Contingency Tables

Ryan Miller

Grinnell College
Statistics

Introduction

- Data visualizations, such as stacked/conditional bar charts, allow us to *qualitatively* assess relationships between two categorical variables
  - **Contingency tables**, or two-way frequency tables, provide a basis for *quantitatively* assessing these relationships:

	HSI	Not HSI
Religious	55	547
Secular	35	311

Conditional Proportions

Conditional proportions are relative frequencies within subgroups defined by one of the variables in a contingency table.

Row proportions:

	HSI	Not HSI
Religious	0.091	0.909
Secular	0.101	0.899

Column proportions:

	HSI	Not HSI
Religious	0.611	0.638
Secular	0.389	0.362

Conditional Proportions and Association

Differences in proportions, also known as **risk difference** in epidemiology, are the most common way to express the association in a contingency table:

	HSI	Not HSI
Religious	0.091	0.909
Secular	0.101	0.899

The chances that a secular institution is an HSI are 1 percentage-point higher (difference in proportions of 0.01) than the chances a religious institution is an HSI.

Which Conditional Proportions?

The contingency table below describes the survival of crew members and first class passengers aboard the Titanic cruise ship:

	Survived	Died
Crew	212	673
1st Class	203	122

1) Which group was more likely to survive the shipwreck?
2) Did you use row or column proportions? Why is the other choice unable to answer this question?

Which Conditional Proportions? (cont.)

- Notice the *proportion of survivors who were crew* is $\frac{212}{212+203} = 0.51$, while the proportion of survivors who were first class passengers is $\frac{203}{212+203} = 0.49$
  - However, using this information is problematic because the *marginal distribution* of 1st class/crew is *concentrated towards crew*
- In other words, most of the survivors were crew members because there were so many more crew members on board, not because the individual crew members were more likely to survive

Relative Risk

- The CDC estimates that the 10-year risk of developing lung cancer for a smoker is 0.483%, while the risk is only 0.045% for a non-smoker
  - This is a risk difference of just 0.4 percentage points (a difference in proportions of 0.004)
  - Does this suggest that smoking is not a meaningful risk factor for lung cancer?

Relative Risk

- For rare events (such as the development of most diseases, including lung cancer), relative comparisons of risk are more informative than absolute ones
  - The **risk ratio**, also known as **relative risk**, is a popular measure of association for contingency tables
- For the lung cancer example, the relative risk is $0.00483/0.00045 = 10.73$, indicating the risk of developing lung cancer is more than 10 times greater among smokers than it is among non-smokers
  - This conveys a very different message than the risk difference of 0.4 percentage points

Shortcomings of Relative Risk

In the 1860s, when the germ theory of disease was in its infancy, Joseph Lister performed an experiment to evaluate sterilization practices during surgical procedures. He randomly assigned his patients to one of two protocols and tracked their survival:

	Survived	Died
Sterile	34	6
Conventional	19	16

- What is the risk of *death* (conditional proportion) among the "sterile" group? What is it among the "conventional" group?
  - What is the relative risk of death for those experiencing the conventional protocol relative to the sterile protocol?
- What is the risk of *survival* among each group?
  - What is the relative risk of *survival* for the sterile protocol vs. the conventional protocol?

Shortcomings of Relative Risk

- Relative risk is an *asymmetric measure* of association
  - Subjects had a 3.05 times greater risk of dying in the conventional surgery group relative to the sterile group
  - Subjects had a 1.57 times greater risk of surviving in the sterile group relative to the conventional group

Odds and Odds Ratios

- The **odds ratio** is a *symmetric measure* of association that is commonly used in the analysis of categorical data
  - As the name implies, an odds ratio is a ratio of two **odds**, which themselves are a ratio of how often an event occurs relative to how often it does not occur
  - An odds of 3, or "3 to 1", means an event can be expected to occur 3 times for every 1 time it doesn't occur (implying a 75% probability of occurrence)

Odds and Odds Ratios

	Survived	Died
Sterile	34	6
Conventional	19	16

- For Lister's experiment, the odds of death are $6/34 = 0.176$ in the sterile group and $16/19 = 0.842$ in the conventional group
  - Thus, the odds ratio comparing the relative chances of *death* in the *conventional group* to the *sterile group* is $0.842/0.176 = 4.78$
    - The odds of a subject dying under the conventional surgery were 4.8 times their odds of death under the sterile protocol
- The odds of survival are $34/6 = 5.667$ in the sterile group and $19/16 = 1.187$ in the conventional group
  - Thus, the relative chances of *survival* in the *sterile group* to the *conventional group* is $5.667/1.187 = 4.78$

Summary

- Two-way frequency tables (contingency tables) are the starting point when analyzing the relationship between two categorical variables
- The difference in conditional proportions (risk difference) or ratio of conditional proportions (relative risk) are numerical ways to quantify the strength of an association
- The odds ratio is often preferred when a relative comparison is beneficial (ie: rare events) due to it being a symmetric measure of association