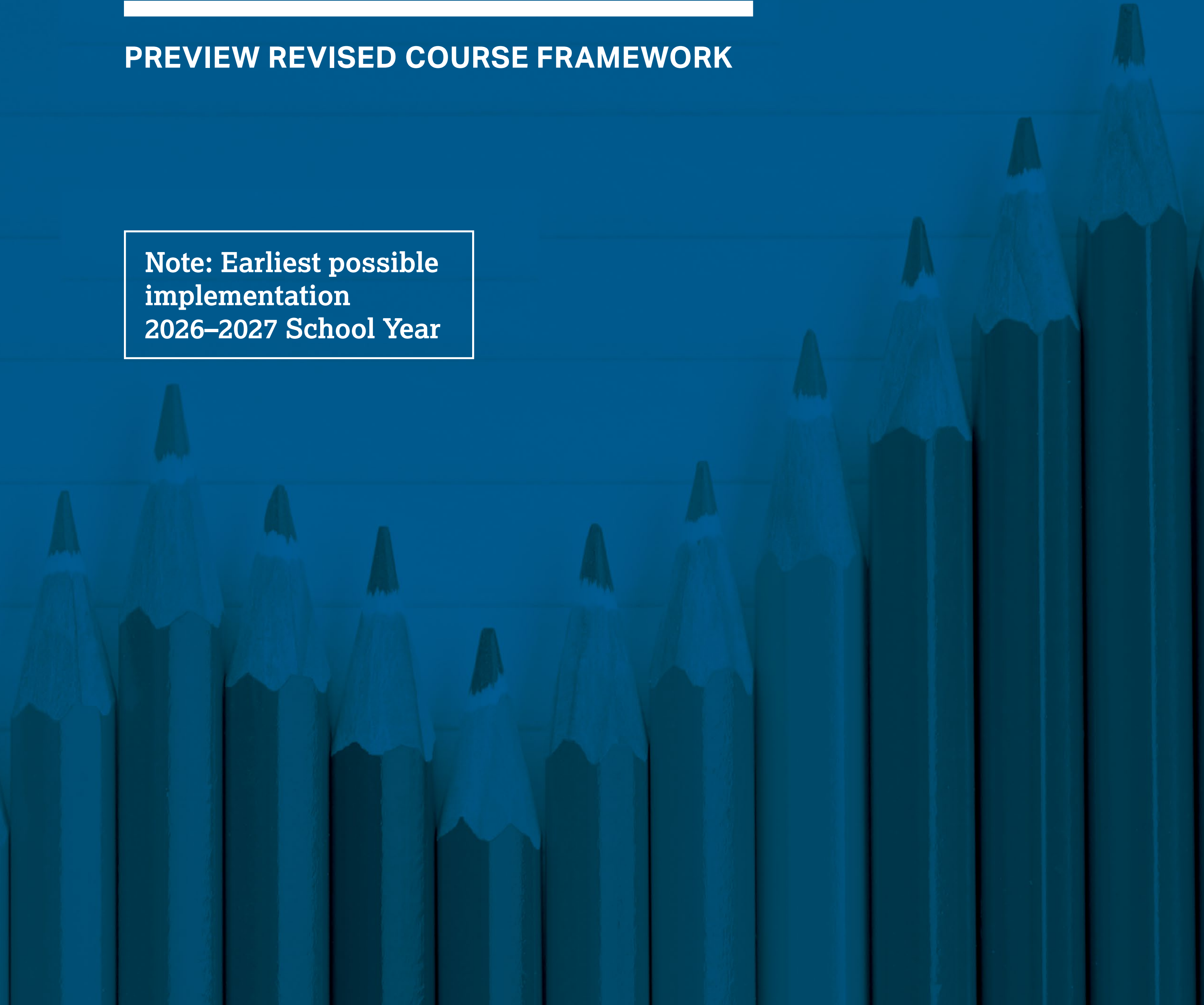
AP Statistics

PREVIEW REVISED COURSE FRAMEWORK

Note: Earliest possible implementation 2026–2027 School Year



What AP® Stands For

Thousands of Advanced Placement teachers have contributed to the principles articulated here. These principles are not new; they are, rather, a reminder of how AP already works in classrooms nationwide. The following principles are designed to ensure that teachers' expertise is respected, required course content is understood, and that students are academically challenged and free to make up their own minds.

1. AP stands for clarity and transparency. Teachers and students deserve clear expectations. The Advanced Placement Program makes public its course frameworks and sample assessments. Confusion about what is permitted in the classroom disrupts teachers and students as they navigate demanding work.
2. AP is an unflinching encounter with evidence. AP courses enable students to develop as independent thinkers and to draw their own conclusions. Evidence and the scientific method are the starting place for conversations in AP courses.
3. AP opposes censorship. AP is animated by a deep respect for the intellectual freedom of teachers and students alike. If a school bans required topics from their AP courses, the AP Program removes the AP designation from that course and its inclusion in the AP Course Ledger provided to colleges and universities. For example, the concepts of evolution are at the heart of college biology, and a course that neglects such concepts does not pass muster as AP Biology.
4. AP opposes indoctrination. AP students are expected to analyze different perspectives from their own, and no points on an AP Exam are awarded for agreement with any specific viewpoint. AP students are not required to feel certain ways about themselves or the course content. AP courses instead develop students' abilities to assess the credibility of sources, draw conclusions, and make up their own minds.

   As the AP English Literature course description states: "AP students are not expected or asked to subscribe to any one specific set of cultural or political values, but are expected to have the maturity to analyze perspectives different from their own and to question the meaning, purpose, or effect of such content within the literary work as a whole."
5. AP courses foster an open-minded approach to the histories and cultures of different peoples. The study of different nationalities, cultures, religions, races, and ethnicities is essential within a variety of academic disciplines. AP courses ground such studies in primary sources so that students can evaluate experiences and evidence for themselves.
6. Every AP student who engages with evidence is listened to and respected. Students are encouraged to evaluate arguments but not one another. AP classrooms respect diversity in backgrounds, experiences, and viewpoints. The perspectives and contributions of the full range of AP students are sought and considered. Respectful debate of ideas is cultivated and protected; personal attacks have no place in AP.
7. AP is a choice for parents and students. Parents and students freely choose to enroll in AP courses. Course descriptions are available online for parents and students to inform their choice. Parents do not define which college-level topics are suitable within AP courses; AP course and exam materials are crafted by committees of professors and other expert educators in each field. AP courses and exams are then further validated by the American Council on Education and studies that confirm the use of AP scores for college credits by thousands of colleges and universities nationwide.

The AP Program encourages educators to review these principles with parents and students so they know what to expect in an AP course. Advanced Placement is always a choice, and it should be an informed one. AP teachers should be given the confidence and clarity that once parents have enrolled their child in an AP course, they have agreed to a classroom experience that embodies these principles.

Contents

COURSE FRAMEWORK

Acknowledgements

College Board would like to acknowledge the following committee members, consultants, and reviewers for their assistance with and commitment to the development of this course. All individuals and their affiliations were current at the time of contribution.

Prince Afriyie, *University of Virginia, Charlottesville, VA*

Prince Allotey, *University of South Carolina, Columbia, SC*

Javier Alvarez, *Phillips Exeter Academy, Exeter, NH*

Oscar Aquirre, *Coral Reef High School, Miami, FL*

Barb Barnet, *University of Wisconsin–Platteville, Platteville, WI*

Tim Bator, *Sayre School, Lexington, KY*

Ellen Breazel, *Clemson University, Clemson, SC*

Carol Chamberlain Hebert, *Barnstable High School, Barnstable, MA*

Holly Deal, *Kennesaw State University, Kennesaw, GA*

Christine Franklin, *University of Georgia, Athens, GA*

Donna LaLonde, *American Statistical Association, Alexandria, VA*

Adam Loy, *Carleton College, Northfield, MN*

Laura Marshall, *Phillips Exeter Academy, Exeter, NH*

Bridget Matamoros-Mota, *John H. Guyer High School, Denton, TX*

Karen McGaughey, *California Polytechnic State University, San Luis Obispo, CA*

Chris Olsen, *Grinnell College, Grinnell, IA*

Michael Posner, *Villanova University, Villanova, PA*

Al Reiff, *The Taft School, Watertown, CT*

Paul Roback, *St. Olaf College, Northfield, MN*

Tom Rohnkohl, *Silverado High School, Las Vegas, NV*

Allan Rossman, *California Polytechnic State University, San Luis Obispo, CA*

Joshua Sawyer, *Elizabeth City–Pasquotank Public Schools, Elizabeth City, NC*

Ashley Schreckengost, *Wheaton High School, Silver Spring, MD*

College Board Staff

Camille Pace, *Director I, AP Statistics Curriculum and Assessment*

Darla Emerson, *Director I, AP Statistics Curriculum and Assessment*

Beth Brown, *Director, Product Development and Editorial*

Chad Hoge, *Senior Director, Course Product Manager*

Dana Kopelman, *Executive Director, AP Content Production and Product Management*

Daniel McDonough, *Senior Director, AP Content and Assessment Publications*

Andrea Pellerito, *Senior Director, Course Product Manager*

Allison Thurber, *Vice President, AP Curriculum and Assessment*

Jason VanBilliard, *Executive Director, AP Math, Computer Science, and Cyber Department Head*

Craig Wright, *Senior Director, Manager, AP Math Curriculum and Assessment*

Return to Table of Contents

About AP

The Advanced Placement® Program (AP®) enables willing and academically prepared students to pursue college-level studies—with the opportunity to earn college credit, advanced placement, or both—while still in high school. Through AP courses in 40 subjects, each culminating in a challenging exam, students learn to think critically, construct solid arguments, and see many sides of an issue—skills that prepare them for college and beyond. Taking AP courses demonstrates to college admission officers that students have sought the most challenging curriculum available to them, and research indicates that students who score a 3 or higher on an AP Exam typically experience greater academic success in college and are more likely to earn a college degree than non-AP students. Each AP teacher's syllabus is evaluated and approved by faculty from some of the nation's leading colleges and universities, and AP Exams are developed and scored by college faculty and experienced AP teachers. Most four-year colleges and universities in the United States grant credit, advanced placement, or both on the basis of successful AP Exam scores—more than 3,300 institutions worldwide annually receive AP scores.

AP Course Development

In an ongoing effort to maintain alignment with best practices in college-level learning, AP courses and exams emphasize challenging, research-based curricula aligned with higher education expectations.

Individual teachers are responsible for designing their own curriculum for AP courses, selecting appropriate college-level readings, assignments, and resources. This course and exam description presents the content and skills that are the focus of the corresponding college course and that appear on the AP Exam. It also organizes the content and skills into a series of units that represent a sequence found in widely adopted college textbooks and that many AP teachers have told us they follow in order to focus their instruction. The intention of this publication is to respect teachers' time and expertise by providing a roadmap that they can modify and adapt to their local priorities and preferences. Moreover, by organizing the AP course content and skills into units, the AP Program is able to provide teachers and students with free formative assessments—Progress Checks—that teachers can assign throughout the year to measure student progress as they acquire content knowledge and develop skills.

Enrolling Students: Access, Opportunity, and Readiness

The AP Program welcomes all students willing to challenge themselves with college-level coursework and career preparation. We strongly encourage educators to invite students into AP classes, including students from ethnic, racial, socioeconomic, geographic, or other groups not broadly participating in a school's AP program. We believe that readiness for AP is attainable, and that educators can expand readiness by opening access to Pre-AP course work. We commit to supporting educators and communities in their efforts to make AP courses widely available, advancing students in their plans for college and careers.

Offering AP Courses: The AP Course Audit

The AP Program unequivocally supports the principle that each school implements its own curriculum that will enable students to develop the content understandings and skills described in the course framework.

While the unit sequence represented in this publication is optional, the AP Program does have a short list of curricular and resource requirements that must be fulfilled before a school can label a course "Advanced Placement" or "AP." Schools wishing to offer AP courses must participate in the AP Course Audit, a process through which AP teachers' course materials are reviewed by college faculty. The AP Course Audit was created to provide teachers and administrators with clear guidelines on curricular and resource requirements for AP courses and to help colleges and universities validate courses marked "AP" on students' transcripts. This process ensures that AP teachers' courses meet or exceed the curricular and resource expectations that college and secondary school faculty have established for college-level courses.

The AP Course Audit form is submitted by the AP teacher and the school principal (or designated administrator) to confirm awareness and understanding of the curricular and resource requirements. A syllabus or course outline, detailing how course requirements are met, is submitted by the AP teacher for review by college faculty.

Please visit the **AP Course Audit** website for more information to support the preparation and submission of materials for the AP Course Audit.

How the AP Program Is Developed

The scope of content for an AP course and exam is derived from an analysis of hundreds of syllabi and course offerings of colleges and universities. Using this research and data, a committee of college faculty and expert AP teachers work within the scope of the corresponding college course to articulate what students should know and be able to do upon the completion of the AP course. The resulting course framework is the heart of this course and exam description and serves as a blueprint of the content and skills that can appear on an AP Exam.

The AP Test Development Committees are responsible for developing each AP Exam, ensuring the exam questions are aligned to the course framework. The AP Exam development process is a multiyear endeavor; all AP Exams undergo extensive review, revision, piloting, and analysis to ensure that questions are accurate, fair, and valid, and that there is an appropriate spread of difficulty across the questions.

Committee members are selected to represent a variety of perspectives and institutions (public and private, small and large schools and colleges), and a range of gender, racial/ethnic, and regional groups. A list of each subject's current AP Test Development Committee members is available on **AP Central®**.

Throughout AP course and exam development, College Board gathers feedback from various stakeholders in both secondary schools and higher education institutions. This feedback is carefully considered to ensure that AP courses and exams are able to provide students with a college-level learning experience and the opportunity to demonstrate their qualifications for advanced placement and/or college credit.

How AP Exams Are Scored

The exam scoring process, like the course and exam development process, relies on the expertise of both AP teachers and college faculty. While multiple-choice questions are scored by machine, the free-response questions and through-course performance assessments, as applicable, are scored by thousands of college faculty and expert AP teachers. Most are scored at the annual AP Reading, while a small portion are scored online. All AP Readers are thoroughly trained, and their work is monitored throughout the Reading for fairness and consistency. In each subject, a highly respected college faculty member serves as Chief Faculty Consultant and, with the help of AP Readers in leadership positions, maintains the accuracy of the scoring standards. Scores on the free-response questions and performance assessments are weighted and combined with the results of the computer-scored multiple-choice questions, and this raw score is converted into a composite AP score on a 1–5 scale.

AP Exams are **not** norm-referenced or graded on a curve. Instead, they are criterion-referenced, which means that every student who meets the criteria for an AP score of 2, 3, 4, or 5 will receive that score, no matter how many students that is. The criteria for the number of points a student must earn on the AP Exam to receive scores of 3, 4, or 5—the scores that research consistently validates for credit and placement purposes—include:

- The number of points successful college students earn when their professors administer AP Exam questions to them.
- Performance that researchers have found to be predictive of an AP student succeeding when placed into a subsequent higher-level college course.
- The number of points college faculty indicate, after reviewing each AP question, that they expect is necessary to achieve each AP grade level.

Using and Interpreting AP Scores

The extensive work done by college faculty and AP teachers in the development of the course and exam and throughout the scoring process ensures that AP Exam scores accurately represent students' achievement in the equivalent college course. Frequent and regular research studies establish the validity of AP scores as follows:

AP Score	Credit Recommendation	College Grade Equivalent
5	Extremely well qualified	A
4	Well qualified	A-, B+, B
3	Qualified	B-, C+, C
2	Possibly qualified	n/a
1	No recommendation	n/a

While colleges and universities are responsible for setting their own credit and placement policies, most private colleges and universities award credit and/or advanced placement for AP scores of 3 or higher. Additionally, most states in the U.S. have adopted statewide credit policies that ensure college credit for scores of 3 or higher at public colleges and universities. To confirm a specific college's AP credit/placement policy, use the search engine available on the **AP Credit Policy Search** page.

BECOMING AN AP READER

Each June, thousands of AP teachers and college faculty members from around the world gather for seven days in multiple locations to evaluate and score the free-response sections of the AP Exams. Ninety-eight percent of surveyed educators who took part in the AP Reading say it was a positive experience.

There are many reasons to consider becoming an AP Reader, including opportunities to:

- **Bring positive changes to the classroom:** Surveys show that the vast majority of returning AP Readers—both high school and college educators—make improvements to the way they teach or score because of their experience at the AP Reading.
- **Gain in-depth understanding of AP Exam and AP scoring standards:** AP Readers gain exposure to the quality and depth of the responses from the entire pool of AP Exam takers and thus are better able to assess their students' work in the classroom.
- **Receive compensation:** AP Readers are compensated for their work during the Reading. Expenses, lodging, and meals are covered for Readers who travel.
- **Score from home:** AP Readers have online distributed scoring opportunities for certain subjects. Check the **AP Reader** site for details.
- **Earn Continuing Education Units (CEUs):** AP Readers earn professional development hours and CEUs that can be applied to PD requirements by states, districts, and schools.

How to Apply

Visit the **Become an AP Reader** site for eligibility requirements and to start the application process.

About the AP Statistics Course

The AP Statistics course introduces students to the major concepts and tools for formulating questions, collecting and analyzing data, and interpreting results from data. The practices and skills for the course have been aligned to help students understand the statistical problem-solving process based on the American Statistical Association (ASA) recommendations.
The content, skills, and assessments in the AP Statistics course focus on exploring data, sampling and experimentation, probability and simulation, and statistical inference. Students use technology, investigations, problem solving, and writing as they build conceptual understanding.

College Course Equivalent

The AP Statistics course is equivalent to a one-semester, introductory, non-calculus-based college course in statistics.

Prerequisites

The AP Statistics course is an excellent option for any secondary school student who possesses both sufficient mathematical maturity and quantitative reasoning ability. Decisions about whether to take AP Statistics and when to take it depend on a student’s plans:

- Students planning to take a science course in their senior year will benefit greatly from taking AP Statistics in their junior year.
- For students who would otherwise take no mathematics in their senior year, AP Statistics allows them to continue to develop their quantitative skills.
- Students who wish to leave open the option of taking calculus in college should include AP Precalculus in their high school program and perhaps take AP Statistics concurrently with AP Precalculus.
- Students with the appropriate mathematical background are encouraged to take both AP Statistics and AP Calculus in high school.
- Students who have an interest in data science are encouraged to take both AP Statistics and an AP Computer Science course.

AP STATISTICS

Course Framework

Course Framework Components

Course Units

Unit 1: Exploring One-Variable Data and Collecting Data

Unit 2: Probability, Random Variables, and Probability Distributions

Unit 3: Inference for Categorical Data: Proportions

Unit 4: Inference for Quantitative Data: Means

Unit 5: Regression Analysis

Curriculum Framework Overview

This course framework provides a clear and detailed description of the course requirements necessary for student success. The framework specifies what students should know and be able to do to qualify for college credit and/or placement.

The course framework includes two essential components:

- **Statistical Practices and Skills**
  The statistical practices and skills are central to the study and problem-solving processes of statistics. Students should develop and apply the described skills on a regular basis over the span of the course.
- **Course Content**
  The course content is organized into commonly taught units of study that provide a suggested sequence for the course and detail required content and conceptual understandings that colleges and universities typically expect students to master to qualify for college credit and/or placement.

Statistical Practices

Practice 1	Practice 2	Practice 3	Practice 4
Formulate Questions	Collect Data	Analyze Data	Interpret Results
Determine an investigative question for a statistical study.	Identify and justify methods for collecting data and conducting statistical inference.	Construct representations of data and calculate numerical statistical outputs.	Interpret results and justify conclusions and methods.
SKILLS			
1.A: Determine a valid investigative question that requires a statistical investigation.	**2.A:** Identify information to answer a question or solve a problem. **2.B:** Justify an appropriate method for ethically gathering and representing data. **2.C:** Identify appropriate statistical inference methods. **2.D:** Identify types of errors and relationships among components in statistical inference methods. **2.E:** Identify the null and alternative hypotheses.	**3.A:** Construct tabular and graphical representations of data and distributions. **3.B:** Calculate summary statistics, relative positions of points within a distribution, and predicted responses. **3.C:** Calculate and estimate expected counts, percentages, probabilities, and intervals. **3.D:** Calculate means, standard deviations, and parameters for probability distributions. **3.E:** Calculate appropriate statistical inference method results.	**4.A:** Describe and compare tabular and graphical representations of data. **4.B:** Justify a claim based on statistical calculations and results. **4.C:** Describe distributions and compare relative positions of points within a distribution. **4.D:** Interpret statistical calculations and results to assess meaning or a claim. **4.E:** Justify the use of a chosen statistical inference method by verifying conditions. **4.F:** Interpret results of statistical inference methods. **4.G:** Justify a claim based on statistical inference method results.

UNIT 1

Exploring One-Variable Data and Collecting Data

THIS PAGE IS INTENTIONALLY LEFT BLANK.

TOPIC 1.1

Introducing Statistics: What Can We Learn from Data?

LEARNING OBJECTIVE

1.1.A

Identify components within a statistical study.

ESSENTIAL KNOWLEDGE

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 N.

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 n.

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

Determine an investigative question within a statistical study.

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.

TOPIC 1.2
Variables

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
1.2.A Identify observational units, variables, parameters, and statistics from a statistical study or data set.	**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 Identify types of variables.	**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 Identify types of quantitative variables.	**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.

TOPIC 1.3

Tabular Representation and Summary Statistics for One Categorical Variable

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
1.3.A Construct categorical one-variable tabular representations.	**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 Describe categorical one-variable tabular representations with summary statistics.	**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.

Return to Table of Contents

TOPIC 1.4

Graphical Representations for One Categorical Variable

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
1.4.A Construct categorical one-variable graphical representations.	**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 Justify a claim using categorical one-variable graphical representations.	**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 Compare multiple categorical one-variable tabular and graphical representations.	**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.

TOPIC 1.5

Graphical Representations for One Quantitative Variable

LEARNING OBJECTIVE

1.5.A

Construct quantitative one-variable graphical representations.

ESSENTIAL KNOWLEDGE

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.

TOPIC 1.6

Descriptions for One Quantitative Variable Distributions

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
1.6.A Describe distributions of quantitative one-variable graphical representations.	**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 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 Justify a claim using distributions of quantitative one-variable graphical representations.	**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.

TOPIC 1.7

Summary Statistics for One Quantitative Variable

LEARNING OBJECTIVE

1.7.A

Calculate measures of center and position for quantitative data.

ESSENTIAL KNOWLEDGE

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: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where x_i represents the *i*th 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 *p*th percentile is the value that has *p*% 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

Calculate measures of variability for quantitative data.

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.

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
1.7.B Calculate measures of variability for quantitative data.	**1.7.B.3** The interquartile range (IQR) is the difference between the third and first quartiles: $\mathrm{Q3}-\mathrm{Q1}$. **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 *s* and calculated by $s=\sqrt{\frac{1}{n-1}\sum\left(x_i-\overline{x}\right)^2}$, where x_i is the data value, $\overline{x}$ is the mean, and *n* is the number of data values in the sample. The square of the sample standard deviation, s^2, is called the sample variance.
1.7.C Calculate different units of measurement for summary statistics.	**1.7.C.1** Changing units of measurement affects the values of the calculated statistics.
1.7.D Calculate outliers for quantitative data.	**1.7.D.1** There are many methods for determining potential outliers. Two methods frequently used are as follows: i. An outlier is a value located more than $1.5\times\mathrm{IQR}$ above the third quartile or more than $1.5\times\mathrm{IQR}$ below the first quartile. ii. An outlier is a value located more than 2 standard deviations above, or below, the mean.
1.7.E Justify the selection of a summary statistic for describing quantitative data.	**1.7.E.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.E.2** Summary statistics of a quantitative variable may reveal information that can be used to justify claims about the variable in context.

TOPIC 1.8

Graphical Representations of Summary Statistics for One Quantitative Variable

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
1.8.A Construct quantitative one-variable graphical representations of summary statistics.	**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 Describe quantitative one-variable graphical representations of summary statistics based on the relationship of the mean and the median.	**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.

TOPIC 1.9

Comparisons of the Distributions for One Quantitative Variable

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
1.9.A Compare multiple quantitative one-variable graphical representations.	**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 Compare multiple quantitative one-variable graphical representations of summary statistics.	**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 Justify a claim using multiple quantitative one-variable graphical representations.	**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 Calculate *z*-scores with population parameters.	**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 $\frac{x_i - \mu}{\sigma}$, 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 Compare *z*-scores as measures of relative position for distributions.	**1.9.E.1** *z*-scores may be used to compare relative positions of individual values within a distribution or between distributions.

TOPIC 1.10

The Investigative Question Revisited and Data Collection

LEARNING OBJECTIVE

1.10.A

Determine the components of an investigative question within a statistical study.

ESSENTIAL KNOWLEDGE

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.

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).

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

Identify a census.

1.10.B.1

A census consists of recording information from all items or individuals in a population.

1.10.C

Identify an experiment.

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.

Return to Table of Contents

LEARNING OBJECTIVE

1.10.C

Identify an experiment.

ESSENTIAL KNOWLEDGE

1.10.C.4

A response variable is an outcome measured on each experimental unit after the treatment has been administered.

LEARNING OBJECTIVE

1.10.D

Identify an observational study.

ESSENTIAL KNOWLEDGE

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 lowering the credibility of the assertion of 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.

LEARNING OBJECTIVE

1.10.E

Justify the appropriateness of generalizations for a statistical study.

ESSENTIAL KNOWLEDGE

1.10.E.1

A sample is considered random when all observational units in the sample have an equal chance of being selected from the population. A random mechanism is any resource used to select the observational units to be included in the sample.

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.

Return to Table of Contents

TOPIC 1.11

Random Sampling

LEARNING OBJECTIVE

1.11.A

Identify a sampling method given a description of a study.

ESSENTIAL KNOWLEDGE

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

Justify the appropriateness of a sampling method.

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.

TOPIC 1.12

Potential Problems with Sampling

LEARNING OBJECTIVE

1.12.A
Identify potential sources of bias in sampling methods.

ESSENTIAL KNOWLEDGE

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.

TOPIC 1.13

Experimental Design

LEARNING OBJECTIVE

1.13.A

Identify elements of a well-designed experiment.

ESSENTIAL KNOWLEDGE

1.13.A.1

A well-designed experiment should include the following:

i. Comparisons of at least two treatment groups, one of which could be a control group

ii. Random assignment of treatments to experimental units

iii. Replication

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 distributed differently among treatment groups and affects the response variable.

1.13.A.9

Replication within an experiment means more than one experimental unit is assigned to each treatment.

LEARNING OBJECTIVE

ESSENTIAL KNOWLEDGE

1.13.A

Identify elements of a well-designed experiment.

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

Identify experimental designs.

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

Justify the appropriateness of a particular experimental design.

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

Justify the appropriateness of the conclusions based on a well-designed experiment.

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.

AP STATISTICS

UNIT 2

Probability, Random Variables, and Probability Distributions

THIS PAGE IS INTENTIONALLY LEFT BLANK.

TOPIC 2.1

Tabular and Graphical Representations for the Distributions of Two Categorical Variables

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
2.1.A Compare tabular and graphical representations for the relationship between two categorical variables.	**2.1.A.1** A two-way table, also called a contingency table, can be used to summarize and compare data for two categorical variables. The entries in the cells of the table can be frequencies (i.e., counts) or relative frequencies (i.e., proportions). **2.1.A.2** Side-by-side bar charts, segmented bar charts, and mosaic plots are examples of graphs used to display the relationship between two categorical variables. In these graphs, the frequency or relative frequency of each category, or level, of one of the categorical variables is displayed for each category of the other categorical variable. **2.1.A.3** Graphical representations of two categorical variables can be used to compare the relationship of one categorical variable across the levels of the other categorical variable and determine whether the two variables are associated.
2.1.B Justify a claim using tabular and graphical representations for the distributions of two categorical variables.	**2.1.B.1** Tabular and graphical representations for the distributions of two categorical variables may reveal information that can be used to justify claims about the variable in context.

TOPIC 2.2

Summary Statistics for Two Categorical Variables

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
2.2.A Calculate summary statistics from two-way tables.	**2.2.A.1** A joint relative frequency in a two-way table is a cell frequency divided by the total for the entire table. **2.2.A.2** A marginal relative frequency in a two-way table is a row total divided by the total for the entire table or a column total divided by the total for the entire table. **2.2.A.3** A conditional relative frequency is a relative frequency computed by restricting to a particular level, or category of interest. A conditional relative frequency can be a cell frequency in a row divided by the total for that row or it can be a cell frequency in a column divided by the total for that column.
2.2.B Compare summary statistics for two categorical variables.	**2.2.B.1** Summary statistics for two categorical variables can be used to compare distributions for evidence of association between the two variables.
2.2.C Justify a claim using summary statistics for two categorical variables.	**2.2.C.1** Summary statistics for two categorical variables may reveal information that can be used to justify claims about the variable in context.

Return to Table of Contents

TOPIC 2.3

Estimating Probabilities Using Simulation

LEARNING OBJECTIVE

2.3.A

Estimate probabilities using simulations.

ESSENTIAL KNOWLEDGE

2.3.A.1

A random process generates results that are determined by chance.

2.3.A.2

An outcome is the result of one trial of a random process.

2.3.A.3

An event is a collection of outcomes.

2.3.A.4

Simulation is a way to model random events such that simulated outcomes closely match real-world outcomes. All possible outcomes are associated with a value to be determined by chance. Record the counts of simulated outcomes and the count total.

2.3.A.5

The probability of an outcome or event is its long-run relative frequency—that is, its relative frequency over a large number of trials.

2.3.A.6

The relative frequency of an outcome or event determined from empirical data can be used to estimate the actual, or true, probability of that outcome or event.

2.3.A.7

The law of large numbers states that for independent trials, as the number of trials increases, the long-run relative frequency of the outcome or event gets closer and closer to a single value.

TOPIC 2.4

Introduction to Probability

LEARNING OBJECTIVE

2.4.A

Calculate probabilities for events and their complements.

ESSENTIAL KNOWLEDGE

2.4.A.1

The sample space of a random process is the set of all possible nonoverlapping outcomes. The probability of the sample space is 1.

2.4.A.2

If all outcomes in the sample space are equally likely, then the theoretical probability an event *E* will occur is $\frac{\text{number of outcomes in event } E}{\text{total number of outcomes in the sample space}}$. The probability of event *E* occurring is written as $P(E)$.

2.4.A.3

The probability of an event is a number between 0 and 1, inclusive.

2.4.A.4

The probability of the complement of an event *E*, which can be written as E', $\overline{E}$, or E^C (i.e., the probability of "not *E*") is equal to $1-P(E)$.

TOPIC 2.5

Mutually Exclusive Events

LEARNING OBJECTIVE

2.5.A

Justify why two events are mutually exclusive (or disjoint) using joint probability.

ESSENTIAL KNOWLEDGE

2.5.A.1

The probability that events *A* and *B* both will occur, sometimes called the joint probability, is the probability of the intersection of *A* and *B*. Joint probability is defined as $P(A \text{ intersect } B)$ or $P(A \cap B)$.

2.5.A.2

Two events are mutually exclusive, or disjoint, if they cannot occur at the same time. This means that if two events are mutually exclusive, then $P(A \cap B) = 0$.

TOPIC 2.6

Conditional Probability

LEARNING OBJECTIVE

2.6.A

Calculate conditional probabilities.

ESSENTIAL KNOWLEDGE

2.6.A.1

The probability that event *A* will occur given that event *B* has occurred is called a conditional probability and is written as $P(A|B)$. Conditional probability is defined as $P(A|B)=\dfrac{P(A\cap B)}{P(B)}$.

2.6.A.2

The general multiplication rule states that the probability that events *A* and *B* will occur is equal to the probability that event *A* will occur multiplied by the conditional probability that event *B* will occur given that event *A* has occurred. The multiplication rule is defined as $P(A\cap B)=P(A)\cdot P(B|A)$.

TOPIC 2.7

Independent Events and Unions of Events

LEARNING OBJECTIVE

2.7.A

Calculate probabilities for independent events and for the union of two events.

ESSENTIAL KNOWLEDGE

2.7.A.1

Events *A* and *B* are independent if and only if knowing whether event *A* has occurred (or will occur) does not change the probability that event *B* will occur. When events *A* and *B* are independent, then $P(A|B)=P(A)$, $P(B|A)=P(B)$, and $P(A\cap B)=P(A)\cdot P(B)$.

2.7.A.2

The probability that event *A* or event *B* (or both) will occur is the probability of *A* union *B*. The probability of the union is defined as $P(A\cup B)$.

2.7.A.3

$P(A \text{ union } B)=P(A)+P(B)-P(A \text{ intersect } B)$, or
$P(A\cup B)=P(A)+P(B)-P(A\cap B)$.

TOPIC 2.8

Introduction to Random Variables and Probability Distributions

LEARNING OBJECTIVE

2.8.A

Construct a probability distribution for a discrete random variable.

ESSENTIAL KNOWLEDGE

2.8.A.1

A random variable is a variable whose values have numerical outcomes that result from a random phenomenon.

2.8.A.2

A probability distribution for a discrete random variable shows the probability associated with every possible value of the random variable. The sum of the probabilities over all possible values of a discrete random variable is 1.

2.8.A.3

A discrete probability distribution can be determined using the rules of probability or estimated with a simulation.

2.8.A.4

A discrete probability distribution can be represented as a graph, table, or function showing the probabilities associated with values of a random variable.

2.8.A.5

A cumulative probability distribution can be represented as a table or function and shows the probability of being less than or equal to each value of the discrete random variable.

TOPIC 2.9

Parameters of Random Variables

LEARNING OBJECTIVE

2.9.A

Calculate the parameter, mean, and standard deviation for a discrete random variable.

ESSENTIAL KNOWLEDGE

2.9.A.1

A numerical value measuring a characteristic of a probability distribution of a random variable, or a population, is a parameter. The value of a parameter is a single, fixed value.

2.9.A.2

The expected value (or mean) of a probability distribution is a parameter and is denoted by $E(X)$ or μ_X. For a discrete random variable *X*, the expected value is calculated as $\mu_X = \sum x_i \cdot P(x_i)$, where x_i is the possible value of the random variable and $P(x_i)$ is the probability of the possible value of the random variable. The expected value can be interpreted as the long-run average outcome of the random variable. The discrete random variable can only take on values that are countable or finite.

2.9.A.3

The standard deviation of a probability distribution is a parameter represented by $SD(X)$ or σ_X. For a discrete random variable *X*, the standard deviation is calculated as $\sigma_X = \sqrt{\sum (x_i - \mu_X)^2 \cdot P(x_i)}$, where x_i is the possible value of the random variable, μ_X is the mean, and $P(x_i)$ is the probability of the possible value of the random variable. The standard deviation can be interpreted as the typical deviation of the values of the random variable from the mean value (or expected value) of the random variable over the long run. The square of the standard deviation of a random variable is called the variance of the random variable and is denoted as $V(X)$ or σ^2_X.

2.9.B

Interpret the parameter, mean, and standard deviation for a discrete random variable.

2.9.B.1

The parameter, mean, and standard deviation for the probability distribution of a discrete random variable should be interpreted in the context of a specific population.

TOPIC 2.10

The Binomial Distribution

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
2.10.A Justify why a random variable is or is not a binomial random variable.	**2.10.A.1** A binomial random variable, *X*, is a discrete random variable that counts the number of successes in repeated independent trials, *n*, that have only two possible outcomes (success or failure), with the probability of success *p* and the probability of failure $1-p$.
2.10.B Calculate the mean and standard deviation for a binomial distribution.	**2.10.B.1** If a random variable is binomial, its mean, μ_X, is np and its standard deviation, σ_X, is $\sqrt{np(1-p)}$.
2.10.C Interpret the mean, standard deviation, and probabilities for a binomial distribution.	**2.10.C.1** The mean, standard deviation, and probabilities for a binomial distribution should be interpreted in context.
2.10.D Estimate probabilities of binomial random variables using data from a simulation.	**2.10.D.1** A probability distribution can be constructed using the rules of probability or estimated with a simulation.
2.10.E Calculate probabilities for a binomial distribution.	**2.10.E.1** The probability that a binomial random variable, *X*, has exactly *x* successes for *n* independent trials, when the probability of success is *p*, is calculated as $P(X=x)=\binom{n}{x}p^{x}(1-p)^{n-x}$, $x=0, 1, 2, \ldots, n$. This is called the binomial probability function.

TOPIC 2.11

The Normal Distribution

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
2.11.A Describe the standard normal distribution.	**2.11.A.1** A continuous random variable is a variable that can take on any value within a specified domain. Every interval within the domain has a probability associated with it. **2.11.A.2** Many continuous random variables are well-modeled by a normal distribution. **2.11.A.3** A normal distribution can be described as a continuous, unimodal, bell-shaped, and symmetric curve. **2.11.A.4** A normal curve can be used to model a distribution of data and a continuous random variable. **2.11.A.5** The normal distribution, or the normal curve, is identified by two parameters, the mean, μ, and the standard deviation, σ. The smaller the standard deviation, the taller and more concentrated the normal curve is around its mean. The larger the standard deviation, the shorter and less concentrated the normal curve is around its mean.
2.11.B Calculate the mean and standard deviation for a normal distribution.	**2.11.B.1** A standard normal distribution is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.
2.11.C Calculate percentages from a normal distribution using the empirical rule.	**2.11.C.1** The empirical rule can be used to estimate the area of a region under the graph of the normal distribution curve. For a normal distribution, approximately 68% of 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, or the 68–95–99.7 rule.
2.11.D Calculate the probability that a particular value lies within a given interval of a normal distribution.	**2.11.D.1** If the distribution of a random variable is approximately normal, the probability that the random variable takes on values within a particular interval of the random variable is determined by the area under the normal curve within that interval. The total probability or area under the normal curve is 1.

LEARNING OBJECTIVE

2.11.E

Calculate the associated intervals and areas of a normal distribution.

ESSENTIAL KNOWLEDGE

2.11.E.1

The boundaries of an interval associated with a given area in a normal distribution can be determined using technology or using *z*-scores and a standard normal table.

2.11.E.2

Intervals associated with a given area in a normal distribution can be determined by assigning appropriate inequalities to the boundaries of the intervals. To determine the intervals, *p* is defined as a number between 0 and 100, x_a is the lower bound, and x_b is the upper bound on a normal distribution.

i. $P(X < x_a) = \frac{p}{100}$ means that the lowest *p*% of the values lie to the left of x_a.

ii. $P(x_a < X < x_b) = \frac{p}{100}$ means that *p*% of the values lie between x_a and x_b.

iii. $P(X > x_b) = \frac{p}{100}$ means that the highest *p*% of the values lie to the right of x_b.

iv. To determine the most extreme *p*% of values on both sides requires dividing the area associated with *p*% into two equal areas on either extreme of the distribution: $P(X < x_a) = \frac{1}{2}\left(\frac{p}{100}\right)$ and $P(X > x_b) = \frac{1}{2}\left(\frac{p}{100}\right)$ mean that half of the *p*% most extreme values lie to the left of x_a and half of the *p*% most extreme values lie to the right of x_b.

2.11.F

Compare measures of relative position for distributions.

2.11.F.1

Percentiles and proportions may be used to compare relative positions of individual values within a normal distribution or between normal distributions.

TOPIC 2.12

Sampling Distributions and the Central Limit Theorem

LEARNING OBJECTIVE

2.12.A

Describe sampling distributions with simulations.

ESSENTIAL KNOWLEDGE

2.12.A.1

A sampling distribution of a statistic is the distribution of values of the statistic for all possible samples of a given size from a given population.

2.12.A.2

The sampling distribution of a statistic can be simulated by repeatedly generating a large number of random samples from the population assuming known value(s) for the parameter(s). The value of the statistic is determined and recorded for each sample. The resulting distribution of the sample statistic values approximates the sampling distribution of the statistic.

2.12.A.3

A randomization distribution is the distribution of a statistic generated by simulation from repeatedly randomly reallocating, or reassigning, the response values to treatment groups. The value of the statistic is determined and recorded for each reallocation, or reassignment. The resulting distribution of the statistic values approximates the sampling distribution of the statistic.

2.12.A.4

The central limit theorem (CLT) states that the sampling distribution of a mean of a random sample has a shape that can be approximated by a normal distribution. The larger the sample is, the better the approximation will be.

THIS PAGE IS INTENTIONALLY LEFT BLANK.

AP STATISTICS

UNIT 3

Inference for Categorical Data: Proportions

THIS PAGE IS INTENTIONALLY LEFT BLANK.

TOPIC 3.1

Estimators

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
3.1.A Justify why an estimator is or is not unbiased.	**3.1.A.1** When estimating a population parameter, an estimator is unbiased if, on average, the value of the estimator does not underestimate or overestimate the population parameter.
3.1.B Calculate estimates for a population parameter.	**3.1.B.1** A sample statistic is a point estimator of the corresponding population parameter and can be thought of as the estimate of the population parameter. For example, the sample proportion $\hat{p}$ is a point estimator for the population proportion *p*.

TOPIC 3.2
Sampling Distributions for Sample Proportions

LEARNING OBJECTIVE

3.2.A

Calculate the mean and standard deviation of a sampling distribution for a sample proportion.

ESSENTIAL KNOWLEDGE

3.2.A.1

For a population with population proportion *p*, when the sampled values are independent, the sampling distribution of a sample proportion $\hat{p}$ has a mean $\mu_{\hat{p}} = p$ and a standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

3.2.B

Justify the appropriateness of conditions for the sampling distribution of a sample proportion.

3.2.B.1

Sampling without replacement requires that two conditions be met:

i. The randomization condition—the data should be collected using a random sample.

ii. The 10% condition—the population size must be at least 10 times larger than the sample size $(n < 10\% N)$, where *N* is the size of the population and *n* is the sample size.

3.2.B.2

The sampling distribution of the sample proportion $\hat{p}$ is approximately normal provided the sample size is large enough. To ensure the sample size is large enough, the following condition must be met: $np \geq 10$ and $n(1-p) \geq 10$, where np is the expected number of successes and $n(1-p)$ is the expected number of failures.

3.2.C

Interpret the mean, standard deviation, and probabilities for a sampling distribution of a sample proportion.

3.2.C.1

The mean, standard deviation, and probabilities for a sampling distribution of a sample proportion should be interpreted in the context of a specific population.

TOPIC 3.3

Constructing a Confidence Interval for a Population Proportion

LEARNING OBJECTIVE

3.3.A

Identify an appropriate confidence interval procedure including the parameter for a population proportion.

ESSENTIAL KNOWLEDGE

3.3.A.1

A confidence interval is an interval estimate for a population parameter. Based on the sample proportion, a confidence interval can be calculated to estimate the value of a single population proportion. The appropriate confidence interval procedure is a one-sample *z*-interval for a population proportion.

3.3.A.2

The parameter for a confidence interval for a population proportion should reference the proportion, the response variable, and the population in context.

3.3.B

Justify the appropriateness of constructing a confidence interval for a population proportion by verifying conditions.

3.3.B.1

A one-sample *z*-interval for a population proportion requires that three conditions be met:

i. The randomization condition—the data should be collected using a random sample.

ii. The 10% condition—when sampling without replacement, the population size must be at least 10 times larger than the sample size $(n < 10\% N)$, where *N* is the size of the population and *n* is the sample size.

iii. The normality condition—the expected number of successes, $n\hat{p}$, and the expected number of failures, $n\left(1-\hat{p}\right)$, should be at least 10.

3.3.C

Calculate an appropriate confidence interval for a population proportion.

3.3.C.1

z^* denotes a critical value, such that $-z^*$ and $+z^*$ represent the boundaries enclosing the middle *C* % of the standard normal distribution, in which *C* % is an approximate confidence level with which the population proportion is estimated.

3.3.C.2

An interval estimate can be constructed as $\text{point estimate} \pm \left(\text{margin of error}\right)$. For a population proportion, the one-sample *z*-interval estimate is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}.$$

LEARNING OBJECTIVE

3.3.D

Calculate the standard error and margin of error of a sample statistic for a confidence interval for a population proportion, and estimate a given sample size from the margin of error.

ESSENTIAL KNOWLEDGE

3.3.D.1

The standard error (*SE*) of a statistic is an estimate of the standard deviation of the sampling distribution of the statistic. The standard error of the sample proportion $\hat{p}$ is $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

3.3.D.2

The standard error quantifies the typical amount that a statistic will vary from the value of the corresponding population parameter.

3.3.D.3

The margin of error of $\hat{p}$ is half the width of the confidence interval and is calculated as the critical value z^* times the standard error (*SE*) of $\hat{p}$, which equals $z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

3.3.D.4

The formula for the margin of error (*MOE*) can be rearranged to solve for *n*, $n = \frac{(z^*)^2(\hat{p})(1-\hat{p})}{(MOE)^2}$, the minimum sample needed to achieve a given margin of error. For this purpose, if $\hat{p}$ is not defined or unable to be calculated, use $\hat{p} = 0.5$ in order to find the upper bound for the sample size that will result in a given margin of error.

TOPIC 3.4

Justifying a Claim Based on a Confidence Interval for a Population Proportion

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
3.4.A Interpret a confidence interval in context for a population proportion.	**3.4.A.1** Because the confidence interval for a population proportion is calculated based on a sample from a population, the computed interval may or may not contain the value of the population proportion. **3.4.A.2** The interpretation of the confidence level is that in repeated random sampling with the same sample size, approximately *C* % of confidence intervals calculated will capture the population proportion, with *C* representing the numerical value of the confidence level used. **3.4.A.3** When interpreting a *C* % confidence interval for a population proportion, we say we are *C* % confident that the interval (a, b) contains the true value of the parameter for the population, where *a* represents the lower limit and *b* represents the upper limit. An interpretation of a confidence interval for a population proportion includes a reference to the parameter with details about the population it represents in the context of the study.
3.4.B Justify a claim based on a confidence interval for a population proportion.	**3.4.B.1** A confidence interval for a population proportion provides a range of plausible values that may serve as convincing evidence to support a particular claim about the population proportion.
3.4.C Identify the relationships among sample size, confidence interval width, confidence level, and margin of error for a population proportion.	**3.4.C.1** For a given sample, increasing the confidence level will result in the following: i. The critical value will increase. ii. The margin of error will increase. iii. The width of the confidence interval will increase. **3.4.C.2** Increasing the sample size decreases the standard error. Thus, when all other things remain the same, the width of the confidence interval for a population proportion tends to decrease as the sample size increases. For a confidence interval for a population proportion with a given confidence level, the width of the interval is approximately proportional to $\frac{1}{\sqrt{n}}$.

TOPIC 3.5

Setting Up a Test for a Population Proportion

LEARNING OBJECTIVE

3.5.A

Identify an appropriate testing method for a population proportion including the parameter for the population proportion.

ESSENTIAL KNOWLEDGE

3.5.A.1

A hypothesis test is a statistical inference procedure that is used to make a decision about the value of a population parameter. The appropriate hypothesis testing procedure is a one-sample *z*-test for a population proportion.

3.5.A.2

The parameter for a hypothesis test for a population proportion should reference the population parameter, the response variable, and the population in context.

3.5.B

Identify the null and alternative hypotheses for a population proportion.

3.5.B.1

In the hypothesis testing procedure, the null hypothesis, H_0, is the statement about a parameter that is assumed to be correct unless there is convincing statistical evidence suggesting otherwise. It is the status quo condition. The alternative hypothesis, H_a, is the claim or belief about a parameter for which evidence is being collected. A researcher's claim or belief about the population parameter is represented by the alternative hypothesis.

3.5.B.2

The null hypothesis contains an equality reference $(=, \geq, \text{or } \leq)$. Although the null hypothesis for a one-sided test may include an inequality symbol, in AP Statistics it is tested at the boundary of equality. The alternative hypothesis with $<$ or $>$ is called one-sided, and the alternative hypothesis with $\neq$ is called two-sided.

3.5.B.3

The null hypothesis for a one-sample *z*-test for a population proportion is as follows: $\mathrm{H}_0 : p = p_0$, where p_0 is the null hypothesized value for the population proportion. A one-sided alternative hypothesis for a one-sample *z*-test for a population proportion is either $\mathrm{H}_a : p < p_0$ or $\mathrm{H}_a : p > p_0$. A two-sided alternative hypothesis is $\mathrm{H}_a : p \neq p_0$.

3.5.C

Justify the appropriateness of a hypothesis test for a population proportion by verifying conditions.

3.5.C.1

A one-sample *z*-test for a population proportion requires that three conditions be met:

i. The randomization condition—the data should be collected using a random sample.

ii. The 10% condition—when sampling without replacement, the population size must be at least 10 times larger than the sample size $(n < 10\% N)$, where *N* is the size of the population and *n* is the sample size.

iii. The normality condition—the expected number of successes, np_0, and the expected number of failures, $n(1 - p_0)$, should be at least 10.

TOPIC 3.6

p-Values

LEARNING OBJECTIVE

3.6.A

Interpret the *p*-value of a hypothesis test for a population proportion.

ESSENTIAL KNOWLEDGE

3.6.A.1

Given the null hypothesis is true, there is a probability distribution of the test statistic called the null distribution. Using the null distribution, the *p*-value is the probability of obtaining a test statistic as extreme or more extreme (i.e., in the direction of the alternative hypothesis) than the test statistic that is observed given that the null hypothesis is true. That is, when *x* is the test statistic, the *p*-value is determined by finding the following:

i. The probability at or above the observed value of the test statistic $\left(P(z \geq x)\right)$, if the alternative is $>$

ii. The probability at or below the observed value of the test statistic $\left(P(z \leq x)\right)$, if the alternative is $<$

iii. The probability less than or equal to the negative of the absolute value of the test statistic plus the probability greater than or equal to the absolute value of the test statistic, $\left(P\left(z \leq -|x|\right)\right)+\left(P\left(z \geq |x|\right)\right)$, if the alternative is $\neq$

3.6.A.2

If the distribution of the test statistic has been simulated, the *p*-value is the proportion of values in the null distribution that are as extreme or more extreme than the observed value of the test statistic. This is as follows:

i. The proportion at or above the observed value of the test statistic, if the alternative is $>$

ii. The proportion at or below the observed value of the test statistic, if the alternative is $<$

iii. The proportion less than or equal to the negative of the absolute value of the test statistic plus the proportion greater than or equal to the absolute value of the test statistic, if the alternative is $\neq$

3.6.A.3

An interpretation of the *p*-value of a hypothesis test for a population proportion should include a statement that the *p*-value is computed by assuming the null hypothesis is true (i.e., by assuming the true population proportion is equal to the particular value stated in the null hypothesis in context).

3.6.A.4

Small *p*-values indicate that the observed value of the test statistic would be unusual if the null hypothesis were true and therefore provide evidence for the alternative hypothesis. The lower the *p*-value, the more convincing the statistical evidence for the alternative hypothesis.

3.6.A.5

p-values that are not small indicate that the observed value of the test statistic would not be unusual if the null hypothesis were true and therefore do not provide convincing statistical evidence for the alternative hypothesis, nor do they provide evidence that the null hypothesis is true.

TOPIC 3.7

Carrying Out a Test for a Population Proportion

LEARNING OBJECTIVE

3.7.A

Calculate an appropriate test statistic and *p*-value for testing a hypothesis about a population proportion.

ESSENTIAL KNOWLEDGE

3.7.A.1

The test statistic for testing a population proportion is $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$. The *z*-statistic has a standard normal distribution when the null hypothesis is true.

3.7.A.2

The distribution of the test statistic assuming the null hypothesis is true (null distribution) can be approximated by a probability model (e.g., a theoretical distribution such as the standard normal distribution).

3.7.A.3

The *p*-value of a one-sample *z*-test for a population proportion is found from the standard normal distribution using a table or technology.

3.7.B

Justify a claim about the population based on the results of a hypothesis test for a population proportion.

3.7.B.1

The significance level of a hypothesis test, denoted by α, is the predetermined probability of rejecting the null hypothesis given that it is true. The significance level may be given or determined by the researcher. The relationship between a test statistic and the significance level of a hypothesis test determines whether a result is statistically significant.

3.7.B.2

A formal decision in a hypothesis test explicitly compares the *p*-value to the significance level, α. If the $p\text{-value} \leq \alpha$, then reject the null hypothesis, $\mathrm{H}_0: p = p_0$. If the $p\text{-value} > \alpha$, then fail to reject the null hypothesis.

3.7.B.3

Rejecting the null hypothesis means there is convincing statistical evidence to support the alternative hypothesis. Failing to reject the null hypothesis means there is not convincing statistical evidence to support the alternative hypothesis.

3.7.B.4

A hypothesis test can lead to rejecting or not rejecting the null hypothesis but can never lead to concluding or proving that the null hypothesis is true. Lack of statistical evidence for the alternative hypothesis is not the same as evidence for the null hypothesis.

3.7.B.5

The results of a hypothesis test for a population proportion can serve as the statistical reasoning to support the answer to an investigative question about the population that was sampled.

3.7.B.6

A conclusion for the hypothesis test for a population proportion is stated in context consistent with, and in terms of, the alternative hypothesis using non-definitive language. The conclusion should contain a reference to the parameter and the population.

TOPIC 3.8

Potential Errors When Performing Tests

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
3.8.A Identify Type I and Type II errors.	**3.8.A.1** A Type I error occurs when there is evidence that the alternative hypothesis is true (due to the small *p*-value), but it is not. **3.8.A.2** A Type II error occurs when there is no evidence that the alternative hypothesis is true (due to the large *p*-value), but it is. **3.8.A.3** The power of a hypothesis test is the probability that a hypothesis test will correctly reject the false null hypothesis.
3.8.B Calculate the probability of Type I and Type II errors.	**3.8.B.1** The probability of making a Type I error is defined as the significance level, α. For a given study and hypothesis test, the probability of making a Type I error is typically set to a small value (e.g., 0.01, 0.05, 0.10) prior to collecting the data. **3.8.B.2** The probability of making a Type II error is $1-\text{power}$.
3.8.C Identify the factors that affect the probability of errors in hypothesis testing.	**3.8.C.1** For a given study and hypothesis test, the probability of a Type II error should ideally be small, and thus, the power will be large (e.g., $P(\text{Type II error})=0.20$ and $\text{power}=0.80$). The probability of a Type II error decreases and the power increases when any one of the following occurs, provided the others do not change: i. Sample size(s) increases. ii. Standard error decreases. iii. True parameter value is farther from the null hypothesis. iv. Significance level (α) of a test increases.
3.8.D Interpret Type I and Type II errors.	**3.8.D.1** In some studies, making a Type I error may have more serious consequences than making a Type II error. In other studies, making a Type II error may have more serious consequences than making a Type I error. The consequences of each error should be considered prior to conducting the study. **3.8.D.2** Because the significance level, α, is the probability of making a Type I error, the consequences of a Type I error influence decisions about a significance level. **3.8.D.3** Because sample size influences the probability of making a Type II error, the consequences of a Type II error influence decisions about how large the sample size should be.

TOPIC 3.9

Sampling Distributions for the Difference Between Sample Proportions

LEARNING OBJECTIVE

3.9.A

Calculate the mean and standard deviation of the sampling distribution for the difference between two sample proportions.

ESSENTIAL KNOWLEDGE

3.9.A.1

For two independent populations, with population proportions p_1 and p_2, when the sampled values are independent, the sampling distribution for the difference in sample proportions, $\hat{p}_1 - \hat{p}_2$, has a mean, $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$, and standard deviation, $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.

3.9.B

Justify the appropriateness of conditions for the sampling distribution for the difference between two sample proportions.

3.9.B.1

When sampling without replacement, two conditions must be met:

i. The randomization condition—the data should be collected using two independent random samples.

ii. The 10% condition—both samples must be less than 10% of the size of their respective populations.

3.9.B.2

If the data come from an experiment, the data only need to meet the randomization condition. The treatments must be randomly assigned to the experimental units to meet the randomization condition.

3.9.B.3

The sampling distribution for the difference between sample proportions, $\hat{p}_1 - \hat{p}_2$, will have an approximately normal distribution provided both sample sizes are large enough. To ensure that both samples are large enough, the data must meet the following conditions: $n_1 p_1 \geq 10$, $n_1(1 - p_1) \geq 10$, $n_2 p_2 \geq 10$, and $n_2(1 - p_2) \geq 10$, where $n_1 p_1$ and $n_2 p_2$ are the expected number of successes and $n_1(1 - p_1)$ and $n_2(1 - p_2)$ are the expected number of failures.

3.9.C

Interpret the mean, standard deviation, and probabilities for the sampling distribution for the difference between two sample proportions.

3.9.C.1

The mean, standard deviation, and probabilities for the sampling distribution for the difference between two sample proportions should be interpreted within the context of two specific populations.

TOPIC 3.10

Constructing a Confidence Interval for the Difference Between Two Population Proportions

LEARNING OBJECTIVE

3.10.A

Identify an appropriate confidence interval procedure including the parameters for the difference between two population proportions.

ESSENTIAL KNOWLEDGE

3.10.A.1

Based on the sample data, a confidence interval can be calculated to estimate the difference between two population proportions. The appropriate confidence interval procedure is a two-sample *z*-interval for a difference between population proportions.

3.10.A.2

The parameters of a confidence interval for the difference between two population proportions should refer to the difference in the proportions, the response variable, and the populations in context.

3.10.B

Justify the appropriateness of constructing a confidence interval for the difference between two population proportions by verifying conditions.

3.10.B.1

A two-sample *z*-interval for a difference between two population proportions requires that three conditions be met:

i. The randomization condition—the data should be collected using two independent random samples or a randomized experiment.

ii. The 10% condition—when sampling without replacement, check that $n_1 < 10\% N_1$ and $n_2 < 10\% N_2$, where N_1 is the size of population 1 and N_2 is the size of population 2. The sample sizes are represented as n_1 and n_2. (Note: This condition is unnecessary when the data are from a randomized experiment.)

iii. The normality condition—the number of successes, $n_1\hat{p}_1$ and $n_2\hat{p}_2$, and number of failures, $n_1\left(1-\hat{p}_1\right)$ and $n_2\left(1-\hat{p}_2\right)$, for both samples are all at least 10.

3.10.C

Calculate an appropriate confidence interval for the difference between two population proportions.

3.10.C.1

The point estimate for the difference between two sample proportions is $\left(\hat{p}_1-\hat{p}_2\right)$.

3.10.C.2

For the difference between two population proportions, the interval estimate can be constructed as $\text{point estimate} \pm \left(\text{margin of error}\right)$. The interval estimate for the difference between two population proportions is

$$\left(\hat{p}_1-\hat{p}_2\right) \pm z^* \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1}+\frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}.$$

LEARNING OBJECTIVE

3.10.D

Calculate the standard error and margin of error for estimating the difference between two population proportions.

ESSENTIAL KNOWLEDGE

3.10.D.1

The standard error (*SE*) for the difference between two population proportions is $\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1}+\frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$.

3.10.D.2

For the difference between two population proportions, the margin of error is the critical value (z^*) times the standard error (*SE*) of the difference between the two proportions, which equals $z^*\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1}+\frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$.

TOPIC 3.11

Justifying a Claim Based on a Confidence Interval for the Difference Between Two Population Proportions

LEARNING OBJECTIVE

3.11.A

Interpret a confidence interval in context for the difference between two population proportions.

ESSENTIAL KNOWLEDGE

3.11.A.1

Because the confidence interval for the difference between two population proportions is calculated based on samples from two populations, the computed interval may or may not contain the value for the difference between those two population proportions.

3.11.A.2

The interpretation of the confidence level is as follows: In repeated random sampling with the same sample sizes from the same populations, approximately *C* % of confidence intervals created will capture the difference between the two population proportions, where *C* represents the numerical value of the confidence level used.

3.11.A.3

When interpreting a *C* % confidence interval for the difference between two population proportions, we say we are *C* % confident that the interval (a, b) contains the parameter for the difference between the populations, where *a* represents the lower limit and *b* represents the upper limit. An interpretation of a confidence interval for the difference between two population proportions includes a reference to the parameter with the details about the populations it represents in the context of the study.

LEARNING OBJECTIVE

3.11.B

Justify a claim based on a confidence interval for the difference between two population proportions.

ESSENTIAL KNOWLEDGE

3.11.B.1

A confidence interval for the difference between two population proportions provides an interval of values that may provide convincing evidence to support a particular claim about the difference between the two population proportions. For example, if the interval contains 0, then there is insufficient evidence to conclude there is a difference between the two population proportions. If the interval does not contain 0, there is sufficient evidence to conclude there is a difference between the two population proportions.

TOPIC 3.12

Setting Up a Test for the Difference Between Two Population Proportions

LEARNING OBJECTIVE

3.12.A

Identify an appropriate testing method for the difference between population proportions including the parameters.

ESSENTIAL KNOWLEDGE

3.12.A.1

The appropriate testing method for the difference between two population proportions is a two-sample *z*-test for the difference between two population proportions.

3.12.A.2

The parameters for a hypothesis test for the difference between two population proportions should reference the population parameters, the response variables, and the populations in context.

3.12.B

Identify the null and alternative hypotheses for the difference between population proportions.

3.12.B.1

For a two-sample *z*-test for the difference between two population proportions, the null hypothesis indicates no difference. The null hypothesis for the difference between two population proportions can be written as either $\mathrm{H}_0: p_1 = p_2$ or $\mathrm{H}_0: p_1 - p_2 = 0$. A one-sided alternative hypothesis for the difference between two population proportions can be written as either $\mathrm{H}_\mathrm{a}: p_1 < p_2$ or equivalently $\mathrm{H}_\mathrm{a}: p_1 - p_2 < 0$, or $\mathrm{H}_\mathrm{a}: p_1 > p_2$ or equivalently $\mathrm{H}_\mathrm{a}: p_1 - p_2 > 0$. A two-sided alternative hypothesis for the difference between two population proportions can be written as either $\mathrm{H}_\mathrm{a}: p_1 \neq p_2$ or equivalently $\mathrm{H}_\mathrm{a}: p_1 - p_2 \neq 0$.

3.12.C

Justify the appropriateness of a hypothesis test for the difference between two population proportions by verifying conditions.

3.12.C.1

A two-sample *z*-test for a difference between two population proportions requires that three conditions be met:

i. The randomization condition—the data should be collected using two independent random samples or a randomized experiment.

ii. The 10% condition—when sampling without replacement, check that $n_1 < 10\% N_1$ and $n_2 < 10\% N_2$, where N_1 is the size of population 1 and N_2 is the size of population 2. The sample sizes are represented as n_1 and n_2. (Note: This condition is unnecessary when the data are from a randomized experiment.)

iii. The normality condition—$n_1\hat{p}_c$, $n_1\left(1-\hat{p}_c\right)$, $n_2\hat{p}_c$, and $n_2\left(1-\hat{p}_c\right)$ must all be at least 10, with $\hat{p}_c = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ being the combined (or pooled) proportion assuming that H_0 is true $\left(\mathrm{H}_0: p_1 = p_2 \text{ or } \mathrm{H}_0: p_1 - p_2 = 0\right)$.

TOPIC 3.13

Carrying Out a Test for the Difference Between Two Population Proportions

LEARNING OBJECTIVE

3.13.A

Calculate an appropriate test statistic and *p*-value for testing a hypothesis for the difference between two population proportions.

ESSENTIAL KNOWLEDGE

3.13.A.1

The test statistic for the difference between two population proportions is $z = \frac{\left(\hat{p}_1 - \hat{p}_2\right) - 0}{\sqrt{\hat{p}_c\left(1-\hat{p}_c\right)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, where $\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$ is the proportion of successes for the two groups combined. The *z*-statistic has a standard normal distribution when the null hypothesis is true.

3.13.A.2

The *p*-value for a two-sample *z*-test for the difference between two population proportions can be found from the standard normal distribution using a table or technology.

3.13.B

Interpret the *p*-value of a hypothesis test for the difference between two population proportions.

3.13.B.1

The *p*-value is the probability of obtaining a test statistic as extreme or more extreme than the test statistic that was observed (i.e., in the direction of the alternative hypothesis) given that the null hypothesis is true. An interpretation of the *p*-value of a hypothesis test for a difference between two population proportions should include a statement that the *p*-value is computed assuming the null hypothesis is true (i.e., by assuming that the true population proportions are equal to each other in context).

3.13.C

Justify a claim about the populations based on the results of a hypothesis test for the difference between two population proportions.

3.13.C.1

A formal decision in a hypothesis test for the difference between two population proportions explicitly compares the *p*-value to the significance level, α. If the $p\text{-value} \leq \alpha$, then reject the null hypothesis, $\mathrm{H}_0: p_1 = p_2$ or $\mathrm{H}_0: p_1 - p_2 = 0$. If the $p\text{-value} > \alpha$, then fail to reject the null hypothesis.

3.13.C.2

The results of a hypothesis test for the difference between two population proportions can serve as the statistical reasoning to support the answer to an investigative question about the two populations that were sampled.

3.13.C.3

A conclusion for the hypothesis test for the difference between two population proportions is stated in context consistent with, and in terms of, the alternative hypothesis using non-definitive language. The conclusion should contain a reference to the parameters and the populations.

TOPIC 3.14

Setting Up a Chi-Square Test for Homogeneity or Independence

LEARNING OBJECTIVE

3.14.A

Describe chi-square distributions.

ESSENTIAL KNOWLEDGE

3.14.A.1

The chi-square statistic measures the distance between observed and expected counts relative to expected counts.

3.14.A.2

Chi-square distributions have positive values and are skewed right. Within this family of density curves, the skew becomes less pronounced with increasing degrees of freedom.

3.14.B

Identify an appropriate testing method for comparing distributions in two-way tables of categorical data including the populations and variables.

3.14.B.1

To determine whether the distributions of a categorical variable for two or more populations are different, the appropriate test is the chi-square test for homogeneity.

3.14.B.2

A chi-square test for homogeneity should reference the categorical variable and the populations in context.

3.14.B.3

To determine whether row and column variables in a two-way table of categorical data might be associated in the single population from which the data were sampled, the appropriate test is the chi-square test for independence.

3.14.B.4

A chi-square test for independence should reference the categorical variables and the population in context.

3.14.C

Identify the null and alternative hypotheses for a chi-square test for homogeneity or independence.

3.14.C.1

The appropriate null hypothesis for a chi-square test for homogeneity is H_0: there is no difference in the distributions of the categorical variable across populations or treatments. The appropriate alternative hypothesis for a chi-square test for homogeneity is H_a: there is a difference in the distributions of the categorical variable across populations or treatments.

3.14.C.2

The appropriate null hypothesis for a chi-square test for independence is H_0: there is no association between two categorical variables in a given population or the two categorical variables in a given population are independent of each other. The appropriate alternative hypothesis for a chi-square test for independence is H_a: there is an association between two categorical variables in a given population or the two categorical variables in a given population are not independent of each other.

LEARNING OBJECTIVE

3.14.D

Justify the appropriateness of a hypothesis test for a chi-square distribution for independence or homogeneity by verifying conditions.

ESSENTIAL KNOWLEDGE

3.14.D.1

A chi-square test for homogeneity or independence requires that three conditions must be met:

i. The randomization condition—the test of independence states that the data should be collected using an independent random sample. The test for homogeneity states that the data should be collected using independent random samples or a randomized experiment.

ii. The 10% condition—when sampling without replacement, check that $n<10\%N$, where *N* is the size of the population and *n* is the sample size. (Note: This condition is unnecessary when the data are from a randomized experiment.)

iii. The expected values condition—all expected values should be greater than 5.

TOPIC 3.15

Carrying Out a Chi-Square Test for Homogeneity or Independence

LEARNING OBJECTIVE

3.15.A

Calculate expected counts for two-way tables of categorical data.

ESSENTIAL KNOWLEDGE

3.15.A.1

The expected values (under the null hypothesis) in a particular cell of a two-way table of categorical data can be calculated using the formula

$$\text{expected value} = \frac{(\text{row total})(\text{column total})}{\text{total table}}.$$

3.15.B

Calculate the appropriate test statistic and *p*-value for a chi-square test for homogeneity or independence.

3.15.B.1

The appropriate test statistic for a chi-square test for homogeneity or independence is the chi-square statistic

$$\chi^2 = \sum \frac{(\text{Observed Count} - \text{Expected Count})^2}{\text{Expected Count}},$$ where the sum is taken over all cells of the two-way table. The chi-square statistics have a chi-square distribution with degrees of freedom equal to $(\text{number of rows} - 1) \cdot (\text{number of columns} - 1)$ when the null hypothesis is true.

3.15.B.2

The *p*-value for a chi-square test for independence or homogeneity is found from a chi-square distribution using a table or technology.

3.15.C

Interpret the *p*-value for the chi-square test for homogeneity or independence.

3.15.C.1

The *p*-value is the probability of obtaining a test statistic as extreme or more extreme than the test statistic that was observed (i.e., in the direction of the alternative hypothesis) given that the null hypothesis is true. An interpretation of the *p*-value for the chi-square test for homogeneity or independence should include a statement that the *p*-value is computed by assuming the null hypothesis is true in context.

3.15.D

Justify a claim about the population based on the results of a chi-square test for homogeneity or independence.

3.15.D.1

A formal decision in a hypothesis test explicitly compares the *p*-value to the significance level, α. If the $p\text{-value} \leq \alpha$, then reject the null hypothesis for the appropriate chi-square test. If the $p\text{-value} > \alpha$, then fail to reject the null hypothesis.

3.15.D.2

The results of a chi-square test for homogeneity or independence can serve as the statistical reasoning to support the answer to an investigative question about the population that was sampled (independence) or the populations that were sampled (homogeneity).

LEARNING OBJECTIVE

3.15.D

Justify a claim about the population based on the results of a chi-square test for homogeneity or independence.

ESSENTIAL KNOWLEDGE

3.15.D.3

A conclusion for a chi-square test for homogeneity or independence is stated in context consistent with, and in terms of, the alternative hypothesis using non-definitive language. The conclusion should contain a reference to the population(s).

THIS PAGE IS INTENTIONALLY LEFT BLANK.

AP STATISTICS

UNIT 4

Inference for Quantitative Data: Means

TOPIC 4.1

Sampling Distributions for Sample Means

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
4.1.A Calculate the mean and standard deviation of a sampling distribution of a sample mean.	**4.1.A.1** For a population with population mean μ and population standard deviation σ, when the sampled values are independent, the sampling distribution of the sample mean has mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
4.1.B Justify the appropriateness of conditions for the sampling distribution of a sample mean.	**4.1.B.1** Sampling without replacement requires that two conditions must be met: i. The randomization condition—the data should be collected using a random sample. ii. The 10% condition—the population size must be at least 10 times larger than the sample size $(n < 10\%N)$, where *N* is the size of the population and *n* is the sample size. **4.1.B.2** For a quantitative variable, if the population distribution can be modeled by a normal distribution, the sampling distribution of the sample mean, $\bar{x}$, can be modeled with a normal distribution regardless of the sample size. **4.1.B.3** For a quantitative variable, if the population distribution cannot be modeled by a normal distribution, the sampling distribution of the sample mean, $\bar{x}$, can be modeled approximately by a normal distribution, provided $n \geq 30$. If the population distribution is extremely skewed, a sample size much larger than 30 may be needed to ensure the sampling distribution is approximately normal.
4.1.C Interpret the mean, standard deviation, and probabilities for the sampling distribution of a sample mean.	**4.1.C.1** The mean, standard deviation, and probabilities for a sampling distribution of a sample mean should be interpreted within the context of a specific population.

TOPIC 4.2

Constructing a Confidence Interval for a Population Mean or Population Mean Difference

LEARNING OBJECTIVE

4.2.A

Describe *t*-distributions.

ESSENTIAL KNOWLEDGE

4.2.A.1

t-distributions, also called Student's *t*-distributions, form a family of symmetric, bell-shaped, standardized distributions with wider tails than that of the standard normal distribution. Specific *t*-distributions are identified using a parameter known as the number of degrees of freedom (*df*), which is based on the sample size(s). When the degrees of freedom are small, the *t*-distribution has a much narrower peak and fatter tails than a normal distribution. As the degrees of freedom increase, the *t*-distribution more closely resembles the standard normal distribution (mean $\mu = 0$ and standard deviation $\sigma = 1$).

4.2.A.2

t-distributions are used for finding critical values and test statistics for inferences about a population mean, μ, when the population standard deviation, σ, is unknown and the sample standard deviation, *s*, must be used instead.

4.2.B

Identify an appropriate confidence interval procedure including the parameter for a population mean or population mean difference.

4.2.B.1

The appropriate confidence interval procedure for estimating the population mean of a quantitative variable for one sample is a one-sample *t*-interval for a population mean. (The population standard deviation, σ, is not typically known for distributions for quantitative variables.)

4.2.B.2

For a matched pairs design with two dependent samples, the appropriate analysis calculates differences between pairs of values to produce one sample of differences. The confidence interval procedure for the matched pairs design is a one-sample *t*-interval for a population mean difference.

4.2.B.3

The parameter for a confidence interval for a population mean or population mean difference should reference the population mean or population mean difference and the response variable, in context. For the population mean difference, it is important to state the order of subtraction for the difference.

LEARNING OBJECTIVE

ESSENTIAL KNOWLEDGE

4.2.C

Justify the appropriateness of constructing a confidence interval for a population mean or population mean difference by verifying conditions.

4.2.C.1

A one-sample *t*-interval for a population mean or population mean difference requires that three conditions be met:

i. The randomization condition—the data should be collected using a random sample or a randomized experiment.

ii. The 10% condition—when sampling without replacement, check that $n < 10\%N$, where *N* is the size of the population and *n* is the sample size.

iii. The sample data condition—it is indicated the population distribution is approximately normal, or $n \geq 30$, or if $n < 30$, the sample data distribution should be free from strong skewness and outliers. For matched pairs, the number of differences should be greater than or equal to 30. If the number of differences is less than 30, the sample of differences should be free from strong skewness and outliers.

4.2.D

Calculate an appropriate confidence interval for a population mean or population mean difference.

4.2.D.1

A point estimate for a population mean is the sample mean, $\bar{x}$, or $\bar{x}_d$ for the sample mean difference.

4.2.D.2

To estimate the population mean for one sample or the population mean difference between values in matched pairs, when the population standard deviation is unknown, the confidence interval is $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$, where t^* is the critical value for the central *C* % of a *t*-distribution with degrees of freedom $n-1$.

4.2.E

Calculate the standard error and margin of error for a sample size for a one-sample *t*-interval.

4.2.E.1

The standard error (*SE*) for a sample mean is given by $s_{\bar{x}} = \frac{s}{\sqrt{n}}$.

4.2.E.2

For a one-sample *t*-interval for a population mean, the margin of error is the critical value (t^*) times the standard error (*SE*), which equals $(t^*)\left(\frac{s}{\sqrt{n}}\right)$.

TOPIC 4.3

Justifying a Claim Based on a Confidence Interval for a Population Mean or Population Mean Difference

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
4.3.A Interpret a confidence interval in context for a population mean or population mean difference.	**4.3.A.1** Because the confidence interval for a population mean or population mean difference is calculated based on a sample from a population, the computed interval may or may not contain the value of the population mean or population mean difference. **4.3.A.2** The interpretation of the confidence level is as follows: In repeated random sampling with the same sample size from the same population, approximately *C* % of confidence intervals created will capture the population mean or population mean difference, where *C* represents the numerical value of the confidence level used. **4.3.A.3** When interpreting a *C* % confidence interval for a population mean or population mean difference, we say we are *C* % confident the interval (a, b) contains the value of the population mean or population mean difference, where *a* represents the lower limit and *b* represents the upper limit. An interpretation of a confidence interval for a population mean or population mean difference includes a reference to the parameter.
4.3.B Justify a claim based on a confidence interval for a population mean or population mean difference.	**4.3.B.1** A confidence interval for a population mean or population mean difference provides an interval of values that may serve as convincing evidence to support a particular claim about the population mean or population mean difference.
4.3.C Identify the relationships among sample size, confidence interval width, confidence level, and margin of error for a population mean or population mean difference.	**4.3.C.1** For a given sample, increasing the confidence level will result in the following: i. The critical value will increase. ii. The margin of error will increase. iii. The width of the confidence interval will increase.

LEARNING OBJECTIVE

4.3.C

Identify the relationships among sample size, confidence interval width, confidence level, and margin of error for a population mean or population mean difference.

ESSENTIAL KNOWLEDGE

4.3.C.2

Increasing the sample size decreases the standard error. Thus, when all other things remain the same, the width of a confidence interval for a population mean or population mean difference tends to decrease as the sample size increases. For a confidence interval for a population mean or population mean difference with a given confidence level, the width of the interval is approximately proportional to $\frac{1}{\sqrt{n}}$.

TOPIC 4.4

Setting Up a Test for a Population Mean or Population Mean Difference

LEARNING OBJECTIVE

4.4.A

Identify an appropriate testing method and parameter for a population mean or population mean difference with unknown σ.

ESSENTIAL KNOWLEDGE

4.4.A.1

The appropriate test for a population mean with unknown population standard deviation σ is a one-sample *t*-test for a population mean.

4.4.A.2

For a matched pairs design with two dependent samples, the appropriate analysis calculates differences between pairs of values to produce one sample of differences. The hypothesis testing procedure for the matched pairs design is a one-sample *t*-test for the population mean difference.

4.4.A.3

The parameter for a hypothesis test for a population mean and population mean difference should reference the population parameter, the response variable, and the population in context.

4.4.B

Identify the null and alternative hypotheses for a population mean or population mean difference with unknown σ.

4.4.B.1

The null hypothesis for a one-sample *t*-test for a population mean is $\mathrm{H}_0: \mu = \mu_0$ in which μ_0 is the null hypothesized value for the population mean. A one-sided alternative hypothesis for a one-sample *t*-test for a population mean is either $\mathrm{H}_a: \mu < \mu_0$ or $\mathrm{H}_a: \mu > \mu_0$. A two-sided alternative hypothesis is $\mathrm{H}_a: \mu \neq \mu_0$.

4.4.B.2

The null hypothesis for a population mean difference is $\mathrm{H}_0: \mu_d = 0$. A one-sided alternative hypothesis for a population mean difference is either $\mathrm{H}_a: \mu_d < 0$ or $\mathrm{H}_a: \mu_d > 0$. A two-sided alternative hypothesis is $\mathrm{H}_a: \mu_d \neq 0$.

4.4.C

Justify the appropriateness of a hypothesis test for a population mean or population mean difference by verifying conditions.

4.4.C.1

A one-sample *t*-test for a population mean or a population mean difference requires that three conditions be met:

i. The randomization condition—the data should be collected using a random sample or a randomized experiment.

ii. The 10% condition—when sampling without replacement, check that $n < 10\% N$, where *N* is the size of the population and *n* is the sample size.

iii. The sample data condition—it is indicated the population distribution is approximately normal, or $n \geq 30$, or if $n < 30$, the sample data distribution should be free from strong skewness and outliers. For matched pairs, the number of differences should be greater than or equal to 30. If the number of differences is less than 30, the sample of differences should be free from strong skewness and outliers.

TOPIC 4.5

Carrying Out a Test for a Population Mean or Population Mean Difference

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
4.5.A Calculate an appropriate test statistic and *p*-value for testing a hypothesis about a population mean or population mean difference.	**4.5.A.1** The test statistic for a one-sample *t*-test for a population mean or population mean difference is $t=\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$, where *t* has degrees of freedom $n-1$. The *t*-statistic has a *t*-distribution with degrees of freedom $n-1$ when the null hypothesis is true. **4.5.A.2** The *p*-value for a one-sample *t*-test for a population mean or population mean difference is found using the appropriate *t*-distribution table or technology.
4.5.B Interpret the *p*-value of a hypothesis test for a population mean or population mean difference.	**4.5.B.1** The *p*-value is the probability of obtaining a test statistic as extreme or more extreme than the test statistic that was observed (i.e., in the direction of the alternative hypothesis) given that the null hypothesis is true. An interpretation of the *p*-value of a hypothesis test for a population mean or population mean difference should include a statement that the *p*-value is computed by assuming that the null hypothesis is true (i.e., by assuming that the population mean is equal to the particular value stated in the null hypothesis in context).
4.5.C Justify a claim about the population based on the results of a hypothesis test for a population mean or population mean difference.	**4.5.C.1** A formal decision explicitly compares the *p*-value to the significance level, α. If the $p\text{-value} \le \alpha$, then reject the null hypothesis, $\mathrm{H}_0: \mu = \mu_0$. If the $p\text{-value} > \alpha$, then fail to reject the null hypothesis. **4.5.C.2** The results of a hypothesis test for a population mean or population mean difference can serve as the statistical reasoning to support the answer to an investigative question about the population that was sampled. **4.5.C.3** A conclusion for the hypothesis test for a population mean or population mean difference is stated in context consistent with, and in terms of, the alternative hypothesis using non-definitive language. The conclusion should contain a reference to the parameter and the population.

TOPIC 4.6

Sampling Distributions for the Difference Between Two Sample Means

LEARNING OBJECTIVE

4.6.A

Calculate the mean and standard deviation of a sampling distribution for the difference between two sample means.

ESSENTIAL KNOWLEDGE

4.6.A.1

For two independent populations with population means μ_1 and μ_2 and population standard deviations σ_1 and σ_2, when the sampled values are independent, the sampling distribution of the difference in sample means $\bar{x}_1 - \bar{x}_2$ has a mean $\mu_{(\bar{x}_1-\bar{x}_2)} = \mu_1 - \mu_2$ and standard deviation $\sigma_{(\bar{x}_1-\bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

4.6.B

Justify the appropriateness of conditions for the sampling distribution of the difference between two sample means.

4.6.B.1

Sampling without replacement requires that two conditions must be met:

i. The randomization condition—the data should be collected using two independent random samples.

ii. The 10% condition—both samples must be less than 10% of the size of their respective populations.

4.6.B.2

If the data come from an experiment, the data only need to meet the randomization condition. The treatments must be randomly assigned to experimental units to meet the randomization condition.

4.6.B.3

The sampling distribution for the difference between sample means, $\bar{x}_1 - \bar{x}_2$, can be modeled with a normal distribution if the two population distributions can each be modeled by a normal distribution.

4.6.B.4

The sampling distribution for the difference between sample means, $\bar{x}_1 - \bar{x}_2$, can be modeled approximately by a normal distribution if the two population distributions cannot be modeled by a normal distribution but $n_1 \geq 30$ and $n_2 \geq 30$.

4.6.C

Interpret the mean, standard deviation, and probabilities for a sampling distribution for the difference between sample means.

4.6.C.1

The mean, standard deviation, and probabilities for a sampling distribution for the difference between sample means should be interpreted within the context of specific populations.

TOPIC 4.7

Constructing a Confidence Interval for the Difference Between Two Population Means

LEARNING OBJECTIVE

4.7.A

Identify an appropriate confidence interval procedure including the parameter for the difference between two population means.

ESSENTIAL KNOWLEDGE

4.7.A.1

Based on the sample data, a confidence interval can be calculated to estimate the difference between two independent population means. The appropriate confidence interval procedure for two independent samples is a two-sample *t*-interval for the difference between population means.

4.7.A.2

The parameter for a confidence interval for a two-sample *t*-interval for the difference between population means should reference the difference in the means, the response variable, and the populations in context.

4.7.B

Justify the appropriateness of constructing a confidence interval for the difference between two population means by verifying conditions.

4.7.B.1

A two-sample *t*-interval for a difference between population means requires that three conditions be met:

i. The randomization condition—the data should be collected using two independent random samples or a randomized experiment.

ii. The 10% condition—when sampling without replacement, the size of each sample should be less than or equal to 10% of the respective population size: $n_1 < 10\% N_1$ and $n_2 < 10\% N_2$, where N_1 is the size of population 1 and N_2 is the size of population 2. The sample sizes are represented as n_1 and n_2. (Note: This condition is unnecessary when the data are from a randomized experiment.)

iii. The sample data condition—both samples should have a sample size greater than or equal to 30 or it is indicated that both population distributions are approximately normal. If either sample size is less than 30, both sample data distributions should be free from strong skewness and outliers.

4.7.C

Calculate an appropriate confidence interval for the difference between two population means.

4.7.C.1

A point estimate for the difference between two population means is the difference in sample means, $\bar{x}_1 - \bar{x}_2$.

4.7.C.2

For the difference between population means when the population standard deviations are unknown, the confidence interval can be constructed as $\text{point estimate} \pm (\text{margin of error})$. The confidence interval for the difference between population means is $\left(\bar{x}_1 - \bar{x}_2\right) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, where t^* are the critical values for the central *C* % of a *t*-distribution with appropriate degrees of freedom that can be found using technology. The degrees of freedom fall between $n_1 + n_2 - 2$ and the smaller of $n_1 - 1$ and $n_2 - 1$.

LEARNING OBJECTIVE

4.7.D

Calculate the standard error and margin of error for estimating the difference between two population means.

ESSENTIAL KNOWLEDGE

4.7.D.1

The standard error for the difference between two sample means is $\sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}}$, where s_1 and s_2 are the sample standard deviations.

4.7.D.2

For the difference between two sample means, the margin of error is the critical value (t^*) times the standard error (SE) of the difference of two sample means, which equals $t^*\sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}}$.

TOPIC 4.8

Justifying a Claim Based on a Confidence Interval for the Difference Between Two Population Means

LEARNING OBJECTIVE

4.8.A

Interpret a confidence interval in context for the difference between two population means.

ESSENTIAL KNOWLEDGE

4.8.A.1

Because the confidence interval for the difference between two population means is calculated based on samples from two populations, the computed interval may or may not contain the value for the difference between the two population means.

4.8.A.2

The interpretation of the confidence level is as follows: In repeated random sampling with the same sample size from the same populations, approximately *C* % of confidence intervals created will capture the difference between the two population means, where *C* represents the numerical value of the confidence level used.

4.8.A.3

When interpreting a *C* % confidence interval for the difference between two population means, we say we are *C* % confident that the interval (a, b) contains the value of the difference in the population means, where *a* represents the lower limit and *b* represents the upper limit. An interpretation of a confidence interval for the difference between two population means includes a reference to the difference in the population means with the details about the populations it represents in the context of the study.

4.8.B

Justify a claim based on a confidence interval for the difference between two population means.

4.8.B.1

A confidence interval for the difference between two population means provides an interval of values that may serve as convincing evidence to support a particular claim about the difference in two population means. For example, if the interval contains 0, then there is insufficient evidence to conclude there is a difference between the two population means. If the interval does not contain 0, then there is sufficient evidence to conclude there is a difference between the two population means.

TOPIC 4.9

Setting Up a Test for the Difference Between Two Population Means

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
4.9.A Identify an appropriate testing method for the difference between two population means including the parameters for the difference between the two population means.	**4.9.A.1** The appropriate test for the difference between two population means is a two-sample *t*-test for a difference between two population means. **4.9.A.2** The parameter for a hypothesis test for the difference between two population means should reference the population parameters, the response variables, and the populations in context.
4.9.B Identify the null and alternative hypotheses for the difference between two population means.	**4.9.B.1** The null hypothesis for a two-sample *t*-test for the difference between two population means, μ_1 and μ_2, can be written as either $\mathrm{H}_0: \mu_1 - \mu_2 = 0$ or $\mathrm{H}_0: \mu_1 = \mu_2$. A one-sided alternative hypothesis for the difference between population means can be written as either $\mathrm{H}_\mathrm{a}: \mu_1 < \mu_2$ (or equivalently $\mathrm{H}_\mathrm{a}: \mu_1 - \mu_2 < 0$) or $\mathrm{H}_\mathrm{a}: \mu_1 > \mu_2$ (or equivalently $\mathrm{H}_\mathrm{a}: \mu_1 - \mu_2 > 0$). A two-sided alternative hypothesis for the difference between population means can be written as $\mathrm{H}_\mathrm{a}: \mu_1 \neq \mu_2$ (or equivalently $\mathrm{H}_\mathrm{a}: \mu_1 - \mu_2 \neq 0$).
4.9.C Justify the appropriateness of a hypothesis test for the difference between two population means by verifying conditions.	**4.9.C.1** A two-sample *t*-test for a difference between population means requires that three conditions be met: i. The randomization condition—the data should be collected using two independent random samples or a randomized experiment. ii. The 10% condition—when sampling without replacement, the size of each sample should be less than or equal to 10% of the respective population size: $n_1 < 10\% N_1$ and $n_2 < 10\% N_2$, where N_1 is the size of population 1 and N_2 is the size of population 2. The sample sizes are represented as n_1 and n_2. (Note: This condition is unnecessary when the data are from a randomized experiment.) iii. The sample data condition—both samples should have a sample size greater than or equal to 30 or it is indicated that both population distributions are approximately normal. If either sample size is less than 30, both sample data distributions should be free from strong skewness and outliers.

TOPIC 4.10

Carrying Out a Test for the Difference Between Two Population Means

LEARNING OBJECTIVE

4.10.A

Calculate an appropriate test statistic and *p*-value for testing a hypothesis for the difference between two population means.

ESSENTIAL KNOWLEDGE

4.10.A.1

The test statistic for a two-sample *t*-test for the difference between two population means is $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\dfrac{s_1^{\,2}}{n_1} + \dfrac{s_2^{\,2}}{n_2}}}$. The *t*-statistic has a *t*-distribution when the null hypothesis is true. The *t*-statistics with the degrees of freedom can be found using technology. The degrees of freedom fall between $n_1 + n_2 - 2$ and the smaller of $n_1 - 1$ and $n_2 - 1$.

4.10.A.2

The *p*-value for a two-sample *t*-test for the difference between two population means can be found using the appropriate *t*-distribution table or from the appropriate *t*-distribution using technology.

LEARNING OBJECTIVE

4.10.B

Interpret the *p*-value of a hypothesis test for the difference between two population means.

ESSENTIAL KNOWLEDGE

4.10.B.1

The *p*-value is the probability of obtaining a test statistic as extreme or more extreme than the test statistic that was observed (i.e., in the direction of the alternative hypothesis) given that the null hypothesis is true. An interpretation of the *p*-value of a hypothesis test for a two-sample test for the difference between two population means should include a statement that the *p*-value is computed by assuming that the null hypothesis is true (i.e., by assuming the population means are equal to each other in context).

LEARNING OBJECTIVE

4.10.C

Justify a claim about the populations based on the results of a hypothesis test for the difference between two population means.

ESSENTIAL KNOWLEDGE

4.10.C.1

A formal decision explicitly compares the *p*-value to the significance level, α. If the $p\text{-value} \le \alpha$, then reject the null hypothesis, $\mathrm{H}_0: \mu_1 - \mu_2 = 0$ or $\mathrm{H}_0: \mu_1 = \mu_2$. If the $p\text{-value} > \alpha$, then fail to reject the null hypothesis.

4.10.C.2

The results of a hypothesis test for a two-sample *t*-test for a difference between two population means can serve as the statistical reasoning to support the answer to an investigative question about the two populations that were sampled.

4.10.C.3

A conclusion for the hypothesis test for the difference between two population means is stated in context consistent with, and in terms of, the alternative hypothesis using non-definitive language. The conclusion should contain a reference to the parameters and the populations.

THIS PAGE IS INTENTIONALLY LEFT BLANK.

AP STATISTICS

UNIT 5

Regression Analysis

THIS PAGE IS INTENTIONALLY LEFT BLANK.

TOPIC 5.1

Graphical Representations Between Two Quantitative Variables

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
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 *x*-axis and one corresponding to the value on the *y*-axis. The explanatory variable is placed on the *x*-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 *y*-axis.
5.1.B Describe the characteristics of a scatterplot.	**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 Justify a claim using scatterplots depicting the distribution of two quantitative variables.	**5.1.C.1** Scatterplots depicting the distribution of two numeric variables may reveal information that can be used to justify claims about the variable in context.

TOPIC 5.2
Correlation

LEARNING OBJECTIVE

5.2.A

Interpret the correlation for a linear relationship.

ESSENTIAL KNOWLEDGE

5.2.A.1

The correlation coefficient, *r*, summarizes the strength and direction of the linear association between two quantitative variables. The correlation coefficient *r* 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.

TOPIC 5.3

Linear Regression Models

LEARNING OBJECTIVE

5.3.A

Calculate a predicted response value using a linear regression model.

ESSENTIAL KNOWLEDGE

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.

TOPIC 5.4

Residuals

LEARNING OBJECTIVE	ESSENTIAL KNOWLEDGE
5.4.A Calculate the differences between the observed and predicted values.	**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: $\text{residual} = y - \hat{y}$ or $(\text{residual} = \text{observed } y - \text{predicted } y)$.
5.4.B Interpret the differences between the observed and predicted values.	**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 Describe the form of association of bivariate data using residual plots.	**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.

Return to Table of Contents

TOPIC 5.5

Least-Squares Regression

LEARNING OBJECTIVE

5.5.A

Calculate the coefficients for the least-squares regression line model.

ESSENTIAL KNOWLEDGE

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 (*LSRL*) 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 square of the correlation coefficient, r^2, is called the coefficient of determination. r^2 is the proportion of variation in the response variable that is explained by the linear relationship with the explanatory variable.

5.5.A.5

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

Interpret coefficients for the least-squares regression line model.

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 or decrease 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.

THIS PAGE IS INTENTIONALLY LEFT BLANK.