Exploration 6.1: Give peas a chance

Learning Goals:

- Review descriptive and inferential methods for comparing groups with a categorical response variable
- Compare and contrast different statistics for evaluating group differences on a binary response variable

Background: Vegetable production is a challenge for both commercial producers as well as home gardeners in places with colder climates, like Wyoming, USA. Researchers Homer and Groose (2015, "Developing Winterhardy Vegetable Pea for Wyoming, USA: Description of Winter Survival in Early Generation Breeding Lines," *International Journal of Scientific and Research Publications,* Vol 5, Issue 6), carried out a study to develop pea cultivars that could be seeded in late summer/early fall to produce fruit in late winter/early spring. Various cultivars were crossed to create new cultivars that researchers hoped would be more likely to survive to spring. In this exploration, we focus on two newly developed crosses of winterhardy parental cultivars: Specter x Common-3 (SC) combination and Windham x Common (WC) combination. For each plant, the researchers recorded whether the plant survived the 2010-2011 Laramie, WY winter.

STEPS 1 and 2: Ask a research question/Design a study and collect data.

1. Identify the observational units, explanatory variable, and response variable. Classify each variable as quantitative or categorical.

2. Give a possible Sources of Variation diagram for this study.

STEP 3: Explore the data.

The data (two categorical variables) have been organized into a *two-way contingency table*.

	Specter x Common-3 (SC) plants	Windham x Common (WC) plants	Total
Survived	33	44	77
Died	11	42	53
Total	44	86	130

3. Use software to produce a *segmented bar graph* (or a *mosaic plot*) for these data, putting the explanatory variable along the horizontal axis. Use the graph and appropriate *conditional proportions* to summarize the observed association between type of cultivar and survival status. Do you think the association will be statistically significant? If so, can we draw a cause-and-effect conclusion from these data?

STEP 4: Draw inferences beyond the data.

How do we convince ourselves that the difference we are observing didn't happen "by chance alone?" Perhaps there is no association between these variables, but the randomness in the growing process alone produced a larger sample proportion of the survivors from one cultivar than the other cultivar. To estimate how large a difference between the conditional proportions we might see when the cultivar is unrelated to survival, we could apply the **3S Strategy**. To do this, we first need a statistic.

Statistic

4. Compute the difference in the conditional proportions of "survival" for each cultivar (SC - WC). Summarize the comparison.

Key Idea: Using the difference in the conditional proportions as the statistic has some limitations. In particular, when the proportions are small, the difference will also be small, even if one proportion is two or three times larger than the other.

Alternative Statistics

Sometimes researchers also consider other possible numerical summaries to compare two conditional proportions.

Definition: The ***relative risk*** is the ratio of the conditional proportions of success. Often the larger proportion is used in the numerator so that the ratio is greater than one and easier to interpret—the ratio is "how many times larger" one conditional proportion is than the other.

5. Calculate and interpret the relative risk for this study. (*Hint*: Keep in mind that now you are talking about a multiplicative change in the proportions.)

Another way to compare the two groups is through the odds of success, where ***odds*** are defined as the ratio of the proportion of successes to the proportion of failures.

Definition: The ***Odds*** of success compare the proportion of successes to the proportion of failures. If the odds are 1 (to 1), then success and failure are equally likely (proportion of successes = 0.5).

$$odds = \frac{proportion\ of successes}{proportion\ of\ failures} = \frac{number\ of\ succeses}{number\ of\ failures}$$

6. Calculate the odds of survival (which we are taking to be "success") for the Specter x Common-3 cross by calculating the ratio of the proportion of successes in the SC group to the proportion of failures in the SC group. Interpret your result in context.

7. Repeat for the Windham x Common plants.

Definition: *Odds ratios* $\frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}}$, compare the odds of success between two groups, usually putting the group with the larger odds in the numerator. Odds ratios are similar to relative risks but technically need to be interpreted in terms of odds rather than likelihood or chances of success.

8. Calculate and interpret the odds ratio for this study. (*Hint*: Make sure you are discussing the *multiplicative* change in the *odds*.)

9. What would be the value of the odds ratio when there is no difference in the odds between the two groups? Is the odds ratio for this study larger or smaller than this value? Do you think by a lot?

Simulate: After choosing a statistic to summarize the comparison, the next step in the 3S strategy is to simulate other values of the statistic that could have arisen from the random process alone.

10. State appropriate null and alternative hypotheses for this study in terms of the population odds ratio.

11. Explain how you could design a simulation study to determine whether the observed odds ratio provides statistically significant evidence against the null hypothesis that the probability of survival is the same for both types of cross cultivars. (*Hint*: Use your answer to #9.)

Open the **Analyzing Two-way Tables** applet. Check the **2x2** box and enter in the cell counts from the table above (as well as row and column headers). Press **Use Table**. Verify the segmented bar graph (and/or toggle to the mosaic plot). Use the **Statistic** pull-down menu to select **Odds ratio**. Verify the calculation of the sample odds ratio. Check the **Show Shuffle Options** box. Select the **Cards** radio button and you will see blue cards for Successes and green cards for Failures, with the number of successes and failures in each group matching the results of the study.

12. Press the **Shuffle** button once and explain what the applet is doing.

13. Set the **Number of Shuffles** to 999 (for 1,000 total shuffles) and press **Shuffle** again. Include a screen capture of the resulting null distribution. Describe the shape, center, and variability of this distribution. Is the center what you would have predicted? Explain why or why not.

Strength of evidence

14. Use the applet to estimate a p-value. Include a screen capture of your results and explain the process you used and why. (*Hint*: What odds ratio values will provide evidence against the null hypothesis and in favor of the alternative hypothesis that the probability of survival is not the same for both types of cross cultivars? That is, that the long-run odds of survival are not the same.)

Because of the right-skewed shape of the odds ratio null distribution, a log transformation often leads to an approximately normal distribution. This allows us to apply our confidence interval methods to the log-odds ratio.

Check the **ln odds ratio** box on the far left.

15. Does the null distribution of the ln odds ratio appear to be approximately normal? What is the standard deviation? Use these facts to approximate a 95% confidence interval for the population ln odds ratio.

16. Calculate a 95% confidence interval for the population odds ratio by taking $e^{\text{left end point}}$ and $e^{\text{right end point}}$. (You can verify your calculations in the applet by checking the 95% CI for odds ratio box.) Provide a one-sentence interpretation of your interval. Is this confidence interval consistent with your p-value? Explain why or why not.

STEP 5: Formulate conclusions.

17. Summarize the conclusions you have drawn for this study (significance, confidence, causation, generalizability).

STEP 6: Look back and ahead.

18. What cautions or concerns do you have about this study? What changes would you make if you were to conduct a similar study?

Alternative Statistics

Another common statistic is the *chi-square statistic*, which compares the observed number of observations in each cell of the two-way table with the number we would expect to see if the null hypothesis was true.

	Specter x Common-3 (SC) plants	Windham x Common (WC) plants	Total
Survived	33	44	77
Died	11	42	53
Total	44	86	130

19. To compute the *expected counts*, first find the overall sample proportion of plants that survived.

20. Under the null hypothesis, we expect the probability of survival to be the same for both types of cross cultivars. So how many of the 44 Specter x Common-3 cross cultivars do you expect will survive? (*Hint*: This does not need to be an integer so you don't need to round it to one.)

21. Repeat for the 86 Windham x Common plants.

22. How many plants in each category do you expect will die if the null hypothesis is true? (*Hint*: Verify that the row totals and the column totals for these expected counts match those of the original table. In fact, once you have computed the first value in the 2x2 table, you could have used subtraction to find the others.)

Definition: The ***chi-squared statistic*** compares the observed counts in the table to the *expected counts* under the null hypothesis by computing

$\mathrm{X}^2 = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}$.

23. Calculate *(observed-expected)^2/expected* for the first cell in the table.

24. Verify your calculation in the Two-way Tables applet by using the Statistic pull-down menu to select X^2 and also checking the **Show $\mathbf{X}^2$ output** box to see the "Chisq Cell Contributions."

25. Describe the resulting null distribution for the chi-square statistic. Does it look like a normal distribution?

26. Enter the observed chi-square value into the Count Samples box and press Count. How does this p-value compare to what you found with the odds ratio?

Key Idea: If you know one of these statistics (difference in conditional proportions, relative risk, odds ratio, chi-square statistic), you can solve for any of the other statistics. Any table that is "more extreme" for one statistic, will also give a more extreme result for all four statistics. In other words, the simulated two-sided p-value will always match among these statistics in a 2x2 table.

Theory-based Approach

The chi-square statistic has some nice properties, including often being well-modelled by the *chi-square distribution*. Just like there are multiple *t* and *F* distributions, there are multiple chi-square distributions.

Key Idea: Like an *F*-statistic, chi-square statistics are bounded below by zero, and larger chi-square values are stronger evidence against the null hypothesis. The degrees of freedom equals one less than the number of parameters you are testing. The expected value (mean) of a chi-square statistic is its degrees of freedom.

27. In the applet, check the box to **Overlay Chi-square distribution**. Does the chi-square distribution fit the simulated distribution reasonably well? Are the simulation and theory-based p-values similar?

The chi-square distribution helps us judge whether we have a large observed chi-square value. Typically chi-square values larger than 4 are considered large (have p-values below 0.05), but this will depend more exactly on the degrees of freedom. But we will always look for the "upper tail" probability as the p-value, similar to an *F*-statistic. One key advantage of the chi-square statistic is its ability to summarize the association for "larger" tables (more rows and columns).