Math 373  Lecture 37  
Comparing several ranked populations 

Suppose you have a dataset of nonnegative integers which sum to a fixed total. For example \{2,2\}, \{1,3\} and \{0,4\} all total to 4. Then the sum of the squares of the integers measures how spread out the set is. \(\{2,2\} \rightarrow 2^2+2^2 = 4+4 = 8\). \(\{1,3\} \rightarrow 1^2+3^2 = 1+9 = 10\). \(\{0,4\} \rightarrow 0^2+4^2 = 0+16 = 16\). \(\{2,2\} \rightarrow 8\) has the least variation and the smallest sum of squares. \(\{0,4\} \rightarrow 16\) has the most variation and the largest sum of squares.

In Lecture 26 we had \(k\) normal quantitative populations. Each population represented a particular treatment. We used the statistic \(F = \text{MST}/\text{MSE}\) to test for a significant difference between any two of the \(k\) populations.

Lecture 26 didn’t apply to highly nonnormal or to nonquantitative populations. Lecture 34 showed how to test for a difference between two populations which could be ranked in increasing order. Here we test for a significant difference between some members of a family of \(k\) populations which can be ranked in increasing order.

We wish to determine if some of the \(k\) populations differ significantly from some of the others. We could test each pair of populations for a significant difference between their means. But if there are 11 populations, the number of possible pairs is \(k(k-1)/2 = 11\times10/2 = 55\). The following method is more efficient and more accurate.

Given \(k\) populations \(P_1, P_2, \ldots, P_k\), choose \(k\) samples, one from each population with \(n_1, n_2, \ldots, n_k\) elements respectively. Each sample should have \(\geq 5\) elements. Let \(n = n_1 + n_2 + \ldots + n_k\) be the total of all sample elements.

Rank the elements of the \(k\) samples together in one long list in ascending order. As in Lecture 33, assign ranks: 1, 2, 3, 4, \ldots to the first, second, third, \ldots item in the list. If several items are equal, average their ranks.

Label elements of \(P_1\) with the number 1, label elements of \(P_2\) with 2, \ldots, elements of \(P_k\) with \(k\). Let \(T_1\) be the total of the ranks of the elements in \(P_1\), \(T_2\) the total of the ranks of elements in \(P_2\), \ldots.

By the argument above, the sum \(T_1^2 + T_2^2 + \ldots + T_k^2\) of the squares should measure how spread out the populations are. However \(T_i\) could be large due to having many elements rather than because it has large elements. \(\{1, 1, 1\} \rightarrow 1^2+1^2+1^2 = 3\) and \(\{2\} \rightarrow 2^2 = 4\) have the same sums but the first has the smaller elements, just more of them. To compensate for the varying number of elements in the samples, divide each square by the number of elements in its sample. This gives the sum \(\sum \frac{T_i^2}{n_i}\).

This sum measures what we want but there is no table for its distribution. However multiplying by the right coefficient and adding the right term gives the following \(H\) statistic which has the \(\chi^2\) distribution.

**Test for a Difference Among Ranked Populations.**

**Definition.** The Kruskal Wallis \(H\) statistic is

\[
H = \frac{12}{nk(n+1)} \left( T_1^2/n_1 + T_2^2/n_2 + \ldots + T_k^2/n_k \right) - 3(n+1)
\]

**Theorem.** \(H\) has the \(\chi^2\) distribution with \(df = k - 1\) degrees of freedom. Table 5 has the critical values for \(\chi^2\).

- You have three species of dogs: \(P_1, P_2, P_3\). You wish to determine if there is a significant difference between the weights of any of the three species. Suppose you take a sample from each of the three populations and record the weights of the dogs in each sample: Three dogs from \(P_1\) with weights \(\{30, 35, 45\}\), four dogs from \(P_2\) with weights \(\{35, 35, 45, 45\}\), and 3 dogs from \(P_3\) with weights \(\{30, 45, 45\}\). There should be at least 5 elements in each sample but we want simple examples here, for homework, and for the final.

Put all the dogs in one list ordered by increasing weight. Above each weight we mark the number of the population from which it came. The first 30 lb. dog came from \(P_1\) and is marked 1; the second 30 lb. dog came from \(P_3\) and is marked 3.

Below each weight, list the rank. 1 for the first, 2 for the second, \ldots. As in Lecture 33, if several items have the same weight, replace their rank with the average of the ranks with that weight.

<table>
<thead>
<tr>
<th>Weight</th>
<th>(P_1)</th>
<th>(P_2)</th>
<th>(P_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>35</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Since the first two items have the same weight, both are assigned the average \(1.5\) of ranks 1 and 2.

**\(H_0: \text{Some difference.} \ H_0: \text{no difference between } P_1, P_2, P_3.\)**

**\(H_0: H > 0\)**  \(H_0: H \leq 0\).

\(T_1 = 1.5 + 4.5 + 8.5 = 14.5\)
\(T_2 = 3\times4.5 + 8.5 = 22\)
\(T_3 = 1.5 + 2\times8.5 = 18.5\) \hspace{1cm} \(n = 10\)

The sum of all ranks is \(n(n+1)/2 = 10(11)/5 = 55\).

Check that the \(T_i\)’s total to this number. \(T_1 + T_2 + T_3 = 55\).

\[
H = \frac{12}{10(11)} \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \ldots + \frac{T_k^2}{n_k} \right) - 3(n+1) = .2909
\]

\(df = k - 1 = 2\)

Acceptance region for \(H\): \([0, \chi^2_{0.05}] = [0, 5.9915]\).

\(H = .2909 \in [0, 5.9915]\).  \(\therefore\) accept \(H_0\).

None of the three species differ significantly when ranked by weight.