10.4 The Kruskal–Wallis One-Way ANOVA Test and Multiple Comparisons

The median test for k samples uses information about the magnitude of each of the N observations relative to a single number, which is the median of the pooled samples. Many popular nonparametric k-sample tests use more of the available information by considering the relative magnitude of each observation when compared with every other observation. This comparison is effected in terms of ranks.

Since under H_0 we have essentially a single sample of size N from the common population, combine the N observations into a single ordered sequence from smallest to largest, keeping track of which observation is from which sample, and assign the ranks $1, 2, \ldots, N$ to the sequence. If adjacent ranks are well distributed among the k samples, which would be true for a random sample from a single population, the total sum of ranks, $\sum_{i=1}^{N} i = N(N+1)/2$, would be divided proportionally according to sample size among the k samples. For the ith sample, which contains n_i observations, the expected sum of ranks would be

$$\frac{n_i}{N}\frac{N(N+1)}{2} = \frac{n_i(N+1)}{2}$$

Equivalently, we can argue that since under H_0 the expected rank for any observation is the average rank $(N+1)/2$, the expected sum of ranks for n_i observations is $n_i(N+1)/2$. Denote the actual sum of ranks assigned to the elements in the ith sample by R_i. A reasonable test statistic could be based on a function of the deviations between these observed and expected rank sums. Since deviations in either direction indicate disparity between the samples and absolute values are not particularly tractable mathematically, the sum of squares of these deviations can be used as

$$S = \sum_{i=1}^{k}\left[R_i - \frac{n_i(N+1)}{2}\right]^2 \tag{10.4.1}$$

The null hypothesis is rejected for large values of S.

In order to determine the null distribution of S, consider the ranked sample data recorded in a table with k columns, where the entries in the ith column are the n_i ranks assigned to the elements in the ith sample. Then R_i is the ith-column sum. Under H_0, the integers $1, 2, \ldots, N$ are assigned at random to the k columns except for the restriction that there be n_i integers in column i. The total number of ways to make the assignment of ranks then is

the number of partitions of N distinct elements into k ordered sets, the ith of size n_i, and this is

$$\frac{N!}{\prod_{i=1}^{k} n_i!}$$

Each of these possibilities must be enumerated and the value of S calculated for each. If $t(s)$ denotes the number of assignments with the particular value s calculated from (10.4.1), then

$$f_S(s) = t(s) \prod_{i=1}^{k} \frac{n_i!}{N!}$$

Obviously, the calculations required are tedious and will not be illustrated here. Tables of exact probabilities for S are available in Rijkoort (1952) for $k = 3$, 4, and 5, but only for n_i equal and very small. Critical values for some larger equal sample sizes are also given there.

A somewhat more useful test criterion is a weighted sum of squares of deviations, with the reciprocals of the respective sample sizes used as weights. This test statistic, due to Kruskal and Wallis (1952), is defined as

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left[R_i - \frac{n_i(N+1)}{2} \right]^2 \tag{10.4.2}$$

The consistency of H is investigated in Kruskal (1952). H and S are equivalent test criteria only for all n_i equal. Exact probabilities for H are given in Table K for $k = 3$, all $n_i \leq 5$. The tables in Iman et al. (1975) also cover $k = 4$, all $n_i \leq 4$ and $k = 5$, all $n_i \leq 3$ for the upper 10% of the exact distribution.

Since there are practical limitations on the range of tables that can be constructed, some reasonable approximation to the null distribution is required if a test based on S or H is to be useful in application.

Under the null hypothesis, the n_i entries in column i were randomly selected from the set $\{1, 2, \ldots, N\}$. They actually constitute a random sample of size n_i drawn without replacement from the finite population consisting of the first N integers. The mean and variance of this population are

$$\mu = \sum_{i=1}^{N} \frac{i}{N} = \frac{N+1}{2}$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{[i - (N+1)/2]^2}{N} = \frac{N^2 - 1}{12}$$

The average rank sum for the ith column, $\overline{R}_i = R_i/n_i$, is the mean of this random sample, and as for any sample mean from a finite population,

$$E(\overline{R}_i) = \mu \quad \text{var}(\overline{R}_i) = \frac{\sigma^2(N - n_i)}{n_i(N - 1)}$$

Here then, under H_0, we have

$$E(\overline{R}_i) = \frac{N + 1}{2} \quad \text{var}(\overline{R}_i) = \frac{(N + 1)(N - n_i)}{12n_i}$$
$$\text{cov}(\overline{R}_i, \overline{R}_j) = -\frac{N + 1}{12}$$

Since $\overline{R}_i$ is a sample mean, if n_i is large, the central limit theorem allows us to approximate the distribution of

$$Z_i = \frac{\overline{R}_i - (N + 1)/2}{\sqrt{(N + 1)(N - n_i)/12n_i}} \tag{10.4.3}$$

by the standard normal. Consequently Z_i^2 is distributed approximately as chi square with one degree of freedom. This holds for $i = 1, 2, \ldots, k$, but the Z_i are clearly not independent random variables since $\sum_{i=1}^{k} n_i\overline{R}_i = N(N + 1)/2$, a constant. Kruskal (1952) showed that under H_0, if no n_i is very small, the random variable

$$\sum_{i=1}^{k} \frac{N - n_i}{N} Z_i^2 = \sum_{i=1}^{k} \frac{12n_i[\overline{R}_i - (N + 1)/2]^2}{N(N + 1)} = H \tag{10.4.4}$$

is distributed approximately as chi square with $k - 1$ degrees of freedom. The approximate size α rejection region is $H \geq \chi^2_{\alpha,k-1}$. Some other approximations to the null distribution of H are discussed in Alexander and Quade (1968) and Iman and Davenport (1976). Andrews (1954) discusses the power of this test.

Under the assumption that the populations are continuous, we do not have to deal with the problem of ties. However, ties can occur in practice. When two or more observations are tied within a column, the value of H is the same regardless of the method used to resolve the ties since the rank sum is not affected. When ties occur across columns, the midrank method is generally used. Alternatively, the ties can be broken in the way that is least conducive to rejection of H_0 for a conservative test.

If ties to the extent t are present and are handled by the midrank method, the variance of the finite population is

$$\sigma^2 = \frac{N^2 - 1}{12} - \frac{\sum t(t^2 - 1)}{12}$$

where the sum is over all sets of ties in the population, and this expression should be used in $\text{var}(\overline{R}_i)$ for the denominator of Z_i. In this case (10.4.4) becomes

$$\sum_{i=1}^{k} \frac{N - n_i}{N} \left\{ \frac{\left[\overline{R}_i - \frac{N(N+1)}{2}\right]^2}{\frac{(N+1)(N-n_i)}{12n_i} - \frac{N-n_i}{n_i(N-1)} \frac{\sum t(t^2-1)}{12}} \right\}$$

$$= \sum_{i=1}^{k} \frac{12n_i \left[\overline{R}_i - \frac{N(N+1)}{2}\right]^2}{N(N+1) - \frac{N \sum t(t^2-1)}{N-1}} = \frac{H}{1 - \frac{\sum t(t^2-1)}{N(N^2-1)}} \tag{10.4.5}$$

Hence the correction for ties is simply to divide H in (10.4.2) by the correction factor $1 - \sum t(t^2 - 1)/N(N^2 - 1)$ where the sum is over all sets of t tied ranks. The details are left as an exercise for the reader.

When the null hypothesis is rejected, we can compare any two groups, say i and j (with $1 \leq i < j \leq k$), by a *multiple comparisons procedure*. We calculate

$$Z_{ij} = \frac{|\overline{R}_i - \overline{R}_j|}{\sqrt{[N(N+1)/12](1/n_i + 1/n_j)}} \tag{10.4.6}$$

and compare it to the $[\alpha/k(k-1)]$st upper standard normal quantile $z^* = z_{\alpha/[k(k-1)]}$. If Z_{ij} exceeds z^*, the two groups are declared to be significantly different. The quantity α is called the *experimentwise error rate* or the *overall significance level*, which is the probability of at least one erroneous rejection among the $k(k-1)/2$ pairwise comparisons. Typically, one takes $\alpha \geq 0.20$ because we are making such a large number of statements. We note that $1 - \alpha$ is the probability that all of the statements are correct. It is not necessary to make all possible comparisons, although we usually do. For convenience, we give the z^* values to three decimal places for a total of $d = k(k-1)/2$ comparisons at $\alpha = 0.20$ as follows:

d	1	2	3	4	5	6	7	8	9	10
z^*	1.282	1.645	1.834	1.960	2.054	2.128	2.189	2.241	2.287	2.326

This multiple comparisons procedure is due to Dunn (1964).

10.4.1 Applications

The Kruskal–Wallis test is the natural extension of the two-sample Wilcoxon test for location to the case of k mutually independent samples from continuous populations. The null hypothesis is that the k populations are the same, but when we assume the location model, this hypothesis can be written in terms of the respective location parameters (or treatment effects) as

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$$
$$H_1 : \text{At least two } \theta\text{'s differ}$$

To perform the test, all $n_1 + n_2 + \cdots + n_k = N$ observations are pooled into a single array and ranked from 1 to N. The test statistic H in (10.4.2) is easier to calculate in the form

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{10.4.7}$$

where R_i is the sum of the ranks from the ith sample. The appropriate rejection region is large values of H. The critical values or P values are found from Table K for $k = 3$, each $n_i \leq 5$. This statistic is asymptotically chi-square distributed with $k - 1$ degrees of freedom; the approximation is generally satisfactory except when $k = 3$ and the sample sizes are five or less. Therefore, Table B can be used when Table K cannot. When there are ties, we divide H by the correction factor, as shown in (10.4.5).

For multiple comparisons, using (10.4.6), we declare treatments i and j to be significantly different in effect if

$$|\overline{R}_i - \overline{R}_j| \geq z^* \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{10.4.8}$$

If $n_i = n_j = N/k$ for all i and j, the right-hand side of (10.4.8) reduces to $z^*\sqrt{k(N+1)/6}$.

Example 10.4.1

For the experiment described in Example 10.2.2, use the Kruskal–Wallis test to see if there is any difference between the medians of the four groups.

SOLUTION

The data are already ranked from 1 to 40 in Table 10.2.1 so we need only calculate the rank sums as $R_1 = 260, R_2 = 122, R_3 = 90, R_4 = 348$. With $n_1 = n_2 = n_3 = n_4 = 10$, we get

$$H = \frac{12}{40(41)(10)}[260^2 + 122^2 + 90^2 + 348^2] - 3(41) = 31.89$$

with 3 degrees of freedom. The P value from Table B is $P < 0.001$, so we reject the null hypothesis that the four medians are the same and do a follow-up analysis by a multiple comparisons test using $\alpha = 0.20$. We have $\overline{R}_1 = 26.0, \overline{R}_2 = 12.2, \overline{R}_3 = 9.0$, and $\overline{R}_4 = 34.8$ and the right-hand side of (10.4.8) is 11.125. The treatments, which have significantly different medians, are 1 and 2, 1 and 3, 2 and 4, and 3 and 4.

The computer solutions to Example 10.4.1 are shown below using the MINITAB, SAS, and STATXACT packages. All of the results for H agree exactly.

```
**********************************
MINITAB SOLUTION TO EXAMPLE 10.4.1
**********************************

   Kruskal-Wallis test: C2 versus C1

      Kruskal-Wallis test: on C2

C1         N     Median     Ave Rank          Z
1         10     25.500         26.0       1.72
2         10     12.500         12.2      -2.59
3         10      8.500          9.0      -3.59
4         10     34.500         34.8       4.47
Overall   40                    20.5
H = 31.89  df = 3  P = 0.000
```

MINITAB shows the value of the test statistics as $H = 31.89$ and the asymptotic P value of 0.000 based on the chi-square approximation with 3 degree of freedom. If there had been ties in the data, MINITAB would have shown H(adjusted), which is calculated from (10.4.5). MINITAB also shows the median, average rank, and Z value for each group. The Z values given are calculated from (10.4.3). This is the standardized value of the deviation between the mean rank $\overline{R}_i$ for the ith group and its expected value $(N + 1)/2$ under the null hypothesis. The sign of the Z statistic indicates whether the mean rank is larger or smaller than expected, and the magnitude measures the relative deviation. The largest absolute Z value is 4.47, which indicates that the mean rank for group 4, which is 34.8, differs from the average rank of 20.5 more than that of any other group. And the smallest absolute Z value, 1.72, shows that the average for group 1, 26.0, differs from the average rank less than that of any other group.

Now we show the program code and the results for SAS and STATXACT.

```
                ******************************
                SAS SOLUTION TO EXAMPLE 10.4.1
                ******************************
Program:

data a;
input group N;
do i = 1 to N;
input battery @@;
output;
end;
cards;
1  10
19  22  25  24  29  26  37  23  27  28
2  10
14  21  2  6  10  16  17  11  18  7
3  10
12  1  5  8  4  13  9  15  3  20
4  10
39  39  40  30  31  32  33  36  34  35
;
proc npar1way wilcoxon;
class group;
run;

Output
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Battery
Classified by Variable Group

Group	−*N*	Sum of Scores	Expected Under H0	Std.-dev. under H0	Mean Score
1	10	260.0	205.0	32.014119	26.00
2	10	122.0	205.0	32.014119	12.20
3	10	90.0	205.0	32.014119	9.00
4	10	348.0	205.0	32.014119	34.80

Average scores were used for ties.

Kruskal–Wallis test

Chi-square	31.8967
df	3
Pr > chi-square	< .0001

```
************************************
STATXACT SOLUTION TO EXAMPLE 10.4.1
************************************

KRUSKAL-WALLIS TEST [That the 4 populations are identically
distributed]

Statistic based on the observed data:

     T(X) = The observed test statistic =        31.89

Asymptotic P value: (based on chi-square distribution with 3 df)
     Pr  { T(X) .GE.       31.89 } =        0.0000

Monte Carlo estimate of P value :
     Pr  {  Statistic  .GE.    31.89  } =     0.0000
     99.00% Confidence interval = (      0.0000,      0.0005)
```

Example 10.4.2

For the experiment described in Example 10.2.1, use the Kruskal–Wallis test to see if there is any difference between the medians of the three groups.

Solution

The first step is to rank the data from 1 to 15, as shown below, where rank 1 is given to the smallest score, which indicates the most effective result.

	Squeaker	Wrist Tie	Chin Strap
	6	15	2
	9	13	3
	10	11	4
	12	14	1
	5	7	8
Sum	42	60	18

We calculate $\sum R^2/n = 5688/5 = 1137.6$ and $H = 12(1137.6)/15(16) - 3(16) = 8.88$. Table K for $k = 3, n_1 = n_2 = n_3 = 5$ shows that $0.001 < P$ value < 0.010, so the null hypothesis of equal treatment effects should be rejected. It appears that the chin strap is the most effective device in reducing snoring since it has the smallest sum of ranks. Since the null hypothesis was rejected, we carry out a multiple comparisons procedure at the 0.20 level. We have $z^* = 1.834$ for $d = 3$ and the right-hand side of (10.4.8) is 5.19. The sample mean ranks are $\overline{R}_1 = 8.4, \overline{R}_2 = 12, \overline{R}_3 = 3.6$. Our conclusion is that only groups 2 and 3 have significantly different treatment effects at the overall 0.20 significance level. Recall that our hand calculations did not lead to a rejection of the null hypothesis by the median test in Example 10.2.1.

The computer solutions to Example 10.4.2 are shown below using the SAS, STATXACT, and MINITAB packages. The results for the value of H agree exactly. The P value using the chi-square approximation is 0.012, which agrees with the outputs. Note that both STATXACT and SAS allow the user the option of computing what they call an exact P value based on the permutation distribution of the Kruskal–Wallis statistic. This can be very useful when the sample sizes are small so that the chi-square approximation could be suspect. However, the exact computation is time consuming even for moderate sample sizes, such as 10 as in Example 10.4.1. For this example, SAS finds this exact P value to be 0.0042, which agrees with our conclusion from Table K. MINITAB does not have an option to calculate an exact P value and it does not provide the correction for ties.

```
*********************************
SAS SOLUTION TO EXAMPLE 10.4.2
*********************************

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Snore
Classified by Variable Group

                 Sum of      Expected      Std.-dev.       Mean
Group      N     Scores      Under H0      Under H0      Score
1          5       42.0          40.0      8.164966       8.40
2          5       60.0          40.0      8.164966      12.00
3          5       18.0          40.0      8.164966       3.60

                Kruskal-Wallis test

        Chi-square                        8.8800
        df                                     2
        Asymptotic  Pr >   chi-square     0.0118
        Exact       Pr >= chi-square     0.0042

  ************************************
  STATXACT SOLUTION TO EXAMPLE 10.4.2
  ************************************

KRUSKAL-WALLIS  TEST  [That  the  three  populations  are
identically distributed]

Statistic based on the observed data:

     T(X) = The Observed test Statistic =    8.880

Asymptotic P value: (based on chi-square distribution with 2 df )
     Pr  {  T(X)  .GE.    8.880  } =    0.0118

Exact P value and point probability:
     Pr  {  Statistic  .GE.    8.880  } =    0.0042
     Pr  {  Statistic  .GE.    8.880  } =    0.0003
```