Statistical Inference for Two-sample Categorical Data

Ryan Miller

Outline

- Video #1
  - Two-sample Categorical Data (example and concepts)
- Video #2
  - The Two-sample Z-test for a Difference in Proportions
- Video #3
  - Odds Ratios and Confidence Intervals

Introduction

- So far in this course all our applications of statistical inference have focused on *one-sample* (one group) settings
  - We've used Normal models (the Z-test) for inference involving one proportion
  - We've used the t-distribution (the T-test) for inference involving one mean
- This week we will learn how to extend these ideas to *two-sample* settings, or situations involving the comparison of two proportions or two means

Surgical Site Infections

- In the 1860's, surgeries often led to infections that resulted in death
- At the time, experts believed these infections were due to "bad air"
  - Hospitals had policies that required their wards open their windows at midday to air out
- It was customary for surgeons to move quickly from patient to patient without any sort of special precautions
  - In fact, many took pride the accumulated stains on their surgical gowns as a measure of experience

Louis Pasteur and Joseph Lister

- In 1862, Louis Pasteur discovered that food spoilage was caused by the growth and proliferation of harmful micro-organisms
- Pasteur identified three methods for eliminating these micro-organisms: heat, filtration, and chemical disinfectants
    - The method of heating became known as pasteurization (named for Pasteur) and is widely applied to milk, beer, and many other food products

Louis Pasteur and Joseph Lister

- In 1862, Louis Pasteur discovered that food spoilage was caused by the growth and proliferation of harmful micro-organisms
- Pasteur identified three methods for eliminating these micro-organisms: heat, filtration, and chemical disinfectants
  - The method of heating became known as pasteurization (named for Pasteur) and is widely applied to milk, beer, and many other food products
- Joseph Lister, a Professor of Surgery at the Glasgow Royal Infirmary, became aware of Pasteur's work and theorized that it might explain the infections that frequently occurred after surgery
  - How would you recommend Lister evaluate his theory?

Lister's Experiment

- Lister proposed a new "sterile" protocol where surgeons were required to wash their hands, wear clean gloves, and disinfect their instruments with a carbolic acid solution
  - He *randomly assigned* 75 patients undergoing surgery to receive either his new "sterile" protocol or be in a control group
  - He then tracked the survival of patients until their discharge from the hospital

	Died	Survived
Control	16	19
Sterile	6	34

Analyzing Lister's Experiment

When evaluating Lister's experiment, we need to rule out possible explanations for the observed differences in survival rates

1. Bias?

Analyzing Lister's Experiment

When evaluating Lister's experiment, we need to rule out possible explanations for the observed differences in survival rates

1. Bias? Probably not, even though double-blinding wasn't possible, it's unlikely the measurement of the outcome (survival) was biased. It's also unlikely that this is a non-representative group of patients (sampling bias)
2. Confounding variables?

Analyzing Lister's Experiment

When evaluating Lister's experiment, we need to rule out possible explanations for the observed differences in survival rates

1. Bias? Probably not, even though double-blinding wasn't possible, it's unlikely the measurement of the outcome (survival) was biased. It's also unlikely that this is a non-representative group of patients (sampling bias)
2. Confounding variables? No, we'd expect random assignment to have balanced the two groups
3. Random chance? ... This is where hypothesis testing is useful

Analyzing Lister's Experiment

- The first step in any hypothesis test is to *determine a null model*
  - In words, what should the null model be for Lister's experiment?

Analyzing Lister's Experiment

- The first step in any hypothesis test is to *determine a null model*
  - In words, what should the null model be for Lister's experiment?
- The null model is that the Lister's proposed sterilization procedure makes no difference
  - That is, equal proportions of the "Sterile" and "Control" groups are expected to die prior to discharge

$$H_0 : p_1 - p_2 = 0$$

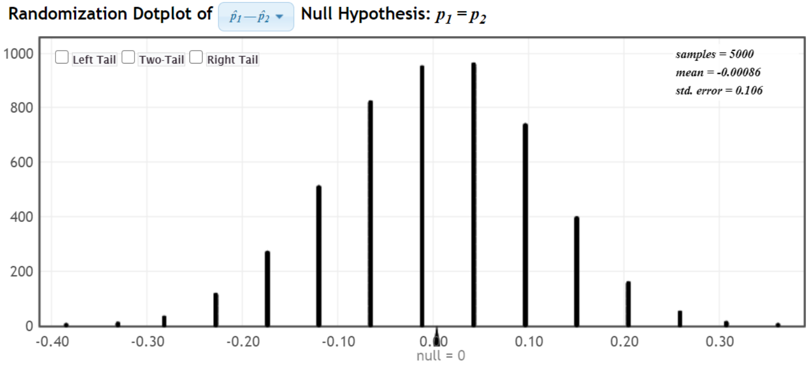
- Here, p_1 denotes the proportion of deaths among the "Control" group, and p_2 is the proportion of deaths among the "Sterile" group

Simulating the Null Distribution

- If the sterilization protocol made no difference, any deaths observed in this study data occurred at random (ie: the assigned group made no difference)
  - Thus, under the null model, we can assume the *overall death rate* (estimated by 22/75, or 29%) applies equally to both groups
- We can use StatKey to simulate possible outcomes that could occur under this null model using sets of $n_1 = 35$ and $n_2 = 40$ "weighted coin-flips", where each flip represents a 29% chance of death

Simulating the Null Distribution

If both groups had the same death rate (29%), we could expect to have observed the following differences in proportions:



The study saw a difference of $\hat{p}_1 - \hat{p}_2 = 16/35 - 6/40 = 0.31$, only 2 of 5000 simulated outcomes were this extreme!

The Two-sample Z-test (categorical data)

- Recognizing the Normality of the null distribution, we can generalize this approach to a *two-sample* Z*-test*

The Two-sample Z-test (categorical data)

- Recognizing the Normality of the null distribution, we can generalize this approach to a *two-sample Z-test*
  - Central Limit theorem (combined with some probability theory and linear algebra) suggests the following:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\tfrac{p_1(1-p_1)}{n_1} + \tfrac{p_2(1-p_2)}{n_2}}\right)$$

The Two-sample Z-test (categorical data)

- Recognizing the Normality of the null distribution, we can generalize this approach to a *two-sample Z-test*
  - Central Limit theorem (combined with some probability theory and linear algebra) suggests the following:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\tfrac{p_1(1-p_1)}{n_1} + \tfrac{p_2(1-p_2)}{n_2}}\right)$$

- One challenge in applying this theoretical result is that our null hypothesis only specifies that $p_1 = p_2$ (which can be satisfied by many different values)
  - The most common solution is to used the *pooled* (overall) proportion in place of both p_1 and p_2
  - In our example, this would be applying the overall death rate of 29% to both groups

The Two-sample Z-test (categorical data)

1. State the null hypothesis (ie: $H_0 : p_1 = p_2$ for two-sample categorical data)
2. Calculate a Z-value using the sample data and an appropriate Normal model (ie: $Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$)
3. Compare the Z-value to the Standard Normal curve to find the p-value
4. Use the p-value to make a conclusion (remember to consider context!)

The Two-sample Z-test for Lister's Experiment

1. $H_0 : p_1 = p_2$, where p_1 is the proportion who died in the control group and p_2 is the proportion who died in the sterile group

The Two-sample Z-test for Lister's Experiment

1. $H_0 : p_1 = p_2$, where p_1 is the proportion who died in the control group and p_2 is the proportion who died in the sterile group
2. We observed $\hat{p}_1 - \hat{p}_2 = 16/35 - 6/40 = 0.31$, the pooled death rate is $\hat{p} = 22/75 = 0.29$, thus:
   $Z = \frac{0.31-0}{\sqrt{0.29*(1-0.29)/35+0.29*(1-0.29)/40}} = 2.94$

The Two-sample Z-test for Lister's Experiment

1. $H_0 : p_1 = p_2$, where p_1 is the proportion who died in the control group and p_2 is the proportion who died in the sterile group
2. We observed $\hat{p}_1 - \hat{p}_2 = 16/35 - 6/40 = 0.31$, the pooled death rate is $\hat{p} = 22/75 = 0.29$, thus:
   $Z = \frac{0.31-0}{\sqrt{0.29*(1-0.29)/35+0.29*(1-0.29)/40}} = 2.94$
3. Using StatKey, a Z-value of 2.94 corresponds to a two-sided p-value of 0.0032
4. We conclude there is overwhelming statistical evidence that the new sterilization procedure leads to lower death rates

Effect Size

- As we've previously discussed, hypothesis testing answers the question "could the observed difference be explained by random chance?"
  - This is a fundamentally different question from "is the observed difference large enough to change how we should act?"

Effect Size

- As we've previously discussed, hypothesis testing answers the question "could the observed difference be explained by random chance?"
  - This is a fundamentally different question from "is the observed difference large enough to change how we should act?"
- Confidence intervals are an important tool for answering the later question
  - The Normal model we've already presented makes for the easy construction of these interval estimates:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Confidence Intervals for a Difference in Proportions

- One subtle difference when comparing the Normal model for confidence intervals with the one we used for hypothesis testing is the standard error
  - For hypothesis testing, the standard error used a *pooled proportion* to be consistent with the null hypothesis (which says $p_1 = p_2$)
  - For confidence interval estimation, the standard error should reflect reality, thus it can use our separate point estimates of $\hat{p}_1$ and $\hat{p}_2$

Confidence Intervals for a Difference in Proportions

- One subtle difference when comparing the Normal model for confidence intervals with the one we used for hypothesis testing is the standard error
    - For hypothesis testing, the standard error used a *pooled proportion* to be consistent with the null hypothesis (which says $p_1 = p_2$)
    - For confidence interval estimation, the standard error should reflect reality, thus it can use our separate point estimates of $\hat{p}_1$ and $\hat{p}_2$
- Thus, for Lister's experiment we can 95% confident that between 11.1% and 50.9% more of a sterile surgery group will survive

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$0.46 - 0.15 \pm 1.96\sqrt{\frac{0.46(1-0.46)}{35} + \frac{0.15(1-0.15}{40}} = (0.111, 0.509)$$

A Second Example

- Public health researchers often look at the proportion of a population who will develop a disease within a fixed time frame
  - Smoking is a well-established risk factor for lung cancer, but how do lung cancer rates compare among the populations of smokers and non-smokers?

A Second Example

- Public health researchers often look at the proportion of a population who will develop a disease within a fixed time frame
  - Smoking is a well-established risk factor for lung cancer, but how do lung cancer rates compare among the populations of smokers and non-smokers?
- Among all smokers, 0.44% are expected to develop lung cancer in a 10-year period
  - Among all non-smokers, only 0.05% are expected to develop lung cancer in a 10-year period
- The difference in proportions is only 0.0039 (0.39%), so is it really worthwhile to get people to quit smoking?

A Second Example

- Public health researchers often look at the proportion of a population who will develop a disease within a fixed time frame
  - Smoking is a well-established risk factor for lung cancer, but how do lung cancer rates compare among the populations of smokers and non-smokers?
- Among all smokers, 0.44% are expected to develop lung cancer in a 10-year period
  - Among all non-smokers, only 0.05% are expected to develop lung cancer in a 10-year period
- The difference in proportions is only 0.0039 (0.39%), so is it really worthwhile to get people to quit smoking?
  - I'd argue "yes", as the cancer risk is nearly *10 times higher*!

Odds Ratios

- The most commonly reported measure of association describing the relationship between two categorical variables is the **odds ratio**
  - The *odds* of an event is the ratio of how often it happens to how often it doesn't happen
  - If a team has a 75% probability of winning a game, the odds of winning are 3, which is often spoken as "3 to 1"

Odds Ratios

- The most commonly reported measure of association describing the relationship between two categorical variables is the **odds ratio**
  - The *odds* of an event is the ratio of how often it happens to how often it doesn't happen
  - If a team has a 75% probability of winning a game, the odds of winning are 3, which is often spoken as "3 to 1"
- In our smoking example, the odds of a smoker developing lung cancer are $\frac{0.00438}{1-0.00438} = 0.00440$
  - Similarly, the odds of a non-smoker developing lung cancer are $\frac{0.00045}{1-0.00045} = 0.00045$

Odds Ratios

- The most commonly reported measure of association describing the relationship between two categorical variables is the **odds ratio**
  - The *odds* of an event is the ratio of how often it happens to how often it doesn't happen
  - If a team has a 75% probability of winning a game, the odds of winning are 3, which is often spoken as "3 to 1"
- In our smoking example, the odds of a smoker developing lung cancer are $\frac{0.00438}{1-0.00438} = 0.00440$
  - Similarly, the odds of a non-smoker developing lung cancer are $\frac{0.00045}{1-0.00045} = 0.00045$
- Thus, the *odds ratio* is $\frac{0.00440}{0.00045} = 9.8$
  - We say that the odds of a smoker developing lung cancer are 9.8 times those of a non-smoker developing lung cancer

Confidence Intervals for Odds Ratios

- We will not cover how to calculate confidence intervals for an odds ratio, but you should recognize that all standard statistical software can do this
  - Instead, I only expect that you're able to properly interpret a CI estimate of an odds ratio

Confidence Intervals for Odds Ratios

- We will not cover how to calculate confidence intervals for an odds ratio, but you should recognize that all standard statistical software can do this
  - Instead, I only expect that you're able to properly interpret a CI estimate of an odds ratio
- For example, the 95% CI for the odds ratio in Lister's experiment is (1.4, 17.2)
  - This means the odds of dying in the control group are estimated (with 95% confidence) to be between 1.4 times and 17.2 times higher than the odds of dying in the sterile surgery group (quite an improvement!)

Conclusion

- This presentation introduced statistical methods for two-sample categorical data
  - Like one-sample categorical data, these methods are built upon a Normal model:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Conclusion

- This presentation introduced statistical methods for two-sample categorical data
  - Like one-sample categorical data, these methods are built upon a Normal model:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\tfrac{p_1(1-p_1)}{n_1} + \tfrac{p_2(1-p_2)}{n_2}}\right)$$

- When using this model for hypothesis testing, we use a *pooled proportion* to calculate the standard error
- When using this model for confidence interval estimation, we use the *sample proportions* to calculate the standard error