Relationships Between Categorical Variables

In 1958, Hammond and Horn published a study investigating the link between lung cancer and smoking. To begin, 187,783 healthy white men from nine states were asked to complete a questionnaire on their smoking habits. Over the course of the next 44 months the men were tracked, and for those that died during that period, their cause of death was recorded. The following two-way contingency table summarizes the data from the study:

	Lung Cancer Death	**Alive (or other cause)**	**Total**
Regular Smoker	397	78,557	78,954
Not Regular Smoker	51	108,778	108,829
Total	448	187,335	187,783

a. Identify the explanatory variable and the response variable.

b. Fill in the table below with the *row percentages* for the two-way table:

	Lung Cancer Death	**Alive (or other cause)**	**Total**
Regular Smoker			100%
Not Regular Smoker			100%
Total			100%

c. This time, fill in the table below with the *column percentages* for the two-way table:

	Lung Cancer Death	**Alive (or other cause)**	**Total**
Regular Smoker			
Not Regular Smoker			
Total	100%	100%	100%

d. We call row and column percentages *conditional proportions*, because they are calculated by conditioning the number of successes upon which category they fall into. If the conditional proportions across rows or columns appear to be significantly different, we can conclude that there is an association between the explanatory and response variables. Based on your answers to (b) and (c), does there appear to be an association between smoking regularity and death due to lung cancer? Why or why not?

e. From the table, calculate the difference in the proportion of "smokers" who died of lung cancer and the proportion of "nonsmokers" who died of lung cancer. Does this difference seem large? Do you think this is a good measure by which to judge if there is an association between smoking and lung cancer? Why or why not?

f. The *risk* that a randomly selected individual within a group falls into the undesirable category is simply the proportion in that cell. The *relative risk* is the ratio of the risks in two different categories:

$$\text{Relative Risk} = \frac{\text{Risk in Category 1}}{\text{Risk in Category 2}}.$$

Compute the relative risk of lung cancer for regular smokers compared to nonsmokers. What does this value tell you?

g. We often treat the denominator in the relative risk as a *baseline risk* since it is often the risk computed for the non-treatment or control case (i.e. nonsmokers). Using this, we can compute the *percent increase in risk* that is observed when comparing a special category to a baseline category:

$$\text{Percent increase in risk} = \frac{\text{Difference in risks}}{\text{Baseline risk}} \times 100\%$$

Compute the percent increase in risk for lung cancer observed when one compares a smoker to a nonsmoker. What does this value tell you?

h. The *odds* of an event compare the chance that the event happens to the chance that it does not:

$$\text{Odds} = \frac{\text{Number of successes in the group}}{\text{Number of failures in the group}}.$$

Define "success" as a death from lung cancer, and compute the odds of dying of lung cancer as a smoker and then as a nonsmoker.

i. To compare the odds of success in two different groups, we use the *odds ratio*, defined as

$$\text{Odds Ratio} = \frac{\text{Odds in Group 1}}{\text{Odds in Group 2}}.$$

Compute the odds ratio of lung cancer between the smoking and nonsmoking groups. What does this value tell you?

j. Write a sentence or two summarizing what you can conclude from this study.

Relationships Between Categorical Variables: Practice Questions

1. The following table shows data of 2858 12th graders for grades usually achieved in school and how often the respondent puts on sunscreen when going out in the sun for more than 1 hour.

	Sunscreen Use			
Grades	**Never/Rarely**	**Sometimes**	**Always/Most Times**	**Total**
A's and B's	1322	450	285	2057
C's	568	83	47	698
D's and F's	85	15	3	103
Total	1975	548	335	2858

   a. Among students who usually get A's and B's in school, what percentage never or rarely use sunscreen? Is this a row percentage or a column percentage?

   b. Among students who sometimes wear sunscreen, what percentage usually gets C's in school? Is this a row percentage or a column percentage?

   c. What percentage of the overall sample usually gets A's and B's in school and also uses sunscreen always or most times?

   d. Determine a complete table of row percentages:

	Sunscreen Use			
Grades	**Never/Rarely**	**Sometimes**	**Always/Most Times**	**Total**
A's and B's				
C's				
D's and F's				
Total				

   e. Do your percentages from part (d) indicate that there is an association between sunscreen use and grades in school? Why or why not?

2. For each of the following measures, give a value that would indicate that there is no difference between the two groups being compared:

   a) Relative risk: b) Odds ratio: c) Percent increase in risk:

3. If the baseline risk of a certain disease for nonsmokers is 1% and the relative risk of the disease is 5 for smokers compared to nonsmokers, what is the risk of the disease for smokers?

4. Using the terminology you've learned today, what term applies to each of the boldface numbers in the following news quotations?

   a. "Fontham found increased risks of lung cancer with increasing exposure to secondhand smoke, whether it took place at home, at work, or in a social setting. A spouse's smoking alone produced an overall **30** percent increase in lung-cancer risk." (*Consumer Reports*, January 1995, p. 28)

   b. "What they found was that women who smoked had a risk of getting lung cancer **27.9** times as great as non-smoking women; in contrast, the risk for men who smoked regularly was only **9.6** times greater than that for male nonsmokers." (Taubes, 1993, p. 1375)

   c. "**One student in five** reports abandoning safe-sex practices when drunk." (*Newsweek*, December 19, 1994, p. 73)