LO 1. 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 observed counts follows the hypothesized distribution, and any ob- served differences are due to chance.

H_A: The distribution of observed counts does not follow the hypothesized distribution.

LO 2. 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 3. Calculate the chi-square test statistic as

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

, where k is the number of cells.

LO 4. Note that the chi-square statistic is always positive, and follows a right skewed distribution with one parameter: degrees of freedom.

LO 5. Note that the degrees of freedom for the chi-square statistic for the goodness of fit test is $df = k - 1$.

LO 6. 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 7. Describe how to use the chi-square table to obtain a p-value.

LO 8. 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 9. Calculate expected counts in two-way tables as

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

LO 10. 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 11. 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.

LO 12. 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 13. 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 14. Use bootstrap methods for confidence intervals for categorical variables with at most two levels.