.13.A.2	A control group is a collection of experimental units that are created for comparison. A control group may be given a treatment different from the treatment of interest to determine if the treatment of interest has an effect (e.g., a treatment with an inactive substance, a placebo, may be given).
1.13.A.3	The placebo effect is the difference between the average response to a placebo and the average response to no treatment.
1.13.A.4	In a single-blind, also called single-masked, experiment, participants do not know which treatment they are receiving, but members of the research team who interact with them know which treatment each participant is receiving, or vice versa.
1.13.A.5	In a double-blind, also called double-masked, experiment, neither the participants nor the members of the research team who interact with them know which treatment each participant is receiving.
1.13.A.6	An extraneous source of variation, also referred to as an extraneous variable, is a variable that is known (or believed) to affect the response but is not an explanatory variable being studied.
1.13.A.7	The purpose of random assignment is to create treatment groups that are as similar as possible with respect to extraneous sources of variation. If random assignment is successful, the respective distributions of each extraneous variable will be approximately the same for all the treatment groups.
1.13.A.8	A confounding variable in an experiment is a variable that is related to the explanatory variable in such a way that it is difficult to determine which variable, explanatory or confounding, is influencing the change in the response variable. However, in a well-designed experiment, the potential for confounding variables is reduced.
1.13.A.9	Replication within an experiment means more than one experimental unit is assigned to each treatment.
1.13.A.10	Direct control in an experiment means keeping the settings of certain potential extraneous sources of variation in the response variable the same from experimental unit to experimental unit.
1.13.B.1	In a completely randomized design, treatments are assigned to experimental units completely at random. Often the number of experimental units assigned to each treatment will be the same, but the sample sizes in each treatment do not have to be the same.
1.13.B.2	A blocking variable is a source of extraneous variation in the response variable. In a randomized block design, the experimental units are first grouped according to similar values of a blocking variable. These groups are called blocks. Units within the same block are homogeneous with respect to the blocking variable. After the blocks are formed, the treatments are randomly assigned to experimental units within each block so that all treatments occur within every block.

1.13.B.3	The purpose of blocking is to separate the variation in the response caused by the blocking variable from the rest of the extraneous variation in the response. Blocking allows for more precise comparisons of the response across the treatments. Within a block, the treatments can be compared without having to worry about variation in the response caused by changes in the blocking variable.
1.13.B.4	A matched pairs design is a randomized block design with only two treatments. Experimental units are arranged in pairs by matching on one or more extraneous sources of variation in the response variable. Each pair receives both treatments by randomly assigning one treatment to one member of the pair and the other treatment to the second member of the pair. Alternatively, each experimental unit may get both treatments while the order of the treatments is randomized.
1.13.C.1	One experimental design may be more appropriate than another experimental design based on the goals of the investigative study, the characteristics of the population, and the sample and variables involved.
1.13.D.1	Using random assignment of treatments to experimental units allows for cause-and-effect conclusions between the explanatory and the response variables because the potential for confounding variables is reduced.
1.13.D.2	Depending on the experimental unit, it may be unethical or difficult to randomly select experimental units to participate in an experiment. In that case, the study's experimental units are obtained from volunteers and will represent the population of experimental units similar to those who participated in the study.
5.1.A	Construct scatterplots depicting the relationship between two quantitative variables.
5.1.A.1	A bivariate quantitative data set consists of observations of ordered pairs from two quantitative variables, collected from the same individuals in a sample or population, and can be used to construct a scatterplot.
5.1.A.2	A scatterplot shows the relationship between two quantitative variables for each observation, one corresponding to the value on the □-axis and one corresponding to the value on the □-axis. The explanatory variable is placed on the □-axis and is the variable whose values are used to explain or predict the corresponding values for the response variable, which is placed on the □-axis.
5.1.B.1	A description of the association shown in a scatterplot includes form, direction, strength, and unusual features.
5.1.B.2	The form of the association shown in a scatterplot, if any, can be described as linear or non-linear.
5.1.B.3	The direction of the association shown in a scatterplot, if any, can be described as positive or negative. A positive association means that as values of the explanatory variable increase, the values of the response variable tend to increase. A negative association means that as values of the explanatory variable increase, the values of the response variable tend to decrease.
5.1.B.4	The strength of the association shown in a scatterplot is how closely the points follow the general pattern. Strength can be described as strong, moderate, or weak.
5.1.B.5	Unusual features of a scatterplot include clusters of individual points or points that don't fit in the general pattern of association between the two variables.
5.1.C.1	Scatterplots depicting the relationship between two numeric variables may reveal information that can be used to justify claims about the variable in context.
5.2.A.1	The correlation coefficient, □, summarizes the strength and direction of the linear association between two quantitative variables. The correlation coefficient □ is unit-free and always between −1 and 1, inclusive. A negative correlation coefficient value indicates a negative association, and a positive correlation coefficient value indicates a positive association.
5.2.A.2	The strength of the linear association is determined by how close the correlation

	coefficient is to −1 or 1. 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.
5.2.A.3	A correlation coefficient close to −1 or 1 does not necessarily mean that a linear model is appropriate.
5.2.A.4	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.
5.3.A.1	If the form of the relationship between x and y appears linear, we can approximate the relationship between x and y using a linear regression model, which is a linear equation that uses an explanatory variable, x, to predict the response variable, y.
5.3.A.2	In a linear regression model, the predicted response value, denoted by $\hat{y}$, is calculated as $\hat{y} = a + bx$, where a is the y-intercept, b is the slope of the regression line, and x is the explanatory variable.
5.3.A.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 the further the estimate is extrapolated.
5.3.A.4	Interpolation is predicting a response value using a value for the explanatory variable that is within the interval of x-values used to determine the regression line.
5.4.A.1	A residual is the difference between the observed response value and the predicted response value for the given value of the explanatory variable: residual $y - \hat{y}$ or (residual = observed y − predicted y).
5.4.B.1	If the residual is positive, the model underpredicts (underestimates) the value of the response variable. If the residual is negative, the model overpredicts (overestimates) the value of the response variable.
5.4.C.1	A residual plot is a scatterplot of the residuals versus the predicted response values (or the explanatory variable values).
5.4.C.2	Residual plots can be used to investigate the appropriateness of the linear regression model for the observed data.
5.4.C.3	The linear regression model should only be fit to the data if the data exhibit a linear trend. Apparent randomness in a residual plot for a linear regression model is confirmation of a linear form in the association between the two variables and indicates that the simple linear regression model is an appropriate model for the data.
5.4.C.4	Curvature in the residual plot for a linear regression model suggests that the linear model is not the most appropriate model for the data.
5.5.A.1	The simple linear regression model is fit to the data by minimizing the sum of the squares of the residuals. Because of this, the resulting equation is often called the least-squares regression line (□□□□) and is calculated using technology. This regression line will pass through the point $(\bar{x}, \bar{y})$.
5.5.A.2	The slope of the regression line, b, is calculated using technology.
5.5.A.3	The y-intercept of the regression line, a, is calculated using technology.
5.5.A.4	In simple linear regression, the correlation coefficient, r, is calculated using technology.
5.5.A.5	In simple linear regression, the square of the correlation coefficient, r^2, is called the coefficient of determination. The value of r^2 is the proportion of variation in the response variable that is explained by the linear relationship with the explanatory variable.
5.5.B.1	The coefficients of the least-squares regression line model (line of best fit) are the slope, b, and the y-intercept, a, because they are based on a sample of values.
5.5.B.2	The slope of the least-squares regression line can be interpreted as the predicted increase or decrease in the response variable for a one-unit increase in the explanatory variable, and it should be interpreted in context.

5.5.B.3	The y-intercept in the least-squares regression line is the predicted value of the response variable when the explanatory variable is equal to 0, and it should be interpreted in context. Sometimes, the y-intercept of the line does not have a reasonable interpretation in context because $x = 0$ might be beyond the interval of x-values used to determine the regression line (extrapolation). At other times, the y-intercept of the line does not have a logical interpretation in context because it might be a negative value for a response variable that has no negative values, such as height.

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
1. Well-designed data collection processes (surveys and experiments) are the foundation of any valid statistical claim; the methodology used to produce data dictates whether conclusions can be generalized to a population or can establish a causal relationship. 2. Graphical displays and numerical summaries reveal underlying patterns, shapes, central trends, and departures from patterns in a dataset that the naked eye or a raw list of numbers cannot capture. 3. Mathematical structures, such as the Normal distribution curve or the Least-Squares Regression Line, serve as probabilistic models to describe reality, simplify complex distributions, and make calculated predictions under uncertainty.	1. How do we design studies and gather data in a way that allows us to confidently make claims about a broader population or prove cause-and-effect? 2. What story does this data display tell us, and how do measures of center and variability help us precisely describe a distribution's behavior? 3. How can we use mathematical models to predict future outcomes or evaluate an individual's relative standing within a dataset, and what are the limitations of those models?

Instructional Focus

Learning Targets

Part 1: Exploring One-Variable Data:

- I can identify the population, sample, individual cases, and the exact investigative question being asked in a statistical study.
- I can classify a variable as either categorical or quantitative (discrete vs. continuous).
- I can represent categorical data using frequency tables, bar charts, and mosaic plots, and use them to calculate marginal and conditional relative frequencies.
- I can construct and interpret quantitative data displays, including stem-and-leaf plots, histograms, and box plots.
- I can write a cohesive, contextual description of a quantitative distribution by addressing its Shape, Outliers, Center, and Variability (CUSS & BS) using comparative language when necessary.
- I can identify outliers mathematically using the 1.5×IQR rule and the mean-plus-standard-deviation boundary.
- I can calculate and interpret a data point's percentile and z-score in context, and analyze the effects of linear transformations (addition, subtraction, multiplication, division) on summary statistics.
- I can model a data distribution using density curves and the Empirical Rule (68-85-99.7), and calculate areas/proportions and boundaries using a graphing calculator (normalcdf and invNorm).

Part 2: Collecting Data

- I can distinguish between an observational study and a controlled experiment, and identify why only an experiment can establish a cause-and-effect relationship.
- I can identify and execute various sampling methodologies, including Simple Random Sampling (SRS), Stratified Sampling, Cluster Sampling, and Systematic Sampling, and describe the operational advantages of each.
- I can identify potential sources of sampling bias (undercoverage, nonresponse, response bias, and convenience sampling) and explain how each bias can alter or systematically misrepresent the true population parameter.
- I can identify the core components of an experimental design: experimental units, explanatory variables (factors), treatments, and response variables.
- I can apply the four principles of experimental design: Comparison, Randomization, Replication, and Control.
- I can design and explain randomized experimental structures, including Completely Randomized Designs, Randomized Block Designs, and Matched Pairs Designs.
- I can evaluate a study's design to state its proper Scope of Inference (determining whether results can be generalized to a broader population, and whether a causal relationship can be claimed).

Part 3: Exploring Two-Variable Data

- I can construct a scatterplot to analyze the relationship between two quantitative variables and describe its Direction, Form, Strength, and Outliers (DUFS) in context.
- I can calculate and interpret the correlation coefficient (r) and explain its mathematical properties (e.g., its vulnerability to outliers and its unitless nature).
- I can construct a Least-Squares Regression Line (LSRL) model, use it to make predictions, and explain the dangers of extrapolation.
- I can calculate and interpret the slope and y-intercept of an LSRL using strict, contextual AP

Statistics template phrasing.

- I can calculate and interpret a residual, construct a residual plot, and evaluate whether a linear model is appropriate based on the plot's randomness.
- I can interpret standard computer output for linear regression to find the regression equation, the standard deviation of the residuals (s), and the coefficient of determination (r^2).
- I can explain what s and r^2 mean in the context of the model's predictive accuracy.
- I can identify influential points and high-leverage outliers on a scatterplot and analyze their specific effects on the slope, y-intercept, correlation coefficient r, and standard deviation of the residuals s.

Prerequisite Skills

Part 1: Exploring One-Variable Data

- Students must seamlessly convert between fractions, decimals, and percentages.
- Students need to confidently substitute values into multi-step formulas.
- Familiarity with reading basic number lines and early-grade displays like standard bar charts, dot plots, and line graphs.
- Prior knowledge of what a "mean" and "median" represent conceptually, even if they haven't yet studied how outliers affect them.
- Understanding intervals on a number line is critical to understanding standard deviation and the boundaries of the normal curve.

Part 2: Collecting Data

- Students must be able to read a dense 5-sentence paragraph and extract the operational core: Who is being studied? What is being measured? What was the goal?
- A baseline intuitive understanding that just because two things happen together does not mean one caused the other.
- An intuitive grasp of chance and fair selection.
- The ability to follow "if-then" scenarios to identify flaws in a process.

Part 3: Exploring Two-Variable Data

- Unhesitating ability to plot (x, y) ordered pairs and read coordinates from a standard four-quadrant scatterplot.
- Students must know what a line is, how to identify the slope (m) and the y-intercept (b), and how to use an x-value to calculate a predicted y-value.
- Understanding that the independent variable (x) is the input/predictor and the dependent variable (y) is the output/outcome.
- Remembering that slope represents a rate.

Common Misconceptions

Part 1: Exploring One-Variable Data

- When looking at a histogram or bar chart, students often confuse the value on the x-axis (the variable itself) with the value on the y-axis (the count/frequency).
    - If a histogram shows a high bar at a score of 10, students might say "the data is highly variable" because the bar is tall, rather than looking at how spread out the values are along the x-axis.

- Students intuitively believe that a wider section of a boxplot contains more data points.
    - Every single section of a boxplot (each whisker and each half of the box) contains exactly 25% of the data. A wider section simply means that 25% of the data is more spread out (variable), not that there are more individuals in it.
- Students use the words "Symmetric" and "Normal" interchangeably.
    - All Normal distributions are symmetric, but not all symmetric distributions are Normal. A uniform distribution (flat line) or a bimodal distribution can be perfectly symmetric but looks nothing like a bell curve. College Board graders will penalize students who call a distribution "Normal" without verifying the bell shape or empirical rule.
- Students write isolated, abstract descriptions like, "The shape is skewed right, there is an outlier at 50, the median is 12, and the IQR is 4."
    - This will score a "Partially Correct" on the AP Exam. They must weave the problem's context into the sentence (e.g., "The distribution of high school student shoe sizes is skewed right...").

Part 2: Collecting Data

- Students mix up these two methods because both involve splitting a population into groups.
    - In Stratified, you sample some individuals from all groups (Homogeneous groups: "Some from all").

      In Cluster, you sample all individuals from some groups (Heterogeneous groups: "All from some").
- Students think bias means the researcher had bad intentions or made a sloppy mistake. Alternatively, they define bias as just "bad sampling."
    - On the AP Exam, defining bias requires explaining the direction of the error. Students must state whether the sampling method will systematically overestimate or underestimate the true population parameter. Simply saying "it's biased because it's a convenience sample" earns no credit.
- Students think any use of randomness accomplishes both goals.
    - Random Selection (surveys) allows you to generalize your findings to the population.

      Random Assignment (experiments) allows you to establish causation between treatments.
- When asked to describe a random sampling process, students will write, "I will pick 20 students randomly from the cafeteria."
    - Human eyes naturally drift toward certain people, which introduces bias. Students must explicitly state a mechanical randomization technique (e.g., "Assign every student a number from 1 to N, use a random number generator to pick 20 unique numbers, and sample those students").

Part 3: Exploring Two-Variable Data

- When interpreting a slope or making a prediction, students write absolute statements like, "For every 1-inch increase in height, a person's weight will increase by 4 pounds."
    - Regression lines are probabilistic models, not algebraic lines. Students must use "wiggle" words: "On average," "the predicted weight increases by," or "we expect." Omitting these words is an automatic point deduction.
- Students assume that if the correlation coefficient $r = 0.95$, the relationship is guaranteed to be linear.
    - A strongly curved relationship (such as a distinct quadratic curve) can still yield a very high

linear correlation. The only way to check if a linear model is appropriate is to look at the residual plot and ensure there is no leftover curved pattern.
- Students frequently mix up the interpretations of r (the strength and direction of a linear relationship) and r^2 (the percentage of variation in y explained by the linear model with x).
    - Force them to stick to Math Medic's rigid script templates during Phase 3 so they don't accidentally blur these two concepts together.

Spiraling For Mastery

Current Unit Content/Skills	Spiral Focus	Activity

Part 1: Exploring One-Variable Statistics
 	
• Constructing/interpreting graphs
  ○ Histograms, Boxplots, Mosaic plots.
• Describing Quantitative Distributions (CUSS & BS)
• Calculating percentiles, z-scores, and Normal curve proportions.
	Part 1: Exploring One-Variable Statistics
	
	• Distinguishing between population vs. sample
	• Properly identifying variable types within context.
	  ○ categorical vs. quantitative
		Part 1: Exploring One-Variable Statistics
		
		• **"What's the Variable?" Snapshot**
		  ○ Provide a brief 3-sentence summary of a study at the start of class. Before analyzing the graphs, give students 2 minutes to identify the population and sample and classify every variable involved. This prevents them from confusing categorical counts with quantitative values during analysis.

Part 2: Collecting Data
 	
• Identifying sampling methods
  ○ SRS, Stratified Cluster, Systematic
• Explaining sources of bias
  ○ Undercoverage, nonresponse, response
	Part 2: Collecting Data
	
	• Using CUSS & BS to justify study design.
	• Reading existing data distributions to identify confounding variables.
		Part 2: Collecting Data
		
		• **"The Treatment Comparison" Drill**
		  ○ When teaching experimental

bias • Designing randomized experiments vs. observational studies.		design, give students summary data (like side-by-side boxplots) from a flawed observational study. Have them write a quick 3-minute paragraph comparing the distributions using CUSS & BS, and use that descriptive analysis to explain why a randomized experiment is necessary to prove causation.

Part 3: Exploring Two-Variable Data
• Interpreting scatterplots
  ○ DUFS & BS
• Calculating and Interpreting LSRL slope, y-intercept, r, r^2, and s.
• Analyzing residuals and computer output.
	Part 3: Exploring Two-Variable Data
	• Assessing how bad sampling impacts a regression model.
	• Analyzing the distribution of residuals using 1-variable tools.
		Part 3: Exploring Two-Variable Data
		• **Residual DUFS & BS and Biased Slopes**
		  ○ Give students a histogram or box plot of the *residuals* from a linear model. Have them describe its shape and check for outliers using the 1.5×IQR rule.
		  ○ Present a scenario where a convenience sample was used to gather two-variable data. Ask students to predict whether the resulting slope will over- or underestimate the true population relationship.

Assessment

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

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

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.