# Hypothesis Testing - Categorical Data (part 2)

Ryan Miller

# Introduction

- Previously, we discussed a variety of different hypothesis testing approaches for scenarios involving *one categorical variable*
  - Generally speaking, these approaches fell into one of three camps: simulation, exact probability calculations, Normal/Chi-square distributions

# Introduction

- Previously, we discussed a variety of different hypothesis testing approaches for scenarios involving *one categorical variable*
  - Generally speaking, these approaches fell into one of three camps: simulation, exact probability calculations, Normal/Chi-square distributions
- This presentation will focus on scenarios involving *two categorical variables*
  - A common example is a categorical response variable and a categorical explanatory variable that defines treatment and control groups

# Surgical Site Infections

- In the 1860's, surgeries often led to infections that resulted in death
- At the time, many experts believed these infections were due to "bad air"
  - Hospitals had policies that required their wards open their windows at midday to air out

# Surgical Site Infections

- In the 1860's, surgeries often led to infections that resulted in death
- At the time, many 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 with out 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 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" procedure or the old standard of care
  - The outcome was how many of each group survived until their discharge from the hospital

| | Died | Survived |
|---|---|---|
| Control | 16 | 19 |
| Sterile | 6 | 34 |

# Analyzing Lister's Experiment

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

1) Bias?

# Analyzing Lister's Experiment

In analyzing 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

In analyzing 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 everything to be balanced in the two groups due to random assignment
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 then simulate possible outcomes under this null model by using sets of $n_1 = 35$ and $n_2 = 40$ "weighted coin-flips" each having a 29% chance of death

# Simulating the Null Distribution

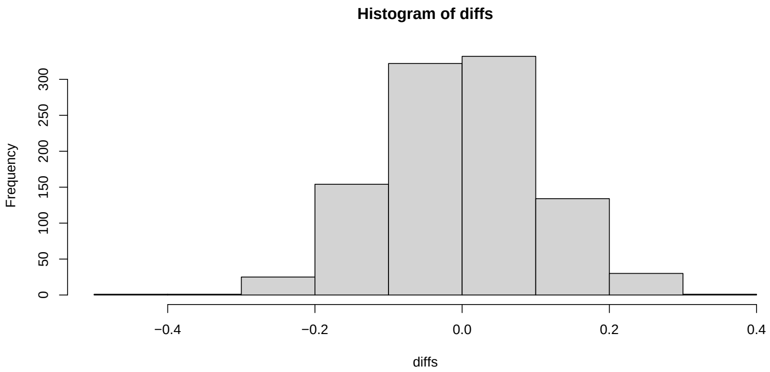
```
## Set seed (for replication purposes)
set.seed(123)

## Simulate deaths for the control group
control_deaths <- rbinom(1000, 35, .29)


## Simulate deaths for the sterile group
sterile_deaths <- rbinom(1000, 40, .29)
```

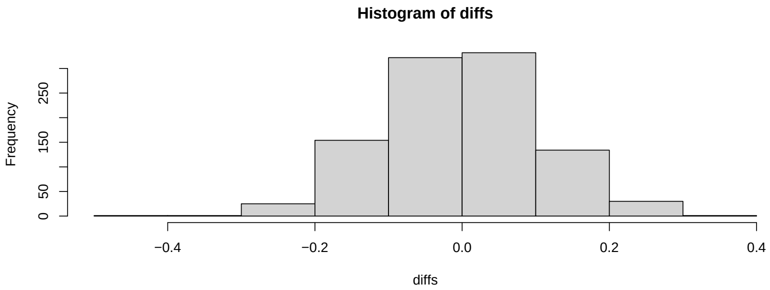
# Null distribution

```r
## Graph the possible differences in proportions
diffs <- control_deaths/35 - sterile_deaths/40
hist(diffs)
```



# Finding the $p$-value

- Recall that we observed $\hat{p}_1 = 16/35$ and $\hat{p}_2 = 6/40$, resulting in a sample difference in proportions of 0.307




- What would you estimate the two-sided $p$-value to be?

# Finding the *p*-value

```r
## Find the p-value
upper <- sum(diffs >= 0.307)/1000
2*upper
```

```
## [1] 0
```

Exactly zero of these 1000 simulations had outcomes as extreme as the actual experimental results. Thus, this experiment provides *overwhelming evidence* that Lister's sterilization protocol improves survival

# $Z$-Test for Differences in Proportions

Previously, we learned Central Limit Theorem (combined with some probability theory for independent random variables) leads to the following distributional result:

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

# $Z$-Test for Differences in Proportions

Previously, we learned Central Limit Theorem (combined with some probability theory for independent random variables) leads to the following distributional result:

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

To use this result, we must substitute for $p_1$ and $p_2$, any ideas?

# $Z$-Test for Differences in Proportions

Previously, we learned Central Limit Theorem (combined with some probability theory for independent random variables) leads to the following distributional result:

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

To use this result, we must substitute for $p_1$ and $p_2$, any ideas?

- $\hat{p}_1 = 6/40 = 0.15$ and $\hat{p}_2 = 16/35 = 0.46$ won't work, they don't satisfy the null hypothesis
- Instead, we must use the *pooled proportion*, or $22/75 = 0.29$, in place of *both* $p_1$ and $p_2$

# $Z$-Test for Differences in Proportions

```r
## Calculate the Z-value
pool <- 22/75
se <- sqrt(pool*(1-pool)/35 + pool*(1-pool)/40)
z <- ((0.46 - .15) - 0)/se

## Find the p-value
upper <- pnorm(z, lower.tail = FALSE)
2*upper
```

```
## [1] 0.00326354
```

# Chi-Squared Test for Association

A third way to approach involves comparing observed vs. expected frequencies, recall we observed:

| | Died | Survived |
|---|---|---|
| Control | 16 | 19 |
| Sterile | 6 | 34 |

If the sterilization procedure made no difference, we'd expect 29% of each group to have died, leading to the following table of *expected counts*:

| | Died | Survived |
|---|---|---|
| Control | $35*.29 = 10.2$ | $35*.71 = 24.9$ |
| Sterile | $40*.29 = 11.6$ | $40*.71 = 28.4$ |

# Chi-Squared Test for Association

Comparing these observed and expected frequencies, we can set up a Chi-squared test:

$$X^2 = \frac{(16-10.2)^2}{10.2} + \frac{(19-24.9)^2}{24.9} + \frac{(6-11.6)^2}{11.6} + \frac{(34-28.4)^2}{28.4} = 8.5$$

# Chi-Squared Test for Association

Comparing these observed and expected frequencies, we can set up a Chi-squared test:

$$X^2 = \frac{(16-10.2)^2}{10.2} + \frac{(19-24.9)^2}{24.9} + \frac{(6-11.6)^2}{11.6} + \frac{(34-28.4)^2}{28.4} = 8.5$$

In a 2*x*2 table with fixed margins, if you know one of the four cells you can calculate the rest, meaning $df = 1$:

```
## p-value from Chi-squared df =1
pchisq(8.5, df = 1, lower.tail = FALSE)
```

```
## [1] 0.003551465
```

```
## Using chisq.test
tab <- data.frame(Died = c(16,6), Survived = c(19,34))
chisq.test(tab, correct = FALSE)$p.value
```

```
## [1] 0.003560924
```

# Comments on Chi-Squared Tests

- Chi-squared testing in this context is known as *testing for association* or *testing for independence*

# Comments on Chi-Squared Tests

- Chi-squared testing in this context is known as *testing for association* or *testing for independence*
- Chi-squared tests can be performed on sample data in any two-way frequency table, not just 2$x$2 tables
  - For these tests, $df = (I - 1) * (J - 1)$, where $I$ is the number of rows and $J$ is the number of columns in the table
  - Because writing the null hypothesis in statistical notation is difficult for large tables, we'll typically just say
    $H_0$ : Independence

# Comments on Chi-Squared Tests

- Chi-squared testing in this context is known as *testing for association* or *testing for independence*
- Chi-squared tests can be performed on sample data in any two-way frequency table, not just 2$x$2 tables
  - For these tests, $df = (I-1)*(J-1)$, where $I$ is the number of rows and $J$ is the number of columns in the table
  - Because writing the null hypothesis in statistical notation is difficult for large tables, we'll typically just say $H_0$ : Independence
- We calculated expected counts using a *pooled proportion* that collapsed the table's rows
  - We could have done something similar using the table's columns and arrived at an identical result
  - Alternatively, many places teach the formula $E_{ij} = \frac{\text{Observed}_{ij}}{\text{Row Total}_i * \text{Column Total}_j}$

# Comments on Chi-Squared Tests (continued)

- Finally, the Chi-squared distribution is related to the standard normal distribution, meaning the $\chi^2$-test is related to the $Z$-test
  - Thus, the $\chi^2$-test is a large-sample approach that is only accurate when every cell has an expected frequency of at least 5
- **Fisher's Exact Test** is an exact test for independence in a two-way table that can be used in small-sample situations

# Fisher's Exact Test

We won't get into the details of **Fisher's Exact test**, but the gist is that if you assume the row and column totals of the table are fixed, the remaining numbers that are free to vary will follow a hypergeometric distribution under the null hypothesis of independence

```
## Using fisher.test
tab <- data.frame(Died = c(16,6), Survived = c(19,34))
fisher.test(tab)$p.value
```

```
## [1] 0.005018047
```

# Comparing the Different Approaches

- We've now evaluated the results of Lister's experiment in four different ways
  - Simulation - $p$-value estimated at $0/1000$
  - $Z$-test - $p$-value of 0.0033
  - $\chi^2$-test - $p$-value of 0.0036
  - Fisher's Exact test - $p$-value of 0.0050

# Comparing the Different Approaches

- We've now evaluated the results of Lister's experiment in four different ways
  - Simulation - $p$-value estimated at $0/1000$
  - $Z$-test - $p$-value of 0.0033
  - $\chi^2$-test - $p$-value of 0.0036
  - Fisher's Exact test - $p$-value of 0.0050
- For large samples, it makes little difference which of these approaches you choose
  - The $Z$-test is only valid when $n_1\hat{p}_{pool} \geq 10$, $n_2\hat{p}_{pool} \geq 10$, $n_1(1-\hat{p}_{pool}) \geq 10$ and $n_2(1-\hat{p}_{pool}) \geq 10$
  - The $\chi^2$-test requires all expected counts are larger than 5
  - Fisher's Exact test and simulation do not have a sample size condition

# Statistical vs. Clinical Significance

- The $\chi^2$ test for independence and Fisher's exact test can both be used to evaluate the strength of an association that exists between two categorical variables
  - The lower the *p*-value, the more strongly the variables are associated (That is, the more incompatible the sample data are with the variables being independent)

# Statistical vs. Clinical Significance

- The $\chi^2$ test for independence and Fisher's exact test can both be used to evaluate the strength of an association that exists between two categorical variables
  - The lower the $p$-value, the more strongly the variables are associated (That is, the more incompatible the sample data are with the variables being independent)
- These methods do not tell us anything about the nature of the association
  - We could report the sample difference in proportions (accompanied by a confidence interval), but this summary measure has a major shortcoming

# Statistical vs. Clinical Significance

- The $\chi^2$ test for independence and Fisher's exact test can both be used to evaluate the strength of an association that exists between two categorical variables
  - The lower the *p*-value, the more strongly the variables are associated (That is, the more incompatible the sample data are with the variables being independent)
- These methods do not tell us anything about the nature of the association
  - We could report the sample difference in proportions (accompanied by a confidence interval), but this summary measure has a major shortcoming
- Consider the proportions of smokers and non-smokers that develop lung cancer in a 10-year period
  - These proportions are estimated at 0.00438 and 0.00045 respectively, or a difference of 0.0039 (far less than 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"

# 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 Interval for an Odds Ratio in R

```r
## 95% CI for an OR (Lister's Experiment)
tab <- data.frame(Died = c(16,6), Survived = c(19,34))
fisher.test(tab, conf.int = TRUE,
            conf.level = .95)$conf.int[1:2]
```

```
## [1]  1.437621 17.166416
```

Thus, we can conclude with 95% the odds of death in the Control group are between 1.4 and 17.2 times higher than the odds of death in the Sterile group

# Summary

- This presentation covered several hypothesis testing approaches for evaluating relationships between two categorical variables
  - Simulation
  - $Z$-test for a difference in proportions
  - $\chi^2$-test for association (independence)
  - Fisher's exact test

# Summary

- This presentation covered several hypothesis testing approaches for evaluating relationships between two categorical variables
  - Simulation
  - $Z$-test for a difference in proportions
  - $\chi^2$-test for association (independence)
  - Fisher's exact test
- In general, Fisher's exact test is the most robust of these methods, but it is very computationally intensive for large tables
  - Because of this, $\chi^2$ tests tend to be most widely used in practice
- We also introduced the **odds ratio** as a measure of association that can be used to gauge *clinical* vs. *statistical* significance