Exploration 4.1-5.1: Does a poor sense of smell indicate of an increased risk of pneumonia?

Part 1: Association and Confounding

LEARNING GOALS

- Calculate and interpret conditional proportions.
- Interpret conditional proportions as to whether they give any indication of an association between the explanatory and response variables.
- Identify which variable is the explanatory variable and which is response in a study involving two variables.
- Identify potential confounding variables and explain how they provide an alternative explanation for the observed association between the explanatory variable and the response variable.
- Draw a diagram to show how the confounding variable provides an alternative explanation for the observed association between the explanatory variable and the response variable.

The loss of smell was a common symptom of those infected with SARS-CoV-2 during the COVID-19 pandemic. This connection inspired researchers Yuan et al. (2021) to investigate the relationship between a poor sense of smell and a future risk of pneumonia. To do this, they used data from the Health, Aging, and Body Composition study. This study started in 1997 with a large sample of older independently-living Americans and followed them for a number of years measuring various factors of health and well-being. Near the beginning of the study, the researchers evaluated the participants' sense of smell and categorized them into those with a poor sense of smell and those with a moderate to good sense of smell. In the following years, the participants' hospitalizations and reasons for these hospitalizations were recorded. Of the 1,773 participants that did not have any history of pneumonia when the study began, 561 were classified as having a poor sense of smell and the other 1,212 were classified as having a moderate to good sense of smell. During the study, 357 had at least one pneumonia hospitalization. For these 357 participants, 135 came from the 561 with a poor sense of smell, while 222 came from the 1,212 that had a moderate to good sense of smell.

1. Identify the observational units and variables in this study. Also classify each variable as categorical (also binary?) or quantitative.

2. For each sense-of-smell group, what proportion had at least one pneumonia hospitalization? (Write the result as a decimal)? Which is larger?

   Poor sense of smell:

   Moderate to good sense of smell:

Definition
Two variables are ***associated*** or related if the value of one variable gives you information about the value of the other variable. When comparing two groups this means that the proportions or averages take different values in the two groups.

3. Do the two variables appear to be associated? How are you deciding?

Often, when a study involves two associated variables, it is natural to consider one the ***explanatory variable*** and the other the ***response variable***.

Definitions
The ***explanatory variable*** is the variable we think is "explaining" the change in the response variable and the ***response variable*** is the variable we think is being impacted or changed by the explanatory variable. The explanatory variable is sometimes called the independent variable and the response variable is sometimes called the dependent variable.

4. Which would you consider the explanatory variable in this study? Which is the response? (That is, what are the *roles* of these variables in this study?)

There are two possible explanations for why those with a poor sense of smell had a higher proportion of pneumonia hospitalizations than the others:

- A poor sense of smell helps *cause* pneumonia.
- A poor sense of smell does *not* help cause pneumonia, but some other issue (variable) explains why those with a poorer sense of smell are more likely to develop pneumonia. In other words, a third variable is at play, which is related to both the sense of smell and pneumonia.

(Of course, another explanation of why one group had a higher rate of pneumonia is random chance. Using methods you will learn later, there is strong evidence that you can rule out random chance in this case.)

5. Consider the second explanation. Suggest a plausible third variable that could explain why those with a poor sense of smell would be more likely to be hospitalized with pneumonia. (Make sure it's clear in your explanation that your variable might affect someone's sense of smell as well as developing pneumonia.)

Definition
A ***confounding variable*** is a variable that is associated both to the explanatory and to the response variable in such a way that its effects on the response variable cannot be separated from those of the explanatory variable.

6. Would you consider your suggested explanation in #5 to be a confounding variable? Explain.

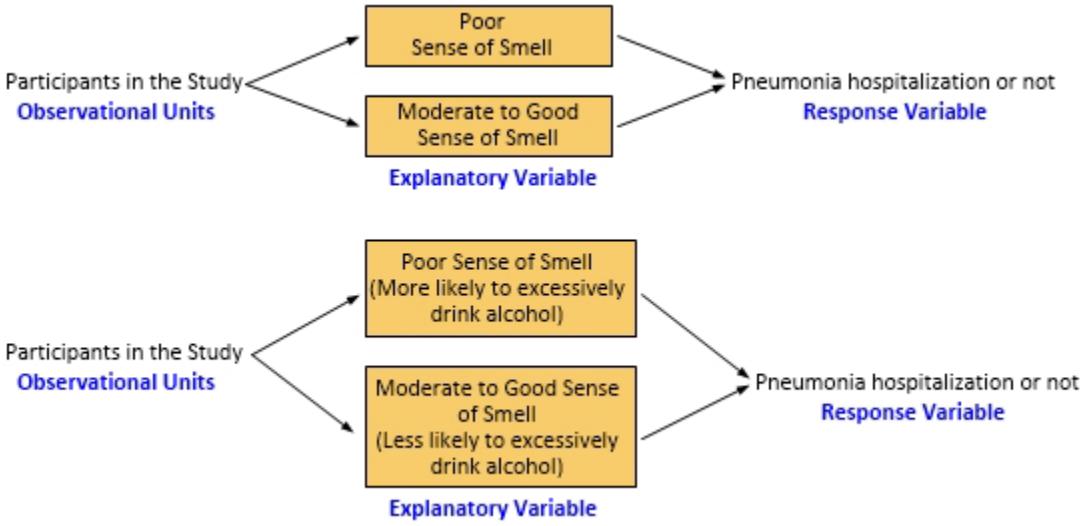
The researchers identified a number of possible confounding variables including age, education, smoking, exercise, body-mass index, and excessive alcohol drinking. Let's take a look at the last one. Near the beginning of the study when they determined the participants' sense of smell, they also asked them whether they had five or more drinks of any kind of alcohol almost every day to determine whether someone was an excessive drinker. Of the 2,494 participants, 244 were identified as excessive drinkers. They found that 109 of these 244 had a poor sense of smell and 699 of those that were not excessive drinkers had a poor sense of smell.

7. Are those identified as excessive drinkers more likely to have a poor sense of smell than those that were not excessive drinkers? Calculate the relevant (conditional) proportions to support your answer.

8. Your answer in #7 should give evidence that excessive drinking is associated with a poor sense of smell. What else is needed to establish that excessive drinking is a confounding variable?

9. According to the Centers of Disease Control and Prevention, people who abuse alcohol are 10 times more likely to develop pneumonia than nondrinkers. Explain how this and your answer to #7 establish that excessive drinking is a confounding variable that prevents drawing a cause-and-effect conclusion between a poor sense of smell and pneumonia hospitalizations.

Confounding explains why you cannot draw a cause-and-effect conclusion from association alone: The groups defined by the explanatory variable could differ in more ways than just the explanatory variable when confounding is present. The diagram below illustrates this. The top panel shows the study design: Observational units (participants in the study) are sorted into groups according to the explanatory variable (whether or not they have a poor sense of smell). Then the response (pneumonia hospitalization or not) was observed. The bottom panel shows the confounding: People with a poor sense of smell were also more likely to excessively drink alcohol. So the two groups, poor smellers and moderate to good smellers are different in another way (likelihood for excessive alcohol drinking).



As stated earlier, the researchers identified a number of possible confounding variables, one of which was excessive alcohol drinking. Through more advanced statistical techniques, they developed models that controlled for some of these possible confounding variables. Controlling for excessive drinking allowed them to estimate the effect of a poor sense of smell that is separate from the influence of excessive alcohol drinking. This was done by adjusting for the difference in the proportion of excessive drinkers between poor smellers and the moderate to good smellers. This helped them move towards establishing a cause-and-effect relationship between the sense of smell and pneumonia. A better way to establish a cause-and-effect relationship, that you will see in the next section, is with an experiment. Many times, however, experiments cannot be done. For an experiment to be done in this case, the researchers would have to be able to control someone's sense of smell throughout their life. Which, of course, they can't.

Part 2: Comparing Two Groups: Categorical Response

LEARNING GOALS

- Organize counts into a two-way table, when data are available on two categorical variables for the same set of observational units.
- Calculate conditional proportion of successes, for different categories of the explanatory variable, and use these conditional proportions to decide whether there is preliminary evidence of an association between the explanatory and response variables.
- Create a segmented bar chart or mosaic plot to display data available on two categorical variables for the same set of observational units.
- Calculate and interpret relative risk.

When analyzing data on two categorical variables, a useful first step is to organize the data into a *two-way table* of counts. When you've compared the variables in this exploration, they all have consisted of just two categories. This means the two-way table you'll look at can also be referred to as a 2 × 2 table. Remember that of the 1,773 participants that did not have any history of pneumonia when the study began, 561 were classified as having a poor sense of smell and the other 1,212 were classified as having a moderate to good sense of smell. During the study, 357 had at least one pneumonia hospitalization. For these 357 participants, 135 came from the 561 with a poor sense of smell, while 222 came from the 1,212 that had a moderate to good sense of smell.

10. Use the results to fill in the counts in the following ***two-way table***. (*Hint*: Start by entering the numbers that were given into the appropriate cells. Then use addition or subtraction to fill in the rest of the cells. Make sure that both rows and both columns add up to the appropriate totals.)

	Poor sense of smell	**Moderate to good sense of smell**	**Total**
At least one pneumonia hospitalization			
No pneumonia hospitalization			
Total			

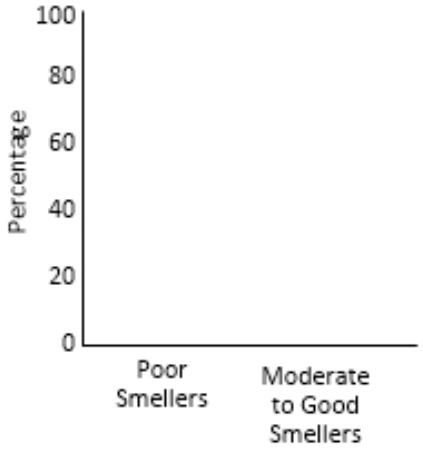
11. A next step in analyzing categorical data in a two-way table is to calculate ***conditional proportions***. A conditional proportion simply means that you consider the counts in only one category of the explanatory variable at a time. (e.g. Only on those with a poor sense of smell.) You should have done this back in #2. Let's do this again using different language.

    a. What is the conditional proportion *of those with a poor sense of smell* that had at least one pneumonia hospitalization?

    b. What is the conditional proportion *of those with a moderate to good sense of smell* that had at least one pneumonia hospitalization?

A ***segmented bar graph*** is an appropriate display for graphing data in a two-way table. Such a graph contains rectangles of total height 100% for each category of the explanatory variable (on the horizontal axis). Then segments divide up each rectangle according to the conditional proportions of the response variable categories.

12. Create a segmented bar graph to display the data from this study: First draw two rectangles, one for those with a poor sense of smell and one for those with a moderate to good sense of smell, each with total height 100%. Then draw a horizontal line in each rectangle corresponding to the conditional proportion for those that had at least one pneumonia hospitalization (on the vertical axis), as you calculated in #11. Label the bottom segment of each bar with "yes" (at least one hospitalization) and the top segments with "no".



13. Do the conditional proportions and segmented bar graph appear to provide some evidence that it is more likely that someone will be hospitalized with pneumonia if they have a poor sense of smell than if they have a moderate to good sense of smell? Explain.

A next step is to produce a single number (statistic) to summarize the data.

14. What arithmetic operation might you perform on the two conditional proportions to obtain a single statistic?

15. Calculate the *difference* in conditional proportions of pneumonia hospitalizations between those with a poor sense of smell and those with a moderate to good sense of smell. Does the value of this difference strike you as noteworthy?

16. Now calculate a different statistic: the *ratio* of conditional proportions of pneumonia hospitalizations between those with a poor sense of smell and those with a moderate to good sense of smell. This ratio is called a ***relative risk***. Write a sentence interpreting this ratio value. Does the value of this ratio strike you as noteworthy?

Definition: ***Relative risk*** is the ratio of two conditional proportions. It indicates how many times greater the risk of an outcome is for one group compared to the risk for the other group.

As stated earlier, using methods you will learn later, there is strong evidence that you can rule out random chance as to why the two conditional proportions in this study are so different. In other words, there is strong evidence of an association between sense of smell and pneumonia in the larger

population. Perhaps the difference in the two sample proportions did not seem that striking to you. But remember, when we have large sample sizes, like we do here, even a fairly small effect can be statistically significant. The small p-value that you would find for these data tells us there is strong evidence of a genuine difference. It does not tell us that the difference is large.

Further Analysis

17. Produce a hypothetical 2 × 2 table with the same marginal totals, but with the property that the data reveal virtually *no association* between the two variables. In other words, produce a table so that the conditional proportions are very similar between the two groups. (*Hint*: For the two conditional proportions to be the same, they must be the same as the overall proportion of people that had at least one pneumonia hospitalization.)

	Poor sense of smell	**Moderate to good sense of smell**	**Total**
At least one pneumonia hospitalization			357
No pneumonia hospitalization			1,416
Total	561	1,212	1,773

18. What would the segmented bar graph look like for a 2 × 2 table with virtually no association between the variables? Explain your answer.

Reference

Yuan, Yaqun, et al., "Poor olfaction and pneumonia hospitalisation among community-dwelling older adults: a cohort study," *The Lancet Healthy Longevity,* Vol 2, Issue 5, E275-E282 (2021) https://doi.org/10.1016/S2666-7568(21)00083-0