Contingency Tables and Association

Ryan Miller

Probability

- *Frequentist statisticians* define probability as the long-run proportion of an outcome over a large number of trials
  - The probability of a coin flip showing "heads" is 1/2, because the proportion of heads *converges* to 1/2 as more and more coin flips are performed
- Obviously an infinite number of trials is impossible, but this definition suggests that we can *estimate* a probability using a proportion
  - For example, we might follow a random sample of 1,000 newly licensed drivers to estimate the probability of getting into a car accident in an individual's first year driving
  - If 88 of these drivers get into accidents, we'd estimate the probability of a new driver getting into an accident is 88/1000 or 8.8%

Conditional Probability

- Probabilities might differ depending upon other events, or the presence/absence of a characteristic
  - We call these *conditional probabilities*, and they can be estimated using *conditional proportions*
- Continuing the prior example, the probability of a new driver getting into an accident might depend upon the type of vehicle they drive:

	Accident	No Accident
Other Vehicle	54	636
Truck	34	276

- The conditional probability of an accident given an individual drives a truck is estimated by $\hat{p}_{\text{acc}|\text{truck}} = \frac{34}{310}$ or 11%

Design Considerations

In this example, most individuals didn't drive trucks, so which estimate do you think has *more uncertainty*?

1. The probability of an accident given an individual drives a truck
2. The probability of an accident given an individual drives another vehicle

Due to the lack of data, #1 has more uncertainty than #2, which might tempt us to collect *two separate samples* (one of truck drivers, one of other drivers)

- What downsides might collecting two samples have?
    - We'd lose the ability to estimate the probability that an individual drives a truck!
    - We will look more closely at the pros/cons of different study designs shortly, for now we will focus on measuring *risk* using proportions to estimate probability

Contingency Tables

Situations involving binary explanatory and response variables are very common, these data are often summarized using **contingency tables** (a specific type of frequency table):

	Event	No Event
Exposure	A	B
No Exposure	C	D

How have we described (summarized) the association shown in a table like this one?

Risk Differences

- A natural choice is the difference in proportions:

$$\hat{p}_{\text{event}|\text{exposed}} - \hat{p}_{\text{event}|\text{not exposed}} = \frac{A}{A+B} - \frac{C}{C+D}$$

- Because conditional proportions can be estimates of probability, this is called the **risk difference** (absolute risk)
  - $\hat{p}_{\text{event}|\text{exposed}}$ is the “risk” (probability) of the event among the exposed
  - $\hat{p}_{\text{event}|\text{not exposed}}$ is the “risk” (probability) of the event among the unexposed
- Risk differences are convenient because we’ve already learned how to construct confidence intervals or perform hypothesis tests with them!

Risk Differences

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

- With your group, find the *risk difference* (for the risk of breast cancer) and discuss whether it provides a good summary of these data

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

Risk Differences

- The risk difference in this study is $\frac{31}{31+1157} - \frac{65}{65+4475} = 0.012$ (1.2%)
  - It seems small, but breast cancer is relatively rare . . .
- Risk differences tend to be used less frequently than **relative risk**:

$$\text{Relative Risk} = \hat{p}_{\text{event}|\text{exposed}} \big/ \hat{p}_{\text{event}|\text{not exposed}} = \frac{A}{A+B} \big/ \frac{C}{C+D}$$

- The *relative risk of breast cancer* is 1.84 times higher (elevated by 84%) for women who gave birth before age 25
  - This seems to tell a different story than the 1.2% risk difference
- You should note that we could also report the relative risk of "no cancer"

Providing a Complete Picture

- A frustrating aspect of statistics is the prevailing opinion that statisticians can manipulate numbers to tell misleading stories
- There is some truth to this sentiment, so we should always strive to provide the most complete picture of the data that we possibly can
- This means reporting *both* the *risk difference* and the *relative risk* (and confidence intervals for each!), not simply whichever better fits the story you're trying to tell

Practice

The contingency table below describes the results of Joseph Lister's sterile surgery experiment:

	Died	Survived
Control	16	19
Sterile	6	34

With your group, I'd like you to:

1. Find the relative risk of death for the control group (relative to the sterile surgery group)
2. Find the relative risk of survival for the sterile surgery group (relative to the control group)
3. Interpret these relative risks, does anything seem troublesome?

Practice (solution)

1. The relative risk of death for the control group is found using: $Pr(\text{death}|\text{sterile}) = 6/40 = 0.15, Pr(\text{death}|\text{control}) = 16/35 = 0.46$
$$\widehat{RR} = 0.46/.15 = 3.1$$
2. The relative risk of surviving for the sterile surgery group is found using: $Pr(\text{survive}|\text{sterile}) = 34/40 = 0.85, Pr(\text{survive}|\text{control}) = 19/35 = 0.54$
$$\widehat{RR} = .85/.54 = 1.6$$
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. One might think that these risks should be the same?

Shortcomings of Relative Risk

- As demonstrated in Lister's experiment, relative risk is *asymmetric*
  - We will discuss a popular, symmetric measure of risk soon (the odds ratio)
- But first we'll explore another problem, which is that relative risk cannot be estimated for certain study designs
  - To understand this, we'll need to discuss three major observational study designs: prospective studies, cross-sectional studies, and retrospective studies

Prospective Studies

- The breast cancer example, which involved following a cohort of 6,168 women born in the 1980s, is an example of a **prospective study** (sometimes called a cohort study)
  - Prospective studies follow a representative sample *forward in time*, waiting for each subject to experience the exposure and experience the event of interest
  - Prospective studies are considered second only to randomized experiments when it comes to the strength of the evidence they provide

Retrospective Studies

- Tracking thousands of individuals for long periods of time is extremely resource intensive (in both time and money)
- An easier way to conduct a study on breast cancer risk factors might:
    - Recruit 100 women with breast cancer (cases)
    - Recruit 100 women without breast cancer (controls)
    - Ask each of these women about their past exposures, such as when they had their first child
- This, which looks backward in time, is called a **retrospective study** (sometimes called a case-control study)

Practice

- 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 probability that an individual in each population develops oral cancer? Can we estimate the relative risk of 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?

Practice (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.
- Nevertheless, 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 this type of design isn't well-equipped to handle

Alternatives to Relative Risk

- Relative risk *cannot* be used to measure association in a retrospective study, but a slightly different measure, the **odds ratio** can
  - The odds ratio is symmetric, so it overcomes one of our previously mentioned shortcomings of relative risk
- The odds ratio is just as it sounds: the odds of the event given the exposure divided by the odds of the event given a lack of the exposure
- 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"

The Odds Ratio for Lister's Experiment

	Died	Survived
Control	16	19
Sterile	6	34

- The odds of dying were $6/34 = 0.176$ for the sterile surgery group, and $16/19 = 0.842$ for the control group
  - Therefore, the odds ratio of death with the control surgery (relative to sterile surgery) are $0.842/0.176 = 4.77$
- The odds of surviving were $34/6 = 5.67$ for the sterile surgery group, and $19/16 = 1.19$ for the control group
  - Therefore, the odds of surviving with sterile surgery (relative to control surgery) are $5.67/1.19 = 4.77$

An Easier Formula for the Odds Ratio

Given:

	Event	No Event
Exposure	A	B
No Exposure	C	D

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

- In its early stages, the odds ratio was sometimes called the *cross-product ratio*

Interpreting the Odds Ratio

- Odds ratios are similar in spirit to relative risk, they provide a relative comparison of the *odds of an event* for different exposures
  - Relative risk provides a relative comparison of the *probability of an event* for different exposures
- The odds ratio and relative risk will always be similar in direction, but often slightly different in magnitude
  - In Lister's Experiment, the relative risk of death (treatment:control) was 3.1, while the odds ratio was 4.8
  - Odds ratios will always be further from 1, the value indicating no association, than relative risks

Hypothesis Testing

- In the Chi-Squared test for association, we typically state the null hypothesis as “no association”
- This is because “no association” implies *all* of the following:
  - The difference in proportions is 0
  - The relative risk is 1
  - Odds ratio is 1
- Because these are equivalent, we use the umbrella statement of “no association”

Practice

- In a previous lecture we discussed the controversy surrounding the heartburn medications *Prilosec* and *Nexium*, and the clinical trial comparing the two drugs
- In this trial, 2430 of the 2624 individuals who were assigned to receive Nexium had their erosive esophagitis heal, compared with 2324 of 2617 individuals in the group assigned to Prilosec.

With your group:

1. Report and interpret the *odds ratio* for the relative odds of healing (Nexium relative to Prilosec)
2. How does the odds ratio compare with the *risk difference*? That is, could reporting one without the other be misleading?

Practice (solution)

For this odds ratio we should set the two-way table as follows:

	Healed	Not Healed
Nexium	2430	194
Prilosec	2324	293

1. Based on the table above, we can calculate the odds ratio as $\widehat{OR} = \frac{2430*283}{2324*194} = 1.53$; so the odds of healing are 53% higher in the Nexium group relative to the Prilosec group
2. The risk difference is $2430/2624 - 2324/2617 = 0.038$, so there was a 3.8% difference in the proportion who were healed on Nexium compared to the proportion healed on Prilosec. In this situation the risk difference is small, but the odds ratio is fairly high.

Prospective vs. Retrospective Studies

Advantages of prospective studies:

- Only a single sample is collected (less room for sampling bias)
- Risk factors and events are directly observed (less potential for recall bias)
- Can be used to estimate probabilities, relative risk, and odds ratios
- More reflective of nature

Advantages of retrospective studies:

- Less expensive and less time consuming
- Easier to use when studying rare events
- No loss to follow-up concerns
- Odds ratios provide a valid measure of association

Cross-Sectional Studies

- The weakest type of observational design is the **cross-sectional study**
- In this design, researchers collect a *single sample* at *a single snapshot in time* and cross-classify individuals in that sample depending upon their exposure/event statuses
  - This differs from a retrospective study, which collects seperate samples of cases and controls (and pays careful attention to the separate challenges of sampling these populations)

Weaknesses of Cross-Sectional Studies

- Cross-sectional studies are the easiest to perform, but because they don't pay attention to time, they struggle to establish cause-effect relationships
- Selection bias is a major issue for cross sectional designs:
  - Consider a cross-sectional sample of factory workers
  - We might want to compare their rate of asthma to the prevalence of asthma in the general public in order to establish an association between factory work and asthma
  - Why might this be problematic?

Weaknesses of Cross-Sectional Studies (cont.)

- Factory workers who develop asthma will likely change jobs, so they will not appear in a cross-sectional sample
  - A cohort, or a case-control study, is less likely to encounter this problem
- It is also nearly impossible to make cause-effect claims from a cross-sectional study
  - If X and Y are measured at the same time, X could cause Y, or Y could cause X, or a confounding variable could cause both!

Practice

- A study surveyed 257 hospitalized individuals, classifying whether suffered from a circulatory disease, a respiratory disease, both, or neither. The results are displayed below:

	Respiratory Disease	No Respiratory Disease
Circulatory Disease	7	29
No Circulatory Disease	13	208

1. With your group, use an appropriate statistical test to determine whether the association between presence of a circulatory disease, and presence of a respiratory disease could be due to random chance
2. Does this mean that you're more likely to get a respiratory disease if you have a circulatory disease?

Practice (solution)

1. $\chi^2 = 7.9$, using $df = 1$, the p-value is 0.005, so it is very unlikely the association is due to random chance
2. No, individuals with both types of disease are more likely to be hospitalized (and biased towards ending up in this sample). The researchers in this study also looked at a sample of non-hospitalized individuals:

	Respiratory Disease	No Respiratory Disease
Circulatory Disease	15	142
No Circulatory Disease	189	2181

- Performing a χ^2 test on this sample, the p-value is 0.48

Conclusion

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

1. You should be familiar 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), how/why they are used, and know how to interpret them