LO 1. Identify variables as numerical and categorical.

- If the variable is numerical, further classify as continuous or discrete based on whether or not the variable can take on an infinite number of values or only non-negative whole numbers, respectively.
- If the variable is categorical, determine if it is ordinal based on whether or not the levels have a natural ordering.

LO 2. Define associated variables as variables that show some relationship with one another. Further categorize this relationship as positive or negative association, when possible.

LO 3. Define variables that are not associated as independent.

* *Reading: Section 1.1 and 1.2 of OpenIntro Statistics*
* *Test yourself: Give one example of each type of variable you have learned.*

LO 4. Identify the explanatory variable in a pair of variables as the variable suspected of affecting the other. However, note that labeling variables as explanatory and response does not guarantee that the relationship between the two is actually causal, even if there is an association identified between the two variables.

LO 5. Classify a study as observational or experimental, and determine and explain why whether the study's results can be generalized to the population and whether they suggest correlation or causation between the variables studied.

- If random sampling has been employed in data collection, the results should be generalizable to the target population.
- If random assignment has been employed in study design, the results suggest causality.

LO 6. Question confounding variables and sources of bias in a given study.

LO 7. Distinguish between simple random, stratified, cluster, and multistage sampling, and recognize the benefits and drawbacks of choosing one sampling scheme over another.

- Simple random sampling: Each subject in the population is equally likely to be selected.
- Stratified sampling: First divide the population into homogenous strata (subjects within each stratum are similar, across strata are different), then randomly sample from within each strata.
- Cluster sampling: First divide the population into clusters (subjects within each cluster are non-homogenous, but clusters are similar to each other), then randomly sample a few clusters, and then sample all cases within those clusters.
- Multistage sampling: First divide the population into clusters, then randomly sample a few clusters, and then randomly sample from within each cluster.

LO 8. Identify the four principles of experimental design and recognize their purposes: control any possible confounders, randomize into treatment and control groups, replicate by using a sufficiently large sample or repeating the experiment, and block any variables that might influence the response.

LO 9. Identify if single or double blinding has been used in a study.

* *Reading: Sections 1.3 - 1.4 of OpenIntro Statistics*

* *Article: How Anecdotal Evidence Can Undermine Scientific Results, Scientific American, 2008*
* *Test yourself:*
    1. *Describe when a study's results can be generalized to the population at large and when causation can be inferred.*
    2. *Explain why random sampling allows for generalizability of results.*
    3. *Explain why random assignment allows for making causal conclusions.*
    4. *Describe a situation where cluster sampling is more efficient than simple random or stratified sampling.*
    5. *Explain how blinding can help eliminate the placebo effect and other biases.*

LO 1. Use scatterplots for describing the relationship between two numerical variables making sure to note the direction (positive or negative), form (linear or non-linear) and the strength of the relationship as well as any unusual observations that stand out.

LO 2. When describing the distribution of a numerical variable, mention its shape, center, and spread, as well as any unusual observations.

LO 3. Note that there are three commonly used measures of center and spread:

- center: mean (the arithmetic average), median (the midpoint), mode (the most frequent observation).
- spread: spread: standard deviation (variability around the mean), range (max-min), interquartile range (middle 50% of the distribution).

LO 4. Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodal, multimodal, or uniform.

LO 5. Use histograms and box plots to visualize the shape, center, and spread of numerical distributions, and intensity maps for visualizing the spatial distribution of the data.

LO 6. Define a robust statistic (e.g. median, IQR) as measures that are not heavily affected by skewness and extreme outliers, and determine when they are more appropriate measured of center and spread compared to other similar statistics.

LO 7. Recognize when transformations (e.g. log) can make the distribution of data more symmetric, and hence easier to model.

* *Reading: Section 2.1 of OpenIntro Statistics*
* *Test yourself:*
  1. *Describe what is meant by robust statistics and when they are used.*
  2. *Describe when and why we might want to apply a log transformation to a variable.*

LO 8. Use frequency tables and bar plots to describe the distribution of one categorical variable.

LO 9. Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

LO 10. Use side-by-side box plots for assessing the relationship between a numerical and a categorical variable.

* *Reading: Section 2.2 of OpenIntro Statistics*
* *Test yourself:*
  1. *Interpret the plot in Figure 1.40 (page 39) of the textbook.*
  2. *You collect data on 100 classmates, 70 females and 30 males. 10% of the class are smokers, and smoking is independent of gender. Calculate how many males and females would be expected to be smokers. Sketch a mosaic plot of this scenario.*

LO 11. Note that an observed difference in sample statistics suggesting dependence between variables may be due to random chance, and that we need to use hypothesis testing to determine if this difference is too large to be attributed to random chance.

LO 12. Set up null and alternative hypotheses for testing for independence between variables, and evaluate data's support for these hypotheses using a simulation technique.

* *Reading: Section 2.3 of OpenIntro Statistics*
* *Test yourself: Explain why difference in sample proportions across two groups does not necessarily indicate dependence between the two variables involved?*

LO 1. Define trial, outcome, and sample space.

LO 2. Explain why the long-run relative frequency of repeated independent events settle down to the true probability as the number of trials increases, i.e. why the law of large numbers holds.

LO 3. Distinguish disjoint (also called mutually exclusive) and independent events.

- If A and B are independent, then having information on A does not tell us anything about B.
- If A and B are disjoint, then knowing that A occurs tells us that B cannot occur.
- Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot.

LO 4. Draw Venn diagrams representing events and their probabilities.

LO 5. Define a probability distribution as a list of the possible outcomes with corresponding probabilities that satisfies three rules:

- The outcomes listed must be disjoint.
- Each probability must be between 0 and 1, inclusive.
- The probabilities must total 1.

LO 6. Define complementary outcomes as mutually exclusive outcomes of the same random process whose probabilities add up to 1.

- If A and B are complementary, $P(A) + P(B) = 1$.

LO 7. Distinguish between union of events (A or B) and intersection of events (A and B).

LO 8. Calculate the probability of union of events using the (general) addition rule.

- If A and B are not mutually exclusive, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.
- If A and B are mutually exclusive, $P(A \text{ or } B) = P(A) + P(B)$, since for mutually exclusive events $P(A \text{ and } B) = 0$.

* *Reading: Section 3.1 of OpenIntro Statistics*
* *Test yourself:*

1. *What is the probability of getting a head on the 6th coin flip if in the first 5 flips the coin landed on a head each time?*
2. *True / False: Being right handed and having blue eyes are mutually exclusive events.*
3. *P(A) = 0.5, P(B) = 0.6, there are no other possible outcomes in the sample space. What is P(A and B)?*

LO 9. Distinguish marginal and conditional probabilities.

LO 10. Calculate the probability of intersection of independent events using the multiplication rule.

- If A and B are dependent, $P(A \text{ and } B) = P(A) \times P(B|A)$.
- If A and B are ~~dependent,~~ independent, $P(A \text{ and } B) = P(A) \times P(B)$, since for independent events $P(B|A) = P(B)$.

LO 11. Construct tree diagrams to calculate conditional probabilities and probabilities of intersection of non-independent events using Bayes' theorem.

* *Reading: Section 3.2 of OpenIntro Statistics*
* *Test yourself: 50% of students in a class are social science majors and the rest are not. 70% of the social science students and 40% of the non-social science students are in a relationship. Create a contingency table and a tree diagram summarizing these probabilities. Calculate the percentage of students in this class who are in a relationship.*

LO 12. Sampling without replacement from a small population means we no longer have independence between our observations.

LO 13. A random variable is a random process or variable with a numerical outcome. Modeling a process using a random variable allows us to apply a mathematical framework and statistical principles for better understanding and predicting outcomes in the real world.

LO 14. We use measures of center and spread to define distributions of random variables.

- Center: Expected value, mean, i.e. average. Denoted as $E(X)$ or μ.
- Variability: Variance (average squared deviation around the expected value). Denoted as $Var(X)$ or σ^2.

LO 15. Expected value and variance of a discrete random variable, X, can be calculated as follows:

$$E(X) = \mu = \sum_{i=1}^{k} x_i P(X = x_i)$$

$$Var(X) = \sigma^2 = \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j)$$

LO 16. Standard deviation is the square root of variance. We use standard deviation also as a measure of the variability of the random variable. Standard deviation is often easier to interpret since it's in the same units of the random variable.

LO 17. Linear combinations of random variables:

- $E(aX + bY) = a \times E(X) + b \times E(Y)$
- $Var(aX + bY) = a^2 \times Var(X) + b^2 \times Var(Y)$

LO 18. Probability density functions represent the distributions of continuous random variables.

* *Reading: Sections 3.3 - 3.5 of OpenIntro Statistics*

Learning Objectives Chapter 4: Distributions of random variables

LO 1. Define the standardized (Z) score of a data point as the number of standard deviations it is away from the mean: $Z = \frac{x-\mu}{\sigma}$.

LO 2. Use the Z score

- if the distribution is normal: to determine the percentile score of a data point (using technology or normal probability tables)
- regardless of the shape of the distribution: to assess whether or not the particular observation is considered to be unusual (more than 2 standard deviations away from the mean)

LO 3. Depending on the shape of the distribution determine whether the median would have a negative, positive, or 0 Z score.

LO 4. Assess whether or not a distribution is nearly normal using the 68-95-99.7% rule or graphical methods such as a normal probability plot.

* *Reading: Section 4.1 of OpenIntro Statistics*
* *Test yourself: True/False: In a right skewed distribution the Z score of the median is positive.*

LO 5. If X is a random variable that takes the value 1 with probability of success p and 0 with probability of success $1-p$, then X is a Bernoulli random variable.

LO 6. The geometric distribution is used to describe how many trials it takes to observe a success.

LO 7. Define the probability of finding the first success in the n^{th} trial as $(1-p)^{n-1}p$.

- $\mu = \frac{1}{p}$
- $\sigma^2 = \frac{1-p}{p^2}$
- $\sigma = \sqrt{\frac{1-p}{p^2}}$

LO 8. Determine if a random variable is binomial using the four conditions:

- The trials are independent.
- The number of trials, n, is fixed.
- Each trial outcome can be classified as a success or failure.
- The probability of a success, p, is the same for each trial.

LO 9. Calculate the number of possible scenarios for obtaining k successes in n trials using the choose function: $\binom{n}{k} = \frac{n!}{k!\ (n-k)!}$.

LO 10. Calculate probability of a given number of successes in a given number of trials using the binomial distribution: $P(k = K) = \frac{n!}{k!\ (n-k)!}\ p^k\ (1-p)^{(n-k)}$.

LO 11. Calculate the expected number of successes in a given number of binomial trials ($\mu = np$) and its standard deviation ($\sigma = \sqrt{np(1-p)}$).

LO 12. When number of trials is sufficiently large ($np \geq 10$ and $n(1-p) \geq 10$), use normal approximation to calculate binomial probabilities, and explain why this approach works.

* *Reading: Section 4.2 and 4.3 of OpenIntro Statistics*

* *Test yourself:*

    1. *True/False: We can use the binomial distribution to determine the probability that in 10 rolls of a die the first 6 occurs on the 8th roll.*
    2. *True / False: If a family has 3 kids, there are 8 possible combinations of gender order.*
    3. *True/ False: When* $n = 100$ *and* $p = 0.92$ *we can use the normal approximation to the binomial to calculate the probability of 90 or more successes.*

LO 1. Define sample statistic as a point estimate for a population parameter, for example, the sample proportion is used to estimate the population proportion, and note that point estimate and sample statistic are synonymous.

LO 2. Recognize that point estimates (such as the sample proportion) will vary from one sample to another, and define this variability as sampling variation.

LO 3. Calculate the sampling variability of the proportion, the standard error, as $SE = \sqrt{\frac{p(1-p)}{n}}$, where p is the population proportion.

- Note that when the population proportion p is not known (almost always), this can be estimated using the sample proportion, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

LO 4. Standard error measures the variability in point estimates from different samples of the same size and from the same population, i.e. measures the sampling variability.

LO 5. Recognize that when the sample size increases we would expect the sampling variability to decrease.

- Conceptually: Imagine taking many samples from the population. When sample sizes are large the sample proportion will be much more consistent across samples than when the sample sizes are small.
- Mathematically: $SE =$, when n increases, SE will decrease since n is in the denominator.

LO 6. Notice that sampling distributions of point estimates coming from samples that don't meet the required conditions for the CLT (about sample size and independence) will not be normal.

* *Reading: Section 5.1 of OpenIntro Statistics*
* *Test yourself:*

1. *For each of the following situations, state whether the variable is categorical or numerical, and whether the parameter of interest is a mean or a proportion.*
   (a) *In a survey, college students are asked whether they agree with their parents' political ideology.*
   (b) *In a survey, college students are asked what percentage of their non-class time they spend studying.*
2. *Explain what is going on in Figures 5.4 and 5.5 of the book (pages 176 and 177).*

LO 7. Define a confidence interval as the plausible range of values for a population parameter.

LO 8. Define the confidence level as the percentage of random samples which yield confidence intervals that capture the true population parameter.

LO 9. Calculate an approximate 95% confidence interval by adding and subtracting 2 standard errors to the point estimate: $point\ estimate \pm 2 \times SE$.

LO 10. Recognize that the Central Limit Theorem (CLT) is about the distribution of point estimates, and that given certain conditions, this distribution will be nearly normal.

- In the case of the proportion the CLT tells us that if

  (1) the observations in the sample are independent, and
  (2) there are at least 10 successes and 10 failures,

then the distribution of the sample proportion will be nearly normal, centered at the true population proportion and with a standard error of $\sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

- When the population proportion is unknown, condition (2) can be checked using the sample proportion.

LO 11. Recall that independence of observations in a sample is provided by random sampling (in the case of observational studies) or random assignment (in the case of experiments).

- In addition, the sample should not be *too* large compared to the population, or more precisely, should be smaller than 10% of the population, since samples that are too large will likely contain observations that are not independent.

LO 12. Recognize that the nearly normal distribution of the point estimate (as suggested by the CLT) implies that a more precise confidence interval can be calculated as

$$point\ estimate \pm z^\star \times SE,$$

where $z^\star$ corresponds to the cutoff points in the standard normal distribution to capture the middle XX% of the data, where XX% is the desired confidence level.

- For proportions this is $\bar{x} \pm Z^\star\sqrt{\frac{p(1-p)}{n}}$.
- Note that $z^\star$ is always positive.

LO 13. Define margin of error as the distance required to travel in either direction away from the point estimate when constructing a confidence interval, i.e. $z^\star \times SE$.

- Notice that this corresponds to half the width of the confidence interval.

LO 14. Interpret a confidence interval as "We are XX% confident that the true population parameter is in this interval", where XX% is the desired confidence level.

- Note that your interpretation must always be in context of the data – mention what the population is and what the parameter is (mean or proportion).

* *Reading: Section 5.2 of OpenIntro Statistics*
* *Test yourself:*
  1. *Explain, in plain English, what is going on in Figure 5.6 of the book (page 182).*
  2. *List the conditions necessary for the CLT to hold. Make sure to list alternative conditions for when we know the population distribution is normal vs. when we don't know what the population distribution is, and the when the sample size is barely over 30 vs. when it's very large.*
  3. *Confirm that $z^\star$ for a 98% confidence level is 2.33. (Include a sketch of the normal curve in your response.)*
  4. *Explain, in plain English, the difference between standard error and margin of error.*

LO 15. Explain how the hypothesis testing framework resembles a court trial.

LO 16. Recognize that in hypothesis testing we evaluate two competing claims:

- the null hypothesis, which represents a skeptical perspective or the status quo, and

- the alternative hypothesis, which represents an alternative under consideration and is often represented by a range of possible parameter values.

LO 17. Construction of hypotheses:

- Always construct hypotheses about population parameters (e.g. population proportion, p) and not the sample statistics (e.g. sample proportion, $\hat{p}$). Note that the population parameter is unknown while the sample statistic is measured using the observed data and hence there is no point in hypothesizing about it.
- Define the null value as the value the parameter is set to equal in the null hypothesis.
- Note that the alternative hypothesis might be one-sided ($\mu <$ or $>$ the null value) or two-sided ($\mu \neq$ the null value), and the choice depends on the research question.

LO 18. Define a p-value as the conditional probability of obtaining a sample statistic at least as extreme as the one observed given that the null hypothesis is true.

$$\text{p-value} = \text{P}(\text{observed or more extreme sample statistic} \mid H_0 \text{ true})$$

LO 19. Calculate a p-value as the area under the normal curve beyond the observed sample proportion (either in one tail or both, depending on the alternative hypothesis). Note that in doing so you can use a Z score, where

$$Z = \frac{\textit{sample statistic} - \textit{null value}}{SE} = \frac{\bar{x} - \mu_0}{SE}$$

- Always sketch the normal curve when calculating the p-value, and shade the appropriate area(s) depending on whether the alternative hypothesis is one- or two-sided.

LO 20. Infer that if a confidence interval does not contain the null value the null hypothesis should be rejected in favor of the alternative.

LO 21. Compare the p-value to the significance level to make a decision between the hypotheses:

- If the p-value $<$ the significance level, reject the null hypothesis since this means that obtaining a sample statistics at least as extreme as the observed data is extremely unlikely to happen just by chance, and conclude that the data provides evidence for the alternative hypothesis.
- If the p-value $>$ the significance level, fail to reject the null hypothesis since this means that obtaining a sample statistics at least as extreme as the observed data is quite likely to happen by chance, and conclude that the data does not provide evidence for the alternative hypothesis.
- Note that we can never "accept" the null hypothesis since the hypothesis testing framework does not allow us to confirm it.

LO 22. Note that the conclusion of a hypothesis test might be erroneous regardless of the decision we make.

- Define a Type 1 error as rejecting the null hypothesis when the null hypothesis is actually true.
- Define a Type 2 error as failing to reject the null hypothesis when the alternative hypothesis is actually true.

LO 23. Choose a significance level depending on the risks associated with Type 1 and Type 2 errors.

- Use a smaller α is Type 1 error is relatively riskier.
- Use a larger α is Type 2 error is relatively riskier.

LO 24. Formulate the framework for statistical inference using hypothesis testing and nearly normal point estimates:

(1) Set up the hypotheses first in plain language and then using appropriate notation.

(2) Identify the appropriate sample statistic that can be used as a point estimate for the parameter of interest.

(3) Verify that the conditions for the CLT holds.

(4) Compute the SE, sketch the sampling distribution, and shade area(s) representing the p-value.

(5) Using the sketch and the normal model, calculate the p-value and determine if the null hypothesis should be rejected or not, and state your conclusion in context of the data and the research question.

LO 25. If the conditions necessary for the CLT to hold are not met, note this and do not go forward with the analysis. (We will later learn about methods to use in these situations.)

LO 26. Distinguish statistical significance vs. practical significance.

* *Reading: Section 5.3 of OpenIntro Statistics*
* *Test yourself:*

1. *List errors in the following hypotheses:* $H_0 : \hat{p} > 0.20$ *and* $H_A : \hat{p} \geq 0.25$
2. *What is wrong with the following statement?*
   "If p-value is large we accept the null hypothesis since a large p-value implies that the observed difference between the null value and the sample statistic is quite likely to happen just by chance."
3. *Suppose a researcher is interested in evaluating the following claim "The proportion of adults in the US is who vote is 40%", and that she believes this is an underestimate.*
   (a) *How should she set up her hypotheses?*
   (b) *Explain to her, in plain language, how she should collect data and carry out a hypothesis test.*
4. *Go back to Section 2.4 and describe the differences and similarities between the hypothesis testing procedure using simulation and using theory. Especially discuss how the calculation of the p-value changes while the definition stays the same.*
5. *If we want to decrease the margin of error, and hence have a more precise confidence interval, should we increase or decrease the sample size?*
6. *In a random sample of 1,017 Americans 60% said they do not trust the mass media when it comes to reporting the news fully, accurately, and fairly. The standard error associated with this estimate is 0.015 (1.5%). What is the margin of error for a 95% confidence level? Calculate a 95% confidence interval and interpret it in context. You may assume that the point estimate is normally distributed (we'll learn how to check this later).*

LO 1. Define population proportion p (parameter) and sample proportion $\hat{p}$ (point estimate).

LO 2. Calculate the sampling variability of the proportion, the standard error, as

$$SE = \sqrt{\frac{p(1-p)}{n}},$$

where p is the population proportion.

- Note that when the population proportion p is not known (almost always), this can be estimated using the sample proportion, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

LO 3. Recognize that the Central Limit Theorem (CLT) is about the distribution of point estimates, and that given certain conditions, this distribution will be nearly normal.

- In the case of the proportion the CLT tells us that if
(1) the observations in the sample are independent,
(2) the sample size is sufficiently large (checked using the success/failure condition: $np \geq 10$ and $n(1-p) \geq 10$),
then the distribution of the sample proportion will be nearly normal, centered at the true population proportion and with a standard error of $\sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

LO 4. Note that if the CLT doesn?t apply and the sample proportion is low (close to 0) the sampling distribution will likely be right skewed, if the sample proportion is high (close to 1) the sampling distribution will likely be left skewed.

LO 5. Remember that confidence intervals are calculated as

$$\text{point estimate} \pm \text{margin of error}$$

and test statistics are calculated as

$$\text{test statistic} = \frac{\text{point estimate - null value}}{\text{standard error}}$$

LO 6. Note that the standard error calculation for the confidence interval and the hypothesis test are different when dealing with proportions, since in the hypothesis test we need to assume that the null hypothesis is true – remember: p-value = P(observed or more extreme test statistic | H_0 true).

- For confidence intervals use $\hat{p}$ (observed sample proportion) when calculating the standard error and when checking the success/failure condition:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- For hypothesis tests use p_0 (null value) when calculating the standard error and checking the success/failure condition:

$$SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$$

- Such a discrepancy doesn't exist when conducting inference for means, since the mean doesn't factor into the calculation of the standard error, while the proportion does.

LO 7. Calculate the required minimum sample size for a given margin of error at a given confidence level, and explain why we use $\hat{p} = 0.5$ if there are no previous studies suggesting a more accurate estimate.

- Conceptually: When there is no additional information, 50% chance of success is a good guess for events with only two outcomes (success or failure).
- Mathematically: Using $\hat{p} = 0.5$ yields the most conservative (highest) estimate for the required sample size.

* *Reading: Section 6.1 of OpenIntro Statistics*
* *Test yourself:*

1. *Suppose 10% of Duke students smoke. You collect many random samples of 100 Duke students at a time, and calculate a sample proportion ($\hat{p}$) for each sample, indicating the proportion of students in that sample who smoke. What would you expect the distribution of these $\hat{p}$s to be? Describe its shape, center, and spread.*
2. *Suppose you want to construct a confidence interval with a margin of error no more than 4% for the proportion of Duke students who smoke. How would your calculation of the required sample size change if you don't know anything about the smoking habits of Duke students vs. if you have a reliable previous study estimating that about 10% of Duke students smoke.*

LO 8. Note that the calculation of the standard error of the distribution of the difference in two independent sample proportions is different for a confidence interval and a hypothesis test.

- confidence interval (and hypothesis test when $H_0 : p_1 - p_2 =$ some value other than 0):

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- hypothesis test when $H_0 : p_1 - p_2 = 0$:

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}},$$

where $\hat{p}_{pool}$ is the overall rate of success:

$$\hat{p}_{pool} = \frac{\text{number of successes in group 1} + \text{number of successes in group 2}}{n_1 + n_2}$$

LO 9. Note that the reason for the difference in calculations of standard error is the same as in the case of the single proportion: when the null hypothesis claims that the two population proportions are equal, we need to take that into consideration when calculating the standard error for the hypothesis test, and use a common proportion for both samples.

* *Reading: Section 6.2 of OpenIntro Statistics*
* *Test yourself:*

1. *Suppose a 95% confidence interval for the difference between the Duke and UNC students who smoke (calculated using $\hat{p}_{Duke} - \hat{p}_UNC$) is (-0.08,0.11). Interpret this interval, making sure to incorporate into your interpretation a comparative statement about the two schools.*

2. *Does the above interval suggest a significant difference between the true proportions of smokers at the two schools?*

3. *Suppose you had a sample of 100 students from Duke where 11 of them smoke, and a sample of 80 students from UNC where 10 of them smoke. Calculate* $\hat{p}_{pool}$.

4. *When and why do we use* $\hat{p}_{pool}$ *in calculation of the standard error for the difference between two sample proportions?*

LO 10. Use a chi-square test of goodness of fit to evaluate if the distribution of levels of a single categorical variable follows a hypothesized distribution.

H_0 : The distribution of the variable follows the hypothesized distribution, and any observed differences are due to chance.

H_A : The distribution of the variable does not follow the hypothesized distribution.

LO 11. Calculate the expected counts for a given level (cell) in a one-way table as the sample size times the hypothesized proportion for that level.

LO 12. Calculate the chi-square test statistic as

$$\chi = \sum_{i=1}^{k} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}},$$

where k is the number of cells.

LO 13. Note that the chi-square distribution is right skewed with one parameter: degrees of freedom. In the case of a goodness of fit test, $df = \#\text{of categories} - 1$.

LO 14. List the conditions necessary for performing a chi-square test (goodness of fit or independence)

(1) the observations should be independent

(2) expected counts for each cell should be at least 5

(3) degrees of freedom should be at least 2 (if not, use methods for evaluating proportions)

LO 15. Describe how to use the chi-square table to obtain a p-value.

* *Reading: Section 6.3 of OpenIntro Statistics*
* *Test yourself:*

1. *Explain the different hypothesis tests one could use when assessing the distribution of a categorical variable (e.g. smoking status) with only two levels (e.g. levels: smoker and non-smoker) vs. more than two levels (e.g. levels: heavy smoker, moderate smoker, occasional smoker, non-smoker).*

2. *Why is the p-value for chi-square tests always "one sided"?*

LO 16. When evaluating the independence of two categorical variables where at least one has more than two levels, use a chi-square test of independence.

H_0 : The two variables are independent.

H_A : The two variables are dependent.

LO 17. Calculate expected counts in two-way tables as

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

LO 18. Calculate the degrees of freedom for chi-square test of independence as $df = (R-1) \times (C-1)$, where R is the number of rows in a two-way table, and C is the number of columns.

LO 19. Note that there is no such thing as a chi-square confidence interval for proportions, since in the case of a categorical variables with many levels, there isn't one parameter to estimate.

* *Reading: Section 6.4 of OpenIntro Statistics*
* *Test yourself:*
    1. *What are the null and alternative hypotheses in chi-square test of independence?*
    2. *Suppose a chi-square test of independence between two categorical variables (one with 5, the other with 3 levels) is yields a test statistic of $\chi^2 = 14$. What's the conclusion of the hypothesis test at 5% significance level?*

LO 20. Use simulation methods when sample size conditions aren't met for inference for categorical variables.

- Note that the t-distribution is only appropriate to use for means, when sample size isn't sufficiently large, and the parameter of interest is a proportion or a difference between two proportions, we need to use simulation.

LO 21. In hypothesis testing

- for one categorical variable, generate simulated samples based on the null hypothesis, and then calculate the number of samples that are at least as extreme as the observed data.
- for two categorical variables, use a randomization test.

LO 22. Use bootstrap methods for confidence intervals for categorical variables with at most two levels.

* *Reading: Sections 6.5 and 6.6 of OpenIntro Statistics*
* *Test yourself:*
    1. *Suppose you want to estimate the proportion of Duke students who smoke. You collect a random sample of 100 students, where only 8 of them smoke. Can you use theoretical methods (Z) to construct a confidence interval based on these data? If not, describe how you could calculate a 95% bootstrap confidence interval.*

Learning Objectives — Chapter 7: Inference for numerical data

LO 1. Use the t-distribution for inference on a single mean, difference of paired (dependent) means, and difference of independent means.

LO 2. Explain why the t-distribution helps make up for the additional variability introduced by using s (sample standard deviation) in calculation of the standard error, in place of σ (population standard deviation).

LO 3. Describe how the t-distribution is different from the normal distribution, and what ?heavy tail? means in this context.

LO 4. Note that the t-distribution has a single parameter, degrees of freedom, and as the degrees of freedom increases this distribution approaches the normal distribution.

LO 5. Use a t-statistic, with degrees of freedom $df = n - 1$ for inference for a population mean:

- Standard error: $SE = \frac{s}{\sqrt{n}}$
- Confidence interval: $\bar{x} \pm t^\star_{df} SE$
- Hypothesis test: $T_{df} = \frac{\bar{x} - \mu}{SE}$

LO 6. Describe how to obtain a p-value for a t-test and a critical t-score ($t^\star_{df}$) for a confidence interval.

* *Reading: Section 7.1 of OpenIntro Statistics*
* *Test yourself:*
  1. *What is the $t^\star$ for a 95% confidence interval for a mean, where the sample size is 13.*
  2. *What is the p-value for a hypothesis test where the alternative hypothesis is two-sided, the sample size is 20, and the test statistic, T, is calculated to be 1.75?*

LO 7. Define observations as paired if each observation in one dataset has a special correspondence or connection with exactly one observation in the other data set.

LO 8. Carry out inference for paired data by first subtracting the paired observations from each other, and then treating the set of differences as a new numerical variable on which to do inference (such as a confidence interval or hypothesis test for the average difference).

LO 9. Calculate the standard error of the difference between means of two paired (dependent) samples as $SE = \frac{s_{diff}}{\sqrt{n_{diff}}}$ and use this standard error in hypothesis testing and confidence intervals comparing means of paired (dependent) groups.

LO 10. Use a t-statistic, with degrees of freedom $df = n_{diff} - 1$ for inference for a population mean:

- Standard error: $SE = \frac{s}{\sqrt{n}}$
- Confidence interval: $\bar{x}_{diff} \pm t^\star_{df} SE$
- Hypothesis test: $T_{df} = \frac{\bar{x}_{diff} - \mu_{diff}}{SE}$. Note that μ_{diff} is often 0, since often $H_0 : \mu_{diff} = 0$.

LO 11. Recognize that a good interpretation of a confidence interval for the difference between two parameters includes a comparative statement (mentioning which group has the larger parameter).

LO 12. Recognize that a confidence interval for the difference between two parameters that doesn?t include 0 is in agreement with a hypothesis test where the null hypothesis that sets the two parameters equal to each other is rejected.

* *Reading: Section 7.2 of OpenIntro Statistics*
* *Test yourself:*
    1. *20 cardiac patients' blood pressure is measured before taking a medication, and after. For a given patient, are the before and after blood pressure measurements dependent (paired) or independent?*
    2. *A random sample of 100 students were obtained and then randomly assigned into two equal sized groups. One group went on a roller coaster while the other in a simulator at an amusement park. Afterwards their blood pressure measurements were taken. Are the measurements dependent (paired) or independent?*

LO 13. Calculate the standard error of the difference between means of two independent samples as $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, and use this standard error in hypothesis testing and confidence intervals comparing means of independent groups.

LO 14. Use a t-statistic, with degrees of freedom $df = min(n_1 - 1, n_2 - 1)$ for inference for a population mean:

- Standard error: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Confidence interval: $(\bar{x}_1 - \bar{x}_2) \pm t^\star_{df} SE$
- Hypothesis test: $T_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE}$. Note that μ_{diff} is often 0, since often $H_0 : \mu_1 - \mu_2 = 0$.

* *Reading: Section 7.3 of OpenIntro Statistics*
* *Test yourself:*
    1. *Describe how the two sample means test is different from the paired means test, both conceptually and in terms of the calculation of the standard error.*
    2. *A 95% confidence interval for the difference between the number of calories consumed by mature and juvenile cats ($\mu_{mat} - \mu_{juv}$) is (80 calories, 100 calories). Interpret this interval, and determine if it suggests a significant difference between the two means.*

LO 15. Calculate the power of a test for a given effect size and significance level in two steps: (1) Find the cutoff for the sample statistic that will allow the null hypothesis to be rejected at the given significance level, (2) Calculate the probability of obtaining that sample statistic given the effect size.

LO 16. Explain how power changes for changes in effect size, sample size, significance level, and standard error.

LO 17. Define analysis of variance (ANOVA) as a statistical inference method that is used to determine if the variability in the sample means is so large that it seems unlikely to be from chance alone by simultaneously considering many groups at once.

LO 18. Recognize that the null hypothesis in ANOVA sets all means equal to each other, and the alternative hypothesis suggest that at least one mean is different.

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

H_A : At least one mean is different

LO 19. List the conditions necessary for performing ANOVA

(1) the observations should be independent within and across groups

(2) the data within each group are nearly normal

(3) the variability across the groups is about equal

and check if they are met using graphical diagnostics.

LO 20. Recognize that the test statistic for ANOVA, the F statistic, is calculated as the ratio of the mean square between groups (MSG, variability between groups) and mean square error (MSE, variability within errors), and has two degrees of freedom, one for the numerator ($df_G = k - 1$, where k is the number of groups) and one for the denominator ($df_E = n - k$, where n is the total sample size).

- Note that you won't be expected to calculate MSG or MSE from the raw data, but you should have a conceptual understanding of how they're calculated and what they measure.

LO 21. Describe why calculation of the p-value for ANOVA is always "one sided".

LO 22. Describe why conducting many t-tests for differences between each pair of means leads to an increased Type 1 Error rate, and we use a corrected significance level (Bonferroni corection, $\alpha^\star = \alpha/K$, where K is the e number of comparisons being considered) to combat inflating this error rate.

LO 23. Describe why it is possible to reject the null hypothesis in ANOVA but not find significant differences between groups as a result of pairwise comparisons.

* *Reading: Section 7.4 and 7.5 of OpenIntro Statistics*
* *Test yourself:*

1. *We would like to compare the average income of Americans who live in the Northeast, Midwest, South, and West. What are the appropriate hypotheses?*
2. *Suppose the sample in the question above has 1000 observations, what are the degrees of freedom associated with the F-statistic?*
3. *Suppose the null hypothesis is rejected. Describe how we would discover which regions' averages are different from each other? Make sure to discuss how many pairwise comparisons we would need to make, and what the corrected significance level would be.*
4. *What visualizations are useful for checking each of the conditions required for performing ANOVA?*