AMS 7
Two-sample Hypothesis Tests
Lecture 12

Department of Applied Mathematics and Statistics, University of California, Santa Cruz

Fall 2013
Query: Do different brands of cookies have the same number of chips on average?

We would have to take into account that there is a lot of variation between cookies.
Let $\mu_1 =$ population mean # of chips per cookie of brand 1
$\mu_2 =$ population mean # of chips per cookie of brand 2
$\bar{x}_1 =$ sample mean of brand 1
$\bar{x}_2 =$ sample mean of brand 2
$\sigma_1, \sigma_2$ population standard deviations
$s_1, s_2$ sample standard deviations

brand 1 = Pantry Essentials
brand 2 = Chewy Chips Ahoy

$\bar{x}_1 = 18.65 \quad \bar{x}_2 = 20.09$
$s_1 = 4.13 \quad s_2 = 4.09$
$n_1 = 147 \quad n_2 = 160$
By CLT, \( \bar{x}_1 \) and \( \bar{x}_2 \) are approximately normal with means \( \mu_1 \) and \( \mu_2 \) and standard deviations \( \frac{\sigma_1}{\sqrt{n_1}} \) and \( \frac{\sigma_2}{\sqrt{n_2}} \).

We want to test the claim that the mean number of chocolate chips in a cookie do not differ between the two brands: \( \mu_1 = \mu_2 \).

It turns out that the sum (or difference) of two normals is also normal. Sooo…. \( \bar{x}_1 - \bar{x}_2 \) is approximately normal with mean and sd:

\[
E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2 \\
SD[\bar{X}_1 - \bar{X}_2] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

In general, \( \sigma_1 \) and \( \sigma_2 \) are unknown, so we estimate with \( s_1 \) and \( s_2 \).

\[
\Rightarrow SD[\bar{X}_1 - \bar{X}_2] = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]
Thus, the test statistic follows a t-distribution:

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

But how many degrees of freedom do we have here?

- If we assume equal population standard deviations: \( \sigma_1 = \sigma_2 \), then \( df = n_1 + n_2 - 2 \)
- Otherwise, we use the smaller value between \( n_1 - 1 \) and \( n_2 - 1 \) (this is the more conservative approach)
- JMP will use a more complicated formula that we do not cover in this class

Key idea: We are comparing two populations to each other, without testing a particular value for either, i.e., just testing \( \mu_1 = \mu_2 \), NOT \( \mu_1 = \mu_2 = 25 \) (which we could do with a two 1-sample t tests)
Testing the claim that the mean \# of chips in a cookie does NOT differ between the two brands.

1. \( H_0 : \mu_1 = \mu_2 \) (claim) vs \( H_1 : \mu_1 \neq \mu_2 \)
   where \( \mu_1 \) and \( \mu_2 \) are the population mean \# of chips per cookie of brand 1 and brand 2, respectively.

2. \( \alpha = 0.05 \)

3. Test statistic: 
   \[
   t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}；
   \]
   a t-distribution with \( 147 - 1 = 146 \) degrees of freedom

4. Compute the test statistic 
   \[
   t = \frac{(18.65 - 20.09)}{\sqrt{\frac{4.13^2}{147} + \frac{4.09^2}{160}}} = -3.07
   \]
   (a) Critical region: \( t < -1.984 \) or \( t > 1.984 \)
   (b) p-value: \( P(t < -3.07) + P(t > 3.07) = 2P(t < -3.07) = \)
   p-value < 0.01 (from table)

5. Reject the null since (a) \(-3.07 < -1.984 \) or (b) p-value < 0.01 < 0.05
6) Conclude that there is sufficient evidence to warrant rejection of the claim that the mean # of chips in a cookie does not differ between brand 1 and brand 2.

Confidence Interval for the difference of population means:

\[ \text{CI for } \mu_1 - \mu_2 \]

* We take \( \bar{x}_1 - \bar{x}_2 \) as the point estimate for the difference, then subtract and add the margin of error \( E \): \( (\bar{x}_1 - \bar{x}_2) \pm E \)

* where the margin of error is given by: \( E = t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \)

* and the degrees of freedom is either \( \min\{n_1 - 1, n_2 - 1\} \) or \( n_1 + n_2 - 2 \).

\[ E = 1.984 \sqrt{\frac{4.13^2}{147} + \frac{4.09^2}{160}} = 0.93 \text{ so the 95\% CI is given by:} \]

\[ (18.65 - 20.09) \pm 0.93 \equiv -1.44 \pm 0.93 \Rightarrow (-2.37, -0.51) \]

The null is that there is no difference, i.e. zero difference, therefore, since 0 is not in the 95\% CI, we reject the null hypothesis.
Oneway Analysis of # chocolate chips By Brand

Quantiles

<table>
<thead>
<tr>
<th>Level</th>
<th>Minimum</th>
<th>10%</th>
<th>25%</th>
<th>Median</th>
<th>75%</th>
<th>90%</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>12</td>
<td>16</td>
<td>19</td>
<td>21</td>
<td>24</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>15</td>
<td>17</td>
<td>20</td>
<td>23</td>
<td>26</td>
<td>31</td>
</tr>
</tbody>
</table>

Means and Std Deviations

<table>
<thead>
<tr>
<th>Level</th>
<th>Number</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Err Mean</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>147</td>
<td>18.6531</td>
<td>4.13249</td>
<td>0.34084</td>
<td>17.979</td>
<td>19.327</td>
</tr>
<tr>
<td>2</td>
<td>160</td>
<td>20.0938</td>
<td>4.09448</td>
<td>0.32370</td>
<td>19.454</td>
<td>20.733</td>
</tr>
</tbody>
</table>

t Test

2-1
Assuming unequal variances

| Difference | 1.44069 | t Ratio | 3.06493 |
| Std Err Dif | 0.47006 | DF | 302.3132 |
| Upper CL Dif | 2.36568 | Prob > t | 0.0024* |
| Lower CL Dif | 0.51569 | Prob > t | 0.0012* |
| Confidence | 0.95 | Prob < t | 0.9988 |
Comparing Two Proportions

†† Example: In a previous example we looked at a poll of 1207 people of whom 53% disapproved of how Bush was handling the economy. In an earlier poll of 1002 people, 48% disapproved. Is this sampling error (assume this is the claim), or has the true population proportion changed?

6-steps

1. Let \( p_1 \) = the population proportion of people who disapproved in the first poll
   and \( p_2 \) = the population proportion of people who disapproved in the second poll
   \( H_0 : p_1 = p_2 \) (claim) vs \( H_1 : p_1 \neq p_2 \)

2. \( \alpha = 0.05 \)

3. test statistic: \( z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \)

where \( \bar{p} = \frac{x_1+x_2}{n_1+n_2} = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2} \)

The test statistic follows a normal distribution
4) \( \bar{p} = \frac{(1002)(0.48)+(1207)(0.53)}{1002+1207} = 0.507 \)

\[ \Rightarrow z = \frac{0.53-0.48}{\sqrt{(0.507)(1-0.507)\left(\frac{1}{1002} + \frac{1}{1207}\right)}} = -2.34 \]

(a) Critical region: \( z < -1.96 \) or \( z > 1.96 \)

(b) p-value: \( P(z < -2.34) + P(z > 2.34) = 2P(z < -2.34) = 2(0.0096) = 0.0192 \)

5) Reject the null since (a) Critical region: \(-2.34 < -1.96 \) OR (b) p-value = 0.0192 < 0.05.

6) Conclude that there is sufficient evidence to warrant rejection of the claim that the proportion of disapproval is due to random variation.
95% CI: \((\hat{p}_1 - \hat{p}_2) \pm E\) where 

\[E = z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\]

\[
\Rightarrow (0.48 - 0.53) \pm 1.96 \sqrt{\frac{(0.48)(0.52)}{1002} + \frac{(0.53)(0.47)}{1207}} \\
\Rightarrow -0.05 \pm 0.0418 \\
\Rightarrow (-0.0918, -0.0082)
\]

\(\star\) 0 is not in the 95% CI, so we reject the null hypothesis.
More two-sample t test examples

†† Example 1: A student working at a fast food restaurant claims that drive-through customers spend more than walk-ups. A random sample of 38 customers who came to the counter spent a mean of $5.19 with a sd of $3.06. Another random sample of 17 drive-through customers spent a mean of $5.94 with a sd $3.25. Test this claim.

1. $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 < \mu_2$ (claim)
   where $\mu_1$ and $\mu_2$ are the population mean amount spend by walk-up and drive-through customers, respectively.

2. $\alpha = 0.05$

3. Test statistic: 
   $$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
   a $t$-distribution with $min\{38 - 1, 17 - 1\} = 16$ degrees of freedom
4) Compute the test statistic $t = \frac{(5.19-5.94)}{\sqrt{\frac{3.06^2}{38} + \frac{3.25^2}{17}}} = \frac{-0.75}{0.931} = -0.805$

(a) Critical region: one-tail test $\Rightarrow t < -1.746$

OR

(b) p-value: $P(t < -0.805) = p-value > 0.10$ (from table)

5) Fail to reject the null since

(a) $-0.805 < -1.746$

OR

(b) p-value $> 0.10 \not< 0.05$

6) There is not sufficient sample evidence to support the claim that drive-through customers spend more than walk-ups.
Example 2: The data, see Berger et al. (1988), consists of survival times of rats in two experimental groups. The first group (Ad libitum group) is comprised of 90 rats who were allowed to eat freely as they desired. This group had a mean survival time of 682.3 days with sd 134.3. The second group (Restricted group) is comprised of 106 rats that were placed. The mean survival time of the Restricted group was 968.7 with sd 284.6. Test the claim, using a 0.02 significance level, that the population mean survival time of rats is the same under both experimental groups.
1. \( H_0 : \mu_1 = \mu_2 \) (claim) vs \\
\( H_1 : \mu_1 \neq \mu_2 \)

where \( \mu_1 \) and \( \mu_2 \) are the population mean survival time of rats under the Ad libitum and Restricted group, respectively.

2. \( \alpha = 0.02 \)

3. Test statistic:

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

a t-distribution with \( \min\{90 - 1, 106 - 1\} = 89 \) degrees of freedom
4) Compute the test statistic:
\[
    t = \frac{(682.3 - 968.7)}{\sqrt{\frac{134.3^2}{90} + \frac{284.6^2}{106}}} = \frac{-286.4}{31.06} = -9.22
\]

(a) Critical region: two-tail test \( t < -2.374 \) and \( t > 2.374 \) OR
(b) p-value: \( P(t < -9.22) + P(t > 9.22) = 2P(t < -9.22) \)
\[\text{p-value} < 0.01 \text{ (from table)}\]

5) Reject the null since
(a) \(-9.22 < -2.374 \)
OR
(b) p-value < 0.01 < 0.02

6) There is not sufficient evidence to warrant rejection of the claim that the population mean survival time of rats is the same under both experimental groups.
Paired t test

When we want to compare two samples that are dependent, we should do a test for matched pairs. In a matched pairs test, we compute the differences

\[ d_i = x_{1i} - x_{2i} \]

then do a one-sample t-test to see if the population mean difference is zero.

†† Example 3: The National Endowment for the Humanities sponsors summer institutes to improve the skills of high school teachers of foreign languages. One such institute at Purdue University about eight years ago hosted a representative sample of 35 Spanish teachers from American high school for 4 weeks. At the beginning of the period, the teachers were given the Modern Language Association’s listening test of understanding of spoken Spanish. After 4 weeks of immersion in Spanish in and out of class, a different version of the same test was given.
The table below gives a summary of the pre-test and post-test scores (the maximum possible was 36). Is it fair to say that if this program were offered to all U.S. Spanish teachers, their language skills would improve on average (i.e. test the claim that the program improved language skills by increasing test scores)?

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Pre-test ((x_1))</th>
<th>Post-test ((x_2))</th>
<th>Difference ((Post - Pre))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>29</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>30</td>
<td>+2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>35</td>
<td>29</td>
<td>32</td>
<td>+3</td>
</tr>
<tr>
<td>Mean</td>
<td>27.3</td>
<td>28.8</td>
<td>+1.5</td>
</tr>
<tr>
<td>SD</td>
<td>5.0</td>
<td>4.7</td>
<td>3.2</td>
</tr>
</tbody>
</table>
1. $H_0 : \mu_d = 0$ vs $H_1 : \mu_d > 0$ (claim)
where $\mu_d$ is the population difference between the post- and pre-test scores (post - pre).

2. $\alpha = 0.05$

3. Test statistic:

$$t = \frac{\bar{x}_d - \mu_d}{s_d \sqrt{n}}$$

a t-distribution with $35 - 1 = 34$ degrees of freedom

4. Compute the test statistic:

$$t = \frac{1.5}{3.2 \sqrt{35}} = 2.77$$

(a) Critical region: one-tail test $\Rightarrow t > 1.691$

OR

(b) p-value: $P(t > 2.77) = p\text{-value} < 0.005$ (from table)
5) Reject the null since
   (a) $2.77 > 1.691$
   OR
   (b) p-value $< 0.005 < 0.05$

6) The sample data support the claim that the program improved language skills.

   → Construct a 95% Confidence interval for the population mean difference (post - pre) of test scores.

   $\bar{x}_d \pm E$, where $E = t_{\alpha/2} \frac{s_d}{\sqrt{n}}$

   $1.5 \pm 2.032 \frac{3.2}{\sqrt{35}}$

   $(0.401, 2.599)$

   • Zero is NOT in the interval, so we reject the null hypothesis.
★ Note: if we were to do an independent two-sample t-test, we would have...

1. \( H_0 : \mu_1 = \mu_2 \) vs
   \( H_1 : \mu_1 < \mu_2 \)
   where \( \mu_1 \) and \( \mu_2 \) are the population pre and post test scores, respectively.

2. \( \alpha = 0.05 \)

3. Test statistic:

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

a t-distribution with \( 35 - 1 = 34 \) degrees of freedom

4. Compute the test statistic:

\[
t = \frac{28.8 - 27.3}{\sqrt{\frac{4.7^2}{35} + \frac{5.0^2}{35}}} = 1.293
\]

Let’s look at the p-value: \( P(t < 1.293) = \text{p-value} > 0.10 \)

5. Fail to reject

6. There is NOT sufficient sample evidence to support the claim...
• Notice that the p-value was significantly smaller in the matched pairs test. In fact, the change was significant enough to lead us to a different conclusion!!!

• Why? Because we are using more information in the matched pairs test by accounting for the dependency within test scores on the same subject.

• We should only do a matched paired test when there is a reason to match observations in pairs.

• If the sample are dependent, $s_d$ may be smaller that the pooling of $s_1$ and $s_2$. 
Key Concepts!!!!!

- 2 sample Hypothesis Tests: for independent populations for comparing means and proportions
- Paired t-test