8 Comparing Two Independent Populations

We’ll study these methods for comparing two independent populations:

1. The Two-Sample T-Test (Normal with Equal but Unknown Variances)
2. The Welch T-Test (Normal with Unequal Unknown Variances)
3. Bootstrap for Two Samples
4. Permutation Test
5. The Wilcoxon Rank Sum or Mann-Whitney Test
6. Comparing Two Population Proportions

8.1 The Two-Sample T-Test (Normal with Equal but Unknown Variances)

e.g. The horned lizard is named for spikes around the back of its head. The spikes may provide the lizard protection from its primary predator, the loggerhead shrike. Researchers wanted to compare dead lizards killed by shrikes and live lizards from the same area. A SRS was taken from each population. The longest spike was measured on each sampled lizard, in mm. Is there a difference in the length of the longest spike across the two populations? Here are the data:

Live: 23.76, 21.17, 26.13, 20.18, 23.01, 24.84, 19.34, 24.94, 27.14, 25.87, 18.95, 22.61

Start by graphing the data. Here are histograms, dotplots, boxplots, and density plots:
Here are some numerical summaries:

<table>
<thead>
<tr>
<th>Group</th>
<th>Sample Size, $n$</th>
<th>Sample Mean</th>
<th>Sample SD</th>
<th>1st Quartile</th>
<th>Sample Median</th>
<th>3rd Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dead</td>
<td>10</td>
<td>20.79</td>
<td>2.22</td>
<td>19.14</td>
<td>20.59</td>
<td>21.72</td>
</tr>
<tr>
<td>Live</td>
<td>12</td>
<td>23.16</td>
<td>2.76</td>
<td>20.92</td>
<td>23.16</td>
<td>25.17</td>
</tr>
</tbody>
</table>

The summaries all show a right shift. Is it statistically significant, or just the result of random variation? The shift of sample means matters only in relation to the spread of the sample data.

We compare two population means, $\mu_1$ and $\mu_2$, by studying their difference, $\mu_1 - \mu_2$. Notation:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Population 1</th>
<th>Population 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$X_1$</td>
<td>$X_2$</td>
</tr>
<tr>
<td>Variance</td>
<td>$\sigma_1^2$</td>
<td>$\sigma_2^2$</td>
</tr>
<tr>
<td>Sample size</td>
<td>$n_1$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>Sample mean</td>
<td>$\bar{X}_1$</td>
<td>$\bar{X}_2$</td>
</tr>
<tr>
<td>Sample variance</td>
<td>$s_1^2$</td>
<td>$s_2^2$</td>
</tr>
</tbody>
</table>

For inference about $\mu_1 - \mu_2$, use the statistic $\bar{X}_1 - \bar{X}_2$, and then...
• test $H_0: \mu_1 - \mu_2 = \delta_0$ ($\delta_0 = 0 \implies \mu_1 = \mu_2$)
• find a confidence interval for $\mu_1 - \mu_2$

To do this, we need the distribution of $\bar{X}_1 - \bar{X}_2$. Recall for independent $X$ and $Y$:

- $E(X - Y) = E(X + (-1)Y) = E(X) + E(-1Y) = E(X) - E(Y)$
- $VAR(X - Y) = VAR(X + (-1)Y) = VAR(X) + VAR(-1Y) = VAR(X) + (-1)^2 VAR(Y)$
- $\sigma^2_{\bar{X}} = VAR(\bar{X}) = \frac{\sigma^2}{n}$
- If $X \sim N(\mu_X, \sigma^2_X)$ and $Y \sim N(\mu_Y, \sigma^2_Y)$, then $X - Y \sim N(\mu_X - \mu_Y, \sigma^2_X + \sigma^2_Y)$
- For normal $X$ and $Y$, $\bar{X} \sim N(\mu_X, \sigma^2_{\bar{X}})$

It follows that, for normal populations 1 and 2, $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2})$.

But we don’t know $\sigma_1$ or $\sigma_2$. Here we’ll assume they are equal and calculate a pooled variance estimate as

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

Now we can state a test and a confidence interval:

• Many hypothesis tests use test statistics of the form

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sim t_{n_1 + n_2 - 2}}$$

• Recall that many confidence intervals have the form

$$(\text{point estimate}) \pm (\text{margin of error}) = (\text{point estimate}) \pm (\text{table value for confidence}) \times [(\text{estimated or true}) \text{ standard deviation of point estimate}]$$
Our confidence 100%(1 – α) confidence interval for $\mu_1 - \mu_2$, assuming normal populations and equal population variances, is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{(n_1 + n_2 - 2, \alpha/2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

e.g. Test that the lizard populations have the same population spike lengths and find a confidence interval for the difference in lengths.

Check normality:

Or take deviations from the means and combine them into one sample:

Check equal variances from plots (above).

Calculate test and interval: ...

$$S_p^2 = \frac{(10-1)2.2^2 + (12-1)2.76^2}{10+12-2}$$

$$= 6.407$$

Test: $H_0: \mu_{\text{live}} - \mu_{\text{dead}} = 0$

$H_A: \mu_{\text{live}} - \mu_{\text{dead}} > 0$
\[ t = \frac{(23.16 - 20.79) - 0}{\sqrt{6.407} \sqrt{\frac{1}{12} + \frac{1}{10}} } \]

\[ = 0.864 \cdot 2.19 \]

\[ p \text{-value} = P(T \geq \frac{2.19}{2.19}) \]

\[ = P(T_{20} > 0.864) \]

\[ = \text{between 0.10 and 0.25 between 0.01 and 0.025 < 0.05} \]

Conclusion:
Reject H0: The data are strong evidence that the population mean spike length is greater in living horned lizards than in (shrike-eaten) dead lizards.

95% Interval for \( \mu_{\text{live}} - \mu_{\text{dead}} \) is

\[ (23.16 - 20.79) \pm t_{20,0.025} \sqrt{6.407} \sqrt{\frac{1}{12} + \frac{1}{10}} \]

\[ = 2.37 \pm 2.086 \cdot 1.08 \]

\[ = 2.37 \pm 2.25 \]