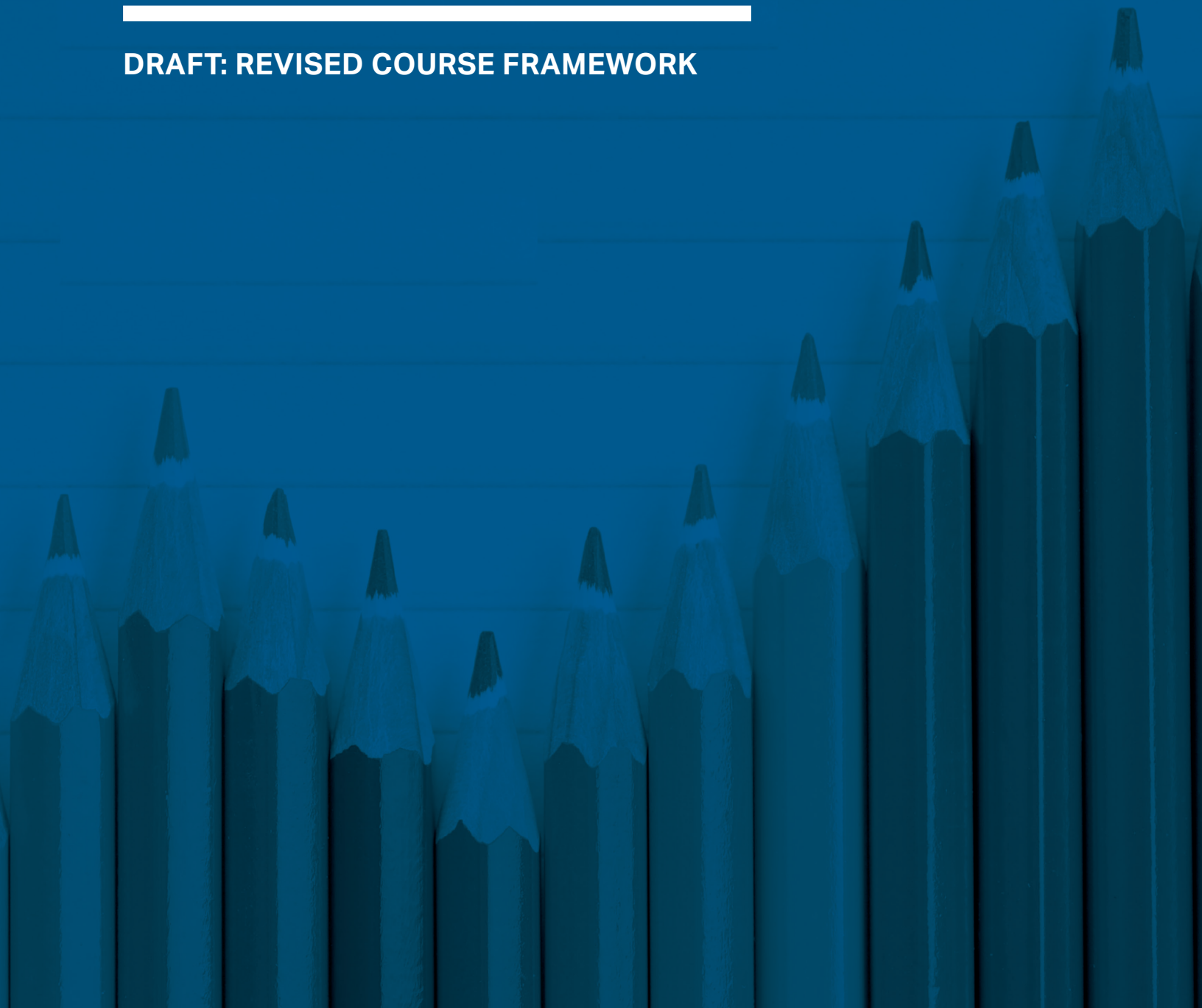
# AP Statistics

## DRAFT: REVISED COURSE FRAMEWORK



THIS PAGE IS INTENTIONALLY LEFT BLANK.

# Contents

# Acknowledgments

**Prince Afriyie,** *University of Virginia, VA*
**Javier Alvarez,** *Phillips Exeter Academy, NH*
**Barb Barnet,** *University of Wisconsin-Platteville, WI*
**Tim Bator,** *Sayre School, KY*
**Ellen Breazel,** *Clemson University, SC*
**Carol Chamberlain Hebert,** *Barnstable High School, MA*
**Holly Deal,** *Kennesaw State University, GA*
**Christine Franklin,** *University of Georgia, GA*
**Donna LaLonde,** *American Statistical Association, VA*
**Laura Marshall,** *Phillips Exeter Academy, NH*
**Bridget Matamoros-Mota,** *John H. Guyer High School, TX*
**Karen McGaughey,** *California Polytechnic State University, CA*
**Chris Olsen,** *Grinnell College, IA*
**Michael Posner,** *Villanova University, PA*
**Al Reiff,** *The Taft School, CT*
**Paul Roback,** *St. Olaf College, MN*
**Tom Rohnkohl,** *Silverado High School, NV*
**Allen Rossman,** *California Polytechnic State University, CA*
**Joshua Sawyer,** *Camden County Schools, NC*

## College Board Staff

**Camille Pace,** *Director, AP Statistics Curriculum and Assessment*
**Jason VanBilliard,** *Senior Director, AP Math and Computer Science Department Head*
**Dana Kopelman,** *Executive Director, AP Instruction and Assessment Production*
**Claire Lorenz,** *Senior Director, AP Classroom Instruction Products*
**Jason Manoharan,** *Vice President, AP Program Development*
**Daniel McDonough,** *Senior Director, AP Content and Assessment Publications*
**Trevor Packer,** *Senior Vice President, AP and Instruction*
**Allison Thurber,** *Vice President, AP Curriculum and Assessment*

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

The curriculum framework includes two essential components:

- **AP Statistical Practices** (p. 8)
  The statistical practices are central to the study and practice of statistics. Students should develop and apply the described practices on a regular basis over the span of the course.
- **Course Content** (p. 9)
  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 a research 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.* |
| **1.A:** Determine a valid research question that requires a statistical investigation. | **2.A:** Identify information to answer a question or solve a problem. | **3.A:** Construct tabular and graphical representations of data and distributions. | **4.A:** Describe and compare tabular and graphical representations of data. |
| | **2.B:** Justify an appropriate method for ethically gathering and representing data. | **3.B:** Calculate summary statistics, relative positions of points within a distribution, and predicted responses. | **4.B:** Justify a claim based on statistical calculations and results. |
| | **2.C:** Identify appropriate statistical inference methods. | **3.C:** Calculate and estimate expected counts, percentages, probabilities, and intervals. | **4.C:** Describe distributions and compare relative positions of points within a distribution. |
| | **2.D:** Identify relationships among components in statistical inference methods. | **3.D:** Calculate parameters for probability distributions. | **4.D:** Interpret statistical calculations and results to assess meaning or a claim. |
| | **2.E:** Identify the null and alternative hypotheses. | **3.E:** Calculate appropriate statistical inference method results. | **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. |

AP STATISTICS

# 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?

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.A | 1.1.A<br>Identify components within a statistical study. | 1.1.A.1<br>A statistical study is a study in which data are collected from a sample to answer a research question about a larger population.<br>1.1.A.2<br>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.<br>1.1.A.3<br>A datum (singular for data) is a piece of information about an item or individual. A collection of data is called a data set.<br>1.1.A.4<br>A population consists of all items or individuals of interest. The population size is represented by the symbol $N$.<br>1.1.A.5<br>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 $n$.<br>1.1.A.6<br>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.A | 1.1.B<br>Determine a research question within a statistical study. | 1.1.B.1<br>A research question for a specific study should have a defined purpose and should not be changed based on the data analysis or results.<br>1.1.B.2<br>A research question should be posed so that the required data can be collected and analyzed. |

# TOPIC 1.2
# Variables

Instructional Periods: 1

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.A | 1.2.A<br>Identify observational units, variables, parameters, and statistics from a statistical study or data set. | 1.2.A.1<br>An observational unit is an item or individual from which a datum is collected.<br>1.2.A.2<br>A variable is a characteristic that may change from one observational unit to another.<br>1.2.A.3<br>Data collected on numerical and categorical variables measured on observational units, which could include photographs, sounds, videos, and text, can be used to convey meaningful information.<br>1.2.A.4<br>A parameter is a numerical attribute or summary of the variable of interest for a population.<br>1.2.A.5<br>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. |
| 2.A | 1.2.B<br>Identify types of variables. | 1.2.B.1<br>A qualitative variable, also called a categorical variable, takes on values that are category names or group labels.<br>1.2.B.2<br>A quantitative variable, also called a numerical variable, takes on numerical values for a measured or counted quantity and generally has units of measure. |
| 2.A | 1.2.C<br>Identify types of quantitative variables. | 1.2.C.1<br>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.<br>1.2.C.2<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.A | 1.3.A<br>Construct categorical one-variable tabular representations. | 1.3.A.1<br>A frequency table gives the number of observational units falling into each category of a categorical variable.<br>1.3.A.2<br>A relative frequency table gives the proportion of observational units falling into each category of a categorical variable. |
| 4.A | 1.3.B<br>Describe categorical one-variable tabular representations. | 1.3.B.1<br>Percentages, relative frequencies, and ratios all provide the same information as proportions.<br>1.3.B.2<br>Counts and relative frequencies of categorical variables reveal information that can be used to justify claims about the variables in context. |

## TOPIC 1.4

# Graphical Representations, Descriptions, and Comparisons for One Categorical Variable

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.A | 1.4.A<br>Construct categorical one-variable graphical representations. | 1.4.A.1<br>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 graph is associated with a category for the categorical variable of interest. The height or length of each bar corresponds to either the frequency or relative frequency of the observational units falling within each category.<br>1.4.A.2<br>Pie charts are used to display frequencies (counts) or relative frequencies (proportions) for categorical data. Each slice on a pie chart is associated with a category for 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 together will equal 1, or 100% of the total area. |
| 4.A | 1.4.B<br>Describe categorical one-variable graphical representations. | 1.4.B.1<br>Graphical representations of a categorical variable reveal information that can be used to justify claims about the variable in context. |
| 4.A | 1.4.C<br>Compare multiple categorical one-variable graphical and tabular representations. | 1.4.C.1<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.A | 1.5.A<br>Construct quantitative one-variable graphical representations. | 1.5.A.1<br>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.<br><br>1.5.A.2<br>A histogram puts the observed values of the quantitative variable into ordered intervals, or bins, along the horizontal axis. A bar is associated with each interval or bin, and the height of each bar shows the frequency or relative frequency of the observations that fall within the interval of the quantitative variable values corresponding to that bar. Altering the interval widths, or bin widths, can change the appearance of the histogram.<br><br>1.5.A.3<br>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(s)). Both stems and leaves are ordered from smallest to largest.<br><br>1.5.A.4<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.A | 1.6.A Describe 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.<br>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 the mirror image of the right half.<br>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.<br>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.<br>1.6.A.5 A gap is a region in a distribution between two values in which there are no observed data.<br>1.6.A.6 Clusters are concentrations of values usually separated by gaps. |

# TOPIC 1.7
# Summary Statistics for One Quantitative Variable

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.B | 1.7.A Calculate measures of center and position for quantitative data. | **1.7.A.1** Two commonly used measures of center in the distribution of a quantitative variable are the mean and median.<br>**1.7.A.2** The mean is the sum of all the values divided by the number of values and can be found 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.<br>**1.7.A.3** The median is the middle value when the data set is ordered from smallest to largest and can be found 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.<br>**1.7.A.4** In an ordered data set, the smallest value is the minimum value, and the largest value is the maximum value.<br>**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.<br>**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. |

# Exploring One-Variable Data and Collecting Data

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.B | 1.7.B<br>Calculate measures of variability for quantitative data. | 1.7.B.1<br>Three commonly used measures of variability (or spread) in the distribution of a quantitative variable are the range, interquartile range, and standard deviation.<br>1.7.B.2<br>The range is the difference between the maximum data value and the minimum data value.<br>1.7.B.3<br>The interquartile range (IQR) is the difference between the third and first quartiles: $Q3 - Q1$.<br>1.7.B.4<br>The standard deviation is a typical deviation of the data values from their mean and can be found using technology. The sample standard deviation is denoted by $s_x$ and calculated as follows: $s_x = \sqrt{\frac{1}{n-1}\Sigma\left(x_i - \overline{x}\right)^2}$, 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. |
| 3.B | 1.7.C<br>Calculate different units of measurement for summary statistics. | 1.7.C.1<br>Changing units of measurement affects the values of the calculated statistics. |
| 3.B | 1.7.D<br>Calculate outliers for quantitative data. | 1.7.D.1<br>There are many methods for determining potential outliers. Two methods frequently used are as follows:<br>i. In a skewed distribution, an outlier is a value located more than $1.5 \times IQR$ above the third quartile or more than $1.5 \times IQR$ below the first quartile.<br>ii. In a symmetric distribution, an outlier is a value located $2$ or more standard deviations above, or below, the mean. |
| 4.B | 1.7.E<br>Justify the selection of a particular measure of center and variability for describing quantitative data. | 1.7.E.1<br>The median and $IQR$ are considered resistant (or robust) measures of center and variability, respectively, because outliers do not greatly (if at all) affect their values. The mean, standard deviation, and range are considered nonresistant (or non-robust) measures because outliers can affect their values greatly. |

## TOPIC 1.8

# Graphical Representations of Summary Statistics for One Quantitative Variable

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.A | 1.8.A<br>Construct quantitative one-variable graphical representations of summary statistics. | 1.8.A.1<br>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.<br>1.8.A.2<br>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. |
| 4.A | 1.8.B<br>Describe quantitative one-variable graphical representations of summary statistics based on the relationship of the mean and the median. | 1.8.B.1<br>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 Distribution for One Quantitative Variable

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.A | 1.9.A<br>Compare multiple quantitative one-variable graphical representations. | 1.9.A.1<br>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). |
| 4.A | 1.9.B<br>Compare multiple quantitative one-variable graphical representations of summary statistics. | 1.9.B.1<br>A comparison of graphical representations for two or more distributions can include any of the numerical summaries (e.g., mean, standard deviation, etc.). |
| 3.B | 1.9.C<br>Calculate $z$-scores with population parameters. | 1.9.C.1<br>A standardized score measures the number of standard deviations a data value falls above or below the mean.<br>1.9.C.2<br>A $z$-score is calculated as $\frac{x_i - \mu}{\sigma}$, where $x_i$ is the data value, $\mu$ is the population mean, and $\sigma$ 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. |
| 4.C | 1.9.D<br>Compare $z$-scores as measures of relative position for distributions. | 1.9.D.1<br>$z$-scores may be used to compare relative positions of individual values within a distribution or between distributions. |

# TOPIC 1.10
# Data Collection

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.A | 1.10.A<br>Identify a census. | 1.10.A.1<br>A census consists of recording information from all items or individuals in a population. |
| 2.A | 1.10.B<br>Identify an observational study. | 1.10.B.1<br>An observational study is one in which the researcher is a passive observer, recording the values of variables of interest in order to investigate a research question of interest.<br>1.10.B.2<br>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.<br>1.10.B.3<br>A retrospective study is one in in which the observational units of study are selected at a point in time and data from the past are gathered.<br>1.10.B.4<br>A survey is an observational study in which the data are collected from humans using a standard set of questions. |
| 2.A | 1.10.C<br>Identify an experiment. | 1.10.C.1<br>An experiment is a study in which a researcher assigns conditions, or treatments, to experimental units to investigate a research question of interest about the population.<br>1.10.C.2<br>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.<br>1.10.C.3<br>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.<br>1.10.C.4<br>A response variable is an outcome measured on each experimental unit after the treatment has been administered. |

# Exploring One-Variable Data and Collecting Data

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.B | 1.10.D<br>Justify the appropriateness of generalizations for a statistical study. | 1.10.D.1<br>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 (e.g., random number table, lottery, etc.) is used to select the observational units to be included in the sample.<br><br>1.10.D.2<br>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.<br><br>1.10.D.3<br>A sample is not randomly selected when observational units are deliberately chosen or volunteer themselves to be in the sample.<br><br>1.10.D.4<br>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. |

# TOPIC 1.11
# Random Sampling

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.A | 1.11.A<br>Identify a sampling method given a description of a study. | 1.11.A.1<br>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 observational unit can be selected again.<br><br>1.11.A.2<br>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 observational unit could be selected again.<br><br>1.11.A.3<br>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 a lottery method.<br><br>1.11.A.4<br>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.<br><br>1.11.A.5<br>A cluster 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.<br><br>1.11.A.6<br>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. |
| 2.B | 1.11.B<br>Justify the appropriateness of a sampling method. | 1.11.B.1<br>Sampling methods can be used to mitigate the consequences of using simple random samples of certain populations for certain research questions. |

# TOPIC 1.12
# Potential Problems with Sampling

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.A | 1.12.A Identify potential sources of bias in sampling methods. | 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.<br>1.12.A.2 Voluntary response bias is a bias that may occur when a sample consists entirely of volunteers or is a convenience sample.<br>1.12.A.3 Undercoverage bias may occur when the sampling method fails to include part of the population.<br>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 would differ significantly in ways that are important for the study.<br>1.12.A.5 Response bias may occur when responses to a survey or measurements of entities 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.<br>1.12.A.6 Nonrandom sampling methods (for example, 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

Instructional Periods: 3

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.A | 1.13.A<br>Identify elements of a well-designed experiment. | 1.13.A.1<br>A well-planned experiment should include the following:<br>i. Comparisons of at least two treatment groups, one of which is a control group<br>ii. Random assignment of treatments to experimental units<br>iii. Replication<br>iv. Direct control of potential extraneous sources of variation in the response<br>1.13.A.2<br>A control group is a collection of experimental units that are constructed 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 (i.e., a treatment with an inactive substance, a placebo, may be given).<br>1.13.A.3<br>The placebo effect is the difference between the average response to a placebo and the average response to no treatment.<br>1.13.A.4<br>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.<br>1.13.A.5<br>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.<br>1.13.A.6<br>An extraneous source of variation, also referred to as an extraneous variable, in the response variable is a variable that is known (or believed) to affect the response but is not an explanatory variable being studied. |

# Exploring One-Variable Data and Collecting Data

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.A | 1.13.A Identify elements of a well-designed experiment. | **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.<br>**1.13.A.8** A confounding variable is a variable that is distributed differently among treatment groups and affects the response variable. A confounding variable provides an alternative explanation for the observed relationship between the response and explanatory 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.<br>**1.13.A.9** Replication within an experiment means more than one experimental unit is assigned to each treatment.<br>**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. |
| 2.A | 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.<br>**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.<br>**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.<br>**1.13.B.4** In randomized complete block designs, one experimental unit is assigned to each treatment within each block.<br>**1.13.B.5** 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. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.B | 1.13.C<br>Justify the appropriateness of a particular experimental design. | 1.13.C.1<br>One experimental design may be more appropriate than another experimental design based on the goals of the research study, the characteristics of the population, and the sample and variables involved. |
| 2.B | 1.13.D<br>Justify the appropriateness of the conclusions based on a well-designed experiment. | 1.13.D.1<br>Using random assignment of treatments to experimental units allows for cause-and-effect conclusions between the explanatory and the response variables if the data show a statistically significant result, because the potential for confounding variables is reduced.<br>1.13.D.2<br>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. |

THIS PAGE IS INTENTIONALLY LEFT BLANK.

AP STATISTICS

# UNIT 2

# Probability, Random Variables, and Probability Distributions

THIS PAGE IS INTENTIONALLY LEFT BLANK.

## TOPIC 2.1

# Graphical and Tabular Representations for the Distributions of Two Categorical Variables

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.A | 2.1.A<br>Compare graphical and tabular representations for the distributions of two categorical variables. | 2.1.A.1<br>Side-by-side bar charts, segmented bar charts, and mosaic plots are examples of graphs used to display and compare 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.<br>2.1.A.2<br>Graphical representations of two categorical variables can be used to compare the distributions of one categorical variable across the levels of the other categorical variable and determine whether the two variables are associated.<br>2.1.A.3<br>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). |

## TOPIC 2.2

# Summary Statistics for Two Categorical Variables

Instructional Periods: 1

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.B | 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.<br>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.<br>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. |
| 4.A | 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. |

## TOPIC 2.3
# Estimating Probabilities

Instructional Periods: 1

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.C | 2.3.A<br>Estimate probabilities. | 2.3.A.1<br>A random process generates results that are determined by chance.<br>2.3.A.2<br>An outcome is the result of one trial of a random process.<br>2.3.A.3<br>An event is a collection of outcomes.<br>2.3.A.4<br>The probability of an outcome or event is its long-run relative frequency, that is, its relative frequency over a large number of trials.<br>2.3.A.5<br>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.<br>2.3.A.6<br>The law of large numbers states that for independent events, 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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.C | 2.4.A Calculate probabilities for events and their complements. | **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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.B | 2.5.A<br>Justify why two events are mutually exclusive (or disjoint) using joint probability. | 2.5.A.1<br>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 $P(A \cap B)$.<br><br>2.5.A.2<br>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 $P(A \text{ intersect } B) = 0$ or $P(A \cap B) = 0$. |

## TOPIC 2.6
# Conditional Probability

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.C | 2.6.A Calculate conditional probabilities. | 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 $P(A \mid B)$. Conditional probability is defined as $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. |
| | | 2.6.A.2 The general multiplication rule states that the probability that events $A$ and $B$ both will occur is equal to the probability that event $A$ will occur multiplied by the conditional probability that event $B$ will occur given that $A$ has occurred. The multiplication rule is defined as $P(A \cap B) = P(A) \cdot P(B \mid A)$. |

## TOPIC 2.7

# Independent Events and Unions of Events

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.C | 2.7.A<br>Calculate probabilities for independent events and for the union of two events. | 2.7.A.1<br>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 $P(A \mid B) = P(A)$, $P(B \mid A) = P(B)$, and $P(A \cap B) = P(A) \cdot P(B)$.<br>2.7.A.2<br>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 $P(A \cup B)$.<br>2.7.A.3<br>The probability that $P(A \text{ union } B) = P(A) + P(B) - P(A \text{ intersect } B)$, or $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. |

## TOPIC 2.8

# Introduction to Random Variables and Probability Distributions

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.A | 2.8.A Construct a probability distribution for a discrete random variable. | 2.8.A.1 A random variable is a variable whose values have numerical outcomes that result from a random phenomenon.<br>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 $1$.<br>2.8.A.3 A discrete probability distribution can be determined using the rules of probability or estimated with a simulation.<br>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.<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.B | 2.9.A<br>Calculate parameters for a discrete random variable. | 2.9.A.1<br>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.<br><br>2.9.A.2<br>The expected value (or mean) of a probability distribution is a parameter and is denoted by $E(X)$ or $\mu_X$. 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.<br><br>2.9.A.3<br>The standard deviation of a probability distribution is a parameter represented by $SD(X)$ or $\sigma_X$. 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, $\mu_X$ 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 $\sigma_X^2$. |
| 4.D | 2.9.B<br>Interpret parameters and probabilities for a discrete random variable. | 2.9.B.1<br>Parameters and probabilities 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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.B | 2.10.A<br>Justify why a random variable is or is not a binomial random variable. | 2.10.A.1<br>A binomial random variable, $X$, is a discrete random variable that counts the number of successes in repeated independent trials that have only two possible outcomes (success or failure), with the probability of success $p$ and the probability of failure $1-p$. |
| 3.D | 2.10.B<br>Calculate parameters for a binomial distribution. | 2.10.B.1<br>If a random variable is binomial, its mean, $\mu_X$, is $np$ and its standard deviation, $\sigma_X$, is $\sqrt{np(1-p)}$. |
| 3.C | 2.10.C<br>Calculate probabilities for a binomial distribution. | 2.10.C.1<br>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. |
| 4.D | 2.10.D<br>Interpret parameters and probabilities for a binomial distribution. | 2.10.D.1<br>Parameters and probabilities for a binomial distribution should be interpreted in context. |

## TOPIC 2.11

# The Normal Distribution

Instructional Periods: 3

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.B | 2.11.A<br>Justify why a random variable is or is not a normal random variable. | 2.11.A.1<br>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.<br>2.11.A.2<br>Many continuous random variables are well-modeled by a normal distribution.<br>2.11.A.3<br>A normal distribution can be described as a continuous, unimodal, bell-shaped, and symmetric curve. |
| 4.C | 2.11.B<br>Describe a normal distribution. | 2.11.B.1<br>The normal distribution, or the normal curve, is identified by two parameters, the mean, $\mu$, and the standard deviation, $\sigma$. 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. |
| 3.D | 2.11.C<br>Calculate parameters for a normal distribution. | 2.11.C.1<br>A standard normal distribution is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. |
| 3.C | 2.11.D<br>Calculate percentages from a normal distribution using the empirical rule. | 2.11.D.1<br>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 the observations are within 1 standard deviation of the mean, approximately 95% of observations are within 2 standard deviations of the mean, and approximately 99.7% of observations are within 3 standard deviations of the mean. This is called the empirical rule, or the $68-95-99.7$ rule. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.C | 2.11.E<br>Calculate the probability that a particular value lies within a given interval of a normal distribution. | 2.11.E.1<br>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. |
| 3.C | 2.11.F<br>Calculate the associated intervals and areas of a normal distribution. | 2.11.F.1<br>The boundaries of an interval associated with a given area in a normal distribution can be determined using $z$-scores and a standard normal table or technology.<br>2.11.F.2<br>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 1, $x_a$ is the lower bound, and $x_b$ is the upper bound on a normal distribution.<br>i. $P(X < x_a) = \frac{p}{100}$ means that the lowest $p\%$ of the values lie to the left of $x_a$.<br>ii. $P(x_a < X < x_b) = \frac{p}{100}$ means that $p\%$ of the values lie between $x_a$ and $x_b$.<br>iii. $P(X > x_b) = \frac{p}{100}$ means that the highest $p\%$ of the values lie to the left of $x_b$.<br>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 $x_b$. |
| 4.C | 2.11.G<br>Compare measures of relative position for distributions. | 2.11.G.1<br>Percentiles and proportions may be used to compare relative positions of individual values within a normal distribution or between normal distributions. |

# TOPIC 2.12
# The Central Limit Theorem

Instructional Periods: 3

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.C | 2.12.A<br>Describe sampling distributions. | 2.12.A.1<br>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.<br>2.12.A.2<br>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.<br>2.12.A.3<br>A randomization distribution is the distribution of a statistic generated 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.<br>2.12.A.4<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.B | 3.1.A<br>Justify why an estimator is or is not unbiased. | 3.1.A.1<br>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.D | 3.1.B<br>Calculate estimates for a population parameter. | 3.1.B.1<br>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

Instructional Periods: 1

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.D | 3.2.A<br>Calculate parameters of a sampling distribution for a sample proportion. | 3.2.A.1<br>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 $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.<br>3.2.A.2<br>When sampling without replacement, two conditions must be met as follows:<br>i. The randomization condition: the data should be collected using a random sample.<br>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.E | 3.2.B<br>Justify why a sampling distribution of a sample proportion can or cannot be described as approximately normal. | 3.2.B.1<br>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 number of successes and $n(1-p)$ is the number of failures. |
| 4.D | 3.2.C<br>Interpret parameters and probabilities for a sampling distribution of a sample proportion. | 3.2.C.1<br>Parameters 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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.C | 3.3.A<br>Identify an appropriate confidence interval procedure including the parameter for a population proportion. | 3.3.A.1<br>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.<br>3.3.A.2<br>The parameter for a confidence interval for a population proportion should reference the proportion, the response variable, and the population in context. |
| 4.E | 3.3.B<br>Justify the appropriateness of constructing a confidence interval for a population proportion by verifying conditions. | 3.3.B.1<br>A one-sample $z$-interval for a population proportion requires that three conditions be met as follows:<br>i. The randomization condition: the data should be collected using a random sample.<br>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.<br>iii. The normality condition: The number of successes, $n\hat{p}$, and the number of failures, $n(1-\hat{p})$, should be at least $10$. |
| 3.E | 3.3.C<br>Calculate an appropriate confidence interval for a population proportion. | 3.3.C.1<br>An interval estimate can be constructed as point estimate $\pm$ (margin of error). For a population proportion, the one-sample $z$-interval estimate is<br>$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. |

# Inference for Categorical Data: Proportions

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.E | 3.3.D<br>Calculate the standard error and margin of error of a sample statistic for a confidence interval for a population proportion. | 3.3.D.1<br>$z^*$ 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.<br>3.3.D.2<br>The standard error 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 $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.<br>3.3.D.3<br>The standard error quantifies the "typical" amount that a statistic will vary from the value of the corresponding population parameter.<br>3.3.D.4<br>The margin of error 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 $z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. |

## TOPIC 3.4

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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.F | 3.4.A<br>Interpret a confidence interval in context for a population proportion. | 3.4.A.1<br>Since 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.<br>3.4.A.2<br>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.<br>3.4.A.3<br>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. 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. |
| 4.G | 3.4.B<br>Justify a claim based on a confidence interval for a population proportion. | 3.4.B.1<br>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. |
| 2.D | 3.4.C<br>Identify the relationships among sample size, confidence interval width, confidence level, and margin of error for a population proportion. | 3.4.C.1<br>For a given sample, increasing the confidence level will result in the following:<br>i. The critical value will increase.<br>ii. The margin of error will increase.<br>iii. The width of the confidence interval will increase.<br>3.4.C.2<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.C | 3.5.A<br>Identify an appropriate testing method for a population proportion including the parameter for the population proportion. | 3.5.A.1<br>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.<br>3.5.A.2<br>The parameter for a hypothesis test for a population proportion should reference the proportion, the response variable, and the population in context. |
| 2.E | 3.5.B<br>Identify the null and alternative hypotheses for a population proportion. | 3.5.B.1<br>In the hypothesis testing procedure, the null hypothesis, $H_0$, is the statement about a parameter that is correct unless evidence suggests 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.<br>3.5.B.2<br>The null hypothesis contains an equality reference (=, ≥, or ≤). 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.<br>3.5.B.3<br>The null hypothesis for a one-sample $z$-test for a population proportion is as follows: $H_0: p = p_0$, 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 $H_a: p < p_0$ or $H_a: p > p_0$. A two-sided alternative hypothesis is $H_a: p \neq p_0$. |
| 4.E | 3.5.C<br>Justify the appropriateness of a hypothesis test for a population proportion by verifying conditions. | 3.5.C.1<br>A one-sample $z$-test for a population proportion requires that three conditions be met as follows:<br>i. The randomization condition: The data should be collected using a random sample.<br>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.<br>iii. The normality condition: The number of successes, $np_0$, and the number of failures, $n(1-p_0)$, should be at least $10$. |

# TOPIC 3.6
# *p*-Values

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.F | 3.6.A<br>Interpret the $p$-value of a hypothesis test for a population proportion. | **3.6.A.1**<br>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:<br>i. The probability at or above the observed value of the test statistic $\left(P(z \geq x)\right)$, if the alternative is ><br>ii. The probability at or below the observed value of the test statistic $\left(P(z \leq x)\right)$, if the alternative is <<br>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, $\left(P(z \leq -\lvert x\rvert)\right) + \left(P(z \geq \lvert x\rvert)\right)$, if the alternative is ≠<br>**3.6.A.2**<br>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:<br>i. The proportion at or above the observed value of the test statistic, if the alternative is ><br>ii. The proportion at or below the observed value of the test statistic, if the alternative is <<br>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 ≠<br>**3.6.A.3**<br>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).<br>**3.6.A.4**<br>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.<br>**3.6.A.5**<br>$p$-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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.E | 3.7.A Calculate an appropriate test statistic and $p$-value for testing a hypothesis about a population proportion. | **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.<br>**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).<br>**3.7.A.3** The $p$-value of a one-sample $z$-test for a population proportion is found from the standard normal distribution table or from the standard normal distribution using technology. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.G | 3.7.B<br>Justify a claim about the population based on the results of a hypothesis test for a population proportion. | 3.7.B.1<br>The significance level of a hypothesis test, denoted by $\alpha$, 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.<br><br>3.7.B.2<br>A formal decision in a hypothesis test explicitly compares the $p$-value to the significance level, $\alpha$. If the $p\text{-value} \leq \alpha$ then reject the null hypothesis, $H_o : p = p_0$. If the $p\text{-value} > \alpha$ then fail to reject the null hypothesis.<br><br>3.7.B.3<br>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.<br><br>3.7.B.4<br>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.<br><br>3.7.B.5<br>The results of a hypothesis test for a population proportion can serve as the statistical reasoning to support the answer to a research question about the population that was sampled.<br><br>3.7.B.6<br>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

# Sampling Distributions for the Difference Between Sample Proportions

Instructional Periods: 1

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.D | 3.8.A<br>Calculate parameters of a sampling distribution for the difference between two sample proportions. | 3.8.A.1<br>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, $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.<br>3.8.A.2<br>When sampling without replacement, two conditions must be met as follows:<br>i. The randomization condition: the data should be collected using two independent random samples.<br>ii. The 10% condition: both samples must be less than 10% of the size of their respective populations.<br>3.8.A.3<br>If the data comes from an experiment, the data only needs to meet the randomization condition. The treatments must be randomly assigned to participants or experimental units to meet the randomization condition. |
| 4.E | 3.8.B<br>Justify why a sampling distribution for the difference between sample proportions can or cannot be described as approximately normal. | 3.8.B.1<br>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 number of successes and $n_1(1-p_1)$ and $n_2(1-p_2)$ are the number of failures. |
| 4.D | 3.8.C<br>Interpret parameters and probabilities for a sampling distribution for the difference between proportions. | 3.8.C.1<br>Parameters and probabilities for a sampling distribution of a sample proportion should be interpreted within the context of a specific population. |

## TOPIC 3.9

# Constructing a Confidence Interval for the Difference Between Two Population Proportions

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.C | 3.9.A<br>Identify an appropriate confidence interval procedure including the parameters for the difference between two population proportions. | 3.9.A.1<br>Based on the sample data, a confidence interval can be calculated to estimate the difference between two independent population proportions. The appropriate confidence interval procedure is a two-sample $z$-interval for a difference between population proportions.<br>3.9.A.2<br>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. |
| 4.E | 3.9.B<br>Justify the appropriateness of constructing a confidence interval for the difference between two population proportions by verifying conditions. | 3.9.B.1<br>A two-sample $z$-interval for a difference between two population proportions requires three conditions to be met as follows:<br>i. The randomization condition: the data should be collected using two independent random samples or a randomized experiment.<br>ii. The 10% condition: When sampling without replacement, check that $n_1 \le 10\% N_1$ and $n_2 \le 10\% N_2$, 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 not necessary when the data are from a randomized experiment.)<br>iii. The normality condition: the number of successes, $n_1\hat{p}_1$ and $n_2\hat{p}_2$, and number of failures, $n_1(1-\hat{p}_1)$ and $n_2(1-\hat{p}_2)$, for both samples are all at least 10. |

# Inference for Categorical Data: Proportions

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.E | 3.9.C Calculate an appropriate confidence interval for the difference between two population proportions. | 3.9.C.1 The point estimate for the difference between two population proportions $\left(\hat{p}_1 - \hat{p}_2\right)$ is the difference in sample proportions.<br>3.9.C.2 For the difference between two population proportions, the interval estimate can be constructed as point estimate ± (margin of error). The interval estimate for the difference between two population proportions is $\left(\hat{p}_1 - \hat{p}_2\right) \pm z^* \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$. |
| 3.E | 3.9.D Calculate the standard error and margin of error for the difference between two population proportions. | 3.9.D.1 The standard error for the difference between two population proportions is $\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$.<br>3.9.D.2 For the difference between two population proportions, the margin of error is the critical value $\left(z^*\right)$ 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 3.10

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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.F | 3.10.A<br>Interpret a confidence interval in context for the difference between two population proportions. | 3.10.A.1<br>Since 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.<br><br>3.10.A.2<br>The interpretation of the confidence level is as follows: In repeated random sampling with the same sample sizes, 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.<br><br>3.10.A.3<br>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 population. 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. |
| 4.G | 3.10.B<br>Justify a claim based on a confidence interval for the difference between two population proportions. | 3.10.B.1<br>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 3.11

# Setting Up a Test for the Difference Between Two Population Proportions

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.C | 3.11.A<br>Identify an appropriate testing method for the difference between population proportions including the parameters. | 3.11.A.1<br>The appropriate testing method for the difference between two population proportions is a two-sample $z$-test for the difference between two population proportions.<br>3.11.A.2<br>The parameters for a hypothesis test for the difference between two population proportions should refer to the difference in the proportions, the response variable, and the populations in context. |
| 2.E | 3.11.B<br>Identify the null and alternative hypotheses for the difference between population proportions. | 3.11.B.1<br>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 $H_0: p_1 = p_2$ or $H_0: p_1 - p_2 = 0$. A one-sided alternative hypothesis for the difference between two population proportions can be written as either:<br>$H_a: p_1 < p_2$, or equivalently $H_a: p_1 - p_2 < 0$, or<br>$H_a: p_1 > p_2$, or equivalently $H_a: p_1 - p_2 > 0$.<br>A two-sided alternative hypothesis for the difference between two population proportions can be written as either $H_a: p_1 \neq p_2$, or equivalently $H_a: p_1 - p_2 \neq 0$. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.E | 3.11.C<br>Justify the appropriateness of a hypothesis test for the difference between two population proportions by verifying conditions. | 3.11.C.1<br>A two-sample $z$-test for a difference between two population proportions requires three conditions to be met as follows:<br>i. The randomization condition: the data should be collected using two independent random samples or a randomized experiment.<br>ii. The $10\%$ condition: When sampling without replacement, check that $n_1 \le 10\% N_1$ and $n_2 \le 10\% N_2$, 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.)<br>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 $\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$ being the combined (or pooled) proportion assuming that $H_0$ is true ($H_0 : p_1 = p_2$ or $H_0 : p_1 - p_2 = 0$). |

## TOPIC 3.12

# Carrying Out a Test for the Difference Between Two Population Proportions

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.E | 3.12.A<br>Calculate an appropriate test statistic and $p$-value for testing a hypothesis for the difference between two population proportions. | 3.12.A.1<br>The test statistic for the difference between two population proportions is as follows: $z = \dfrac{\left(\hat{p}_1 - \hat{p}_2\right) - 0}{\sqrt{\hat{p}_c\left(1-\hat{p}_c\right)}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$, where $\hat{p}_c = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$ is the proportion of successes for the two groups combined. The $z$-statistic has a standard normal distribution when the null hypothesis is true.<br>3.12.A.2<br>The $p$-value for a two-sample $z$-test for the difference between two population proportions can be found using the standard normal table or from the standard normal distribution using technology. |
| 4.F | 3.12.B<br>Interpret the $p$-value of a hypothesis test for the difference between two population proportions. | 3.12.B.1<br>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). |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.G | 3.12.C<br>Justify a claim about the population based on the results of a hypothesis test for the difference between two population proportions. | 3.12.C.1<br>A formal decision in a hypothesis test for the difference between two population proportions explicitly compares the $p$-value to the significance level, $\alpha$. If the $p\text{-value} \leq \alpha$ then reject the null hypothesis, $H_0 : p_1 = p_2$ or $H_0 : p_1 - p_2 = 0$. If the $p\text{-value} > \alpha$ then fail to reject the null hypothesis.<br>3.12.C.2<br>The results of a hypothesis test for the difference between two population proportions can serve as the statistical reasoning to support the answer to a research question about the two populations that were sampled.<br>3.12.C.3<br>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 3.13

# Setting Up a Chi-Square Test for Homogeneity or Independence

Instructional Periods: 3

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.C | 3.13.A<br>Describe chi-square distributions. | 3.13.A.1<br>The chi-square statistic measures the distance between observed and expected counts relative to expected counts.<br>3.13.A.2<br>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. |
| 2.C | 3.13.B<br>Identify an appropriate testing method for comparing distributions in two-way tables of categorical data. | 3.13.B.1<br>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.<br>3.13.B.2<br>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. |
| 2.E | 3.13.C<br>Identify the null and alternative hypotheses for a chi-square test for homogeneity or independence. | 3.13.C.1<br>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.<br>3.13.C.2<br>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 dependent on each other. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.E | 3.13.D<br>Justify the appropriateness of a hypothesis test for a chi-square distribution for independence or homogeneity by verifying conditions. | 3.13.D.1<br>For a chi-square test for homogeneity or independence, three conditions must be met as follows:<br>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.<br>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.)<br>iii. The expected values condition: all expected values should be greater than $5$. |

## TOPIC 3.14

# Carrying Out a Chi-Square Test for Homogeneity or Independence

Instructional Periods: 3

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.C | 3.14.A<br>Calculate expected counts for two-way tables of categorical data. | 3.14.A.1<br>The expected values (under the null hypothesis) in a particular cell of a two-way table of categorical data can be calculated using the following formula: $\text{expected value} = \frac{(\text{row total})(\text{column total})}{\text{total table}}$. |
| 3.E | 3.14.B<br>Calculate the appropriate test statistic and $p$-value for a chi-square test for homogeneity or independence. | 3.14.B.1<br>The appropriate test statistic for a chi-square test for homogeneity or independence is the chi-square statistic: $\chi^2 = \sum \frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}}$, 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.<br>3.14.B.2<br>The $p$-value for a chi-square test for independence or homogeneity is found using a chi-square critical values table or from a chi-square distribution using technology. |
| 4.F | 3.14.C<br>Interpret the $p$-value for the chi-square test for homogeneity or independence. | 3.14.C.1<br>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. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.G | 3.14.D<br>Justify a claim about the population based on the results of a chi-square test for homogeneity or independence. | 3.14.D.1<br>A formal decision in a hypothesis test explicitly compares the $p$-value to the significance level, $\alpha$. 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.<br>3.14.D.2<br>The results of a chi-square test for homogeneity or independence can serve as the statistical reasoning to support the answer to a research question about the population that was sampled (independence) or the populations that were sampled (homogeneity).<br>3.14.D.3<br>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

THIS PAGE IS INTENTIONALLY LEFT BLANK.

# TOPIC 4.1
# Sampling Distributions for Sample Means

Instructional Periods: 1

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.D | 4.1.A<br>Calculate parameters of a sampling distribution of a sample mean. | 4.1.A.1<br>For a population with population mean $\mu$ and population standard deviation $\sigma$, when the sampled values are independent, the sampling distribution of the sample mean has mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.<br>4.1.A.2<br>When sampling without replacement, two conditions must be met as follows:<br>i. The randomization condition: the data should be collected using a random sample.<br>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.E | 4.1.B<br>Justify why the sampling distribution of a sample mean can or cannot be described as approximately normal. | 4.1.B.1<br>For a quantitative variable, if the population distribution can be modeled with a normal distribution, the sampling distribution of the sample mean, $\bar{x}$, can be modeled with a normal distribution regardless of the sample size.<br>4.1.B.2<br>For a quantitative variable, if the population distribution is not normal, 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.D | 4.1.C<br>Interpret parameters and probabilities for the sampling distribution of a sample mean. | 4.1.C.1<br>Parameters 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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.C | 4.2.A<br>Describe $t$-distributions. | 4.2.A.1<br>$t$-distributions, also called Student's $t$-distributions, make up 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 $\sigma = 1$).<br>4.2.A.2<br>$t$-distributions are used for finding critical values and test statistics for inferences about a population mean, $\mu$, when the population standard deviation, $\sigma$, is unknown and the sample standard deviation, $s$, must be used instead. |
| 2.C | 4.2.B<br>Identify an appropriate confidence interval procedure including the parameter for a population mean, including matched pairs. | 4.2.B.1<br>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, $\sigma$, is not typically known for distributions for quantitative variables).<br>4.2.B.2<br>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.<br>4.2.B.3<br>The parameter for a confidence interval for a population mean, including matched pairs, 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. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.E | 4.2.C<br>Justify the appropriateness of constructing a confidence interval for a population mean, including the mean difference between values in matched pairs, by verifying conditions. | 4.2.C.1<br>A one-sample $t$-interval for a population mean or population mean difference requires that three conditions be met as follows:<br>i. The randomization condition: the data should be collected using a random sample or a randomized experiment.<br>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.<br>iii. The sample data condition: $n \geq 30$ , 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. |
| 3.E | 4.2.D<br>Calculate an appropriate confidence interval for a population mean, including the mean difference between values in matched pairs. | 4.2.D.1<br>A point estimate for a population mean is the sample mean, $\bar{x}$, or $\bar{x}_d$ for the sample mean difference.<br>4.2.D.2<br>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 $\pm t^*$ is the critical value for the central C% of a $t$-distribution with degrees of freedom, $n-1$. |
| 3.E | 4.2.E<br>Calculate the standard error and margin of error for a sample size for a one-sample $t$-interval for a population mean. | 4.2.E.1<br>The standard error for a sample mean is given by $s_{\bar{x}} = \frac{s}{\sqrt{n}}$.<br>4.2.E.2<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.F | 4.3.A<br>Interpret a confidence interval in context for a population mean, including the mean difference between values in matched pairs. | 4.3.A.1<br>Since 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.<br><br>4.3.A.2<br>The interpretation of the confidence level is as follows: In repeated random sampling with the same sample size, 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.<br><br>4.3.A.3<br>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. An interpretation of a confidence interval for a population mean includes a reference to the parameter. |
| 4.G | 4.3.B<br>Justify a claim based on a confidence interval for a population mean, including the mean difference between values in matched pairs. | 4.3.B.1<br>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. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.D | 4.3.C<br>Identify the relationships among sample size, confidence interval width, confidence level, and margin of error for a population mean. | 4.3.C.1<br>For a given sample, increasing the confidence level will result in the following:<br>i. The critical value will increase.<br>ii. The margin of error will increase.<br>iii. The width of the confidence interval will increase.<br>4.3.C.2<br>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 tends to decrease as the sample size increases. For a confidence interval for a population mean 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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.C | 4.4.A<br>Identify an appropriate testing method and parameter for a population mean with unknown $\sigma$, including the mean difference between values in matched pairs. | 4.4.A.1<br>The appropriate test for a population mean with unknown population standard deviation $\sigma$ is a one-sample $t$-test for a population mean.<br>4.4.A.2<br>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 mean difference.<br>4.4.A.3<br>The parameter for a hypothesis test for a population mean, including matched pairs, should reference the population mean or population mean difference and the response variable in context. |
| 2.E | 4.4.B<br>Identify the null and alternative hypotheses for a population mean with unknown $\sigma$, including the mean difference between values in matched pairs. | 4.4.B.1<br>The null hypothesis for a one-sample $t$-test for a population mean is $H_0: \mu = \mu_0$, in which $\mu_0$ 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 $H_a: \mu < \mu_o$ or $H_a: \mu > \mu_o$. A two-sided alternative hypothesis is $H_a: \mu \neq \mu_o$.<br>4.4.B.2<br>The null hypothesis for a matched pairs design is $H_0: \mu_d = 0$. A one-sided alternative hypothesis for a matched pair design is either $H_a: \mu_d < 0$ or $H_a: \mu_d > 0$. A two-sided alternative hypothesis is $H_a: \mu_d \neq 0$. |
| 4.E | 4.4.C<br>Justify the appropriateness of a hypothesis test for a population mean by verifying conditions. | 4.4.C.1<br>A one-sample $t$-test for a population mean requires that three conditions be met as follows:<br>i. The randomization condition: the data should be collected using a random sample or a randomized experiment.<br>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.<br>iii. The sample data condition: $n \geq 30$, 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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.E | 4.5.A<br>Calculate an appropriate test statistic and $p$-value for a population mean, including the mean difference between values in matched pairs. | 4.5.A.1<br>The test statistic for a one-sample $t$-test for a population mean or population mean difference is $t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$, 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.<br>4.5.A.2<br>The $p$-value for a one-sample $t$-test for a population mean is found using the appropriate $t$-distribution table or from the appropriate $t$-distribution using technology. |
| 4.F | 4.5.B<br>Interpret the $p$-value of a hypothesis test for a population mean, including the mean difference between values in matched pairs. | 4.5.B.1<br>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 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). |
| 4.G | 4.5.C<br>Justify a claim about the population based on the results of a hypothesis test for a population mean, including the mean difference between values in matched pairs. | 4.5.C.1<br>A formal decision explicitly compares the $p$-value to the significance level, $\alpha$. If the $p\text{-value} \leq \alpha$ then reject the null hypothesis, $H_0 : \mu = \mu_0$. If the $p\text{-value} < \alpha$ then fail to reject the null hypothesis.<br>4.5.C.2<br>The results of a hypothesis test for a population mean can serve as the statistical reasoning to support the answer to a research question about the population that was sampled.<br>4.5.C.3<br>A conclusion for the hypothesis test for a population mean 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

Instructional Periods: 1

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.D | 4.6.A<br>Calculate parameters of a sampling distribution for the difference between two sample means. | 4.6.A.1<br>For two independent populations with population means $\mu_1$ and $\mu_2$ and population standard deviations $\sigma_1$ and $\sigma_2$, when the sampled values are independent, the sampling distribution of the difference in sample means $\overline{x}_1 - \overline{x}_2$ has a mean $\mu_{(\overline{x}_1 - \overline{x}_2)} = \mu_1 - \mu_2$ and standard deviation $\sigma_{(\overline{x}_1 - \overline{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.<br><br>4.6.A.2<br>When sampling without replacement, the data must meet two conditions as follows:<br>i. The randomization condition: the data should be collected using two independent random samples.<br>ii. The $10\%$ condition: both samples must be less than $10\%$ of the size of their respective populations.<br><br>4.6.A.3<br>If the data come from an experiment, the data only need to meet the randomization condition. The treatments must be randomly assigned to participants or experimental units to meet the randomization condition. |
| 4.E | 4.6.B<br>Justify why a sampling distribution for the difference between sample means can or cannot be described as approximately normal. | 4.6.B.1<br>The sampling distribution for the difference between sample means, $\overline{x}_1 - \overline{x}_2$, can be modeled with a normal distribution if the two population distributions can each be modeled by a normal distribution.<br><br>4.6.B.2<br>The sampling distribution for the difference between sample means, $\overline{x}_1 - \overline{x}_2$, 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 $n_2 \geq 30$. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.D | 4.6.C<br>Interpret parameters and probabilities for a sampling distribution for the difference between sample means. | 4.6.C.1<br>Parameters 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 Sample Means

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.C | 4.7.A<br>Identify an appropriate confidence interval procedure including the parameter for the difference between two population means. | 4.7.A.1<br>The sample data can be used to calculate a confidence interval 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.<br>4.7.A.2<br>The parameter for a confidence interval for a two-sample $t$-interval for the difference between population means should reference the population mean and the response variable in context. |
| 4.E | 4.7.B<br>Justify why a sampling distribution for the difference between sample means can or cannot be described as approximately normal. | 4.7.B.1<br>A two-sample $t$-interval for a difference between population means requires that three conditions be met as follows:<br>i. The randomization condition: the data should be collected using two random samples or a randomized experiment.<br>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.)<br>iii. The sample data condition: Both samples should have a sample size greater than or equal to $30$. If either sample size is less than $30$, both sample data distributions should be free from strong skewness and outliers. |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.E | 4.7.C<br>Calculate an appropriate confidence interval for the difference between two population means. | 4.7.C.1<br>A point estimate for the difference between two population means is the difference in sample means, $\overline{x}_1 - \overline{x}_2$.<br>4.7.C.2<br>For the difference between population means when the population standard deviations are unknown, the confidence interval can be constructed as point estimate ± (margin of error). The confidence interval for the difference between population means is $\left(\overline{x}_1 - \overline{x}_2\right) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, where $\pm t^*$ 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 $n_2 - 1$. |
| 3.E | 4.7.D<br>Calculate the standard error and margin of error for the difference between two population means. | 4.7.D.1<br>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.<br>4.7.D.2<br>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

# Constructing a Confidence Interval for the Difference Between Two Sample Means

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.F | 4.8.A<br>Interpret a confidence interval in context for the difference between two population means. | 4.8.A.1<br>Since 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.<br><br>4.8.A.2<br>The interpretation of the confidence level is as follows: In repeated random sampling with the same sample size, 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.<br><br>4.8.A.3<br>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. 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.G | 4.8.B<br>Justify a claim based on a confidence interval for the difference between two population means. | 4.8.B.1<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 2.C | 4.9.A<br>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<br>For two independent samples, the appropriate test for the difference between two population means is a two-sample $t$-test for a difference between two population means.<br><br>4.9.A.2<br>The parameter for the two-sample $t$-test for the difference between two population means should reference the difference in means of the populations and the response variable in context. |
| 2.E | 4.9.B<br>Identify the null and alternative hypotheses for the difference between two population means. | 4.9.B.1<br>The null hypothesis for a two-sample $t$-test for the difference between two population means, $\mu_1$ and $\mu_2$, can be written as either: $H_0: \mu_1 - \mu_2 = 0$ or $H_0: \mu_1 = \mu_2$. A one-sided alternative hypothesis for the difference between population means can be written as either $H_a: \mu_1 < \mu_2$ (or equivalently $H_a: \mu_1 - \mu_2 < 0$) or $H_a: \mu_1 > \mu_2$ (or equivalently $H_a: \mu_1 - \mu_2 > 0$). A two-sided alternative hypothesis for the difference between population means can be written as $H_a: \mu_1 \neq \mu_2$ (or equivalently $H_a: \mu_1 - \mu_2 \neq 0$). |
| 4.E | 4.9.C<br>Justify the appropriateness of a hypothesis test for the difference between two population means by verifying conditions. | 4.9.C.1<br>A two-sample $t$-test for a difference between population means requires that three conditions be met as follows:<br>i. The randomization condition: the data should be collected using two random samples or a randomized experiment.<br>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.)<br>iii. The sample data condition: Both samples should have a sample size greater than or equal to 30. 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 Sample Means

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.E | 4.10.A<br>Calculate an appropriate test statistic and $p$-value for testing a hypothesis for the difference between two population means. | 4.10.A.1<br>The test statistic for a two-sample $t$- test for the difference between two population means is $t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$. The $t$-statistic has a $t$-distribution when the null hypothesis is true. The $t$-statistics and the degrees of freedom, which fall between $n_1 + n_2 - 2$ and the smaller of $n_1 - 1$ and $n_2 - 1$, can be found with technology.<br>4.10.A.2<br>The $p$-value for a two-sample $t$-test for the difference between two population means is found using the appropriate $t$-distribution table or from the appropriate $t$-distribution using technology. |
| 4.F | 4.10.B<br>Interpret the $p$-value of a hypothesis test for the difference between two population means. | 4.10.B.1<br>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). |

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.G | 4.10.C<br>Justify a claim about the population based on the results of a hypothesis test for the difference between two population means in context. | 4.10.C.1<br>A formal decision explicitly compares the $p$-value to the significance level, $\alpha$. If the $p\text{-value} \leq \alpha$ then reject the null hypothesis, $H_0 : \mu_1 - \mu_2 = 0$ or $H_0 : \mu_1 = \mu_2$. If the $p\text{-value} > \alpha$ then fail to reject the null hypothesis.<br>4.10.C.2<br>The results of a hypothesis test for a two-sample test for a difference between two population means can serve as the statistical reasoning to support the answer to a research question about the populations that were sampled.<br>4.10.C.3<br>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

Instructional Periods: 1

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.A | 5.1.A<br>Construct scatterplots depicting the distribution of two numeric variables. | 5.1.A.1<br>A bivariate quantitative data set consists of observations of ordered pairs from two quantitative variables, acquired from the same individuals in a sample or in a population, that can be used to construct a scatterplot.<br>5.1.A.2<br>A scatterplot shows the values of 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. |
| 4.A | 5.1.B<br>Describe the characteristics of a scatterplot. | 5.1.B.1<br>A description of the association shown in a scatterplot includes form, direction, strength, and unusual features.<br>5.1.B.2<br>The form of the association shown in a scatterplot, if any, can be described as linear or nonlinear.<br>5.1.B.3<br>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.<br>5.1.B.4<br>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.<br>5.1.B.5<br>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. |

# TOPIC 5.2
# Correlation

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 4.D | 5.2.A<br>Interpret the correlation for a linear relationship. | 5.2.A.1<br>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.<br>5.2.A.2<br>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.<br>5.2.A.3<br>A correlation coefficient close to $-1$ or $1$ does not necessarily mean that a linear model is appropriate.<br>5.2.A.4<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.B | 5.3.A<br>Calculate a predicted response value using a linear regression model. | 5.3.A.1<br>If the form of the relationship between $x$ and $y$ appears linear, we can approximate the relationship between $x$ and $y$ using a simple linear regression model, which is a linear equation that uses an explanatory variable, $x$, to predict the response variable, $y$.<br><br>5.3.A.2<br>In a simple linear regression model, the predicted response value, denoted by $\hat{y}$, is calculated as $\hat{y} = a + bx$, where $a$ is the $y$-intercept and $b$ is the slope of the regression line and $x$ is the explanatory variable.<br><br>5.3.A.3<br>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.<br><br>5.3.A.4<br>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

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.B | 5.4.A<br>Calculate the differences between actual and predicted values. | 5.4.A.1<br>A residual is the difference between the observed response value and the predicted response value for the given value of the explanatory variable: residual = $y - \hat{y}$ or (residual = observed $y$ – predicted $y$). |
| 4.D | 5.4.B<br>Interpret the differences between actual and predicted values. | 5.4.B.1<br>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. |
| 4.A | 5.4.C<br>Describe the form of association of bivariate data using residual plots. | 5.4.C.1<br>A residual plot is a scatterplot of the residuals versus the predicted response values (or the explanatory variable values).<br>5.4.C.2<br>Residual plots can be used to investigate the appropriateness of the simple linear regression model for the observed data.<br>5.4.C.3<br>The simple linear regression model should only be fit to the data if the data exhibits a linear trend. Apparent randomness in a residual plot for a simple 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.<br>5.4.C.4<br>Curvature in the residual plot for a simple linear regression model suggests that the linear model is not the most appropriate model for the data. |

# TOPIC 5.5
# Least-Squares Regression

Instructional Periods: 2

| SKILL | LEARNING OBJECTIVE | ESSENTIAL KNOWLEDGE |
|---|---|---|
| 3.B | 5.5.A<br>Calculate the coefficients for the least-squares regression line model. | 5.5.A.1<br>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 $(\bar{x}, \bar{y})$.<br>5.5.A.2<br>The slope of the regression line, $b$, is calculated using technology.<br>5.5.A.3<br>The $y$-intercept of the regression line, $a$, is calculated using technology.<br>5.5.A.4<br>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. |
| 4.D | 5.5.B<br>Interpret coefficients for the least-squares regression line model. | 5.5.B.1<br>The coefficients of the least-squares regression line model (line of best fit) are the estimated slope, $b$, and the estimated $y$-intercept, $a$, because they are based on a sample of values.<br>5.5.B.2<br>The slope coefficient 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.<br>5.5.B.3<br>The $y$-intercept coefficient 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 since $x = 0$ 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 since it might be a negative value for a response variable that has no negative values, such as height. |