Permutation Test
for the Two Sample Problem

• we wish to compare results for two groups of experimental units

• the first group could be some subjects who have been given a treatment, whereas the second group has not

• in some cases we are unable to assume that
  – the two samples of sizes \( n_1 \) and \( n_2 \) are from normal populations and/or
  – the populations have the same variance

• however we may be able to assume that the groups were obtained by randomly splitting the subjects \( n = n_1 + n_2 \) into two groups

• with only this assumption, we are able to base the test on the permutation distribution, described below

• the hypotheses are

\[
H_0 : \text{no effect of the treatment} \\
H_a : \text{there is an effect}
\]

• a reasonable test statistic is

\[
T = \bar{X}_1 - \bar{X}_2
\]

which measures the effect of the treatment

• if \( H_0 \) is true the observed differences in the data are due only to variation among the subjects

• with a different random allocation of subjects, a different value for \( T \) would be obtained
• there are exactly
\[
\binom{n_1 + n_2}{n_1} = \frac{(n_1 + n_2)!}{n_1! n_2!}
\]
ways of randomly allocating \(n_1\) of the subjects to group 1 and the remaining \(n_2\) to group 2

• each of these is equally likely, and each can lead to a different value of the test statistic \(T\)

• the permutation distribution describes the possible values for \(T\) for all possible allocations of the subjects

• the P value is the fraction of values for \(T\) which are as least as extreme as contrary to the null hypothesis as is the observed value \(T_{obs}\)

• for a one-sided alternative the P value is the proportion in one tail of the permutation distribution

• for a two-sided alternative the P value is double the probability in one tail of the permutation distribution

• If the alternative is that the population 2 measurements are smaller than in population 1, and if the test statistic is \(T = \bar{X}_1 - \bar{X}_2\), then the p-value is the proportion of possible values of \(T\) which are at least as large as \(T_{obs}\). (If your test statistic was \(T = \bar{X}_2 - \bar{X}_1\) then the p-value would be the proportion of possible values of \(T\) which are at least as small as \(T_{obs}\).)

• If the alternative is that the population 2 measurements are greater than in population 1, and if the test statistic is \(T = \bar{X}_1 - \bar{X}_2\), then the p-value is the proportion of possible values of \(T\) which are at least as small as \(T_{obs}\). (If your test statistic was \(T = \bar{X}_2 - \bar{X}_1\) then the p-value would be the proportion of possible values of \(T\) which are at least as large as \(T_{obs}\).)
• If the alternative is two sided - that the distribution in the two populations are different, then the test statistic is \( T = |\bar{X}_1 - \bar{X}_2| \), and the p-value is the proportion of possible values of \( T \) which are at least as large as \( T_{obs} \).

Example: A simple study has only \( n_1 = n_2 = 3 \) subjects in each group

\[
\begin{array}{ccc}
\text{Treatment} & 175 & 250 & 260 & x_1 = 228.33 \\
\text{Control} & 255 & 275 & 300 & x_2 = 276.67 \\
\end{array}
\]

Two of the three largest smallest observations are in the treatment group, so it looks as though the treatment may be effective. What is the p-value?

• the test statistic is \( T = 228.33 - 276.67 = -48.33 \)

• there are only \( \binom{3+3}{3} = 20 \) possible allocations of subjects to the two groups

• these are shown in the table below, along with the value for \( T \)
|   |   |   |   |   |   | 175 | 250 | 255 | 260 | 275 | 300 |  \( \bar{X}_1 - \bar{X}_2 \) |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|   |   |   |   |   |   |     | 51.67 | 38.33 | 21.67 | 18.33 | 35 | 5 |     |  |  |  |
| 1 | 1 | 2 | 1 | 2 | 2 | -51.67 | 51.67 | 48.33 | 21.67 | 18.33 | 35 | 5 |     |  |  |  |
| 1 | 2 | 1 | 2 | 2 | 1 | -51.67 | 51.67 | 48.33 | 21.67 | 18.33 | 35 | 5 |     |  |  |  |
| 1 | 2 | 1 | 1 | 2 | 2 | -48.33 | 38.33 | 35 | 5 |     |  |  |  |  |  |  |
| 1 | 2 | 1 | 1 | 2 | 2 | -48.33 | 38.33 | 35 | 5 |     |  |  |  |  |  |  |
| 1 | 2 | 1 | 1 | 2 | 1 | -21.67 | 21.67 | 18.33 | 15 | 5 |     |  |  |  |  |  |
| 2 | 1 | 1 | 2 | 2 | 2 | 18.33 | 18.33 | 0 | 5 |     |  |  |  |  |  |  |
| 2 | 1 | 1 | 2 | 2 | 2 | 18.33 | 18.33 | 0 | 5 |     |  |  |  |  |  |  |
| 2 | 1 | 1 | 2 | 2 | 1 | -51.67 | 51.67 | 48.33 | 31.67 | 5 |     |  |  |  |  |  |
| 2 | 2 | 1 | 1 | 2 | 2 | 21.67 | 21.67 | 18.33 | 15 | 5 |     |  |  |  |  |  |
| 2 | 2 | 1 | 1 | 2 | 1 | 38.33 | 38.33 | 0 | 5 |     |  |  |  |  |  |  |
| 2 | 2 | 1 | 1 | 2 | 1 | 38.33 | 38.33 | 0 | 5 |     |  |  |  |  |  |  |
| 2 | 2 | 1 | 1 | 1 | 1 | 51.67 | 51.67 | 48.33 | 35 | 5 |     |  |  |  |  |  |
| 2 | 2 | 1 | 1 | 1 | 1 | 51.67 | 51.67 | 48.33 | 35 | 5 |     |  |  |  |  |  |
| 2 | 2 | 1 | 1 | 1 | 1 | 51.67 | 51.67 | 48.33 | 35 | 5 |     |  |  |  |  |  |
| 2 | 2 | 1 | 1 | 1 | 1 | 51.67 | 51.67 | 48.33 | 35 | 5 |     |  |  |  |  |  |

- For the one sided alternative (treatment leads to smaller observations), \( T_{\text{obs}} = -48.33 \), and there is 1 possible sample (the configuration [1,1,1,2,2,2]) which provides greater evidence against the null hypothesis than \( T_{\text{obs}} \). Therefore, the p-value is 2/20 = .1.

- For the two sided alternative (unspecified difference between treatment and control), \( T_{\text{obs}} = 48.33 \), and there are 4 samples which provide at least as much evidence against \( H_0 \) than does \( T_{\text{obs}} \), and so the p-value is 4/20 = .2.
Example: The data below is from the example of soil surface pH which was used to illustrate the (pooled) two sample \( t \) test.

<table>
<thead>
<tr>
<th>Location 1</th>
<th>8.53</th>
<th>8.52</th>
<th>8.01</th>
<th>7.99</th>
<th>7.93</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.89</td>
<td>7.85</td>
<td>7.82</td>
<td>7.80</td>
<td></td>
</tr>
<tr>
<td>Location 2</td>
<td>7.85</td>
<td>7.73</td>
<td>7.58</td>
<td>7.40</td>
<td>7.35</td>
</tr>
<tr>
<td></td>
<td>7.30</td>
<td>7.27</td>
<td>7.27</td>
<td>7.23</td>
<td></td>
</tr>
</tbody>
</table>

- the test statistic is

\[
T_{obs} = 8.038 - 7.442 = .596
\]

- note that only one value (7.85) from Location 2 is larger than two of the values from Location 1

- exchanging this value with one of the smaller values in Location 1 increases the mean for Location 1 and decreases the mean for Location 2, giving a larger \( T = \bar{X}_1 - \bar{X}_2 \)

- the same value for \( T_{obs} \) is obtained if the value 7.85 from Location 2 is switched with the value 7.85 from Location 1

- so there are 4 permutations (including the original data) for which \( T \) is as large or larger than \( T_{obs} \), and 8 permutations for which \( T \) is as extreme or more extreme

- there are

\[
\binom{18}{9} = \frac{18!}{9!9!} = 48620
\]

permutations in total
• if we test the hypotheses

\[ H_0 : \text{no difference between locations} \]
\[ H_a : \text{there is a difference} \]

using the permutation test, the P value is \( P = 8/48620 = .0001645 \)

• so there is very strong evidence of a difference in the mean surface soil pH at the two locations

• this is consistent with the result obtained earlier using the \( t \) distribution, which requires the assumptions of normality and equal variances

• in this example we are fortunate that it is straightforward to determine how extreme \( T_{\text{obs}} \) is relative to the permutation distribution

• it would be difficult to list all 48620 possible permutations

• one approach in this situation is to approximate the permutation distribution using random permutations chosen by the computer

• 50,000 such permutations give the following histogram, for this example
50,000 randomly chosen permutations
• one can see that there are very few values of $T$ beyond $T_{obs}$
• the computer found 5 cases as extreme or more extreme
• the approximate P value using this approach is $P = \frac{5}{50000} = .0001$
• this is quite close to the exact value