Categorical Variables - Measuring Association and Study Design

Ryan Miller

Inference on Categorical Variables

- Consider a 2x2 two-way frequency table relating two binary categorical variables
  - The data summarized in the table could be analyzed using a difference in proportions test, or using a χ^2 test of association
- In practice, χ^2 tests are used far more frequently than difference in proportions tests
  - χ^2 tests are not restricted to 2x2 tables
  - χ^2 tests make sense for a lot of common observational study designs, including: *prospective studies*, *retrospective studies*, and *cross-sectional studies*
- To understand why this is, we'll need to take a brief look each of these different observational designs

Study Design

- Randomized controlled experiments are the gold standard of study designs, but they aren't always feasible
- Usually, the next best design is a **prospective study** (sometimes called a cohort study)
  - In this type of study researchers recruit a large sample that is representative of a population
  - They follow the sample forward in time, classifying individuals into groups based upon their exposure to a risk factor
  - They then wait and observe the outcomes in each exposure group

Prospective Study Example

- CDC researchers tracked a cohort of 6,168 women born in the 1960s in hopes of finding risk factors that led to breast cancer
  - One risk factor they studied was the age at which each woman gave birth to their first child:

	Didn't Develop Cancer	Developed Breast Cancer
Before Age 25	4475	65
After Age 25	1157	31

1. Based upon this study's design, do you believe these data can be used to estimate the overall proportion of this population that develops breast cancer?
2. Enter the data from this table into Minitab and perform a χ^2 test of association (Use Stat -> Tables -> Chi-Square Test for Association). What do you conclude from the test?

Prospective Study Example (solution)

	Developed Breast Cancer	Didn't Develop Cancer
Before Age 25	4475	65
After Age 25	1157	31

- Yes, these data can be used to estimate the proportion of the population that develops breast cancer. We'd estimate that 1.6% of women represented by the cohort will develop breast cancer
- The test statistic is: $\chi^2 = 7.8$ and the p-value is 0.005
- There is evidence that the age at which a woman has their first child is associated with breast cancer risk, although this association could be due to confounding factors

Retrospective Studies

- Tracking a cohort of several thousand people for decades is very expensive and time consuming
- Oftentimes a more feasible approach is to collect a sample of people that experienced an outcome of interest, a second sample of people that did not experience the outcome, and then look backward in time and see which individuals in the two samples were exposed to the risk factor
  - This is design is called a **retrospective** or **case-control** study
  - For example, the CDC could have studied breast cancer risk by recruiting 100 women with breast cancer and 100 women without breast cancer and then asking each woman what age they had their first child at

Retrospective Studies

- Retrospective studies are popular because they are cheap and easy to conduct
- Unfortunately, they are much more prone to sampling biases than prospective studies
  - In a prospective study, the study's population is determined at it's onset, so all individuals with/without the outcome of interest necessarily come from the same population
  - In a retrospective study, the cases and controls are recruited separately after having developed or not developed the outcome, which makes it challenging to ensure that both samples actually came from the same population
- Recall bias can be particularly problematic in case-control studies, individuals experiencing more severe outcomes are more likely to remember certain details from their past
  - For example, individuals sick with food poisoning are more likely to be able to recall their meals prior to getting sick

Retrospective Studies Example

- In a 1986 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	49	46
$\geq$ 16 cigarettes per day	255	93

1. Based upon this study design, do you believe these data can be used to estimate the proportion of the population that develops oral cancer?
2. Enter the data from this table into Minitab and perform a χ^2 test of association. What do you conclude from the test?

Retrospective Studies Example (solution)

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

- No, in this design the subjects were recruited after they developed the outcome. There is no way that 69% of the population develops oral cancer.
- The χ^2 test statistic is 16.3 and the p-value is nearly 0
- There appears to be strong evidence that smoking is related with oral cancer, although we must be cautious because of possible biases and confounding that might be present with this study's design

Cross-sectional Studies

- The weakest type of observational study design is the **cross-sectional study**
- In this design, researchers gather a single sample at snapshot in time and cross-classify the sample based upon who has the risk factor and who has the outcome of interest
- To illustrate the weakness of this design, consider a study looking at the prevalence of asthma among factory workers and the general public
  - Workers suffering from asthma are likely to quit their factory jobs and not be included in a cross-sectional sample
  - A prospective, or even a retrospective study, could identify this trend, but a cross-sectional design cannot not

Measuring Association

- I mentioned previously that the χ^2 is popular in part because it works for each of these study designs
  - The χ^2 test provides a measure of evidence against the null hypothesis that two variables aren't associated
  - But we've learned that effect size is an important consideration in addition to statistical significance
- We've seen that differences in proportions are related to the χ^2 test for 2x2 tables
  - However, using this metric as a measure of effect size has its shortcomings

Measuring Association

- It has been estimated that over a 10-year timespan, smokers have a 0.483% probability of developing lung cancer, while non-smokers have a 0.045% probability of developing lung cancer
  - The **risk difference** here is just 0.004
  - Quantifying association using a difference in proportions makes it look like smoking has a very minor impact on lung cancer
- In many scenarios, in particular rare events, associations are best expressed using ratios
  - In the smoking and lung cancer example, the **relative risk** is $0.483/0.045 = 10.7$
  - That is, smokers are 10.7 times more likely to develop lung cancer than non-smokers

Absolute vs. Relative Risks

- In our example, smoking has a very large relative effect on the risk of lung cancer, but a very small absolute effect
  - Even among smokers lung cancer is rare
- A contrary example is the relationship between smoking and coronary artery disease (CAD)
  - For a 10-year timespan, an estimated 2.947% of smokers are expected to develop CAD, while 1.695% of non-smokers are expected to develop the disease
  - With coronary artery disease, the risk difference is 1.25% but the relative risk is only 1.7

Relative Risk and Study Design

A well-known case-control study published in 1969 examined the relationship between oral contraceptive (OC) use and the risk of blood clots. Data from the study is summarized in the table below:

	Cases (blood clots)	Controls (no clots)
Didn't use OC	42	145
Used OC	42	23

- Calculate the conditional proportion for developing blood clots given OC use
  - Does this seem like it really is the probability of developing blood clots for an OC user?
  - Can we calculate the relative risk of developing blood clots? How about the risk difference?

Relative Risk and Study Design

- The conditional proportion of 65% is nowhere close to the actual probability of an OC user developing blood clots (the actual probability is less than 1%)
- It turns out we cannot use conditional proportions to estimate the risk of an outcome given an exposure using a retrospective study
    - This means differences in proportions and relative risks cannot be used to measure association in retrospective studies

Odds

- A slightly different measure of association, the **odds ratio**, does work for retrospective studies
- Instead of the ratio of two probabilities, the odds ratio is exactly what its name suggests, the ratio of two **odds**
- The odds of an event is the number of times that event occurs relative to the number of times that event doesn't occur
  - Suppose the probability of an event is 50%, the odds here are 1 (.5/.5), which people tend to express as "1 to 1 odds"
  - Suppose the probability of an event is 75%, the odds here are 3 (.75/.25), or "3 to 1 odds"

Odds Ratios

Using the ratio of two odds (odds ratio) to measure association comes with two major advantages:

1. The odds ratio is symmetric
2. The odds ratio can be used retrospective studies

Odds Ratios - Example #1

	Cases (blood clots)	Controls (no clots)
Didn't use OC	42	145
Used OC	42	23

1. Find the odds of blood clots for OC users
2. Find the odds of blood clots for those not using OC
3. Find the odds ratio for the risk of blood clots given OC use

Odds Ratios - Example #1 (solution)

	Cases (blood clots)	Controls (no clots)
Didn't use OC	42	145
Used OC	42	23

- The odds of blood clots were 1.83 (42/23) for OC users
- The odds of blood clots were 0.29 (45/145) for those not using OC
- Thus, the odds ratio for blood clots given OC use is 1.83/0.29 = 6.31

Odds Ratios - Example #1 (notes)

- We could also present the odds ratio for blood clots given no OC use: $0.29/1.83 = 0.16$
  - There are always two odds ratios for any 2x2 table, but notice that $1/0.16 = 6.3$
  - Thus we could say that using OC increases the odds of blood clots a factor of 6.31
  - Or we could say that not using OC decreases the odds of blood clots by 84%

The Cross-product Ratio

For a 2x2 table that looks like:

a	b
c	d

We can quickly calculate the odds ratio:

$$\widehat{OR} = \frac{a * d}{b * c}$$

Example #2 - Lister's Experiment

For Lister's sterile surgery experiment:

1. Estimate the conditional probability of death for each group using conditional proportions, use these probabilities to calculate the relative risk of death for the control group
2. Estimate the relative risk of survival for the sterile surgery group
3. Interpret each of these relative risks (from 1 and 2)
4. Estimate the odds ratio of death for the control vs. sterile surgery group
5. Estimate the odds ratio of survival for the sterile surgery vs. control group

	Died	Survived
Control	16	19
Sterile	6	34

Example #2 - Lister's Experiment (solution)

1. $Pr(\text{death}|\text{sterile}) = 6/40 = 0.15$, \$Pr(death | control) = 16/35 = 0.46 \$
$$\widehat{RR} = 0.46/.15 = 3.1$$
2. $Pr(\text{survive}|\text{sterile}) = 34/40 = 0.85$, \$Pr(survive | control) = 19/35 = 0.54 \$
$$\widehat{RR} = .85/.54 = 1.6$$
3. Patients in the control group are 3.1 times more likely to die than patients in the sterile surgery group. Patients in the sterile surgery group are 1.6 times more likely to survive than patients in the control group
4. $\widehat{OR} = \frac{16*34}{19*6} = 4.8$
5. $\widehat{OR} = \frac{34*16}{6*19} = 4.8$

Example #2 - Lister's Experiment (comments)

- Relative risk easy to interpret, but it not a symmetric metric, which opens the door for manipulative reporting
  - Patients in the control group are 3.1 times more likely to die than patients in the sterile surgery group
  - Patients in the sterile surgery group are 1.6 times more likely to survive than patients in the control group
  - Both are accurate statements, but can be spun differently
- Odds ratios are tougher to interpret, but they are symmetric

Summary

- Odds ratios are a popular measure of association for categorical data
  - Odds ratios are symmetric
  - Odds ratios can be used on data from retrospective studies
- Another reason for the popularity of odds ratios is their relationship with **Logistic Regression**
  - We won't cover logistic regression in this class, but it is an extremely popular method for modeling categorical outcomes
  - In fact, the Minitab help documentation indicates the way to find odds ratios with the software is to use logistic regression

Conclusion

These notes are supplemental to the Ch 7 of the textbook.

1. You should be familar with the importance of study design and how it influences the conclusions you can reach using a data set.
2. You should also be aware of different measures of association (risk differences, relative risks, and odds ratios) and how/why they are used.