Independent Samples

Inference on $\mu_1 - \mu_2$

We learned how to make inference on one population mean, $\mu$, through a confidence interval and then through a hypothesis test. We then learned to make inference on the difference between two population means, $\mu_1 - \mu_2$ when we are in the special case of the paired (dependent) samples model.

We would like to make inference on the difference between two population means, $\mu_1 - \mu_2$ when we have independent samples.

Recall, in the dependent samples setting, we took the difference of the columns of data and worked only with the “difference” column. This collapsed the two sample problem into a one sample problem. In the independent samples case, we cannot employ this same strategy (we might not even have equal sample sizes!).

For the hypothesis test on $\mu_1 - \mu_2$, we’ll need our point estimate and standard error of the point estimate. For the confidence interval, we’ll also need a critical point.

**Point Estimate of $\mu_1 - \mu_2$**

$\bar{Y}_1 - \bar{Y}_2$ is unbiased for $\mu_1 - \mu_2$

**Standard Error of the Point Estimate**

In the special case where we know we have equal population variances ($\sigma_1^2 = \sigma_2^2$), we’ll calculate the $SE(\bar{Y}_1 - \bar{Y}_2)$ a little differently. Recall that in the case where each population can have its own variance, we have the following

**General Case**

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

In the case where we know $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we’ll estimate that common variance by weighting the sample variances by the degrees of freedom available for each column in the following way:

$$\sigma_1^2 = \sigma_2^2 = \sigma^2 \quad SE(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{s_{pool}^2}{n_1} + \frac{s_{pool}^2}{n_2}} \quad \text{where} \quad s_{pool}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$
**Critical Point**

**General Case**
We cannot find the exact degrees of freedom, so we use a highly accurate approximation to degrees of freedom, called the Welsh-Satterthwaite approximation.

\[
df_{ws} = \frac{\left( \frac{SE^2(\bar{Y}_1)}{n_1 - 1} + \frac{SE^2(\bar{Y}_2)}{n_2 - 1} \right)^2}{\frac{SE^4(\bar{Y}_1)}{n_1 - 1} + \frac{SE^4(\bar{Y}_2)}{n_2 - 1}} = \frac{\frac{S^2_1}{n_1} + \frac{S^2_2}{n_2}}{\frac{S^4_1}{n_1(n_1-1)} + \frac{S^4_2}{n_2(n_2-1)}}
\]

- If normality of the sample mean holds for each sample, then this approximation to degrees of freedom yields a highly accurate *approximate* confidence interval.
- If normality does not hold, the central limit theorem may still apply, but sample sizes must grow much larger \((n_1, n_2 \geq 20)\).

**Equal Population Variances** \((\sigma^2_1 = \sigma^2_2 = \sigma^2)\)

\[df = n_1 + n_2 - 2\]

- If normality of the sample mean holds for each sample and \(\sigma^2_1 = \sigma^2_2 = \sigma^2\), then using the standard error with pooled standard deviation yields an *exact* confidence interval.

---

**TI-83/84**

STAT -> TESTS ->2SampTInt (or 2SampTTest for a hypothesis test) -> choose Data or Stats -> Enter appropriate information.

The option for Pooled: No Yes is asking if you want to pool variances (only check yes if you have reason to believe the two populations have equal variances)

**TI-89**

Apps -> Stats/List Editor -> F6 or F7 (according to whether you want a confidence interval or hypothesis test)

The option for Pooled: No Yes is asking if you want to pool variances (only check yes if you have reason to believe the two populations have equal variances)
Example Exercise 7.15 A study was conducted to determine the psychoactive effect of the drug Pargyline on the feeding behavior of the black bowfly, *Phormia regina*. The experimenters used two groups of flies to conduct this experiment – one group injected with Pargyline and a control group injected with saline. The accompanying table summarizes the sucrose consumption (mg in 30 minutes).

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Pargyline</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{Y}$</td>
<td>14.9</td>
<td>46.5</td>
</tr>
<tr>
<td>$s$</td>
<td>5.4</td>
<td>11.7</td>
</tr>
<tr>
<td>$n$</td>
<td>900</td>
<td>905</td>
</tr>
</tbody>
</table>

Construct and interpret a 99% confidence interval for the difference in population means. You may proceed as though the assumptions were checked and deemed acceptable.
Example  A researcher investigated the effect of green light, in comparison to red light on the growth rate of bean plants. The following table shows data on the growth of plants (inches), from the soil to the first branching stem, two weeks after germination.

<table>
<thead>
<tr>
<th></th>
<th>Red</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \bar{Y}=8.36, s=1.5, n=17 )</td>
<td>( \bar{Y}=8.94, s=1.78, n=25 )</td>
</tr>
<tr>
<td></td>
<td>8.4 10.4 8.4 10.0 8.8</td>
<td>8.6 5.9 4.6 9.1 9.8</td>
</tr>
<tr>
<td></td>
<td>7.1 9.4 8.8 4.3 9.0</td>
<td>10.1 6.0 10.4 10.8 9.6</td>
</tr>
<tr>
<td></td>
<td>8.4 7.1 9.6 9.3 8.6</td>
<td>10.5 9.0 8.6 10.5 9.9</td>
</tr>
<tr>
<td></td>
<td>6.1 8.4</td>
<td>11.1 5.5 8.2 8.3 10.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.7 9.8 9.5 11.0 8.0</td>
</tr>
</tbody>
</table>

Use the QQplots to check the assumption of normality.

Test whether there is evidence that the mean growth is different under the green and red light conditions.