AP Statistics

Gloucester County Public Schools
Curriculum Reference and Pacing Guide

Based on the Collegeboard AP Statistics Course and Exam Description
(Effective Fall 2020)

GCPS Curriculum Guide AP Statistics

Table of Contents

GCPS Curriculum Guide AP Statistics

Introduction

The Gloucester County Public Schools *Curriculum Reference and Pacing Guide* serves as a companion document to the Advanced Placement (AP) Statistics Course and Exam Description and delineates in greater specificity the content that all teachers should teach and all students should learn. It serves as a guide for teachers when planning instruction and assessments.

The format of the *Curriculum Reference and Pacing Guide* facilitates teacher planning by identifying the key concepts, knowledge, and skills that should be the focus of instruction for each objective. The Curriculum Guide is divided by unit and ordered to match the established GCPS pacing. Each unit is divided into three parts: Standards, Content, and Instruction.

The Standards Information section includes the AP topics for the unit, as well as the anticipated pacing.

The Content section includes the essential skills along with strategies and resources to support them. It also includes the foundational objectives and/or future skills correlated to each SOL (Vertical Articulation), key vocabulary, essential questions, and key concepts that support successful instruction of the standard.

The Instruction section contains information to assist teachers with planning and implementing effective lesson plans. This section includes suggested assessment tools, instructional resources, common student misconceptions, and strategies for differentiating instruction. It also contains suggestions for incorporating the Mathematical Process Goals for Students: becoming mathematical problem solvers, communicating mathematically, reasoning mathematically, making mathematical connections, and using mathematical representations to model and interpret practical situations.

AP Statistics Course Description

The AP Statistics course is composed of the seven units aligned with the Stats Modeling The World textbook and throughout it the three big ideas are emphasized: Variation and Distribution (VAR); Patterns and Uncertainty (UNC); Data-Based Predictions, Decisions, and Conclusions (DAT).
Specific course skills will be developed throughout the course to support learning in the following categories: Selecting Statistical Methods, Data Analysis, Using Probability and Simulation, and Statistical Argumentation. The content and conceptual understandings meet the typical expectations that colleges and universities expect students to master to qualify for college credit and/or placement.

BIG IDEA 1: VARIATION AND DISTRIBUTION (VAR) The distribution of measures for individuals within a sample or population describes variation. The value of a statistic varies from sample to sample. How can we determine whether differences between measures represent random variation or meaningful distinctions? Statistical methods based on probabilistic reasoning provide the basis for shared understandings about variation and about the likelihood that variation between and among measures, samples, and populations is random or meaningful.

BIG IDEA 2: PATTERNS AND UNCERTAINTY (UNC) Statistical tools allow us to represent and describe patterns in data and to classify departures from patterns. Simulation and probabilistic reasoning allow us to anticipate patterns in data and to determine the likelihood of errors in inference.

BIG IDEA 3: DATA-BASED PREDICTIONS, DECISIONS, AND CONCLUSIONS (DAT) Data-based regression models describe relationships between variables and are a tool for making predictions for values of a response variable. Collecting data using random sampling or randomized experimental design means that findings may be generalized to the part of the population from which the selection was made. Statistical inference allows us to make data-based decisions.

Source: AP Statistics Course and Exam Description - Fall 2020, Collegeboard © 2020

GCPS Curriculum Guide AP Statistics

GCPS AP Statistics Course Overview

Textbook: Bock, Velleman, & DeVeaux Stats Modeling the World, 2nd ed, 2007

High School Pacing Outline (Semester-long)

First Quarter / Third Quarter	Second Quarter / Fourth Quarter
Unit 1: Exploring and Understanding Data **Unit 2:** Exploring Relationships Between Variables **Unit 3:** Gathering Data **Unit 4:** Randomness and Probability	**Unit 4:** Randomness and Probability **Unit 5:** From the Data at Hand to the World at Large **Unit 6**: Learning About the World **Unit 7:** Inference When Variables are Related

Exam Weighting for the Multiple-Choice Section of the AP Exam

Exploring One-Variable Data	15-23%
Exploring Two-Variable Data	5-7%
Collecting Data	12-15%
Probability, Random Variables, and Probability Distributions	10-20%
Sampling Distributions	7-12%
Inference for Categorical Data: Proportions	12-15%
Inference for Quantitative Data: Means	10-18%
Inference for Categorical Data: Chi-Square	2-5%
Inference for Quantitative Data: Slopes	2-5%

GCPS Curriculum Guide AP Statistics

GCPS AP Statistics Pacing Guide

Unit	Suggested # of days Semester	Enduring Understanding	Learning Objective Content	Textbook Correlation
Unit 1 **Exploring and Understanding Data**	17	VAR-1 UNC-1 VAR-2	Given that variation may be random or not, conclusions are uncertain. Graphical representations and statistics allow us to identify and represent key features of data. The normal distribution can be used to represent some population distributions.	Ch. 2-6
Unit 2 **Exploring Relationships Between Variables**	11	VAR-1 UNC-1 DAT-1	Given that variation may be random or not, conclusions are uncertain. Graphical representations and statistics allow us to identify and represent key features of data. Regression models may allow us to predict responses to changes in an explanatory variable	Ch. 7-10
Unit 3 **Gathering Data**	9	VAR-1 DAT-2 VAR-3	Given that variation may be random or not, conclusions are uncertain. The way we collect data influences what we can and cannot say about a population. Well-designed experiments can establish evidence of causal relationships.	Ch. 11-13
Unit 4 **Randomness and Probability**	8	VAR-1 UNC-2 VAR-4 VAR-5 UNC-3	Given that variation may be random or not, conclusions are uncertain. Simulation allows us to anticipate patterns in data. The likelihood of a random event can be quantified. Probability distributions may be used to model variation in populations. Probabilistic reasoning allows us to anticipate patterns in data.	Ch. 14-17
Unit 5 **From the Data at Hand to the World at Large**	10	VAR-1 VAR-6	Given that variation may be random or not, conclusions are uncertain. The normal distribution may be used to model variation.	Ch. 18-22

		UNC-3 UNC-4 DAT-3 UNC-5	Probabilistic reasoning allows us to anticipate patterns in data An interval of values should be used to estimate parameters, in order to account for uncertainty. Significance testing allows us to make decisions about hypotheses within a particular context. Probabilities of Type I and Type II errors influence inference.	
Unit 6 Learning About the World	10	VAR-1 VAR-7 UNC-4 DAT-3	Given that variation may be random or not, conclusions are uncertain. The t-distribution may be used to model variation. An interval of values should be used to estimate parameters, in order to account for uncertainty. Significance testing allows us to make decisions about hypotheses within a particular context.	Ch. 23-25
Unit 7 Inference When Variables are Related	10	VAR-1 VAR-8 DAT-3 UNC-4 VAR-7	Given that variation may be random or not, conclusions are uncertain. The chi-square distribution may be used to model variation. Significance testing allows us to make decisions about hypotheses within a particular context. An interval of values should be used to estimate parameters, in order to account for uncertainty The t-distribution may be used to model variation.	Ch. 26-27
Review & Testing			Unit Tests, AP Test, Final Exam	

GCPS Curriculum Guide AP Statistics

Unit 1 Exploring and Understanding Data	Essential Knowledge and Skills and Key Instructional Information

Standards	The student will use <u>problem solving</u>, <u>mathematical communication</u>, <u>mathematical reasoning</u>, <u>connections</u>, and <u>representations</u> to:
Focus: Displaying, Interpreting, Analyzing, and Comparing Data **Enduring Understanding:** VAR-1 Given that variation may be random or not, conclusions are uncertain. UNC-1 Graphical representations and statistics allow us to identify and represent key features of data. VAR-2 The normal distribution can be used to represent some population distributions. **Anticipated Pacing:** Semester: 17 days	**Content** **Learning Objective** VAR-1 Given that variation may be random or not, conclusions are uncertain. ● Identify questions to be answered, based on variation in one-variable data ● Identify variables in a set of data. ● Classify types of variables. UNC-1 Graphical representations and statistics allow us to identify and represent key features of data. ● Represent categorical data using frequency or relative frequency tables. ● Describe categorical data represented in frequency or relative tables. ● Represent categorical data graphically. ● Describe categorical data represented graphically. ● Compare multiple sets of categorical data. ● Classify types of quantitative variables ● Represent quantitative data graphically. ● Describe the characteristics of quantitative data distributions. ● Calculate measures of center and position for quantitative data. ● Calculate measures of variability for quantitative data ● Explain the selection of a particular measure of center and/or variability for describing a set of quantitative data. ● Represent summary statistics for quantitative data graphically ● Describe summary statistics of quantitative data represented graphically. ● Compare graphical representations for multiple sets of quantitative data. ● Compare summary statistics for multiple sets of quantitative data.

- Compare numerical and graphical representations for two categorical variables.
- Calculate statistics for two categorical variables
- Compare statistics for two categorical variables.

VAR-2 The normal distribution can be used to represent some population distributions.

- Compare a data distribution to the normal distribution model.
- Determine proportions and percentiles from a normal distribution.
- Compare measures of relative position in data sets

Essential Knowledge:

- A variable is a characteristic that changes from one individual to another.
- A categorical variable takes on values that are category names or group labels.
- A quantitative variable is one that takes on numerical values for a measured or counted quantity
- A frequency table gives the number of cases falling into each category. A relative frequency table gives the proportion of cases falling into each category.
- Percentages, relative frequencies, and rates all provide the same information as proportions.
- Counts and relative frequencies of categorical data reveal information that can be used to justify claims about the data in context.
- Bar charts (or bar graphs) are used to display frequencies (counts) or relative frequencies (proportions) for categorical data.
- The height or length of each bar in a bar graph corresponds to either the number or proportion of observations falling within each category
- There are many additional ways to represent frequencies (counts) or relative frequencies (proportions) for categorical data.
- Graphical representations of a categorical variable reveal information that can be used to justify claims about the data in context.
- Frequency tables, bar graphs, or other representations can be used to compare two or more data sets in terms of the same categorical variable.
- A discrete variable can take on a countable number of values. The number of values may be finite or countably infinite, as with the counting numbers.
- A continuous variable can take on infinitely many values, but those values cannot be counted. No matter how small the interval between two values of a continuous variable, it is always possible to determine another value between them.

- In a histogram, the height of each bar shows the number or proportion of observations that fall within the interval corresponding to that bar. Altering the interval widths can change the appearance of the histogram.
- In a stem and leaf plot, each data value is split into a “stem” (the first digit or digits) and a “leaf” (usually the last digit).
- A dotplot represents each observation by a dot, with the position on the horizontal axis corresponding to the data value of that observation, with nearly identical values stacked on top of each other.
- A cumulative graph represents the number or proportion of a data set less than or equal to a given number.
- There are many additional ways to graphically represent distributions of quantitative data.
- Descriptions of the distribution of quantitative data include shape, center, and variability (spread), as well as any unusual features such as outliers, gaps, clusters, or multiple peaks.
- Outliers for one-variable data are data points that are unusually small or large relative to the rest of the data.
- A distribution is skewed to the right (positive skew) if the right tail is longer than the left. A distribution is skewed to the left (negative skew) if the left tail is longer than the right. A distribution is symmetric if the left half is the mirror image of the right half.
- Univariate graphs with one main peak are known as unimodal. Graphs with two prominent peaks are bimodal. A graph where each bar height is approximately the same (no prominent peaks) is approximately uniform.
- A gap is a region of a distribution between two data values where there are no observed data.
- Clusters are concentrations of data usually separated by gaps.
- Descriptive statistics does not attribute properties of a data set to a larger population, but may provide the basis for conjectures for subsequent testing.
- A statistic is a numerical summary of sample data.
- The mean is the sum of all the data values divided by the number of values. For a sample, the mean is denoted by x-bar: $\bar{x} = \frac{1}{n}\sum_{i-1}^{n} x_i$, where x_i represents the i th data point in the sample and n represents the number of data values in the sample.
- The median of a data set is the middle value when data are ordered. When the number of data points is even, the median can take on any value between the two middle values. In AP Statistics, the most commonly used value for the median of a data set with an even number of values is the average of the two middle values
- The first quartile, Q1, is the median of the half of the ordered data set from the minimum to the position of the median. The third quartile, Q3, is the median of the half of the ordered data set from the position of the median to the maximum. Q1 and Q3 form the boundaries for the middle 50% of values in an ordered data set.
- The pth percentile is interpreted as the value that has p% of the data less than or equal to it.
- Three commonly used measures of variability (or spread) in a distribution are the range, interquartile range, and standard deviation.
- The range is defined as the difference between the maximum data value and the minimum data value. The

interquartile range (IQR) is defined as the difference between the third and first quartiles: Q3 − Q1. Both the range and the interquartile range are possible ways of measuring variability of the distribution of a quantitative variable.

- Standard deviation is a way to measure variability of the distribution of a quantitative variable. For a sample, the standard deviation is denoted by s: $s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$. The square of the sample standard deviation, s^2 , is called the sample variance.
- Changing units of measurement affects the values of the calculated statistics.
- There are many methods for determining outliers. Two methods frequently used in this course are:

  i. An outlier is a value greater than 1.5 × IQR above the third quartile or more than 1.5 × IQR below the first quartile.

  ii. An outlier is a value located 2 or more standard deviations above, or below, the mean.

- The mean, standard deviation, and range are considered nonresistant (or non-robust) because they are influenced by outliers. The median and IQR are considered resistant (or robust), because outliers do not greatly (if at all) affect their value.
- Taken together, the minimum data value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum data value make up the five-number summary.
- A boxplot is a graphical representation of the five-number summary (minimum, first quartile, median, third quartile, maximum). The box represents the middle 50% of data, with a line at the median and the ends of the box corresponding to the quartiles. Lines (“whiskers”) extend from the quartiles to the most extreme point that is not an outlier, and outliers are indicated by their own symbol beyond this.
- Summary statistics of quantitative data, or of sets of quantitative data, can be used to justify claims about the data in context.
- If a distribution is relatively symmetric, then the mean and median are relatively close to one another. If a distribution is skewed right, then the mean is usually to the right of the median. If the distribution is skewed left, then the mean is usually to the left of the median.
- Any of the graphical representations, e.g., histograms, side-by-side boxplots, etc., can be used to compare two or more independent samples on center, variability, clusters, gaps, outliers, and other features.
- Any of the numerical summaries (e.g., mean, standard deviation, relative frequency, etc.) can be used to compare two or more independent samples.
- A parameter is a numerical summary of a population.
- Some sets of data may be described as approximately normally distributed. A normal curve is mound-shaped

and symmetric. The parameters of a normal distribution are the population mean, μ, and the population standard deviation, σ.

- 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
- Many variables can be modeled by a normal distribution.
- A standardized score for a particular data value is calculated as (data value – mean)/(standard deviation), and measures the number of standard deviations a data value falls above or below the mean.
- One example of a standardized score is a z-score, which is calculated as $z - score = (\frac{x_i - \mu}{\sigma})$. A z-score measures how many standard deviations a data value is from the mean.
- Technology, such as a calculator, a standard normal table, or computer-generated output, can be used to find the proportion of data values located on a given interval of a normally distributed random variable.
- Given the area of a region under the graph of the normal distribution curve, it is possible to use technology, such as a calculator, a standard normal table, or computer-generated output, to estimate parameters for some populations.
- Percentiles and z-scores may be used to compare relative positions of points within a data set or between data sets.
- Side-by-side bar graphs, segmented bar graphs, and mosaic plots are examples of bar graphs for one categorical variable, broken down by categories of another categorical variable.
- Graphical representations of two categorical variables can be used to compare distributions and/or determine if variables are associated.
- A two-way table, also called a contingency table, is used to summarize two categorical variables. The entries in the cells can be frequency counts or relative frequencies.
- A joint relative frequency is a cell frequency divided by the total for the entire table.
- The marginal relative frequencies are the row and column totals in a two-way table divided by the total for the entire table.
- A conditional relative frequency is a relative frequency for a specific part of the contingency table (e.g., cell frequencies in a row divided by the total for that row).
- Summary statistics for two categorical variables can be used to compare distributions and/or determine if variables are associated.

Vocabulary

- Variable
- Quantitative
- Categorical
- Identifier
- Data
- Sample
- Population
- Frequency Table
- Bar Chart
- Pie Chart
- Contingency Table
- Marginal Distribution
- Conditional Distribution
- Independence
- Simpson's Paradox
- Dotplot
- Histogram
- Stem and Leaf Plot
- Modes
- Symmetric
- Skewed
- Outliers
- Gaps
- Median
- Mean
- Range
- Variance
- Standard Deviation
- Boxplot
- Z-score
- Distribution
- Model

Instruction

Assessment Tools

Pre/Post Assessments:

- Unit Test

Textbook

- Ch. 2-6

Investigative Tasks

- Race and the Death Penalty
- Dollars for Students
- Auto Safety
- SUV Insurance
- Normal Models

Unit 2 Exploring Relationships Between Variables	Essential Knowledge and Skills and Key Instructional Information

<table>
<tr><th>Standards</th><th>The student will use <u>problem solving</u>, <u>mathematical communication</u>, <u>mathematical reasoning</u>, <u>connections</u>, and <u>representations</u> to:</th></tr>
<tr><td>Focus:
Scatterplots and Modeling Data

Enduring Understanding:
Var-1 Given that variation may be random or not, conclusions are uncertain.

UNC-1 Graphical representations and statistics allow us to identify and represent key features of data.

DAT-1 Regression models may allow us to predict responses to changes in an explanatory variable

Anticipated Pacing:
Semester: 11 days</td>
<td>
<table>
<tr><th>Content</th></tr>
<tr><th>Learning Objective</th></tr>
<tr><td>Var-1 Given that variation may be random or not, conclusions are uncertain.
Identify questions to be answered about possible relationships in data.</td></tr>
<tr><td>UNC-1 Graphical representations and statistics allow us to identify and represent key features of data.
Represent bivariate quantitative data using scatterplots.</td></tr>
<tr><td>DAT-1 Regression models may allow us to predict responses to changes in an explanatory variable
Describe the characteristics of a scatter plot.
Determine the correlation for a linear relationship.
Interpret the correlation for a linear relationship.
Calculate a predicted response value using a linear regression model.
Represent differences between measured and predicted responses using residual plots.
Describe the form of association of bivariate data using residual plots.
Estimate parameters for the least-squares regression line model.
Interpret coefficients for the least-squares regression line model.
Identify influential points in regression.
Calculate a predicted response using a leastsquares regression line for a transformed data set.</td></tr>
</table>
</td></tr>
</table>

Essential Knowledge:

- Apparent patterns and associations in data may be random or not.
- A bivariate quantitative data set consists of observations of two different quantitative variables made on individuals in a sample or population.
- A scatterplot shows two numeric values for each observation, one corresponding to the value on the x-axis and one corresponding to the value on the y-axis.
- An explanatory variable is a variable whose values are used to explain or predict corresponding values for the response variable
- A description of a scatter plot includes form, direction, strength, and unusual features.
- 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 one variable increase, the values of the other variable tend to increase. A negative association means that as values of one variable increase, values of the other variable tend to decrease.
- The form of the association shown in a scatterplot, if any, can be described as linear or non-linear to varying degrees.
- The strength of the association is how closely the individual points follow a specific pattern, e.g., linear, and can be shown in a scatterplot. Strength can be described as strong, moderate, or weak.
- Unusual features of a scatter plot include clusters of points or points with relatively large discrepancies between the value of the response variable and a predicted value for the response variable.
- The correlation, r, gives the direction and quantifies the strength of the linear association between two quantitative variables.
- The correlation coefficient can be calculated by: $r = \frac{1}{n-1}\sum(\frac{x_i - \bar{x}}{s_x})(\frac{y_1 - \bar{y}}{s_y})$. However, the most common way to determine r is by using technology
- A correlation coefficient close to 1 or −1 does not necessarily mean that a linear model is appropriate.
- The correlation, r, is unit-free, and always between −1 and 1, inclusive. A value of r 0 indicates that there is no linear association. A value of r 1 or r 1 indicates that there is a perfect linear association.
- 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.

	• A simple linear regression model is an equation that uses an explanatory variable, x, to predict the response variable, y. • 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 value of the explanatory variable • Extrapolation is predicting a response value using a value for the explanatory variable that is beyond the interval of x-values used to determine the regression line. The predicted value is less reliable as an estimate the further we extrapolate • The residual is the difference between the actual value and the predicted value: residual = y − $\hat{y}$. • A residual plot is a plot of residuals versus explanatory variable values or predicted response values. • Apparent randomness in a residual plot for a linear model is evidence of a linear form to the association between the variables. • Residual plots can be used to investigate the appropriateness of a selected model. • The least-squares regression model minimizes the sum of the squares of the residuals and contains the point ($\bar{x} - \bar{y}$) • The slope, b, of the regression line can be calculated as $b = r(\frac{s_y}{s_x})$where r is the correlation between x and y, s y is the sample standard deviation of the response variable, y, and sx is the sample standard deviation of the explanatory variable, x. • Sometimes, the y-intercept of the line does not have a logical interpretation in context. • In simple linear regression, $r^2$is the square of the correlation, r. It is also called the coefficient of determination. r^2 is the proportion of variation in the response variable that is explained by the explanatory variable in the model. • The coefficients of the least-squares regression model are the estimated slope and y-intercept. • The slope is the amount that the predicted y-value changes for every unit increase in x. • The y-intercept value is the predicted value of the response variable when the explanatory variable is equal to 0. The formula for the y-intercept, a, is $a = \bar{y} - b\bar{x}$ • An outlier in regression is a point that does not follow the general trend shown in the rest of the data and has a large residual when the Least Squares Regression Line (LSRL) is calculated. • A high-leverage point in regression has a substantially larger or smaller x-value than the other observations have. • An influential point in regression is any point that, if removed, changes the relationship substantially. Examples include much different slope, y-intercept, and/or correlation. Outliers and high leverage points are often influential. • Transformations of variables, such as evaluating the natural logarithm of each value of the response variable or squaring each value of the explanatory variable, can be used to create transformed data sets, which may be

more linear in form than the untransformed data.

- Increased randomness in residual plots after transformation of data and/or movement of r^2 to a value closer to 1 offers evidence that the least-squares regression line for the transformed data is a more appropriate model to use to predict responses to the explanatory variable than the regression line for the untransformed data.

Vocabulary

- Association
- Direction
- Positive Direction
- Negative Direction
- Linear
- Curved
- Strength
- Outliers
- Clusters
- Subgroups
- Variables
- Scatterplot
- Explanatory/predictor variable
- Response variable
- Correlation
- Regression
- Prediction
- Correlation coefficient
- Quantitative variables condition
- Straight enough condition
- Outlier condition
- Lurking variable
- Standardize
- Linear model
- Residual
- Model
- Line of best fit/least squares line
-
- Residuals
- Equal variance assumption
- Extrapolation
- Influential points

Instruction

Assessment Tools

Pre/Post Assessments:
- Unit Test

Textbook
- Ch. 7-10

Investigative Tasks

	• Smoking • Olympic Long Jumps • Alligators

Unit 3 Gathering Data	Essential Knowledge and Skills and Key Instructional Information

Standards	The student will use problem solving, mathematical communication, mathematical reasoning, connections, and representations to:
Focus: Collecting data, conducting experiments, and observations **Enduring Understanding:** Var-1 Given that variation may be random or not, conclusions are uncertain. DAT-2 The way we collect data influences what we can and cannot say about a population. VAR-3 Well-designed experiments can establish evidence of causal relationships. **Anticipated Pacing:** Semester: 9 days	**Content** **Learning Objective** Var-1 Given that variation may be random or not, conclusions are uncertain. ● Identify questions to be answered about data collection methods. DAT-2 The way we collect data influences what we can and cannot say about a population. ● Identify the type of a study. ● Identify appropriate generalizations and determinations based on observational studies. ● Identify a sampling method, given a description of a study. ● Explain why a particular sampling method is or is not appropriate for a given situation. ● Identify potential sources of bias in sampling methods. VAR-3 Well-designed experiments can establish evidence of causal relationships. ● Identify the components of an experiment. ● Describe elements of a well-designed experiment. ● Compare experimental designs and methods. ● Explain why a particular experimental design is appropriate. ● Interpret the results of a well-designed experiment. **Essential Knowledge:** ● Methods for data collection that do not rely on chance result in untrustworthy conclusions. ● A population consists of all items or subjects of interest.

- A sample selected for study is a subset of the population.
- In an observational study, treatments are not imposed. Investigators examine data for a sample of individuals (retrospective) or follow a sample of individuals into the future collecting data (prospective) in order to investigate a topic of interest about the population. A sample survey is a type of observational study that collects data from a sample in an attempt to learn about the population from which the sample was taken.
- In an experiment, different conditions (treatments) are assigned to experimental units (participants or subjects).
- It is only appropriate to make generalizations about a population based on samples that are randomly selected or otherwise representative of that population.
- A sample is only generalizable to the population from which the sample was selected.
- It is not possible to determine causal relationships between variables using data collected in an observational study
- When an item from a population can be selected only once, this is called sampling without replacement. When an item from the population can be selected more than once, this is called sampling with replacement.
- A simple random sample (SRS) is a sample in which every group of a given size has an equal chance of being chosen. This method is the basis for many types of sampling mechanisms. A few examples of mechanisms used to obtain SRSs include numbering individuals and using a random number generator to select which ones to include in the sample, ignoring repeats, using a table of random numbers, or drawing a card from a deck without replacement.
- A stratified random sample involves the division of a population into separate groups, called strata, based on shared attributes or characteristics (homogeneous grouping). Within each stratum a simple random sample is selected, and the selected units are combined to form the sample.
- A cluster sample involves the division of a population into smaller groups, called clusters. Ideally, there is heterogeneity within each cluster, and clusters are similar to one another in their composition. A simple random sample of clusters is selected from the population to form the sample of clusters. Data are collected from all observations in the selected clusters.
- 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.
- A census selects all items/subjects in a population.
- There are advantages and disadvantages for each sampling method depending upon the question that is to be answered and the population from which the sample will be drawn.
- Bias occurs when certain responses are systematically favored over others.
- When a sample is comprised entirely of volunteers or people who choose to participate, the sample will typically not be representative of the population (voluntary response bias).
- When part of the population has a reduced chance of being included in the sample, the sample will typically not be representative of the population (undercoverage bias).

- Individuals chosen for the sample for whom data cannot be obtained (or who refuse to respond) may differ from those for whom data can be obtained (nonresponse bias).
- Problems in the data gathering instrument or process result in response bias. Examples include questions that are confusing or leading (question wording bias) and self-reported responses.
- Non-random sampling methods (for example, samples chosen by convenience or voluntary response) introduce potential for bias because they do not use chance to select the individuals.
- The experimental units are the individuals (which may be people or other objects of study) that are assigned treatments. When experimental units consist of people, they are sometimes referred to as participants or subjects.
- An explanatory variable (or factor) in an experiment is a variable whose levels are manipulated intentionally. The levels or combination of levels of the explanatory variable(s) are called treatments.
- A response variable in an experiment is an outcome from the experimental units that is measured after the treatments have been administered.
- A confounding variable in an experiment is a variable that is related to the explanatory variable and influences the response variable and may create a false perception of association between the two.
- A well-designed experiment should include the following: a. Comparisons of at least two treatment groups, one of which could be a control group. b. Random assignment/allocation of treatments to experimental units. c. Replication (more than one experimental unit in each treatment group). d. Control of potential confounding variables where appropriate.
- In a completely randomized design, treatments are assigned to experimental units completely at random. Random assignment tends to balance the effects of uncontrolled (confounding) variables so that differences in responses can be attributed to the treatments.
- Methods for randomly assigning treatments to experimental units in a completely randomized design include using a random number generator, a table of random values, drawing chips without replacement, etc.
- In a single-blind experiment, subjects do not know which treatment they are receiving, but members of the research team do, or vice versa.
- In a double-blind experiment neither the subjects nor the members of the research team who interact with them know which treatment a subject is receiving.
- A control group is a collection of experimental units either not given a treatment of interest or given a treatment with an inactive substance (placebo) in order to determine if the treatment of interest has an effect.
- The placebo effect occurs when experimental units have a response to a placebo
- For randomized complete block designs, treatments are assigned completely at random within each block.
- Blocking ensures that at the beginning of the experiment the units within each block are similar to each other with respect to at least one blocking variable. A randomized block design helps to separate natural variability from differences due to the blocking variable.

- A matched pairs design is a special case of a randomized block design. Using a blocking variable, subjects (whether they are people or not) are arranged in pairs matched on relevant factors. Matched pairs may be formed naturally or by the experimenter. Every pair receives both treatments by randomly assigning one treatment to one member of the pair and subsequently assigning the remaining treatment to the second member of the pair. Alternately, each subject may get both treatments.
- There are advantages and disadvantages for each experimental design depending on the question of interest, the resources available, and the nature of the experimental units.
- Statistical inference attributes conclusions based on data to the distribution from which the data were collected.
- Random assignment of treatments to experimental units allows researchers to conclude that some observed changes are so large as to be unlikely to have occurred by chance. Such changes are said to be statistically significant.
- Statistically significant differences between or among experimental treatment groups are evidence that the treatments caused the effect.
- If the experimental units used in an experiment are representative of some larger group of units, the results of an experiment can be generalized to the larger group. Random selection of experimental units gives a better chance that the units will be representative.

Vocabulary

- Simulation
- Component
- Random
- Trial
- Randomness
- Bias
- Undercoverage
- Convenience sample
- Voluntary response
- Nonresponse
- Response bias
- SRS (simple random sample
- Stratified sample
- Cluster sample
- Multi-stage sample
- Systematic sample
- Observational study
- Retrospective study
-
- Prospective study
- Experiment
- Factor
- Response
- Experimental units
- level
- Treatment
- Control
- Randomize
- Replicate
- Block
- Control group
- Single-blinding
- Double-blinding
- Placebo
- Causation
- Lurking variable
- Confounding variable

	Instruction
	Assessment Tools
	Pre/Post Assessments: ● Data Collection Project ● Unit Test **Textbook** ● Ch. 11-13 **Investigative Tasks** ● ESP ● Backhoes and Forklifts

GCPS Curriculum Guide AP Statistics

<table>
<tr><th>Unit 4
Randomness and Probability</th><th>Essential Knowledge and Skills and Key Instructional Information</th></tr>
<tr><td>Standards</td><td>The student will use <u>problem solving</u>, <u>mathematical communication</u>, <u>mathematical reasoning</u>, <u>connections</u>, and <u>representations</u> to:</td></tr>
<tr><td>Focus:
Randomness, simulations, and probability calculations and distributions

Enduring Understanding:
VAR-1 Given that variation may be random or not, conclusions are uncertain.

UNC-2 Simulation allows us to anticipate patterns in data.

VAR-4 The likelihood of a random event can be quantified.

VAR-5 Probability distributions may be used to model variation in populations.

UNC-3 Probabilistic reasoning allows us to anticipate patterns in data.

Anticipated Pacing:
Semester: 8 days</td>
<td>
<table>
<tr><th>Content</th></tr>
<tr><th>Learning Objective</th></tr>
<tr><td>VAR-1 Given that variation may be random or not, conclusions are uncertain.
Identify questions suggested by patterns in data.</td></tr>
<tr><td>UNC-2 Simulation allows us to anticipate patterns in data.
Estimate probabilities using simulation.</td></tr>
<tr><td>VAR-4 The likelihood of a random event can be quantified.
Calculate probabilities for events and their complements.
Interpret probabilities for events.
Explain why two events are (or are not) mutually exclusive.
Calculate conditional probabilities.
Calculate probabilities for independent events and for the union of two events.</td></tr>
<tr><td>VAR-5 Probability distributions may be used to model variation in populations.
Represent the probability distribution for a discrete random variable.
Interpret a probability distribution.
Calculate parameters for a discrete random variable.
Interpret parameters for a discrete random variable.
Calculate parameters for linear combinations of random variables.
Describe the effects of linear transformations of parameters of random variables.</td></tr>
</table>
</td></tr>
</table>

UNC-3 Probabilistic reasoning allows us to anticipate patterns in data.

- Estimate probabilities of binomial random variables using data from a simulation.
- Calculate probabilities for a binomial distribution.
- Calculate parameters for a binomial distribution.
- Interpret probabilities and parameters for a binomial distribution.
- Calculate probabilities for geometric random variables.
- Calculate parameters of a geometric distribution.
- Interpret probabilities and parameters for a geometric distribution.

Essential Knowledge:

- Patterns in data do not necessarily mean that variation is not random.
- A random process generates results that are determined by chance.
- An outcome is the result of a trial of a random process.
- An event is a collection of outcomes.
- Simulation is a way to model random events, such that simulated outcomes closely match real-world outcomes. All possible outcomes are associated with a value to be determined by chance. Record the counts of simulated outcomes and the count total.
- The relative frequency of an outcome or event in simulated or empirical data can be used to estimate the probability of that outcome or event.
- The law of large numbers states that simulated (empirical) probabilities tend to get closer to the true probability as the number of trials increases.
- The sample space of a random process is the set of all possible non-overlapping outcomes.
- If all outcomes in the sample space are equally likely, then the probability an event E will occur is defined as the fraction: $\frac{number\ of\ outcomes\ in\ event\ E}{total\ number\ of\ outcomes\ in\ sample\ space}$
- The probability of an event is a number between 0 and 1, inclusive.
- The probability of the complement of an event E, E′ or E^c , (i.e., not E) is equal to 1 - P(E)
- Probabilities of events in repeatable situations can be interpreted as the relative frequency with which the event will occur in the long run.
- 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, denoted P(A ∩B).
- Two events are mutually exclusive or disjoint if they cannot occur at the same time. So P(A∩B)=0.
- The probability that event A will occur given that event B has occurred is called a conditional probability and

denoted $P(A|B) = \frac{P(A \cap B)}{P(B)}$

- The 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 probability that event B will occur, given that A has occurred. This is denoted $P(A \cap B) = P(A) * P(B|A)$.
- 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
- If, and only if, events A and B are independent, then $P(A|B) = P(A),\ P(B|A) = P(B),\ and\ P(A \cap B) = P(A) * P(B)$
- The probability that event A or event B (or both) will occur is the probability of the union of A and B, denoted P(A ∪ B) .
- The addition rule states that the probability that event A or event B or both will occur is equal to the probability that event A will occur plus the probability that event B will occur minus the probability that both events A and B will occur. This is denoted $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- The values of a random variable are the numerical outcomes of random behavior.
- A discrete random variable is a variable that can only take a countable number of values. Each value has a probability associated with it. The sum of the probabilities over all of the possible values must be 1.
- A probability distribution can be represented as a graph, table, or function showing the probabilities associated with values of a random variable.
- A cumulative probability distribution can be represented as a table or function showing the probability of being less than or equal to each value of the random variable.
- An interpretation of a probability distribution provides information about the shape, center, and spread of a population and allows one to make conclusions about the population of interest.
- A numerical value measuring a characteristic of a population or the distribution of a random variable is known as a parameter, which is a single, fixed value.
- The mean, or expected value, for a discrete random variable X is $\mu_x = \sum x_i \cdot P(x_i)$
- The standard deviation for a discrete random variable X is $\sigma^2 = \sqrt{\sum (x_i - \mu_x)^2 \cdot P(x_i)}$
- Parameters for a discrete random variable should be interpreted using appropriate units and within the context of a specific population.
- For random variables X and Y and real numbers a and b, the mean of $aX + bY\ is\ a\mu_x + b\mu_y$
- Two random variables are independent if knowing information about one of them does not change the probability distribution of the other

- For independent random variables X and Y and real numbers a and b, the mean of aX+bY is $a\mu_x + b\mu_y$and the variance of aX+bY is $a^2\sigma^2_x + b^2\sigma^2_y$
- For Y =a+bX, the probability distribution of the transformed random variable, Y, has the same shape as the probability distribution for X, so long as a > 0 and b > 0. The mean of Y is $\mu_y = a + b\mu_x$.The standard deviation of Y is $\sigma_y = |b|\sigma_x$
- A probability distribution can be constructed using the rules of probability or estimated with a simulation using random number generators.
- A binomial random variable, X, counts the number of successes in n repeated independent trials, each trial having two possible outcomes (success or failure), with the probability of success p and the probability of failure 1 – p.
- 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 $P(X = x) = (n\, x)p^x(1 - p)^{n-x},\ x = 0, 1, 2,..., n$. This is the binomial probability function.
- If a random variable is binomial, its mean, µx, is np and its standard deviation, σx, is $\sqrt{np(1 - p)}$
- Probabilities and parameters for a binomial distribution should be interpreted using appropriate units and within the context of a specific population or situation.
- For a sequence of independent trials, a geometric random variable, X, gives the number of the trial on which the first success occurs. Each trial has two possible outcomes (success or failure) with the probability of success p and the probability of failure 1 – p.
- The probability that the first success for repeated independent trials with probability of success p occurs on trial x is calculated as $P(X = x) = (1 - p)^{x-1}P,\ x = 1, 2, 3,...$This is the geometric probability function.
- If a random variable is geometric, its mean, µx, is $\frac{1}{p}$and its standard deviation, σ_x , is $\frac{\sqrt{(1-p)}}{p}$
- Probabilities and parameters for a geometric distribution should be interpreted using appropriate units and within the context of a specific population or situation.

Vocabulary

- Probability
- Law of large numbers
- Sample space
- Disjoint (mutually exclusive)
- Conditional probability
- General addition
- Variance
- Standard deviation
- Variance

- Law of averages
- Random phenomenon
- Trial
- Outcome
- Event
- independent
- Addition rule
- Multiplication rule
- Complement rule
- Venn diagram
- rule
- General multiplication rule
- Tree diagram
- Probability model
- Expected value
- Bernoulli trials
- Geometric model
- Binomial model
- 10% condition
- success/failure condition

Instruction

Assessment Tools

Pre/Post Assessments:

- Unit Test

Textbook

- Ch. 14-17

Unit 5 From the Data at Hand to the World at Large	Essential Knowledge and Skills and Key Instructional Information

Standards	The student will use problem solving, mathematical communication, mathematical reasoning, connections, and representations to:
Focus: Normal distribution, significance testing for proportions, confidence intervals for proportions **Enduring Understanding:** VAR-1 Given that variation may be random or not, conclusions are uncertain. VAR-6 The normal distribution may be used to model variation. UNC-3 Probabilistic reasoning allows us to anticipate patterns in data UNC-4 An interval of values should be used to estimate parameters, in order to account for uncertainty. DAT-3 Significance testing allows us to make decisions about hypotheses within a particular	**Content** **Learning Objective** VAR-1 Given that variation may be random or not, conclusions are uncertain. ● Identify questions suggested by variation in statistics for samples collected from the same population. ● Identify questions suggested by variation in the shapes of distributions of samples taken from the same population. VAR-6 The normal distribution may be used to model variation. ● Calculate the probability that a particular value lies in a given interval of a normal distribution. ● Determine the interval associated with a given area in a normal distribution. ● Determine the appropriateness of using the normal distribution to approximate probabilities for unknown distributions. ● Identify the null and alternative hypotheses for a population proportion. ● Identify an appropriate testing method for a population proportion. ● Verify the conditions for making statistical inferences when testing a population proportion. ● Calculate an appropriate test statistic and p-value for a population proportion. ● Identify the null and alternative hypotheses for a difference of two population proportions. ● Identify an appropriate testing method for the difference of two population proportions. ● Verify the conditions for making statistical inferences when testing a difference of two population proportions. ● Calculate an appropriate test statistic for the difference of two population proportions.

context. UNC-5 Probabilities of Type I and Type II errors influence inference. **Anticipated Pacing:** Semester: 10 days	UNC-3 Probabilistic reasoning allows us to anticipate patterns in data • Estimate sampling distributions using simulation. • Explain why an estimator is or is not unbiased. • Calculate estimates for a population parameter. • Determine parameters of a sampling distribution for sample proportions. • Determine whether a sampling distribution for a sample proportion can be described as approximately normal. • Interpret probabilities and parameters for a sampling distribution for a sample proportion. • Determine parameters of a sampling distribution for a difference in sample proportions • Determine whether a sampling distribution for a difference of sample proportions can be described as approximately normal. • Interpret probabilities and parameters for a sampling distribution for a difference in proportions. • Determine parameters for a sampling distribution for sample means. • Determine whether a sampling distribution of a sample mean can be described as approximately normal. • Interpret probabilities and parameters for a sampling distribution for a sample mean. • Determine parameters of a sampling distribution for a difference in sample means. • Determine whether a sampling distribution of a difference in sample means can be described as approximately normal. • Interpret probabilities and parameters for a sampling distribution for a difference in sample means.
	UNC-4 An interval of values should be used to estimate parameters, in order to account for uncertainty. • Identify an appropriate confidence interval procedure for a population proportion. • Verify the conditions for calculating confidence intervals for a population proportion. • Determine the margin of error for a given sample size and an estimate for the sample size that will result in a given margin of error for a population proportion. • Calculate an appropriate confidence interval for a population proportion. • Calculate an interval estimate based on a confidence interval for a population proportion. • Interpret a confidence interval for a population proportion. • Justify a claim based on a confidence interval for a population proportion. • Identify the relationships between sample size, width of a confidence interval, confidence level, and margin of error for a population proportion. • Identify an appropriate confidence interval procedure for a comparison of population proportions. • Verify the conditions for calculating confidence intervals for a difference between population proportions. • Calculate an appropriate confidence interval for a comparison of population proportions.

- Calculate an interval estimate based on a confidence interval for a difference of proportions.
- Interpret a confidence interval for a difference of proportions.
- Justify a claim based on a confidence interval for a difference of proportions.

DAT-3 Significance testing allows us to make decisions about hypotheses within a particular context.

- Interpret the p-value of a significance test for a population proportion.
- Justify a claim about the population based on the results of a significance test for a population proportion.
- Interpret the p-value of a significance test for a difference of population proportions.
- Justify a claim about the population based on the results of a significance test for a difference of population proportions.

UNC-5 Probabilities of Type I and Type II errors influence inference.

- Identify Type I and Type II errors.
- Calculate the probability of a Type I and Type II errors.
- Identify factors that affect the probability of errors in significance testing.
- Interpret Type I and Type II errors.

Essential Knowledge:

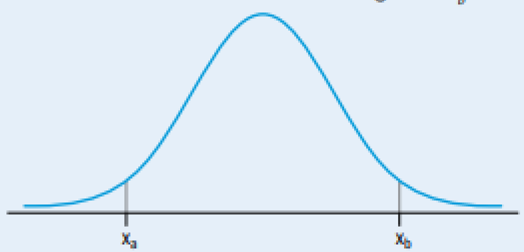
- Variation in statistics for samples taken from the same population may be random or not.
- A continuous random variable is a variable that can take on any value within a specified domain. Every interval within the domain has a probability associated with it.
- A continuous random variable with a normal distribution is commonly used to describe populations. The distribution of a normal random variable can be described by a normal, or “bell-shaped,” curve.
- The area under a normal curve over a given interval represents the probability that a particular value lies in that interval.
- The boundaries of an interval associated with a given area in a normal distribution can be determined using z-scores or technology, such as a calculator, a standard normal table, or computer-generated output.
- Intervals associated with a given area in a normal distribution can be determined by assigning appropriate inequalities to the boundaries of the intervals:

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

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

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

d. To determine the most extreme $p\%$ of values requires dividing the area associated with $p\%$ into two equal areas on either extreme of the distribution: $P(X < x_a) = \frac{1}{2}\frac{p}{100}$ and $P(X > x_b) = \frac{1}{2}\frac{p}{100}$ means 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.



- Normal distributions are symmetrical and “bell-shaped.” As a result, normal distributions can be used to approximate distributions with similar characteristics.
- A sampling distribution of a statistic is the distribution of values for the statistic for all possible samples of a given size from a given population.
- The central limit theorem (CLT) states that when the sample size is sufficiently large, a sampling distribution of the mean of a random variable will be approximately normally distributed.
- The central limit theorem requires that the sample values are independent of each other and that n is sufficiently large.
- A randomization distribution is a collection of statistics generated by simulation assuming known values for the parameters. For a randomized experiment, this means repeatedly randomly reallocating/reassigning the

response values to treatment groups.
- The sampling distribution of a statistic can be simulated by generating repeated random samples from a population.
- When estimating a population parameter, an estimator is unbiased if, on average, the value of the estimator is equal to the population parameter.
- When estimating a population parameter, an estimator exhibits variability that can be modeled using probability
- A sample statistic is a point estimator of the corresponding population parameter.
- For independent samples (sampling with replacement) of a categorical variable from a population with population proportion, p, the sampling distribution of the sample proportion, $\hat{p}$, has a mean, $\mu_{\hat{p}} = p$ and a standard deviation, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- If sampling without replacement, the standard deviation of the sample proportion is smaller than what is given by the formula above. If the sample size is less than 10% of the population size, the difference is negligible.
- For a categorical variable, the sampling distribution of the sample proportion, pˆ, will have an approximate normal distribution, provided the sample size is large enough: np ≥10 and n(1-p) ≥10
- Probabilities and parameters for a sampling distribution for a sample proportion should be interpreted using appropriate units and within the context of a specific population.
- For a categorical variable, when randomly sampling with replacement from two independent populations with population proportions $p_1 and\ p_2$, the sampling distribution of the difference in sample proportions $\widehat{p_1} - \widehat{p_2}$has mean, $\mu_{\widehat{p_1}-\widehat{p_2}} = p_1 - p_2$and standard deviation $\sigma_{\widehat{p_1}-\widehat{p_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
- If sampling without replacement, the standard deviation of the difference in sample proportions is smaller than what is given by the formula above. If the sample sizes are less than 10% of the population sizes, the difference is negligible.
- The sampling distribution of the difference in sample proportions $\widehat{p_1} - \widehat{p_2}$will have an approximate normal distribution provided the sample sizes are large enough:
$n_1p_1 \geq 10,\ n_1(1-p_1) \geq 10,\ n_2p_2 \geq 10,\ n_2(1-p_2) \geq 10$
- Parameters for a sampling distribution for a difference of proportions should be interpreted using appropriate units and within the context of a specific populations.
- For a numerical variable, when random sampling with replacement from a population with mean μ and standard deviation, σ , the sampling distribution of the sample mean has mean $\mu_{\bar{x}} = \mu$and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- If sampling without replacement, the standard deviation of the sample mean is smaller than what is given by the

formula above. If the sample size is less than 10% of the population size, the difference is negligible.

- For a numerical 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.
- For a numerical variable, if the population distribution cannot be modeled with a normal distribution, the sampling distribution of the sample mean,$\bar{x}$, can be modeled approximately by a normal distribution, provided the sample size is large enough, e.g., greater than or equal to 30.
- Probabilities and parameters for a sampling distribution for a sample mean should be interpreted using appropriate units and within the context of a specific population.
- For a numerical variable, when randomly sampling with replacement from two independent populations with population means μ1 and μ2 and population standard deviations σ 1 and σ 2, the sampling distribution of the difference in sample means $\bar{x}_1 - \bar{x}_2$ has mean $\mu_{(\bar{x}_1 - \bar{x}_2)} = \mu_1 - \mu_2$ and standard deviation

  $$\sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
- If sampling without replacement, the standard deviation of the difference in sample means is smaller than what is given by the formula above. If the sample sizes are less than 10% of the population sizes, the difference is negligible.
- The sampling distribution of the difference in sample means $\bar{x}_1 - \bar{x}_2$ can be modeled with a normal distribution if the two population distributions can be modeled with a normal distribution.
- The sampling distribution of the difference in sample means$\bar{x}_1 - \bar{x}_2$ can be modeled approximately by a normal distribution if the two population distributions cannot be modeled with a normal distribution but both sample sizes are greater than or equal to 30
- Probabilities and parameters for a sampling distribution for a difference of sample means should be interpreted using appropriate units and within the context of a specific populations.
- Variation in shapes of data distributions may be random or not.
- The appropriate confidence interval procedure for a one-sample proportion for one categorical variable is a one sample z-interval for a proportion.
- In order to make assumptions necessary for inference on population proportions, means, and slopes, we must check for independence in data collection methods and for selection of the appropriate sampling distribution.
- In order to calculate a confidence interval to estimate a population proportion, p, we must check for independence and that the sampling distribution is approximately normal.

  a. To check for independence:

    i. Data should be collected using a random sample or a randomized experiment.

    ii. When sampling without replacement, check that n N ≤10% , where N is the size of the population.

b. To check that the sampling distribution of $\hat{p}$is approximately normal (shape):

i. For categorical variables, check that both the number of successes, n$\hat{p}$, and the number of failures, $n(1-\hat{p})$are at least 10 so that the sample size is large enough to support an assumption of normality

- Based on sample data, the standard error of a statistic is an estimate for the standard deviation for the statistic. The standard error of $\hat{p}$ is$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- A margin of error gives how much a value of a sample statistic is likely to vary from the value of the corresponding population parameter
- For categorical variables, the margin of error is the critical value (z*) times the standard error (SE) of the relevant statistic, which equals $z^{*}\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}$for a one sample proportion.
- The formula for margin of error can be rearranged to solve for n, the minimum sample size needed to achieve a given margin of error. For this purpose, use a guess for $\hat{p}$or use $\hat{p}$= 0.5 in order to find an upper bound for the sample size that will result in a given margin of error.
- In general, an interval estimate can be constructed as point estimate ± (margin of error). For a one-sample proportion, the interval estimate is $\hat{p} \pm z^{*}\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}$
- Critical values represent the boundaries encompassing the middle C% of the standard normal distribution, where C% is an approximate confidence level for a proportion
- Confidence intervals for population proportions can be used to calculate interval estimates with specified units.
- A confidence interval for a population proportion either contains the population proportion or it does not, because each interval is based on random sample data, which varies from sample to sample.
- We are C% confident that the confidence interval for a population proportion captures the population proportion.
- In repeated random sampling with the same sample size, approximately C% of confidence intervals created will capture the population proportion.
- Interpreting a confidence interval for a onesample proportion should include a reference to the sample taken and details about the population it represents.
- A confidence interval for a population proportion provides an interval of values that may provide sufficient evidence to support a particular claim in context.
- 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 population proportion, the width of the interval is proportional to $\frac{1}{\sqrt{n}}$

- For a given sample, the width of the confidence interval for a population proportion increases as the confidence level increases.
- The width of a confidence interval for a population proportion is exactly twice the margin of error.
- The null hypothesis is the situation that is assumed to be correct unless evidence suggests otherwise, and the alternative hypothesis is the situation for which evidence is being collected.
- For hypotheses about parameters, the null hypothesis contains an equality reference (=, ≥, or ≤), while the alternative hypothesis contains a strict inequality (<, >, or ≠). The type of inequality in the alternative hypothesis is based on the question of interest. Alternative hypotheses with < or > are called one-sided, and alternative hypotheses with ≠ are called twosided. Although the null hypothesis for a onesided test may include an inequality symbol, it is still tested at the boundary of equality.
- The null hypothesis for a population proportion is: $H_0: p = p_0,\ where\ p_0$is the null hypothesized value for the population proportion
- A one-sided alternative hypothesis for a proportion is either $H_a: p < p_0 or\ H_a: p > p_0$. A two-sided alternate hypothesis is $H_a: p_1 \neq p_2$.
- For a one-sample z-test for a population proportion, the null hypothesis specifies a value for the population proportion, usually one indicating no difference or effect.
- For a single categorical variable, the appropriate testing method for a population proportion is a one-sample z-test for a population proportion.
- In order to make statistical inferences when testing a population proportion, we must check for independence and that the sampling distribution is approximately normal:
  a. To check for independence:
  i. Data should be collected using a random sample or a randomized experiment.
  ii. When sampling without replacement, check that n N ≤10% .
  b. To check that the sampling distribution of $\hat{p}$is approximately normal (shape):
  i. Assuming that $H_0$is true $(p = p_0)$, verify that both the number of successes, np_0, and the number of failures, $n(1 - p_0)$are at least 10 so that that the sample size is large enough to support an assumption of normality
- The distribution of the test statistic assuming the null hypothesis is true (null distribution) can be either a randomization distribution or when a probability model is assumed to be true, a theoretical distribution (z).
- When using a z-test, the standardized test statistic can be written:
  $test\ statistic = \frac{sample\ statistic - null\ value\ of\ the\ parameter}{standard\ deviation\ of\ the\ statistic}$This is called a z-statistic for proportions.
- The test statistic for a population proportion is: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

- A p-value is the probability of obtaining a test statistic as extreme or more extreme than the observed test statistic when the null hypothesis and probability model are assumed to be true. The significance level may be given or determined by the researcher.
- The p-value is the proportion of values for the null distribution that are as extreme or more extreme than the observed value of the test statistic. This is:
  a. The proportion at or above the observed value of the test statistic, if the alternative is >.
  b. The proportion at or below the observed value of the test statistic, if the alternative is <.
  c. 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 ≠.
- An interpretation of the p-value of a significance test for a one-sample proportion should recognize that the p-value is computed by assuming that the probability model and null hypothesis are true, i.e., by assuming that the true population proportion is equal to the particular value stated in the null hypothesis.
- The significance level, , is the predetermined probability of rejecting the null hypothesis given that it is true.
- A formal decision explicitly compares the p-value to the significance level, α. If the p-value $\leq \alpha$, reject the null hypothesis. If the p-value $> \alpha$, fail to reject the null hypothesis.
- Rejecting the null hypothesis means there is sufficient statistical evidence to support the alternative hypothesis. Failing to reject the null means there is insufficient statistical evidence to support the alternative hypothesis.
- The conclusion about the alternative hypothesis must be stated in context
- A significance 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.
- Small p-values indicate that the observed value of the test statistic would be unusual if the null hypothesis and probability model were true, and so provide evidence for the alternative. The lower the p-value, the more convincing the statistical evidence for the alternative hypothesis.
- p-values that are not small indicate that the observed value of the test statistic would not be unusual if the null hypothesis and probability model were true, so do not provide convincing statistical evidence for the alternative hypothesis nor do they provide evidence that the null hypothesis is true.
- A formal decision explicitly compares the p-value to the significance α. If the p-value $\leq \alpha$, then reject the null hypothesis, $H_0: p = p_0$. If the p-value $> \alpha$, then fail to reject the null hypothesis.
- The results of a significance 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.
- A Type I error occurs when the null hypothesis is true and is rejected (false positive).
- A Type II error occurs when the null hypothesis is false and is not rejected (false negative).

Table of Errors

		Actual Population Value	
		H_0 true	H_a true
Decision	Reject H_0	Type I Error	Correct Decision
	Fail to Reject H_0	Correct Decision	Type II Error

- The significance level, α, is the probability of making a Type I error, if the null hypothesis is true.
- The power of a test is the probability that a test will correctly reject a false null hypothesis.
- The probability of making a Type II error = 1- power.
- The probability of a Type II error decreases when any of the following occurs, provided the others do not change:
  i. Sample size(s) increases.
  ii. Significance level (α) of a test increases.
  iii. Standard error decreases.
  iv. True parameter value is farther from the null.
- Whether a Type I or a Type II error is more consequential depends upon the situation.
- Since the significance level, α, is the probability of a Type I error, the consequences of a Type I error influence decisions about a significance level.
- The appropriate confidence interval procedure for a two-sample comparison of proportions for one categorical variable is a two-sample z-interval for a difference between population proportions.
- In order to calculate confidence intervals to estimate a difference between proportions, we must check for independence and that the sampling distribution is approximately normal:

a. To check for independence:
   i. Data should be collected using two independent, random samples or a randomized experiment.
   ii. When sampling without replacement, check that $n_1 \le 10\% N_1$ and $n_2 \le 10\% N_2$.
b. To check that sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal (shape).
   i. For categorical variables, check that $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ are all greater than or equal to some predetermined value, typically either 5 or 10.

- For a comparison of proportions, the interval estimate is $(\widehat{p_1} - \widehat{p_2} \pm z^* \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$
- Confidence intervals for a difference in proportions can be used to calculate interval estimates with specified units.
- In repeated random sampling with the same sample size, approximately C% of confidence intervals created will capture the difference in population proportions.
- Interpreting a confidence interval for difference between population proportions should include a reference to the sample taken and details about the population it represents.
- A confidence interval for difference in population proportions provides an interval of values that may provide sufficient evidence to support a particular claim in context.
- For a two-sample test for a difference of two proportions, the null hypothesis specifies a value of 0 for the difference in population proportions, indicating no difference or effect
- The null hypothesis for a difference in proportions is: $H_0: p_1 = p_2 \; or \; H_0: p_1 - p_2 = 0$
- A one-sided alternative hypothesis for a difference in proportions is$H_a: p_1 < p_2 \; or \; H_a: p_1 > p_2$. A two-sided alternative hypothesis for a difference of proportions is$H_a: p_1 \neq p_2$
- For a single categorical variable, the appropriate testing method for the difference of two population proportions is a two-sample z-test for a difference between two population proportions.
- In order to make statistical inferences when testing a difference between population proportions, we must check for independence and that the sampling distribution is approximately normal:
  a. To check for independence:

i. Data should be collected using two independent, random samples or a randomized experiment.
ii. When sampling without replacement, check that $n_1 \leq 10\% N_1$ and $n_2 \leq 10\% N_2$.

b. To check that the sampling distribution of $\widehat{p_1} - \widehat{p_2}$ is approximately normal (shape):

i. For the combined sample, define the combined (or pooled) proportion, $\widehat{p_c} = \frac{n_1\widehat{p_1}+n_2\widehat{p_2}}{n_1+n_2}$. Assuming that H_0 is true ($p_1 - p_2 = 0$ or $p_1 = p_2$). check that $n_1\widehat{p_c}$, $n_1(1 - \widehat{p_c})$, and $n_2(1 - \widehat{p_c})$ are all greater than or equal to some predetermined value, typically either 5 or 10.

- The test statistic for a difference in proportions is:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_c(1-\hat{p}_c)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$
where $\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$.

- An interpretation of the p-value of a significance test for a difference of two population proportions should recognize that the p-value is computed by assuming that the null hypothesis is true, i.e., by assuming that the true population proportions are equal to each other.
- A formal decision explicitly compares the p-value to the significance α. If the p-value ≤ α, then reject the null hypothesis, $H_0: p_1 = p_2$, or $H_0: p_1 = p_2 = 0$. If the p-value > α, then fail to reject the null hypothesis.
- The results of a significance test for a difference of two population proportions can serve as the statistical reasoning to support the answer to a research question about the two populations that were sampled.

Vocabulary

- Proportion
- Sampling distribution
- Central limit theorem
- Sample size
- Margin of error
- Standard error
- Confidence interval
- One proportion z-interval
- Plausible independence condition
- Hypothesis
- P-value
- Null hypothesis
- Alternative hypothesis
- Alpha level
- Type I error
- Type II error
- Critical value
- Random sampling condition

Instruction

Assessment Tools

Pre/Post Assessments:
- Unit Test

Textbook
- Ch. 18-22

Investigative Tasks
- Simulated Coins
- Life After High School?

Unit 6 Learning About the World	Essential Knowledge and Skills and Key Instructional Information

Standards	The student will use problem solving, mathematical communication, mathematical reasoning, connections, and representations to:
Focus: T-distributions, significance testing for means, confidence intervals for means **Enduring Understanding:** VAR-1 Given that variation may be random or not, conclusions are uncertain. VAR-7 The t-distribution may be used to model variation. UNC-4 An interval of values should be used to estimate parameters, in order to account for uncertainty. DAT-3 Significance testing allows us to make decisions about hypotheses within a particular context. **Anticipated Pacing:** Semester: 10 days	**Content** **Lesson Objective** VAR-1 Given that variation may be random or not, conclusions are uncertain. ● Identify questions suggested by probabilities of errors in statistical inference. VAR-7 The t-distribution may be used to model variation. ● Describe t-distributions. ● Identify an appropriate testing method for a population mean with unknown σ, including the mean difference between values in matched pairs. ● Identify the null and alternative hypotheses for a population mean with unknown σ, including the mean difference between values in matched pairs. ● Verify the conditions for the test for a population mean, including the mean difference between values in matched pairs. ● Calculate an appropriate test statistic for a population mean, including the mean difference between values in matched pairs. ● Identify an appropriate selection of a testing method for a difference of two population means. ● Identify the null and alternative hypotheses for a difference of two population means. ● Verify the conditions for the significance test for the difference of two population means. ● Calculate an appropriate test statistic for a difference of two means. UNC-4 An interval of values should be used to estimate parameters, in order to account for uncertainty. ● Identify an appropriate confidence interval procedure for a population mean, including the mean difference

between values in matched pairs.
- Verify the conditions for calculating confidence intervals for a population mean, including the mean difference between values in matched pairs.
- Determine the margin of error for a given sample size for a one-sample t-interval.
- Calculate an appropriate confidence interval for a population mean, including the mean difference between values in matched pairs.
- Interpret a confidence interval for a population mean, including the mean difference between values in matched pairs.
- Justify a claim based on a confidence interval for a population mean, including the mean difference between values in matched pairs.
- Identify the relationships between sample size, width of a confidence interval, confidence level, and margin of error for a population mean.
- Identify an appropriate confidence interval procedure for a difference of two population means.
- Verify the conditions to calculate confidence intervals for the difference of two population means.
- Determine the margin of error for the difference of two population means.
- Calculate an appropriate confidence interval for a difference of two population means.
- Interpret a confidence interval for a difference of population means.
- Justify a claim based on a confidence interval for a difference of population means. [
- Identify the effects of sample size on the width of a confidence interval for the difference of two means.

DAT-3 Significance testing allows us to make decisions about hypotheses within a particular context.

- Interpret the p-value of a significance test for a population mean, including the mean difference between values in matched pairs.
- Interpret the p-value of a significance test for a difference of population means.
- Justify a claim about the population based on the results of a significance test for a difference of two population means in context.
- Justify a claim about the population based on the results of a significance test for a population mean.

Essential Knowledge

- Random variation may result in errors in statistical inference.
- When s is used instead of σ to calculate a test statistic, the corresponding distribution, known as the t-distribution, varies from the normal distribution in shape, in that more of the area is allocated to the tails of

the density curve than in a normal distribution.

- As the degrees of freedom increase, the area in the tails of a t-distribution decreases.
- Because σ is typically not known for distributions of quantitative variables, the appropriate confidence interval procedure for estimating the population mean of one quantitative variable for one sample is a one-sample t-interval for a mean.
- For one quantitative variable, X, that is normally distributed, the distribution of $t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}}$ is a t-distribution with n −1 degrees of freedom.
- Matched pairs can be thought of as one sample of pairs. Once differences between pairs of values are found, inference for confidence intervals proceeds as for a population mean
- In order to calculate confidence intervals to estimate a population mean, we must check for independence and that the sampling distribution is approximately normal:
  a. To check for independence:
  i. Data should be collected using a random sample or a randomized experiment.
  ii. When sampling without replacement, check that n ≤10%N , where N is the size of the population.
  b. To check that the sampling distribution of $\bar{x}$ is approximately normal (shape):
  i. If the observed distribution is skewed, n should be greater than 30.
  ii. If the sample size is less than 30, the distribution of the sample data should be free from strong skewness and outliers.
- The critical value t* with n −1 degrees of freedom can be found using a table or computer-generated output.
- The standard error for a sample mean is given by $SE = \frac{s}{\sqrt{n}}$ where s is the sample standard deviation.
- For a one-sample t-interval for a mean, the margin of error is the critical value (t*) times the standard error (SE), which equals $t^{*}(\frac{s}{\sqrt{n}})$
- The point estimate for a population mean is the sample mean, $\bar{x}$
- For the population mean for one sample with unknown population standard deviation, the confidence interval is $\bar{x} \pm t^{*}\frac{s}{\sqrt{n}}$
- A confidence interval for a population mean either contains the population mean or it does not, because each interval is based on data from a random sample, which varies from sample to sample.
- We are C% confident that the confidence interval for a population mean captures the population mean.
- An interpretation of a confidence interval for a population mean includes a reference to the sample taken and details about the population it represents.
- A confidence interval for a population mean provides an interval of values that may provide sufficient evidence to support a particular claim in context.

- 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 single mean, the width of the interval is proportional to $\frac{1}{\sqrt{n}}$
- For a given sample, the width of the confidence interval for a population mean increases as the confidence level increases.
- The appropriate test for a population mean with unknown σ is a one-sample t-test for a population mean.
- Matched pairs can be thought of as one sample of pairs. Once differences between pairs of values are found, inference for significance testing proceeds as for a population mean.
- The null hypothesis for a one-sample t-test for a population mean is $H_0: \mu = \mu_0$, where μ_0 is the hypothesized value. Depending upon the situation, the alternative hypothesis is $H_a: \mu < \mu_0,\ or\ H_a: \mu > \mu_0,\ or\ H_a: \mu \neq \mu_0$
- When finding the mean difference, μ_d, between values in a matched pair, it is important to define the order of subtraction.
- In order to make statistical inferences when testing a population mean, we must check for independence and that the sampling distribution is approximately normal:
  a. To check for independence:
  i. Data should be collected using a random sample or a randomized experiment.
  ii. When sampling without replacement, check that n ≤10% N.
  b. To check that the sampling distribution of $\bar{x}$ is approximately normal (shape):
  i. If the observed distribution is skewed, n should be greater than 30.
  ii. If the sample size is less than 30, the distribution of the sample data should be free from strong skewness and outliers.
- For a single quantitative variable when random sampling with replacement from a population that can be modeled with a normal distribution with mean μ and standard deviation σ , the sampling distribution of $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ has a t-distribution with n – 1 degrees of freedom
- An interpretation of the p-value of a significance test for a population mean should recognize that the p-value is computed by assuming that the null hypothesis is true, i.e., by assuming that the true population mean is equal to the particular value stated in the null hypothesis.
- A formal decision explicitly compares the p-value to the significance α. If the p-value ≤ α, then reject the null hypothesis, $H_0: \mu = \mu_0$. If the p-value > α, then fail to reject the null hypothesis.
- The results of a significance 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.
- Consider a simple random sample from population 1 of size n_1 , mean μ_1 , and standard deviation σ_1 and a

second simple random sample from population 2 of size n_2, mean μ_2, and standard deviation σ_2. If the distributions of populations 1 and 2 are normal or if both n_1 and n_2 are greater than 30, then the sampling distribution of the difference of means, $\overline{x}_1 - \overline{x}_2$ is also normal. The mean for the sampling distribution of $\overline{x}_1 - \overline{x}_2$ is $\mu_1 - \mu_2$. The standard deviation of $\overline{x}_1 - \overline{x}_2$ is $\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$

- The appropriate confidence interval procedure for one quantitative variable for two independent samples is a two-sample t-interval for a difference between population means.
- In order to calculate confidence intervals to estimate a difference of population means, we must check for independence and that the sampling distribution is approximately normal:
  a. To check for independence:
  i. Data should be collected using two independent, random samples or a randomized experiment.
  ii. When sampling without replacement, check that $n_1 \leq 10\%N_1$ and $n_2 \leq 10\%N_2$.
  b. To check that the sampling distribution of $\overline{x}_1 - \overline{x}_2$ should be approximately normal (shape):
  i. If the observed distributions are skewed, both $n_1 \, and \, n_2$ should be greater than 30.
- For the difference of two sample means, the margin of error is the critical value (t*) times the standard error (SE) of the difference of two means.
- The standard error for the difference in two sample means with sample standard deviations, $s_1 and \, s_2 is \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$
- The point estimate for the difference of two population means is the difference in sample means, $\overline{x}_1 - \overline{x}_2$
- For a difference of two population means where the population standard deviations are not known, the confidence interval is $(\overline{x}_1 - \overline{x}_2) \pm t^* \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$ where ± t* are the critical values for the central C% of a t-distribution with appropriate degrees of freedom that can be found using technology
- In repeated random sampling with the same sample size, approximately C% of confidence intervals created will capture the difference of population means.
- An interpretation for a confidence interval for the difference of two population means should include a reference to the samples taken and details about the populations they represent.
- A confidence interval for a difference of population means provides an interval of values that may provide sufficient evidence to support a particular claim in context.
- When all other things remain the same, the width of the confidence interval for the difference of two means

tends to decrease as the sample sizes increase.

- For a quantitative variable, the appropriate test for a difference of two population means is a two-sample t-test for a difference of two population means.
- The null hypothesis for a two-sample t-test for a difference of two population means $\mu_1 and\ \mu_2\ is\ H_0: \mu_1 - \mu_2 = 0\ or\ H_0: \mu_1 = \mu_2$The alternative hypothesis is $H_a: \mu_1 - \mu_2 < 0,\ or\ H_a: \mu_1 - \mu_2 > 0,\ or\ H_a: \mu_1 - \mu_2 \neq 0\ or\ H_a: \mu_1 > \mu_2 or\ H_a: \mu_1 < \mu_2\ or\ H_a: \mu_1 \neq \mu_2$
- In order to make statistical inferences when testing a difference between population means, we must check for independence and that the sampling distribution is approximately normal:
  a. Individual observations should be independent:
  i. Data should be collected using simple random samples or a randomized experiment.
  ii. When sampling without replacement, check that $n_1 \leq 10\%N_1$and $n_2 \leq 10\%N_2$.
  b. The sampling distribution of $\overline{x_1} - \overline{x_2}$should be approximately normal (shape).
  i. If the observed distribution is skewed, both $n_1 and\ n_2$ should be greater than 30.
  ii. If the sample size is less than 30, the distribution of the sample data should be free from strong skewness and outliers. This should be checked for BOTH samples.
- For a single quantitative variable, data collected using independent random samples or a randomized experiment from two populations, each of which can be modeled with a normal distribution, the sampling distribution of $t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$t-distribution with degrees of freedom that can be found using technology.
  The degrees of freedom fall between the smaller of $n_1 - 1$and $n_2 - 1$ and $n_1 + n_2 - 2$
- An interpretation of the p-value of a significance test for a two-sample difference of population means should recognize that the p-value is computed by assuming that the null hypothesis is true, i.e., by assuming that the true population means are equal to each other.
- A formal decision explicitly compares the p-value to the significance α. If the p-value ≤ α, then reject the null hypothesis, $H_0: \mu_1 - \mu_2 = 0\ or\ H_0: \mu_1 = \mu_2$.If the p-value >$\alpha$, then fail to reject the null hypothesis
- The results of a significance 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.

Vocabulary

- Standard error
- P-value
- Means
- Significance
- Hypotheses
- T-test
- Degrees of freedom
- Confidence interval
- Margin of error
- Variance
- Null hypothesis
- Alternative hypothesis
- Nearly normal condition
- Independent groups assumption
- Randomization condition
- Paired data

Instruction

Assessment Tools

Pre/Post Assessments:

- Unit Test
- Inference Review

Textbook

- Ch. 23-25

Investigative Tasks

- SAT Performance
- SAT Performance (Part II)

Unit 7 Inference When Variables are Related	Essential Knowledge and Skills and Key Instructional Information

Standards	The student will use <u>problem solving</u>, <u>mathematical communication</u>, <u>mathematical reasoning</u>, <u>connections</u>, and <u>representations</u> to:
Focus: Chi-square distributions, chi-square significance testing, chi-square confidence intervals, t-distribution, slope of regression model testing, slope of regression model confidence intervals **Enduring Understanding** VAR-1 Given that variation may be random or not, conclusions are uncertain. VAR-8 The chi-square distribution may be used to model variation. DAT-3 Significance testing allows us to make decisions about hypotheses within a particular context. UNC-4 An interval of values should be used to estimate parameters, in order to account	**Content** **Learning Objective** VAR-1 Given that variation may be random or not, conclusions are uncertain. ● Identify questions suggested by variation between observed and expected counts in categorical data. ● Identify questions suggested by variation in scatter plots. VAR-8 The chi-square distribution may be used to model variation. ● Describe chi-square distributions. ● Identify the null and alternative hypotheses in a test for a distribution of proportions in a set of categorical data. ● Identify an appropriate testing method for a distribution of proportions in a set of categorical data. ● Calculate expected counts for the chi-square test for goodness of fit. ● Verify the conditions for making statistical inferences when testing goodness of fit for a chi-square distribution. ● Calculate the appropriate statistic for the chi-square test for goodness of fit. ● Determine the p-value for chi-square test for goodness of fit significance test. ● Calculate expected counts for two-way tables of categorical data. ● Identify the null and alternative hypotheses for a chi-square test for homogeneity or independence. ● Identify an appropriate testing method for comparing distributions in two-way tables of categorical data. ● Verify the conditions for making statistical inferences when testing a chi-square distribution for independence or homogeneity ● Calculate the appropriate statistic for a chi-square test for homogeneity or independence. ● Determine the p-value for a chi-square significance test for independence or homogeneity.

<table>
<tr>
<td>
for uncertainty

VAR-7 The t-distribution may be used to model variation.

Anticipated Pacing:
Semester: 10 days
</td>
<td>
DAT-3 Significance testing allows us to make decisions about hypotheses within a particular context.

Interpret the p-value for the chi-square test for goodness of fit
Justify a claim about the population based on the results of a chi-square test for goodness of fit.
Interpret the p-value for the chi-square test for homogeneity or independence.
Justify a claim about the population based on the results of a chi-square test for homogeneity or independence.
Interpret the p-value of a significance test for the slope of a regression model.
Justify a claim about the population based on the results of a significance test for the slope of a regression model.

UNC-4 An interval of values should be used to estimate parameters, in order to account for uncertainty

Identify an appropriate confidence interval procedure for a slope of a regression model.
Verify the conditions to calculate confidence intervals for the slope of a regression model.
Determine the given margin of error for the slope of a regression model.
Calculate an appropriate confidence interval for the slope of a regression model.
Interpret a confidence interval for the slope of a regression model.
Justify a claim based on a confidence interval for the slope of a regression model.
Identify the effects of sample size on the width of a confidence interval for the slope of a regression model.

VAR-7 The t-distribution may be used to model variation.

Identify the appropriate selection of a testing method for a slope of a regression model.
Identify appropriate null and alternative hypotheses for a slope of a regression model.
Verify the conditions for the significance test for the slope of a regression model.
Calculate an appropriate test statistic for the slope of a regression model.

Essential Knowledge:

Variation between what we find and what we expect to find may be random or not.
Expected counts of categorical data are counts consistent with the null hypothesis. In general, an expected count is a sample size times a probability
The chi-square statistic measures the distance between observed and expected counts relative to expected counts.

</td>
</tr>
</table>

- Chi-square distributions have positive values and are skewed right. Within a family of density curves, the skew becomes less pronounced with increasing degrees of freedom.
- For a chi-square goodness-of-fit test, the null hypothesis specifies null proportions for each category, and the alternative hypothesis is that at least one of these proportions is not as specified in the null hypothesis.
- When considering a distribution of proportions for one categorical variable, the appropriate test is the chi-square test for goodness of fit.
- Expected counts for a chi-square goodness-of-fit test are (sample size) (null proportion).
- In order to make statistical inferences for a chi-square test for goodness of fit we must check the following:
  a. To check for independence:
  i. Data should be collected using a random sample or randomized experiment.
  ii. When sampling without replacement, check that n ≤10%N .
  b. The chi-square test for goodness of fit becomes more accurate with more observations, so large counts should be used (shape).
  i. A conservative check for large counts is that all expected counts should be greater than 5.
- The test statistic for the chi-square test for goodness of fit is $\chi^2 = \sum \frac{(Observed\ count - Expected\ count)^2}{Expected\ count}$ with degrees of freedom number of categories = −1.
- The distribution of the test statistic assuming the null hypothesis is true (null distribution) can be either a randomization distribution or, when a probability model is assumed to be true, a theoretical distribution (chi-square).
- The p-value for a chi-square test for goodness of fit for a number of degrees of freedom is found using the appropriate table or computer generated output.
- An interpretation of the p-value for the chi-square test for goodness of fit is the probability, given the null hypothesis and probability model are true, of obtaining a test statistic as, or more, extreme than the observed value.
- A decision to either reject or fail to reject the null hypothesis is based on comparison of the p-value to the significance level, α.
- The results of a chi-square test for goodness of fit can serve as the statistical reasoning to support the answer to a research question about the population that was sampled.
- The expected count in a particular cell of a two-way table of categorical data can be calculated using the formula: $expected\ count = \frac{(row\ total)(column\ total)}{table\ total}$
- The appropriate hypotheses for a chi-square test for homogeneity are: H0: There is no difference in distributions of a categorical variable across populations or treatments. Ha : There is a difference in distributions of a categorical variable across populations or treatments.

- The appropriate hypotheses for a chi-square test for independence are: H0: There is no association between two categorical variables in a given population or the two categorical variables are independent. Ha : Two categorical variables in a population are associated or dependent.
- When comparing distributions to determine whether proportions in each category for categorical data collected from different populations are the same, the appropriate test is the chi-square test for homogeneity.
- To determine whether row and column variables in a two-way table of categorical data might be associated in the population from which the data were sampled, the appropriate test is the chi-square test for independence.
- VAR-8.K.1 In order to make statistical inferences for a chi-square test for two-way tables (homogeneity or independence), we must verify the following:
  a. To check for independence:
  i. For a test for independence: Data should be collected using a simple random sample.
  ii. For a test for homogeneity: Data should be collected using a stratified random sample or randomized experiment.
  iii. When sampling without replacement, check that $n \le 10\%N$.
  b. The chi-square tests for independence and homogeneity become more accurate with more observations, so large counts should be used (shape).
  i. A conservative check for large counts is that all expected counts should be greater than 5.
- The appropriate test statistic for a chi-square test for homogeneity or independence is the chi-square statistic:
  $\chi^2 = \sum \frac{(Observed\ count - Expected\ count)^2}{Expected\ count}$ with degrees of freedom equal to: (number of rows − 1)(number of columns −1).
- The p-value for a chi-square test for independence or homogeneity for a number of degrees of freedom is found using the appropriate table or technology
- For a test of independence or homogeneity for a two-way table, the p-value is the proportion of values in a chi-square distribution with appropriate degrees of freedom that are equal to or larger than the test statistic.
- An interpretation of the p-value for the chi-square test for homogeneity or independence is the probability, given the null hypothesis and probability model are true, of obtaining a test statistic as, or more, extreme than the observed value.
- A decision to either reject or fail to reject the null hypothesis for a chi-square test for homogeneity or independence is based on comparison of the p-value to the significance level, α .
- 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).
- Variation in points' positions relative to a theoretical line may be random or non-random.
- Consider a response variable, y, that is linearly related to an explanatory variable, x. For a simple random sample

of n observations, the sample regression line, $\hat{y} = a + bx$, is an estimate of the population regression line $\mu_y = \alpha + \beta x$For a particular observation, (xi , yi), the residual from the sample regression line, $y_i - \hat{y}_i = y_i - (a + bx_i)$, is an estimate of $y_i - (\alpha + \beta x_i)$, the deviation of the response variable from the population regression line. For all points (x, y) in the population, the standard deviation of all of the deviations of the response variable from the population regression line, σ, can be estimated by the standard deviation of the residuals from the sample regression line, $s = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$

- For a simple random sample of n observations, let b represent the slope of a sample regression line. Then the mean of the sampling distribution for b equals the population slope: $\mu_b = \beta$The standard deviation of the sampling distribution for b is $\sigma_b = \frac{\sigma}{\sigma_x\sqrt{n}}$, where $\sigma_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}$
- The appropriate confidence interval for the slope of a regression model is a t-interval for the slope.
- In order to calculate a confidence interval to estimate the slope of a regression line, we must check the following:
  a. The true relationship between x and y is linear. Analysis of residuals may be used to verify linearity.
  b. The standard deviation for y, σ_y , does not vary with x. Analysis of residuals may be used to check for approximately equal standard deviations for all x.
  c. To check for independence:
  i. Data should be collected using a random sample or a randomized experiment.
  ii. When sampling without replacement, check that n ≤10%N .
  d. For a particular value of x, the responses (y-values) are approximately normally distributed. Analysis of graphical representations of residuals may be used to check for normality.
  i. If the observed distribution is skewed, n should be greater than 30.
- For the slope of a regression line, the margin of error is the critical value (t*) times the standard error (SE) of the slope.
- The standard error for the slope of a regression line with sample standard deviation, s, is$SE = \frac{s}{s_x\sqrt{n-1}}$where s is the estimate of σ and x_s is the sample standard deviation of the x values.
- The point estimate for the slope of a regression model is the slope of the line of best fit, b.
- For the slope of a regression model, the interval estimate is $b \pm t^*(SE_b)$.

- In repeated random sampling with the same sample size, approximately C% of confidence intervals created will capture the slope of the regression model, i.e., the true slope of the population regression model.
- An interpretation for a confidence interval for the slope of a regression line should include a reference to the sample taken and details about the population it represents.
- A confidence interval for the slope of a regression model provides an interval of values that may provide sufficient evidence to support a particular claim in context.
- When all other things remain the same, the width of the confidence interval for the slope of a regression model tends to decrease as the sample size increases.
- The appropriate test for the slope of a regression model is a t-test for a slope.
- The null hypothesis for a t-test for a slope is: $H_0: \beta = \beta_0$, where β_0 is the hypothesized value from the null hypothesis. The alternative hypothesis is $H_0: \beta < \beta_0$ or $H_0: \beta > \beta_0$ or $H_0: \beta \neq \beta_0$
- In order to make statistical inferences when testing for the slope of a regression model, we must check the following:
  a. The true relationship between x and y is linear. Analysis of residuals may be used to verify linearity.
  b. The standard deviation for y, σ_y , does not vary with x. Analysis of residuals may be used to check for approximately equal standard deviations for all x.
  c. To check for independence:
  i. Data should be collected using a random sample or a randomized experiment.
  ii. When sampling without replacement, check that n ≤10%N .
  d. For a particular value of x, the responses (y-values) are approximately normally distributed. Analysis of graphical representations of residuals may be used to check for normality.
  i. If the observed distribution is skewed, n should be greater than 30.
  ii. If the sample size is less than 30, the distribution of the sample data should be free from strong skewness and outliers.
- The distribution of the slope of a regression model assuming all conditions are satisfied and the null hypothesis is true (null distribution) is a t-distribution.
- For simple linear regression when random sampling from a population for the response that can be modeled with a normal distribution for each value of the explanatory variable, the sampling distribution of $t = \frac{b-\beta}{SE_b}$ has a t-distribution with degrees of freedom equal to n-2. When testing the slope in a simple linear regression model with one parameter, the slope, the test for the slope has df n-1.
- An interpretation of the p-value of a significance test for the slope of a regression model should recognize that the p-value is computed by assuming that the null hypothesis is true, i.e., by assuming that the true population slope is equal to the particular value stated in the null hypothesis.
- A formal decision explicitly compares the p-value to the significance α. If the p-value ≤α , then reject the null

hypothesis, $H_0: \beta = \beta_0$. If the p-value $>\alpha$, then fail to reject the null hypothesis.

- The results of a significance test for the slope of a regression model can serve as the statistical reasoning to support the answer to a research question about that sample.

Vocabulary

- Degrees of freedom
- Goodness-of-fit test
- Test of homogeneity
- Test of independence
- Association
- Null hypothesis
- Alternative hypothesis
- Independence
- Large enough sample
- Expected value
- Categorical data
- component
- Regression
- Slope
- y-intercept
- t-model
- Residuals
- Normal model
- Association
- Strength
- Confidence interval
- Standard error
- Variability
- Standard deviation

Instruction

Assessment Tools

Pre/Post Assessments:

- Unit Test

Textbook

- Ch. 26-27

Investigative Tasks

- ‘97 AP Stat Scores