Chapter 6: Inference for categorical variables
Learning Objectives

Reading: Section 6.1 of OpenIntro Statistics

Video: Sample proportions, YouTube (3:09)

Video: Confidence intervals for population proportions, YouTube (4:18)

Video: Calculating required sample size to estimate population proportions, YouTube (2:45)

Video: One sample Z test for a proportion, YouTube (6:08)

1. Define population proportion p (parameter) and sample proportion $\hat{p}$ (point estimate).
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, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

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 $\sqrt{\frac{p(1-p)}{n}}$.

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

4. Remember that confidence intervals are calculated as

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

and test statistics are calculated as

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

5. Th 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 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.

6. Explain why when calculating the required minimum sample size for a given margin of error at a given confidence level, we use $\hat{p} = 0.5$ if there are no previous studies suggesting a more accurate estimate. Reasons for this are:

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

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.

Reading: Section 6.2 of OpenIntro Statistics

Video: Sampling distributions for proportions, Dr. Çetinkaya-Rundel, YouTube (16:16)

Video: Comparing proportions via hypothesis testing, YouTube (4:05)

Video: Comparing proportions via confidence intervals, YouTube (1:58)

7. 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:

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

8. 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.

Test yourself:

1. Suppose a 95% confidence interval for the difference between the Duke and UNC students who smoke (calculated using $\hat{p}_{Duke} - \hat{p}_U NC$) 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?

Reading: Section 6.3 of OpenIntro Statistics

Video: Chi-square test of goodness of fit, YouTube (4:00)

9. 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 follow the hypothesized distribution, and any observed differences are due to chance.

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

10. 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.

11. Calculate the chi-square test statistic as

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

where k is the number of cells.

12. The chi-square distribution is right skewed with one parameter: degrees of freedom. In the case of a goodness of fit test, $df = k - 1$.

13. 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)

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

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 represented by the upper tail of the chi-square distribution and never the lower tail?

Reading: Section 6.4 of OpenIntro Statistics

15. 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.

16. Calculate expected counts in two-way tables as

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

17. 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.

18. We do not use the chi-square distribution to calculate confidence intervals, because there are several different parameters to consider.

Test yourself:

1. What are the null and alternative hypotheses in a 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) yields a test statistic of $X^2 = 14$. What's the conclusion of the hypothesis test at 5% significance level?

Reading: Sections 6.5 of OpenIntro Statistics

19. 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.

20. To conduct a hypothesis test for a proportion when the success-failure condition is not met 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.

Test yourself: 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?

Reading: Section 6.6 of OpenIntro Statistics

21. To conduct a hypothesis test for comparing two proportions when the success-failure condition is not met and for two-way tables when the sample size condition for the chi-square test is not met, use a randomization test (as described in Section 1.8).