Classical Approaches to Statistical Inference (part II)

Ryan Miller

Two-sample data

- In many situations analyzing a variable for a single group is useful, but a more common occurrence is to collect data for two (or more) groups
- We've seen this type of data many times already, for example in the Iowa City Home Sales Data:

AC	sale.amount
Yes	172500
Yes	90000
Yes	168500
Yes	205000
Yes	...

- The binary variable "AC" tells us the two groups (the two samples) and "sale.amount" represents our variable of interest

Two-sample data

- With two-sample data, we are most interested in how the samples (groups) differ
  - For example, how does the mean sale price differ for homes with and without air conditioning?
- If the variable of interest is quantitative, we can assess the difference across groups by looking for a difference in means:

$$\mu_{\text{group 1}} - \mu_{\text{group 2}}$$

- For the variable of interest is categorical, we can assess the difference via the difference in proportions:

$$p_{\text{group 1}} - p_{\text{group 2}}$$

Case Study - Joseph Lister's Experiment

- In the 1860s, it was not customary for surgeons to wash their hands or surgical instruments prior to operating on patients
  - At the time most people believed infections were due to exposure to bad air, and hospitals were frequently aired out at midday
  - Some surgeons even took pride in the accumulated stains on their operating gowns as a display of experience
- A paper published by the French chemist, Louis Pasteur, showed that food spoilage occurred due to the proliferation of harmful micro-organisms under certain conditions
  - Pasteur suggested three methods for eliminating these micro-organisms: heat, filtration, and chemical solutions
  - Lister became aware of this paper and theorized that similar micro-organisms were responsible for the infections frequently occurred after surgery

Case Study - Joseph Lister's Experiment

- Lister proposed a new protocol in which surgeons were required to wash their hands, wear clean gloves, and disinfect their instruments with a carbolic acid solution
  - He performed an experiment, randomly assigning 75 patients to receive either his new "sterile" procedure or the old standard of care and recording whether each patient survived until their discharge from the hospital
  - Here are 6 rows of Lister's data:

	Group	Survival
19	Sterile	Survived
39	Sterile	Died
8	Sterile	Survived
65	Control	Died
60	Control	Died
30	Sterile	Survived

Case Study - Joseph Lister's Experiment

- This experiment is an example of *two-sample categorical data*
  - The variable "Group" defines the two samples (groups)
  - The variable of interest "Survived" is categorical
- We've seen that this type of data can be summarized using two-way frequency tables:

	Died	Survived
Control	16	19
Sterile	6	34

- By convention, statisticians like to use the grouping variable to define the rows and the variable of interest to define the columns

Case Study - Joseph Lister's Experiment (Review)

- One way of analyzing these data, which we've already learned, is to construct separate confidence intervals for the proportion who survived within group
- With your group, construct separate 99% confidence intervals for the proportion of sterile and non-sterile surgery patients who died

	Died	Survived
Control	16	19
Sterile	6	34

Hint: $SE(p) = \sqrt{\frac{p(1-p)}{n}}$

Case Study - Joseph Lister's Experiment (Review)

For the sterile surgery group: $\hat{p} = 6/40 = 0.15$, leading to the 99% CI:

$$0.15 \pm 2.576 * \sqrt{\frac{0.15(1-0.15)}{40}} = (.005, .295)$$

For the non-sterile (control) group: $\hat{p} = 16/35 = 0.46$, leading to the 99% CI:

$$0.46 \pm 2.576 * \sqrt{\frac{0.46(1-0.46)}{35}} = (.243, .677)$$

Notice that these confidence intervals overlap, but this doesn't mean the two groups aren't different...

Differences in Proportions

- We know that values near the boundaries of a confidence interval are considerably less plausible than those near the center
- When two confidence intervals barely overlap it doesn't necessarily indicate that a difference of zero is plausible
- A better approach would be to look at the difference in proportions using a single interval

The Distribution of a Difference in Proportions

- The standard error of a difference in proportions, $\hat{p}_1 - \hat{p}_2$, is given by:

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- Notice how this formula relates to that of a single proportion
  - The difference has more variability than either proportion does by itself
  - Variability accumulates across the samples, but it is always less than comparing the samples seperately
- To use this standard error with a normal approximation, we need to check:

$$n_1 p_1 \geq 10, n_1(1-p_1) \geq 10 \text{ and } n_2 p_2 \geq 10, n_2(1-p_2) \geq 10$$

Two-sample Categorical Data - Example #1

1. Construct a 99% confidence interval for the difference in the proportions of patients that died in Lister's experiment

	Died	Survived
Control	16	19
Sterile	6	34

2. Based upon your confidence interval, what do you conclude about the effectiveness of the sterile surgery procedure?

Two-sample Categorical Data - Example #1 (solution)

1. $\hat{p}_{\text{sterile}} - \hat{p}_{\text{non-sterile}} = 0.15 - 0.46 = -0.31$

$$SE = \sqrt{\frac{.15(1-.15)}{40} + \frac{.46(1-.46)}{35}} = 0.101$$

99% CI: $-0.31 \pm 2.576 * 0.101 = (-0.57, -0.05)$

2. We conclude that the sterile procedure improves patient survival, “no difference” is not a plausible value according to our 99% confidence interval

Two-sample Categorical Data - Hypothesis Testing

- When constructing a confidence interval we use the most likely values as suggested by our sample ($\hat{p}_{\text{sterile}}$ and $\hat{p}_{\text{non-sterile}}$)
- When hypothesis testing we operate in the hypothetical world of the null hypothesis, so want to use the values that are most likely in this world
  - In other words, we want to use the most likely proportions that satisfy the hypothesis $H_0 : p_1 - p_2 = 0$
  - What makes these values difficult to determine?

Two-sample Categorical Data - Hypothesis Testing

- There are many different ways for $p_1 - p_2$ to equal zero, but not all of them are equally plausible given our data
  - If we really believe the null, it makes sense to combine both groups into one big sample and find a single proportion
  - This is called the **pooled proportion**, which our textbook denotes as $\hat{p}$, but I will denote it $\hat{p}_{1+2}$
- For Lister's experiment, $\hat{p}_{1+2} = (6 + 15)/(40 + 35) = 0.28$
  - The pooled proportion is our best guess for each group in the world of the null hypothesis, so we use it to calculate the standard error when hypothesis testing

Two-sample Categorical Data - Example #2

With your group, conduct a hypothesis test (using $\alpha = 0.05$) to investigate whether there is no difference in the proportion that died for the two groups in Lister's experiment. Some relevant information is provided below:

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

	Died	Survived
Control	16	19
Sterile	6	34

Two-sample Categorical Data - Example #2 (solution)

$$\hat{p}_{\text{sterile}} - \hat{p}_{\text{non-sterile}} = 0.15 - 0.46 = -0.31$$

$$\hat{p}_{1+2} = (6 + 15)/(40 + 35) = 0.28$$

$$SE = \sqrt{\frac{0.28(1-0.28)}{49} + \frac{0.28(1-0.28)}{35}} = 0.104$$

$$z_{\text{test}} = \frac{-0.31 - 0}{0.104} = -2.98$$

With a two-sided *p*-value of 0.0028, there is strong evidence that Lister's sterilization protocol leads to fewer patient deaths following surgery.

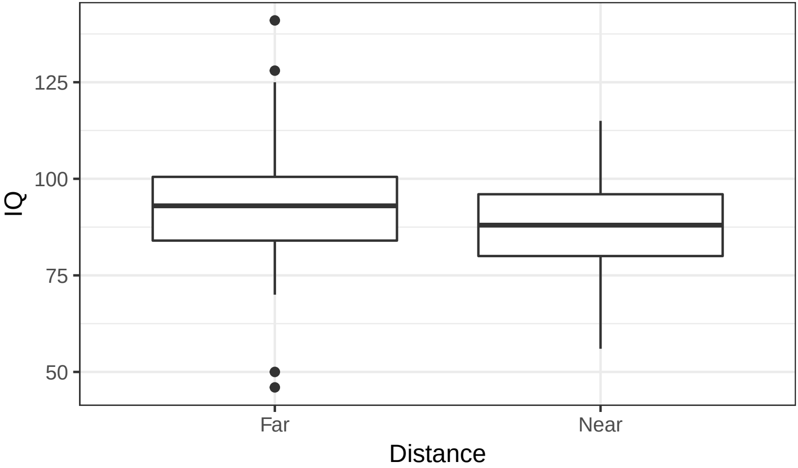
Two-sample Quantitative Data

- In Lister's experiment the outcome variable was categorical, we'll now shift our focus to quantitative outcome variables
- Our case study will be a study investigating relationship between lead exposure and neurological development
    - Researchers in El Paso, TX measured the IQ scores (age-adjusted) of 57 children who lived within 1 mile of a lead smelter and 67 children who lived at least 1 mile away
    - Here is a portion of Lead Exposure data:

	Distance	IQ
11	Far	118
51	Far	86
63	Far	82
49	Far	107
27	Far	87
35	Far	89

Case Study - Lead Exposure and IQ

1. Do the data for each group appear to be normally distributed?
2. Does there appear to be a relationship between distance from the smelter and IQ?



Case Study - Lead Exposure and IQ

- Like Lister's experiment, we could analyze these data using separate confidence intervals for each group:

$$\bar{x}_{\text{near}} \pm t^*_{df=56} * \frac{s_{\text{near}}}{\sqrt{n_{\text{near}}}} = (86.0, 92.4)$$

$$\bar{x}_{\text{far}} \pm t^*_{df=66} * \frac{s_{\text{far}}}{\sqrt{n_{\text{far}}}} = (88.8, 96.6)$$

- There is a lot of overlap between these intervals, but we've learned that looking at each group separately isn't a very powerful approach
- It is better to look at the difference: $\bar{x}_{\text{near}} - \bar{x}_{\text{far}}$

Case Study - Lead Exposure and IQ

The standard error of a difference in means is given by:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Similar to one-sample quantitative data, we need to use the t-distribution with this standard error
- Finding the correct degrees of freedom turns out to be quite messy
  - Programs like Minitab will figure it out for us
  - When working by hand, we will use the smaller of $n_1 - 1$ and $n_2 - 1$
  - This provides a conservative approach (smaller df leading to slightly wider confidence intervals and slightly larger p-values)

Two-sample Quantitative Data - Example #1

1. Load the "LeadIQ" dataset into Minitab (it's posted on the course website)
2. Using Minitab to calculate the necessary summary statistics, conduct a two-sample t-test (by hand) to determine whether the mean IQ differs for children living near or far from a lead smelter
3. Perform the same test using Minitab (Stat -> Basic Statistics -> Two-sample t-test)

Two-sample Quantitative Data - Example #1 (solution)

$$H_0 : \mu_{\text{near}} - \mu_{\text{far}} = 0, H_A : \mu_{\text{near}} - \mu_{\text{far}} \neq 0$$

$$\bar{x}_{\text{near}} - \bar{x}_{\text{far}} = -3.49$$

$$SE_{\text{diff}} = \sqrt{\tfrac{12.2^2}{57} + \tfrac{16.0^2}{67}} = 2.54$$

$$t_{\text{test}} = \frac{-3.49 - 0}{2.54} = -1.374$$

- Using our "by hand" rule, $df = 56$, and this test statistic results in a p-value of 0.174
- We fail to reject the null hypothesis that there is no difference in IQ, though there is a trend towards children living near the smelter having slightly lower IQ levels.
- Minitab calculates $df = 120$, leading to a similar p-value of 0.170

Two-sample Quantitative Data - Example #2

- At the 2008 Olympics, several swimming world records were broken and controversy arose over new swimsuit designs providing an unfair competitive advantage
- In 2010, new international rules were implemented regulating swimsuit coverage and material
    - These rules naturally prompt the question "Do certain swimsuits really make swimmers faster?"
- Data from a study looking at the 1500m swim velocity of 12 competitive swimmers is shown in the table below:

Wetsuit	1.57	1.47	1.42	1.35	1.22	1.75	1.64	1.57	1.56	1.53	1.49	1.51
NoWetsuit	1.49	1.37	1.35	1.27	1.12	1.64	1.59	1.52	1.50	1.45	1.44	1.41

Two-sample Quantitative Data - Example #2

1. Download the "Wetsuits" dataset from http://www.lock5stat.com/datapage.html
2. By hand, constuct a 95% confidence interval for the average difference in 1500m swim velocity
3. Use Minitab to test whether there is a difference in mean 1500m swim velocity when swimming in a wetsuit vs. swimming without a wetsuit

Two-sample Quantitative Data - Example #2 (solution)

$$H_0 : \mu_{\text{Wetsuit}} - \mu_{\text{NoWetsuit}} = 0, H_A : \mu_{\text{Wetsuit}} - \mu_{\text{NoWetsuit}} \neq 0$$

$$\bar{x}_{\text{Wetsuit}} - \bar{x}_{\text{NoWetsuit}} = 0.078$$

- The 95% CI is given by: $0.078 \pm 2.201\sqrt{\frac{0.136^2}{12} + \frac{0.141^2}{12}} = (-0.046, 0.202)$
- The test statistic (provided by Minitab) is 1.37 and the *p*-value 0.186
- There is insufficient evidence to conclude that wearing a wetsuit has an impact on the average 1500m swim velocity

Paired Quantitative Data

- There is something special about the "Wetsuits" data
  - Each participant is measured twice, so our two groups actually consist of the same cases!
  - This setup is referred to as **paired data**
- Paired data provides a huge advantage in terms of reducing extraneous variability
  - In example #2, the two-sample t-test naively treated the groups as unrelated
  - In other words, the test looked at how different *all* of the swimmers in the Wetsuit group compared to *all* of the swimmers in the NoWetsuit group
  - We really are interested in how different *each* swimmer's velocity is relative to themselves

Paired Quantitative Data

To take advantage of paired design in the "Wetsuits" dataset we need to look at each individual's difference in swim velocity:

Wetsuit	1.57	1.47	1.42	1.35	1.22	1.75	1.64	1.57	1.56	1.53	1.49	1.51
NoWetsuit	1.49	1.37	1.35	1.27	1.12	1.64	1.59	1.52	1.50	1.45	1.44	1.41
Difference	0.08	0.10	0.07	0.08	0.10	0.11	0.05	0.05	0.06	0.08	0.05	0.10

Notice:

1. *Every swimmer* was faster when wearing a wetsuit
2. The variability of these differences is *much lower* than the amount of variability across different swimmers:

$$s_{\text{Wetsuit}} = 0.136 \text{ and } s_{\text{NoWetsuit}} = 0.141 \text{ versus } s_{\text{Difference}} = 0.022$$

Paired Quantitative Data

- As you might expect, statistical inference when the data is paired should be done using the “Difference” variable
  - Sometimes we might need to create this variable ourselves using a Minitab formula
  - Using the difference variable, our inference uses the exact same procedures we learned for one-sample quantitative data, for example:

$$\text{P\% CI: } \bar{x}_d \pm t^* \tfrac{s_d}{\sqrt{n_d}}$$

$$t_{\text{stat}} = \frac{\bar{x}_d - \text{Null Value}}{s_d/\sqrt{n_d}}$$

- $\bar{x}_d$ is the average of the paired differences, and n_d is the number of differences, or 12 in our Wetsuit example. Our t-distribution has 11 degrees of freedom in this example.

Paired Quantitative Data - Example

1. Use Minitab to test whether there is a difference in mean 1500m swim velocity when swimming in a wetsuit vs. swimming without a wetsuit
2. Compare the results of your paired test with those you previously got from the naive two-sample *t*-test

Paired Quantitative Data - Example (solution)

$$\bar{x}_d = 0.078, s_d = 0.022$$

$$t_{test} = \frac{0.078 - 0}{0.022/\sqrt{12}} = 12.3$$

- Using a t-distribution with 11 degrees of freedom the *p*-value for this test is nearly zero. There is overwhelming evidence that wearing a wetsuit improves 1500m swim velocity
- This is the opposite conclusion of the naive two-sample test, thus using the paired *t*-test greatly improved our power

Summary

- Procedurally, two-sample data is analyzed very similar to one-sample data with a focus on differences in means or proportions rather than single means or proportions
- When testing with two-sample categorical data we need to be careful to use the pooled proportion $\hat{p}_{1+2}$
  - We also need to be sure that $n_i p_i \geq 10$ and $n_i(1 - p_i) \geq 10$ for $i \in \{1, 2\}$
- With two-sample quantitative data we need to use the correct degrees of freedom
  - We also need to be sure that $n_1 \geq 30$ and $n_2 \geq 30$ or that both samples are roughly normal
- Finally, we should be on the lookout for paired data and know that it should be analyzed differently

Conclusion

Right now you should…

1. Be able to construct confidence intervals and perform hypothesis tests for two-sample categorical data (differences of proportions)
2. Be able to construct confidence intervals and perform hypothesis tests for two-sample quantitative data (differences of means)
3. Be able to identify paired data scenarios and use procedures that take advantage of the data's paired structure

These notes cover Sections 6.3, 6.4, and 6.5 of the textbook, I encourage you to read through those sections and their examples