MATH 3339 - 03 15951
Statistics for the Sciences
Chapter 10: Inferences on Two Groups or Populations

Wendy Wang
wwang60@central.uh.edu

Lecture 33 - 3339
Outline

1. Matched Pairs Test

2. Two-population inference introduction

3. Comparing Two Means
Popper Set Up

- Fill in all of the proper bubbles.
- Make sure your ID number is correct.
- Make sure the filled in circles are very dark.
- This is popper number 29.
Matched Pairs

Matched Pairs T test
Matched Pairs

In low-speed crash test of five BMW cars, the repair costs were computed for a factory-authorized repair center and an independent repair facility. The results are as follows.

<table>
<thead>
<tr>
<th>Authorized repair center</th>
<th>$797</th>
<th>$571</th>
<th>$904</th>
<th>$1147</th>
<th>$418</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Repair center</td>
<td>$523</td>
<td>$488</td>
<td>$875</td>
<td>$911</td>
<td>$297</td>
</tr>
</tbody>
</table>

We want to estimate the mean of the difference between the two repair centers.
Inference for Matched Pairs

- The previous question is a matched pair.
- We are looking at the same car. The subject units are exactly the same for both responses.
- We calculate the differences first and find the mean and standard deviation of the differences.
- Then this problem is the same as a one-sample confidence interval.
  - We first find the differences from each observation.
  - The point estimate is $\bar{x}_d = \text{mean of the differences}$.
  - The standard deviation is $s_d = \text{the standard deviation of the differences}$.
  - Then the margin of error is $m = t^* \left( \frac{s_d}{\sqrt{n}} \right)$.
  - The confidence interval is $\bar{x}_d \pm t^* \left( \frac{s_d}{\sqrt{n}} \right)$.
- If we want a hypothesis test, the test statistic is: $t = \frac{\bar{x}_d - \mu_d}{s_d/\sqrt{n}}$. 

Wendy Wang  wwang60@central.uh.edu  MATH 3339  Lecture 33 - 3339  6 / 28
Matched Pairs Assumptions

- Matched pairs is a special test when we are comparing corresponding values in data.

- This test is used only when our data samples are DEPENDENT upon one another (like before and after results).

- Matched pairs t – test assumptions:
  1. Each sample is an SRS of size n from the same population.
  2. The test is conducted on paired data (the samples are NOT independent).
  3. Unknown population standard deviation.
  4. Either a Normal population or large samples ($n \geq 30$).

- Hypotheses - $H_0 : \mu_d = 0$ and $H_a : \mu_d \neq 0$ or $\mu_d < 0$ or $\mu_d > 0$. Where $\mu_d$ is the mean of the differences.
Crash Test Repair Costs

We want to determine a 95% confidence interval for the difference in the repair cost of the authorized repair center and the independent repair center.

<table>
<thead>
<tr>
<th>Authorized repair center</th>
<th>$797</th>
<th>$571</th>
<th>$904</th>
<th>$1147</th>
<th>$418</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Repair center</td>
<td>$523</td>
<td>$488</td>
<td>$875</td>
<td>$911</td>
<td>$297</td>
</tr>
<tr>
<td>Differences</td>
<td>$274</td>
<td>$83</td>
<td>$29</td>
<td>$236</td>
<td>$121</td>
</tr>
</tbody>
</table>

Wendy Wang  wwang60@central.uh.edu  MATH 3339  Lecture 33 - 3339  8 / 28
> auth=c(797,571,904,1147,418)
> indep=c(523,488,875,911,297)
> t.test(auth,indep,conf.level = 0.95, paired = TRUE)

Paired t-test

data:  auth and indep
t = 3.2148, df = 4, p-value = 0.03244
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
20.26155 276.93845
sample estimates:
mean of the differences
148.6
Example 2

A new law has been passed giving city police greater powers in apprehending suspected criminals. For six neighborhoods, the numbers of reported crimes one year before and one year after the new law were given. Does this indicate that the number of reported crimes have dropped?

<table>
<thead>
<tr>
<th>Neighborhood</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>18</td>
<td>35</td>
<td>44</td>
<td>28</td>
<td>22</td>
<td>37</td>
</tr>
<tr>
<td>After</td>
<td>21</td>
<td>23</td>
<td>30</td>
<td>19</td>
<td>24</td>
<td>29</td>
</tr>
</tbody>
</table>
R code

> t.test(before, after, alternative="greater", paired=TRUE)

Paired t-test

data: before and after
t = 2.1624, df = 5, p-value = 0.04147
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.4316912 Inf
sample estimates:
mean of the differences
6.333333
Two-population inference

- Is the mean miles per gallon of automobiles significantly different depending on the manufacturer of the automobile?

- Is the mean price of a business textbook significantly lower than the mean price of a general course textbook?

- What is the difference between the mean height of men and mean height of women?

- We want to estimate: $\mu_1 - \mu_2$
### Notations Used

- **For the population**

<table>
<thead>
<tr>
<th>Population variable</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$\mu_1$</td>
<td>$\sigma_1$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$\mu_2$</td>
<td>$\sigma_2$</td>
</tr>
</tbody>
</table>

- **For the sample**

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample Size</th>
<th>Sample Mean</th>
<th>Sample Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$</td>
<td>$\bar{x}_1$</td>
<td>$s_1$</td>
<td></td>
</tr>
<tr>
<td>$n_2$</td>
<td>$\bar{x}_2$</td>
<td>$s_2$</td>
<td></td>
</tr>
</tbody>
</table>
Two Population problems

- The goal of inference is to compare the responses in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group.
Assumptions for Difference of Two Means

1. Both samples must be independent SRSs from the populations of interest.

2. Both sets of data must come from normally distributed populations.
Two-sample \( t \)

- If the population standard deviations \( \sigma_1 \) and \( \sigma_2 \) is unknown the sample standard deviations \( s_1 \) and \( s_2 \) is used.

- When we use the sample standard deviations we use the **two-sample \( t \) statistic**

\[
t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}
\]

with \( k \) degrees of freedom approximated by software or the smaller value of \( n_1 - 1 \) or \( n_2 - 1 \).
Approximate Degrees of Freedom

- The reality is that the previous model is not really Student’s $t$, but only something close.

- So the calculators and other software such as R uses an approximate degrees of freedom called Satterthwaite degrees of freedom.

- Calculated

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

- This is only to show what degrees of freedom R and the calculators are using. If we do this by hand use the smaller of $n_1 - 1$ or $n_2 - 1$. 
Interval Estimation of $\mu_1 - \mu_2$

1. **Point Estimate:** $\bar{x}_1 - \bar{x}_2$

2. **Confidence level:** $1 - \alpha$

3. **Critical value:** $t^*$ with degrees of freedom of $n_1 - 1$ or $n_2 - 1$ whichever is smaller. In R: $t^* = qt(C + \alpha/2, df)$.

4. **Margin of Error:**

\[ E = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]

5. **Confidence Interval:** point estimate $\pm$ margin of error
Example: Check out

A well known grocery store chain performed a study to determine whether the average purchase through a self-checkout facility was less than the average purchase at the traditional checkout stand. To conduct the test, a random sample of 125 customer transactions at the self-checkout was obtained and a second random sample of 125 transactions from customers using traditional checkout process was obtained. The following statistics were computed from each sample:

<table>
<thead>
<tr>
<th>Self-Checkout</th>
<th>Traditional Checkout</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{x}_1 = $45.68</td>
<td>$\bar{x}_2 = $78.49</td>
</tr>
<tr>
<td>$s_1 = $58.20</td>
<td>$s_2 = $62.45</td>
</tr>
<tr>
<td>$n_1 = 125$</td>
<td>$n_2 = 125$</td>
</tr>
</tbody>
</table>

Develop a 90% confidence interval of the difference between the different checkout systems.
In R: \((\bar{x}_1 - \bar{x}_2) - qt((1 + C)/2) \times \sqrt{s_1^2/n_1 + s_2^2/n_2}\) for lower value of confidence interval, \((\bar{x}_1 - \bar{x}_2) + qt((1 + C)/2) \times \sqrt{s_1^2/n_1 + s_2^2/n_2}\) for upper value of confidence interval.

```r
> (45.68-78.49)-qt(1.9/2,124)*sqrt(58.2^2/125+62.45^2/125)
[1] -45.4635
> (45.68-78.49)+qt(1.9/2,124)*sqrt(58.2^2/125+62.45^2/125)
[1] -20.1565
```
Two - Sample $t$-Test

- Compare the responses to two treatments or characteristics of two populations.
- These tests are different than the matched pairs $t$-test.
- Hypotheses
  - Null - $H_0: \mu_1 = \mu_2$
  - Alternative - $H_a: \mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$
Assumptions for a Two-Sample $t$-Test

The goal of inference is to compare the responses in two groups.

1. Each group is considered to be a **simple random sample** from two **distinct** populations.

2. The responses in each group are **independent** of those in the other group.

3. The distribution of the variables are **Normal** or have a large sample $n_1 \geq 30$ and $n_2 \geq 30$.

Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

With degrees of freedom equal to the smaller of $n_1 - 1$ or $n_2 - 1$. 
Comparing mean MPG

From a random sample of 45 Prius automobiles and 45 Civic automobiles we get the following statistics:

<table>
<thead>
<tr>
<th>Automobile</th>
<th>n</th>
<th>Sample mean $\bar{x}$</th>
<th>Sample SD s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prius</td>
<td>45</td>
<td>47.62</td>
<td>2.430</td>
</tr>
<tr>
<td>Civic</td>
<td>45</td>
<td>49.4</td>
<td>7.226</td>
</tr>
</tbody>
</table>

Can we say from this information that the Civic has a different mean mpg than the Prius?
Is the mean MPG for Prius automobiles different from mean MPG for Civic automobiles?

- Null hypothesis: $H_0 : \mu_{\text{Prius}} = \mu_{\text{Civic}}$
- Alternative hypothesis: $H_A : \mu_{\text{Prius}} \neq \mu_{\text{Civic}}$
Two-sample \( t \) test statistic

Formula:

\[
t = \frac{\text{estimate} - \text{hypothesized mean of estimate}}{\text{SE of estimate}}
\]

\[
= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}
\]

\[
= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}
\]

\[
= \frac{47.62 - 49.4}{\sqrt{\frac{2.430^2}{45} + \frac{7.226^2}{45}}} = -1.5662
\]
Two-sample $t$ test statistic

Formula:

$$t = \frac{\text{estimate} - \text{hypothesized mean of estimate}}{\text{SE of estimate}}$$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{47.62 - 49.4}{\sqrt{\frac{2.430^2}{45} + \frac{7.226^2}{45}}} = -1.5662$$
Two-sample $t$ test statistic

Formula:

$$ t = \frac{\text{estimate} - \text{hypothesized mean of estimate}}{\text{SE of estimate}} $$

$$ = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} $$

$$ = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} $$

$$ = \frac{47.62 - 49.4}{\sqrt{\frac{2.430^2}{45} + \frac{7.226^2}{45}}} = -1.5662 $$
**P-value and Conclusion**

\[ P - \text{value} = 2P(T < -1.5663) \]

In R:

```r
2*pt(-1.5663, df=44)
[1] 0.1244429
```

*P*-value = 0.1244, which is greater than 0.1 (10%). Thus we fail to reject the null hypothesis. Thus we **cannot conclude** that the mean MPG of Honda Civic automobiles is significantly different than the mean MPG of a Toyota Prius automobile.
Solid fats are more likely to raise blood cholesterol levels than liquid fats. Suppose a nutritionist analyzed the percentage of saturated fat for a sample of 6 brands of stick margarine (solid fat) and for a sample of 6 brands of liquid margarine and obtained the following results:

Liquid: [16.5, 17.1, 17.5, 17.3, 17.2, 16.7]

We want to determine if there a significant difference in the average amount of saturated fat in solid and liquid fats.
> stick = c(25.5, 26.7, 26.5, 26.6, 26.3, 26.4)
> liquid = c(16.5, 17.1, 17.5, 17.3, 17.2, 16.7)
> t.test(stick, liquid, mu = 0, alternative = "two.sided")

Welch Two Sample t-test

data: stick and liquid
t = 39.604, df = 9.8276, p-value = 3.608e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 8.759808 9.806859
sample estimates:
mean of x mean of y
26.33333  17.05000