Lecture 17

Matched Pairs (cont.) and Two-Sample Proportions

Today's Updates / Reminders

- Lab 7 is due Friday. You will use Excel and practice inference on two samples.
- Homework 8 due end of the day, next time we meet.
- Exam 2 review has been posted. Next class is exam review day. Exam is one week from today!
- STAT Hub is open for business! From 5:30-8pm, M-Th in SAS1101. You can park right in front of SAS if driving!

Review of Independent Means

ST312: Exam 2A Fall 2024

USE THE FOLLOWING TO ANSWER THE NEXT FOUR QUESTIONS. Is there evidence that the mean blood sugar level of adults in the US is less than the mean blood sugar level of young adults in the US? Below are two ***different*** sets of R output corresponding to the analysis of the collected data to be used in this problem. You can assume that the difference of means is calculated as $\bar{x}_{adult} - \bar{x}_{young\_adult}$.

```
        Two Sample t-test

data:  bloodsugar$adult and bloodsugar$young_adult
t = -1.5802, df = 82, p-value = 0.05896
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 0.216155
sample estimates:
mean of x mean of y
 20.69800  24.78972
```

```
        Welch Two Sample t-test

data:  bloodsugar$adult and bloodsugar$young_adult
t = -2.2381, df = 55.449, p-value = 0.01463
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.033431
sample estimates:
mean of x mean of y
 20.69800  24.78972
```

14. (4 pts) The appropriate test to answer the above question is:
    - ○ Chi-squared test for independence
    - ○ Two-sample comparison of proportions t-test
    - ○ Matched pairs t-test
    - ○ Two-sample comparison of means t-test

15. (4 pts) Upon reviewing the data, the researchers noticed the sample standard deviations were very different… $s_1 = 13.65$ and $s_2 = 0.62$. Based on this, they wisely decided to NOT pool the variance. Given this decision, calculate the standard error, $\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}$, given that $n_1 = 56$ and $n_2 = 28$. Write the formula used with the proper symbols and substitute the values. Round your answer to 3 decimal places.

ANS 15:

Review (continued)

16. (4 pts) Given $\bar{x}_1 = 20.7$ and $\bar{x}_2 = 24.8$, calculate the value of the test statistic and write the probability statement that we would use to test the hypothesis. Include the correct symbol for the test statistic in your answer. Round your answer to 2 decimal places.

ANS 16:

17. (4 pts) Again, assuming we are not pooling the variance, and based on the output given at the beginning of the problem, at the 5% significance level:

- ○ There is evidence that the mean blood sugar level of adults in the US is less than the mean blood sugar level of young adults in the US, because P-val = 0.05896.
- ○ There is evidence that the mean blood sugar level of adults in the US is less than the mean blood sugar level of young adults in the US, because P-val = 0.01463.
- ○ There is no evidence that the mean blood sugar level of adults in the US is less than the mean blood sugar level of young adults in the US, because P-val = 0.05896.
- ○ There is no evidence that the mean blood sugar level of adults in the US is less than the mean blood sugar level of young adults in the US, because P-val = 0.01463.

Chapter 9.3

Analysis of Paired Data

Does running make you lose weight?

- This summer, I did what anyone else would do with their time off… I did a scientific study that involved statistics. 🤪
- I meticulously weighed myself just before I went on my run. I kept the distance the same, the route the same, and I ate/drank nothing and didn't use the bathroom between scale measurements. I weighed immediately after I got back from my run. Then I weighed again 20 minutes later.
- How much weight did I lose from my run?

pre_post_run_weight.csv

Matched Pairs Hypothesis Test

- 5 men participated in a study to determine if exercise would reduce their LDL cholesterol reading. The data is show below.

$s_d = 6.496153$

Subject	A	B	C	D	E
BEFORE	182	121	133	221	198
AFTER	174	115	131	205	199
DIFF					

- Conduct a hypothesis test that the cholesterol was reduced by exercise using a significance level of 0.05.
- Could you use a CI to make a hypothesis test conclusion?

Hypothesis Test

$\bar{d}$ = -6.2

s_d = 6.496153

1. Write the null and alternative hypotheses
2. Calculate the test statistic
3. Find the p-value given df = n – 1
4. Write the conclusions

Matched Pairs or Not?

- To test the effectiveness of the Atkins diet, 36 randomly selected subjects are weighed before the diet and six months after treatment with the diet. The two samples consist of the before/after weights.

- To determine whether smoking affects memory, 50 randomly selected smokers are given a test of word recall and 50 randomly selected nonsmokers are given the same test. Sample data consist of the scores from the two groups.

- IQ scores are obtained from a random sample of 75 wives and IQ scores are obtained from their husbands.

- Annual incomes are obtained from a random sample of 1200 residents of Alaska and another random sample of 1200 residents of Hawaii.

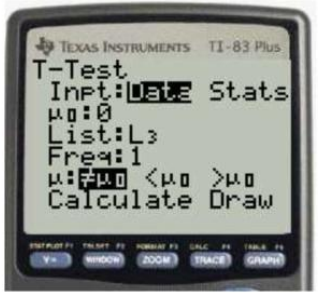
- Scores from a standard test of mathematical reasoning are obtained from a random sample of statistics students and another random sample of sociology students.

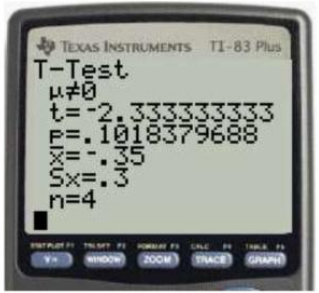
Appendix: Matched Pairs Calculator

CALCULATOR STEPS – TWO-SAMPLE T-TEST: DEPENDENT SAMPLES

Though this calculation seems like it would be almost identical to the calculations from 9.3, it is VERY different:

1) Press STAT and ENTER twice to start entering the data.
2) Then arrow up to the top of the column, L_3. Notice that L_3 is highlighted. Now also notice your cursor is flashing at the bottom. Now we will type in the expression, $L_1 - L_2$. Press 2^{nd}, the button for 1. Press minus. Press 2^{nd}, the button for 2. This effectively takes all values in list 1 and subtracts list 2 from it.
3) Press STAT and arrow right twice to TESTS
4) Choose option 2, T-Test…
5) We leave μ_0 as zero. We are wishing to test our data against the premise that there is no difference -- or said another way, that the difference is zero. We also adjust the List to point to L_3, which is where our difference values are. We choose the option of $\neq\mu_0$ in this case. Then move to Calculate and press ENTER.





Inference for Two Means

Also known as Welch t-Test

• Pooled variance

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$df = (n_1-1)+(n_2-1)$$

• Unequal variance

No pooling – this step not required

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Decimal df

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Confidence interval: $(\bar{x}_1 - \bar{x}_2) \pm t_{df} \cdot \hat{\sigma}_{\bar{X}_1-\bar{X}_2}$ -find t_{df}, given α and df

Hypothesis test: $t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\hat{\sigma}_{\bar{X}_1-\bar{X}_2}}$ Usually 0, or D_0

Then go to df on the table, find where t belongs on that row to find p-value

• Matched pairs

No pooling – this step not required

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$$

*this is just plain old standard deviation of $\bar{d}$

$$df = n-1$$

Confidence interval: $\bar{d} \pm t_{df}\frac{s_d}{\sqrt{n}}$

Hypothesis test: $t = \frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}$ Usually 0, or D_0

Then go to df on the table, find where t belongs on that row to find p-value

When is it ok to use the t procedures?

Sample size	Use t procedures
n < 15	Ok if your **data look Normal distributed** (you can check this with a Normal Quantile Plot)
15 ≤ n < 40	Ok if data shows **some skewness**. Except in the presence of outliers or strong skewness.
n ≥ 40	Ok even if data are **strongly skewed**. Except in the presence of outliers.

Use different procedure or increase sample size

Investigate the cause: data wrongly recorded, equipment malfunction, Response bias

Important! No matter the sample size, t-test is not recommended if there are outliers in our data set!!

Chapter 9.4

Inferences Concerning a Difference Between Population Proportions

Set-Up

- We assume population 1 has a proportion p_1 and standard error, $\sqrt{\frac{p_1 q_1}{n_1}}$ and population 2 has a proportion p_2 and standard error, $\sqrt{\frac{p_2 q_2}{n_2}}$.
- Since neither p_1 or p_2 is known, we collect data.
- Note, as with means, that we could say:
  - $p_1 > p_2$ but instead we will say, $p_1 - p_2 > 0$
  - $p_1 < p_2$ but instead we will say, $p_1 - p_2 < 0$
  - $p_1 \neq p_2$ but instead we will say, $p_1 - p_2 \neq 0$
- If we know that one is greater than the other from the outset, we may wish to state $p_1 - p_2$ is <, >, ≠ to D_0, where D_0 represents some presumed difference between the two populations.

Standard Error

Recall:
$\hat{p}$ = sample proportion
p = population proportion

- It can easily be shown that the best estimate of the mean is $p_1 - p_2$.
  - $\mu_{\hat{p}_1-\hat{p}_2} = \mu_{x_1/n_1} - \mu_{x_2/n_2} = \frac{1}{n_1} \cdot n_1 p_1 - \frac{1}{n_2} \cdot n_2 p_2 = p_1 - p_2$
- For standard deviation, we must start by using variance:
  - $\sigma^2_{\hat{p}_1-\hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \left(\sqrt{\frac{p_1 q_1}{n_1}}\right)^2 + \left(\sqrt{\frac{p_2 q_2}{n_2}}\right)^2$
  - So then, $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$
  - Under the null hypothesis, $p_1 = p_2$. So, we can consider these as two samples from the same population. Combining the information from each sample gives us one sample with $n_1 + n_2$ observations whose proportion is some value p. We can estimate p using the **pooled estimate**, $\bar{p}$.
  - $\bar{p} = \frac{x_1+x_2}{n_1+n_2}$ or if you don't have x_1, x_2, you can use $\bar{p} = \left(\frac{n_1}{n_1+n_2}\right)\hat{p}_1 + \left(\frac{n_2}{n_1+n_2}\right)\hat{p}_2$
  - So finally, $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Inferential Statistics

- Recall that if we don't know anything about the data being tested, we start with a confidence interval.
- If a claim has been made, we can do a hypothesis test.
- Confidence intervals are all the same:

$$point\ estimate\ \pm confidence\ factor * standard\ error\ of\ the\ mean$$

- In terms of this problem, we will say:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Note that to use the confidence interval, which is based on the Normal approximation, we need n_1p_1, n_1q_1, n_2p_2, n_2q_2 all ≥ 10

- Hypothesis tests are built as before $(sample - null)/standard\ error)$:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Usually assumed to be zero

Example

- A researcher believes that young adults who are active on social media platforms (like TikTok, Instagram, or X/Twitter) are more likely to participate in political activities (such as voting, attending rallies, or contacting elected officials) than those who are not active on social media.
- The researcher conducts a survey of 500 college students, randomly selecting 250 who are active on social media and 250 who are not. They ask the students whether they have participated in any political activities in the past year.
- Among the 250 social media users, 180 report having participated in at least one political activity. Among the 250 non-social media users, 150 report having participated in at least one political activity.
- Run a hypothesis test, confirming conditions have been met.

Example

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } \bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- $n_1 = 250$, $x_1 = 180$
- $n_2 = 250$, $x_2 = 150$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

CI Example

- A sample of 244 men and 232 women was collected and the cholesterol levels of each subject were measured. Of the men, 73 had elevated cholesterol and, of the women, 44 had elevated cholesterol.
- Calculate a 95% confidence interval for the difference in the proportion of men and women with elevated cholesterol levels.

- Use the CI to make a hypothesis test claim.

Sample Size Determination

- Continuing with the previous example, if we want a 95% confidence interval with a margin of error no greater than 5%, what sample size do we need?

$$n = \left(\frac{z_{\alpha/2}}{m}\right)^2 (p_1^* q_1^* + p_2^* q_2^*)$$

- We always round up.
- Also if no estimates are known (usually the case), we replace p_1^* and p_2^* with 0.5.
- Note that to use the confidence interval, which is based on the Normal approximation, we still need n_1p_1, n_1q_1, n_2p_2, n_2q_2 all ≥ 5
- Finally sample size calculation is applied to each population… $n = n_1 = n_2$.

Formulas Summary

- Always use z (not t)!
- Confidence Interval (use $\hat{p}_1$ and $\hat{p}_2$!)
  - $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
- Hypothesis Test
  - $z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Usually assumed to be zero

  - $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$ or if you don't have x_1, x_2, you can use $\bar{p} = \left(\frac{n_1}{n_1 + n_2}\right)\hat{p}_1 + \left(\frac{n_2}{n_1 + n_2}\right)\hat{p}_2$
- Sample Size Determination
  - $n = \left(\frac{z_{\alpha/2}}{m}\right)^2 (p_1^* q_1^* + p_2^* q_2^*)$ where p_1^* and p_2^* are assumed to be 0.5 unless otherwise given. Also, always round up.

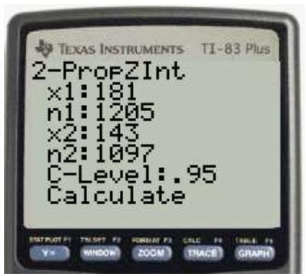
Appendix: Two-Sample Proportion

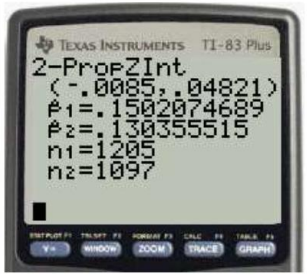
21. Tattoos The Harris Poll conducted a survey in which they asked, "How many tattoos do you currently have on your body?" Of the 1205 males surveyed, 181 responded that they had at least one tattoo. Of the 1097 females surveyed, 143 responded that they had at least one tattoo. Construct a 95% confidence interval to judge whether the proportion of males that have at least one tattoo differs significantly from the proportion of females that have at least one tattoo. Interpret the interval.

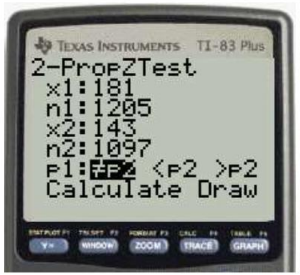
CALCULATOR STEPS – TWO-SAMPLE PROPORTION

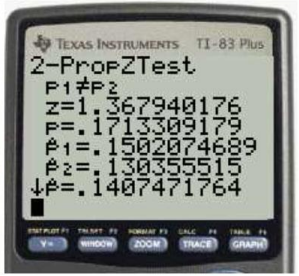
If you've been tracking with everything else we've done thus far in the course, this is simple.

1) As with all other tests, click STAT, arrow right twice to TESTS and choose B: 2-PropZInt.
2) Enter the males as x1 and n1 (though it makes no difference which is which since we are asked if they differ… not if one is greater than the other). Enter the females as x2 and n2. Leave C-Level as 0.95. Then choose Calculate.
3) We get the results on the last screen. The interval contains zero… so we would say that there is not sufficient evidence to conclude that the proportion of males is different than the proportion of females.









4) Note that we can run it as a 2-PropZTest and we arrive at the same conclusion. As stated, we are only looking for a difference, so we set the test to be ≠ p2 in the second picture below. We find a p-value of 0.1713. Keep in mind that a 95% confidence interval results in an α = 0.025. (Recall this is because a confidence interval splits the alpha among both tails. Take 0.05 / 2.)