Two-Sample Inference

Javier E. Flores

March 13, 2019

A Quick Story...

Ever since my first year in graduate school, a couple of friends and I have made an annual trip to Las Vegas for a weekend of amazing buffets, music, comedy, and a bit of gambling*.

*Disclaimer: I do not endorse gambling.

A Quick Story...

The first year I took this trip, I had a string of good luck at a video blackjack machine.

Starting with only \$20, I managed to rack up \$200 in winnings (after sitting at the same machine for probably 4 hours).

A Quick Story...

Riding high on my winnings, I decided to take a risk and throw down a $100 bet (mind you, I was/am a poor grad student).

And, as fate would have it, the dealer hit 21 and my $100 went down the drain.

You'd think that after such a devastating loss I would have learned my lesson, but I immediately threw down a second $100 bet confident that the first loss was only an anomaly.

And yet again, I lost.

A Quick Story...

Thinking back on this story, I wondered whether it's common for people to take risks after experiencing heavy loss.

So for this lecture, the first set of data we will consider is from a study investigating risk taking behavior after large real-world losses following a natural disaster.

Brisbane Floods of 2010-11

In late 2010, early 2011, Australia experienced one of the worst floods in its history.

An estimated 35 people died in the flooding, and property damage and other economic losses amounted in the billions.

Some observers and rescue workers even compared the amount of flooding and economic costs to be on the scale of what was seen in the US after Hurricane Katrina in 2005.

Risk Study

The authors of today's motivating study compared the risk taking behavior of homeowners who had been affected by these floods to homeowners who were not affected.

Residential homeowners in affected areas were sampled and offered the choice between receiving

- a fixed sum of \$10,
- or a lottery scratch card (\$10 face value) potentially worth \$500,000.

The study data are provided in the table below:

Flood Victim	Choice	Count
Yes	Fixed Sum	19
Yes	Lotto	75
No	Fixed Sum	54
No	Lotto	53

Two-Sample Categorical Data

These data are an example of **two-sample categorical data**.

The variable *Flood Victim* defines the two samples, or groups.

The binary variable *Choice* is our outcome of interest.

This type of data should be familiar to us as we learned (during the first lab) that these data may be summarized via two-way frequency tables:

	Fixed Sum	**Lotto**
Affected by Flood	19	75
Unaffected by Flood	54	53

In creating these tables, the convention is to set the grouping variable to define each row and the outcome to define each column.

Review

Aside from learning how to summarize these data, we've also learned of a few ways to analyze these data.

For example, if we wanted to determine the proportion of risk-takers in each group, we might think to construct separate confidence intervals for the proportions of lotto-choosers in each group.

As practice, construct separate 99% confidence intervals for the proportion of flood-affected and flood-unaffected homeowners who chose to receive a lotto ticket. If possible, use the normal approximation.

	Fixed Sum	**Lotto**
Affected by Flood	19	75
Unaffected by Flood	54	53

Solution

For the flood-affected group, $\hat{p} = 75/94 = 0.798$.

Since $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$, we may use a normal approximation to obtain the 99% CI:

$$0.798 \pm 2.576\sqrt{\frac{0.798(1-0.798)}{94}} = (0.691, 0.905)$$

The normal approximation also holds for the flood-unaffected group ($\hat{p} = 53/107 = 0.495$) yielding the 99% CI:

$$0.495 \pm 2.576\sqrt{\frac{0.495(1-0.495)}{107}} = (0.370, 0.620)$$

Solution

Comparing the endpoints of both intervals, it is clear that there is no overlap.

Without any overlap, we could comfortably conclude that the proportion of risk-takers among each group is different. More specifically, it appears that flood-affected homeowners are more likely to take a risk.

Question: If the intervals did over lap slightly, would that mean that the proportions in these two groups are likely the same?

Difference in Proportions

Remember that values near the endpoints of a confidence interval are less plausible than those near the center.

This considered, if we were to observe some slight overlap in confidence intervals, we should not necessarily conclude that the difference in proportions is 0.

To avoid the inferential uncertainty of such scenarios, a better approach to analyzing these data would be to construct a single interval for the difference in proportion.

Difference in Proportions

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

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

Note the similarity between this formula and that for the standard error of a single proportion:

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

Difference in Proportions

While the standard error of a difference in proportions is greater than that of the standard error of a single proportion, it is less than the standard error of two separate proportions:

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

It is for this reason that finding a confidence interval for the difference in proportions, rather than for two proportions separately, is the preferred approach for these data.

In order to use this standard error formula, we first must check that the following conditions are met:

- $n_1 p_1 \geq 10$ and $n_1(1-p_1) \geq 10$
- $n_2 p_2 \geq 10$ and $n_2(1-p_2) \geq 10$

Practice

In the 1860's Joseph Lister performed an experiment investigating the effects of following a sterile protocol on patient mortality after surgery. The data are provided in the table below.

	Died	**Survived**
Control	16	19
Sterile	6	34

Construct separate 99% confidence intervals for the proportion of sterile and non-sterile surgery patients who died. Comparing the intervals, what can you conclude?

Construct a single 99% confidence interval for the difference in proportions of patients that died. What can you conclude using this interval?

Solution

For the sterile surgery group, the 99% CI is:

$$0.15 \pm 2.576 * \sqrt{\frac{0.15(1-0.15)}{40}} = (0.005, 0.295)$$

For the non-sterile surgery group, the 99% CI is:

$$0.46 \pm 2.576 * \sqrt{\frac{0.46(1-0.46)}{35}} = (0.243, 0.677)$$

Comparing these two intervals, we see that they overlap. However, it seems that a smaller proportion of sterile patients die.

Computing the 99% CI for the difference in proportions ($\hat{p}_{\text{sterile}} - \hat{p}_{\text{non-sterile}} = 0.15 - 0.46 = -0.31$), we obtain:

$$-0.31 \pm 2.576 * 0.101 = (-0.57, -0.05)$$

Based on this confidence interval, it is much more apparent that using the sterile procedure improves patient survival.

Hypothesis Testing

Confidence intervals aside, an alternative analytic approach to these data would be hypothesis testing.

We've learned how to conduct hypothesis tests for a difference in proportions using a randomization distribution, but we can also conduct this test using a normal approximation.

When conducting a hypothesis test for a difference in proportions, our null hypothesis is often specified as: $H_0 : p_1 - p_2 = 0$.

Similar to what was seen for a single proportion, the appropriate test statistic for this hypothesis is:

$$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE}$$

Standard Error

In this case, what do we use for the standard error (SE)?

Question: Do we use the same standard error formula for confidence intervals? Why or why not?

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

Answer: Remember that, in hypothesis testing, we assume that H_0 is true. Therefore, the standard error that we use should correspond to the standard error under the null distribution.

The formula for the standard error that we use for confidence intervals does not correspond to the null distribution standard error.

Standard Error

Under H_0 we are assuming that $p_1 - p_2 = 0$, or that $p_1 = p_2$.

Therefore rather than compute $\hat{p}_1$ and $\hat{p}_2$ separately, we combine both groups into a single group and compute a single, **pooled proportion**. This proportion is denoted as $\hat{p}_{\text{pooled}}$.

Using this pooled proportion, the correct standard error formula for our test statistic is:

$$SE = \sqrt{\frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_2}}$$

Practice

Conduct a hypothesis test to investigate whether there is a difference in the proportion of homeowners who chose the lotto ticket between those affected by the flood and those unaffected by the flood. Use a significance level of 0.05.

$$z_{\text{test}} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}}$$

	Fixed Sum	**Lotto**
Affected by Flood	19	75
Unaffected by Flood	54	53

Solution

$$\hat{p}_{\text{flood aff.}} - \hat{p}_{\text{flood unaff.}} = 0.798 - 0.495 = 0.303$$

$$\hat{p}_{\text{pooled}} = (75 + 53)/(94 + 107) = 0.637$$

$$SE = \sqrt{\frac{0.637(1-0.637)}{94} + \frac{0.637(1-0.637)}{107}} = 0.068$$

$$z_{\text{test}} = \frac{0.303-0}{0.068} = 4.456$$

With a two-sided p-value of 0.000008, there is substantial evidence that flood-affected homeowners are more likely to choose the lotto ticket over the fixed sum.

Two-Sample Quantitative Data

Now that we've sufficiently discussed and practiced methods for analyzing two-sample categorical data, we'll shift our focus to learning methods appropriate for two-sample quantitative data.

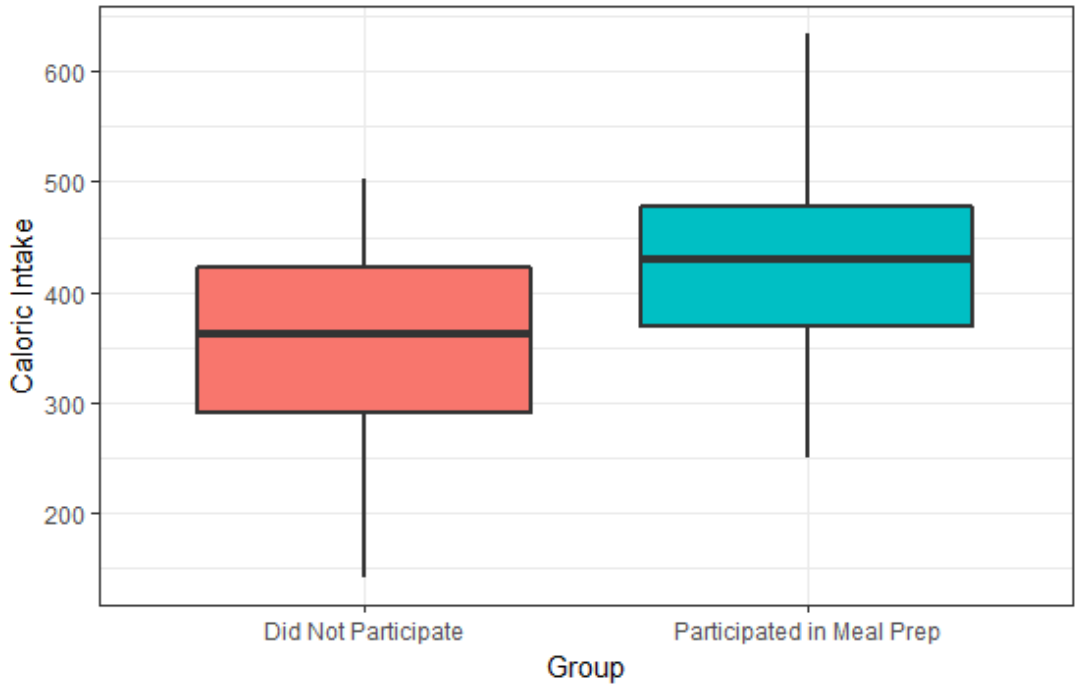
Since I've been binge-watching episodes of MasterChef Junior, we'll be looking at a dataset obtained from a study investigating the effect of child-participation in meal prep on their caloric intake.

Caloric Intake Study

The researchers behind this study recruited a total of 47 children, aged 6 to 10, and assigned each to either help their parents prepare a meal or to allow their parent(s) to prepare a meal without their assistance.

Boxplots of caloric intake for each group are provided below.

Question: Do the data appear normal? Do you suspect that there is an association between caloric intake and meal prep participation?



Caloric Intake Study

Given that the boxplots indicate that the distributions of caloric intake among each group are symmetric, it would be reasonable to assume that the data are normally distributed.

Furthermore, it seems that there might be an association considering the observed disparity in median (and mean) between the two groups.

Using our current toolset, we might think to go beyond this visual assessment and compute separate confidence intervals for each group:

$$\bar{x}_p \pm t_{crit,df=24} \frac{s_p}{\sqrt{n_p}} = (213.24, 649.56)$$

$$\bar{x}_{np} \pm t_{crit,df=21} \frac{s_{np}}{\sqrt{n_{np}}} = (139.88, 553.72)$$

Caloric Intake Study

Comparing these intervals, there is a substantial degree of overlap. If we were using the degree of overlap as a basis for claiming a difference between the groups, we might conclude that the groups are the same.

However, when discussing two sample inference for proportions, we learned that looking at the difference is often more *powerful* than looking at the individual components.

Here, rather than look at a difference in proportions, we should look at the difference in means: $\bar{x}_p - \bar{x}_{np}$

Difference in Means

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

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

As with one-sample quantitative data, we need to use the t-distribution when using this standard error.

Unlike with one-sample data, finding the degrees of freedom is not as straightforward a task.

For one-sample data, to find the degrees of freedom we simply subtracted one from the total sample size. With two samples, this same procedure does not apply.

Difference in Means

Given the complexity involved with determining the degrees of freedom for two-sample inference, we need to rely on software (Minitab) to determine the degrees of freedom for us.

If you don't have access to Minitab or are working by hand, a *conservative* approach to finding the degrees of freedom would be to use the smaller of $n_1 - 1$ and $n_2 - 1$.

By "conservative" I mean that this approach will lead to slightly wider confidence intervals and therefore slightly larger p-values.

Practice

Load the "Kid Calories" dataset into Minitab. Note that a value of "1" under the variable "Trt" denotes children who participated in meal prep.

Conduct a two-sample t-test by hand to determine whether the mean caloric intake differs between children who participated in meal prep and those that didn't.

Perform the same test using Minitab by clicking **Stat** -> **Basic Statistics** -> **Two-sample t-test**. How different is your p-value from what you obtained in Minitab?

Solution

$$H_0 : \mu_p - \mu_{np} = 0; \quad H_A : \mu_p - \mu_{np} \neq 0$$

$$\bar{x}_p - \bar{x}_{np} = 84.60$$

$$SE = \sqrt{\frac{105.70^2}{25} + \frac{99.5^2}{22}} = 29.95$$

$$t_{test} = \frac{84.60 - 0}{29.95} = 2.82$$

Using the conservative approach we find $df = 21$ and compute a p-value of 0.01.

We reject the null hypothesis that there is no difference in caloric intake. Children participating in meal prep on average have higher caloric intake than those who do not.

Performing this in Minitab, we obtain $df = 44$ and a p-value of 0.007.

2008 Olympics

In the 2008 Olympics a new full body swimsuit was debuted in competition.

That same year, several swimming world records were broken.

Controversy subsequently arose leading some to claim that the new swimsuit designs were providing an unfair advantage to those competitors wearing them.

2008 Olympics

In response to this controversy, international rules were changed in 2010 so that swimsuit coverage and material was more heavily regulated.

Was this rule change necessary? Does the full body swimsuit really make swimmers faster?

The "Wetsuits" dataset contains data from a study comparing the 1500m swim velocity of 12 competitive swimmers while wearing and not wearing full body swimsuits.

Practice

Download the "Wetsuits" dataset from http://www.lock5stat.com/datapage.html.

By hand, construct a 95% confidence interval for the average difference in 1500m swim velocity.

Perform a two-sample t-test in Minitab to test whether there is a difference in mean 1500m swim velocity while wearing and not wearing full body swimsuits.

Solution

$$H_0 : \mu_{suit} - \mu_{nosuit} = 0; \quad H_A : \mu_{suit} - \mu_{nosuit} \neq 0$$

$$\bar{x}_{suit} - \bar{x}_{nosuit} = 0.078$$

Using the standard error formula for two independent samples, the 95% CI is given by:

$$0.078 \pm 2.201\sqrt{\frac{0.136^2}{12} + \frac{0.141^2}{12}} = (-0.046, 0.202)$$

In Minitab, we obtain a test statistic of 1.37 and p-value of 0.186. Assuming $\alpha = 0.05$, we conclude that there is insufficient evidence to reject the null hypothesis. Based on this test it is not likely that wearing a wetsuit provides a competitive advantage...

BUT!

Are these data really from two *independent* samples?

Remember that these data were collected by computing the velocities for each of 12 swimmers while they wore the wetsuit AND when they didn't wear the wetsuit.

Therefore these data are not from two independent samples, but are instead from a single sample of **paired data**.

> Each case (i.e. an individual swimmer) is measured twice under different conditions.

Paired Data

In order to leverage the paired nature of the data for our inference, we need to look at the individual differences in swim velocity as opposed to the group difference in swim velocity.

Looking at the data below, we see that the variability among the individual differences is much lower than the variability across the different swimmers.

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
No Wetsuit	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

$s_{suit} = 0.136; \quad s_{nosuit} = 0.141; \quad s_{diff} = 0.022$

Paired Data

Given this massive reduction in variability, inference for paired data is based on using the average difference (i.e. $\bar{x}_{diff}$) as opposed to the difference in averages (i.e. $\bar{x}_{suit} - \bar{x}_{nosuit}$).

When using the average difference, we simply apply the same formulas we used for one-sample quantitative data:

$$\text{P\% CI:} \quad \bar{x}_{diff} \pm t_{crit} \frac{s_{diff}}{\sqrt{n_{diff}}}$$

$$t_{test} = \frac{\bar{x}_{diff} - \text{Null Value}}{\frac{s_{diff}}{\sqrt{n_{diff}}}}$$

Practice

In Minitab, create a variable which computes the difference in swim velocity for each swimmer.

Using this variable, test whether there is a difference in mean swim velocity when swimming in a wetsuit.

Compare the results of this test with those obtained previously from the naive two-sample t-test.

Solution

$$\bar{x}_{diff} = 0.078, \quad s_{diff} = 0.022$$

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

Using a t-distribution with $df = 12 - 1 = 11$, we obtain a p-value of nearly 0. This is substantial evidence against the null. We therefore conclude that there is indeed a difference in swim velocity when wearing a wetsuit.

This conclusion is exactly the opposite of what we obtained with the naive two-sample test. Using the paired t-test resulted in a massive increase in power due to the substantial reduction in variability observed when using individual differences as opposed to group differences.

Wrap-Up

Right now, you should...

- Know the assumptions behind the formulas used for each two-sample inferential procedure (i.e. the conditions should you check before using each)
- Be able to perform hypothesis tests and construct confidence intervals for two-sample categorical or quantitative data.
- Identify paired data and understand how it may be leveraged for more powerful inference.

These notes cover sections 6.3, 6.4, and 6.5 of the textbook. Please read through the section and its examples along with any links provided in this lecture.