Analysis Method: kruswal [Kru-52]

Brian T. Luke (lukeb@ncifcrf.gov)

The Kruskal-Wallis one-way analysis of variance uses the rank of the intensities in a feature to remove the requirement of a normal distribution. All intensities are ranked from lowest to highest so that $K_{n,l}$ is the ranking of subject n for the feature l. The first step is to determine the average ranking across all samples in state s.

$$\overline{K}_{s,l} = \frac{1}{n_s} \sum_{m \in Es} K_{m,l}$$

If $\overline{K}_l$ is the average ranking across all N samples [$\overline{K}_l = (N+1)/2$] then the statistic is given by the following.

$$K_l = \frac{12 \sum_{s=1}^{S} n_s \left(\overline{K}_{s,l} - \overline{K}_l \right)^2}{N(N+1)}$$

The features should be ranked from highest to lowest value of K_l.

Note that when only two States are present in the dataset (i.e. Cases and Controls) the final ordering of the features is the same as for a Mann-Whitney (Wilcoxon) U-test. Since this procedure is also applicable for datasets with more than two States, it is selected.

The results examining 10,000 features representing either Feature-a or Feature-b, and comparing their scores against the maximum possible score obtained from features with no information is shown in the following table.

Eack	Thresh	10a	10b	15a	15b	20a	20b	25a	25b	30a	30b	35a	35b	40a	40b
30	17.26	1	1	2	3	6	5	25	15	82	40	218	97	451	195
45	17.85	4	2	10	5	29	16	80	63	298	183	761	338	1430	725
60	15.5	8	10	55	23	179	134	490	332	1210	819	2597	1586	4395	2801
90	13.64	50	39	227	196	765	641	2027	1612	4198	3113	6538	4981	8548	7101
150	13.58	115	121	700	617	2414	1975	5225	4375	8059	6935	9528	8850	9903	9647
300	16.56	236	224	1902	1554	5628	4859	8993	8274	9905	9690	9997	9981	10000	9999

As stated earlier, the first column represents the number of Cases and the number of Controls in each dataset. The second column represents the maximum K_l obtained from 10,000 features where the intensities for both Cases and Controls are randomly assigned within the range of 0.0 to 100.0. The remaining columns show the number of times in 10,000 randomly generated feature intensities that a feature has a K_l that is above this threshold. The headings for these column show whether the features represent Feature-a

or Feature-b, described previously, and the value of Za or 2Zb. For example, the column labeled 10a is for features that represent Feature-a with Za=10, while the column labeled 10b is for features that represent Feature-b with 2Zb=10 (Zb=5).

This procedure produces nearly comparable results for features represented by Feature-a and Feature-b if there are a large number of samples. In the largest dataset (300 Cases and 300 Controls) a feature of type Feature-a with Za=20 (the range of one State is only 80% as high as the other) has a 56.3% chance of having a K_l value that is higher than is found in 10,000 features with identical ranges. A comparable feature representing Feature-b has a 48.6% chance of having a K_l value above the threshold. As the size of the dataset decreases, the values of Za and 2Zb must increase in order to have a 50% chance of having a higher value of K_l, and the value of Za does not need to be as large as 2Zb. In the smallest dataset (30 Cases and 30 Controls), Za must be at least 65 for a feature of type Feature-a to exceed a 50% probability of having a sufficiently large value of K_l. For these features, the intensity range for one State can be at most 35% as large as the other and 32.5% of all samples have intensity in a range that is not covered by samples in the other State. If the feature is of type Feature-b, 2Zb must be at least 80 for the probability that its K_l value is greater than the threshold to exceed 50%. For these features, the shift in ranges must be at least 40% of the range and 40% of all samples will have intensity in a range that is not covered by samples in the other State.

(Last updated 4/4/07)