Hypothesis testing : Paired Data

• Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be measurements of samples from populations 1 and 2, respectively.
• Assume each $x_i$ is paired with each $y_i$, that is, $x_i$ is not independent of the corresponding $y_i$.
• Let $\mu_1, \mu_2$ be the means of populations 1 and 2, respectively, and let $\mu = \mu_1 - \mu_2$.
• To test $H_0 : \mu = 0$ against one of the three possible $H_a$, first compute the test statistic

$$t = \frac{\bar{d} - 0}{s/\sqrt{n}}, df = n - 1.$$ 

where $\bar{d} =$ the mean of the differences $x_1 - y_1, \ldots, x_n - y_n$, and $s$ is the standard deviation of these differences. Then find (or estimate) the $p$-value for $t$.

Example 1, continued Test the hypothesis that pleasant scents improve the time required to complete a maze. See the Minitab worksheet “May 4 lecture” in the public directory.
Math 223  Section 1  Lecture Notes
May 4, 2004
Section 6.4
Inferences about $\mu_1 - \mu_2$: Paired Data

- Suppose that we have 2 samples from normal populations.
- Also suppose that the data in one sample is “matched” or “paired” with the data in the other sample.

Example 1: Do pleasant odors improve students’ performance on tests? To test this idea, 21 subjects worked a paper-and-pencil maze while wearing a mask. The mask was either unscented or carried a floral scent. The response variable is their average time on three trials. Each subject worked the maze with both masks, in a random order. The randomization is important because subjects tend to improve their times as they work on a maze repeatedly. The table on the next page gives the subjects’ average times with both masks:

**Average time to complete a maze**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Unscented (seconds)</th>
<th>Scented (seconds)</th>
<th>Difference</th>
<th>Subject</th>
<th>Unscented (seconds)</th>
<th>Scented (seconds)</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.60</td>
<td>37.97</td>
<td>-7.37</td>
<td>12</td>
<td>58.93</td>
<td>83.50</td>
<td>-24.57</td>
</tr>
<tr>
<td>2</td>
<td>48.43</td>
<td>51.57</td>
<td>-3.14</td>
<td>13</td>
<td>54.47</td>
<td>38.30</td>
<td>16.17</td>
</tr>
<tr>
<td>3</td>
<td>60.77</td>
<td>56.67</td>
<td>4.10</td>
<td>14</td>
<td>43.53</td>
<td>51.37</td>
<td>-7.84</td>
</tr>
<tr>
<td>4</td>
<td>36.07</td>
<td>40.47</td>
<td>-4.40</td>
<td>15</td>
<td>37.93</td>
<td>29.33</td>
<td>8.60</td>
</tr>
<tr>
<td>5</td>
<td>68.47</td>
<td>49.00</td>
<td>19.47</td>
<td>16</td>
<td>43.50</td>
<td>54.27</td>
<td>-10.77</td>
</tr>
<tr>
<td>6</td>
<td>32.43</td>
<td>43.23</td>
<td>-10.80</td>
<td>17</td>
<td>87.70</td>
<td>62.73</td>
<td>24.97</td>
</tr>
<tr>
<td>7</td>
<td>43.70</td>
<td>44.57</td>
<td>-0.87</td>
<td>18</td>
<td>53.53</td>
<td>58.00</td>
<td>-4.47</td>
</tr>
<tr>
<td>8</td>
<td>37.10</td>
<td>28.40</td>
<td>8.70</td>
<td>19</td>
<td>64.30</td>
<td>52.40</td>
<td>11.90</td>
</tr>
<tr>
<td>9</td>
<td>31.17</td>
<td>28.23</td>
<td>2.94</td>
<td>20</td>
<td>47.37</td>
<td>53.63</td>
<td>-6.26</td>
</tr>
<tr>
<td>10</td>
<td>51.23</td>
<td>68.47</td>
<td>-17.24</td>
<td>21</td>
<td>53.67</td>
<td>47.00</td>
<td>6.67</td>
</tr>
<tr>
<td>11</td>
<td>65.40</td>
<td>51.10</td>
<td>14.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This is an example of paired data.
(more space for example 1)
Confidence intervals for $\mu_1 = \mu_2$: Paired data
A $C\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{d} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}, \ df = n - 1.$$ 

Example 1, continued Find a 90\% confidence interval for $\mu_1 - \mu_2$. 
7.3 Estimation and Tests for Computing Two Population Variances
To carry out inference procedures concerning two population means, we often have to know whether or not \( \sigma_1 = \sigma_2 \).

Testing the hypothesis \( \sigma_1 = \sigma_2 \).
• To test whether or not \( \sigma_1 = \sigma_2 \), we begin with measurements of random samples of sizes \( n_1 \) and \( n_2 \) from two populations.
• The ratio

\[
F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}
\]

is a random variable with an \( F \)-distribution, with \( df_1 = n_1 - 1 \) and \( df_2 = n_2 - 1 \). \( df_1 \) is also called the degrees of freedom in the numerator and \( df_2 \) is called the degrees of freedom in the denominator. This means that the probability distribution function of \( F \) is

\[
f(x) = \frac{\Gamma \left( \frac{\nu_1 + \nu_2}{2} \right) \left( \frac{\nu_1}{\nu_2} \right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2} - 1}}{\Gamma \left( \frac{\nu_1}{2} \right) \Gamma \left( \frac{\nu_2}{2} \right) \left( 1 + \frac{\nu_1 x}{\nu_2} \right)^{\frac{\nu_1 + \nu_2}{2}}}
\]

where

\[
\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} \, dt.
\]

and

\[
\nu_1 = n_1 - 1, \nu_2 = n_2 - 1.
\]

We will never use this formula, it’s just here in case you were curious. The \( F \)-distribution is skewed to the right: it has a shape somewhat like

This complicates the testing procedure. More on this later.

In this class, when testing population variance for equality, the null hypothesis is always

\[
H_0 : \sigma_1^2 = \sigma_2^2.
\]

Now, in a hypothesis test, we always start out by assuming that \( H_0 \) is true. But then, the \( F \)-statistic reduces to

\[
F = \frac{s_1^2}{s_2^2}.
\]

This is the test statistic we will use to test \( H_0 \) against \( H_a \).