Descriptive Statistics

Part 1 - Categorical variables and contingency tables

Ryan Miller

Grinnell College
Statistics

Review

- We'll work with data organized into "tidy" format
  - Each row records data for a **case**/**observation** with each column corresponding to a **variable**
- *Data visualization* is one approach used to understand the contents and patterns contained within a data set
  - A **univariate graph** shows the *distribution* of a single variable
  - A **Bivariate graph** shows the relationship between variables
    - We often want to assess whether an **explanatory variable** is associated with a **response variable**

Descriptive Statistics

- Data visualizations convey a lot of information, but they lend themselves towards *qualitative* descriptions of a distribution or association
  - Example: "X and Y have a weak, positive, linear relationship"

Descriptive Statistics

- Data visualizations convey a lot of information, but they lend themselves towards *qualitative* descriptions of a distribution or association
  - Example: "X and Y have a weak, positive, linear relationship"
- **Descriptive statistics** (or numerical summaries) are numerical values calculated from the data that *quantitatively* summarize a distribution or an association
  - Example: "X and Y have a correlation coefficient of $r = 0.34$"

Descriptive Statistics for Categorical Variables

- Univariate statistics:
    - **Frequencies**: counts of how many cases belong to a particular category
    - **Proportions**: fractions based upon frequencies, sometimes called *relative frequencies*
- One-way table (frequencies):

	Frequency
Private	647
Public	448

- One-way table (proportions):

	Proportion
Private	0.591
Public	0.409

Descriptive Statistics for Categorical Variables

- Bivariate/multivariate statistics:
  - **Conditional Proportions**: relative frequencies within a subgroup of data defined by other variables in the data
- Two-way table (frequencies):

	Private	Public
Far West	59	45
Great Lakes	125	64
Mid East	126	72
New England	44	27
Plains	84	42
Rocky Mountains	8	22
South East	163	130
South West	38	46

Conditional Proportions

In a two-way table there are 2 sets of conditional proportions

1. *Row proportions* (conditioning on the `Region` variable):

	Private	Public
Far West	0.567	0.433
Great Lakes	0.661	0.339
Mid East	0.636	0.364
New England	0.620	0.380
Plains	0.667	0.333
Rocky Mountains	0.267	0.733
South East	0.556	0.444
South West	0.452	0.548

56.7% of colleges in the Far West region are private schools

Conditional Proportions

2. *Column proportions* (conditioning on the `Type` variable):

	Private	Public
Far West	0.091	0.100
Great Lakes	0.193	0.143
Mid East	0.195	0.161
New England	0.068	0.060
Plains	0.130	0.094
Rocky Mountains	0.012	0.049
South East	0.252	0.290
South West	0.059	0.103

9.1% of all private colleges are located in the Far West region

Practice

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?

Practice (solution)

1) Using *row proportions*, $\frac{212}{623+212} = 0.24$, or 24% of the crew survived; while $\frac{203}{122+203} = 0.62$, or 62% of first class passengers survived.
2) This question cannot be answered using column proportions. 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 $\frac{203}{212+203} = 0.49$
   - Conditioning on the column variable is problematic here because the *marginal distribution* of 1st class/crew is *skewed towards crew*
   - In other words, most of the survivors were crew members because there were so many more crew members, not because the individual crew members were more likely to survive

Contingency Tables

Scenarios with a binary explanatory and binary response variable are often summarized using a **contingency table** (a special case of a two-way frequency table):

	Event	No Event
Exposure	A	B
No Exposure	C	D

Contingency tables are widely used in fields like epidemiology to relate risk factors or exposures to the occurrence of an event or onset of a disease.

Summarizing Risk

In a contingency table, **risk** is estimated as the *relative frequency of the event*. This leads to 3 ways to describe the relationship between exposure and risk:

1. **Risk difference** - The risk among the exposed minus the risk among the non-exposed (ie: difference in conditional proportions)
2. **Risk ratio (relative risk)** - Ratio of the risk among the exposed over the risk among the non-exposed (ie: ratio of conditional proportions)
3. **Odds Ratio** - A ratio of *odds* (we'll explain these later today)

Example

Consider the following study, which tracked a cohort of 6,168 women born in the 1980s in search of risk factors for breast cancer

	Breast Cancer	No Cancer
Birth Before Age 25	65	4475
Birth After Age 25	31	1157

1. What is the *risk difference* observed in this study?
2. What is the *relative risk* observed in this study?
3. Which descriptive statistic do you think is more useful?

Note: Some women in this cohort never had children and are not included in this contingency table

Example (solution)

1. The *risk difference in this study* is $\frac{31}{31+1157} - \frac{65}{65+4475} = 0.012$ (1.2%)
   - This seems to suggest a slightly elevated risk
2. The *relative risk of breast cancer* is 1.84 times higher (elevated by 84%) for women who gave birth after age 25
   - This seems to tell a different story than the 1.2% risk difference
3. If you had to report only one, relative risk is more useful for a rare event. However, it's always prudent to report as much information as possible to paint a complete picture.

Another example

In 1986, a case-control study investigating the relationship between smoking and oral cancer, researchers collected the smoking history of 304 cases with oral cancer and 139 controls without oral cancer. Data from the study are summarized below:

	Cases	Controls
≥ 16 cigarettes per day	255	93
< 16 cigarettes per day	49	46

1. Calculate a relative risk for this contingency table. Based upon your knowledge of cancer, does this seem reasonable?
2. What might be problematic about trying to calculate a relative risk in this study?

Another example (solution)

- We'd calculate the relative risk as $0.733/0.516 = 1.42$
- However, 51.6% of the low exposure subjects and 73.3% of the high exposure subjects had cancer
    - This is not reasonable, particularly for the control group
- It turns out this type of study (case-control design) is incompatible with relative risk
    - As an illustration, think about what would happen if we recruited more cases without changing the number of controls

Odds Ratios

- Relative risks are only applicable to data collected using certain study designs
  - Consequently, **odds ratios** tend to be more widely used since they can be used for a wider variety of study designs

Odds Ratios

- Relative risks are only applicable to data collected using certain study designs
  - Consequently, **odds ratios** tend to be more widely used since they can be used for a wider variety of study designs
- The *odds* of an event is a ratio itself, it is how many times an event occurs relative to how many times it doesn't occur
  - If there is a 50% probability of an event, the odds are 1, which we often express as "1 to 1"
  - If there is a 75% probability of an event, the odds are 3, which we often express as "3 to 1"
- An *odds ratio* is a ratio of the odds of an event for one group relative to the odds of that event for another group

Odds Ratios

Let's revisit our case-control data:

	Cases	Controls
$\geq$ 16 cigarettes per day	255	93
< 16 cigarettes per day	49	46

- The odds of cancer (being a case) among high-frequency smokers is $255/93 = 2.742$
- The odds of cancer (being a case) among low-frequency smokers is $49/46 = 1.065$
- Thus, the odds ratio is $\frac{(255/93)}{(49/46)} = \frac{2.742}{1.065} = 2.57$
    - So the odds of a high-frequency smoker having cancer are 2.57 times (or 157% higher) than the odds of a low-frequency smoker having cancer

An Easier Odds Ratio Formula

Given a generic contigency table:

	Event	No Event
Exposure	A	B
No Exposure	C	D

$$\text{Odds Ratio} = \frac{a*d}{b*c}$$

- This formula also makes it clear why odds ratios work for case-control studies
  - If we recruited more cases, the fraction a/c should remain roughly the same
  - The same goes for the fraction b/d if more controls are recruited

Conclusion

- **Descriptive statistics** are numerical summaries of a distribution or an association
  - For categorical variables, descriptive statistics typically stem from *frequency tables*, with conditional proportions being particularly useful
- For the special case of two *binary* categorical variables we can use contingency tables
  - Risk difference, relative risk, and odds ratio are all ways to summarize the association present in a **contingency table**
  - We'll need to carefully consider study design (a future topic) when interpreting these measures (for now we'll focus on calculating them)