Analysis Method: vartest

Brian T. Luke (lukeb@ncifcrf.gov)

This procedure, based on the relevance index proposed by Yip and coworkers [Yip-03], attempts to find those features where the variance of intensities for each category of samples is smallest relative to the total variance of all intensities. If s_i^2 is the estimated variance of the samples in category *i*, and s_T^2 is the estimated variance for all samples, they are defined by the following formulas.

$$s_i^2 = \sum_{k \in i} (I_k - \bar{I}_i)^2 \Big/ (N_i - 1)$$

$$s_T^2 = \sum_{k=1}^{N} (I_k - \bar{I})^2 \Big/ (N - 1)$$

In the top equation I_k is the intensity of the k^{th} sample in category *i*, $\bar{I}_k$ is the average intensity for all samples in this category and N_i is the number of samples in this category. The second equation sums over all N samples and $\bar{I}$ is the average intensity for this feature. The *vartest* score for a given feature, *V*, containing *C* categories is given by the following.

$$V = \frac{\sum_{i=1}^{C} s_i^2}{s_T^2}$$

The features are then ranked from lowest to highest *V* since the objective it to find the feature with the smallest intra-category variance. In many instances with equal numbers of Cases and Controls, this procedure yields an identical ordering of features as found using *student*.

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.

Each	Thresh	10a	10b	15a	15b	20a	20b	25a	25b	30a	30b	35a	35b	40a	40b
30	1.5699	15	16	35	28	85	72	247	169	606	322	1277	605	2435	1030
45	1.6193	2	4	13	12	49	28	175	88	510	239	1349	532	2883	1138
60	1.7698	22	22	112	75	390	258	1137	664	2640	1480	5028	2777	7460	4327
90	1.7686	3	4	18	15	145	80	622	265	2164	954	4885	2310	7893	4461
150	1.8802	33	25	281	177	1481	915	4537	2805	8040	5591	9700	8098	9982	9491
300	1.9251	560	461	3463	2522	8104	6613	9866	9338	9997	9949	10000	10000	10000	10000

As stated earlier, the first column represents the number of Cases and the number of Controls in each dataset. The second column represents the minimum value of *V* 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 value of *V* that is below 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 recognizes putative biomarkers represented by Feature-a slightly better than those for Feature-b. For datasets with 300 cases and 300 controls, approximately 81% of the features with Za=20 produced a lower *V* value than any observed feature with no information. In contrast, if 2Zb=20, 66.1% of the features had lower *V* values. As with the other methods examined, the ability to identify a weak putative biomarker is much better if the dataset contains more samples. If there are only 30 Cases and 30 Controls and the features have the form of Feature-a, there is at least a 50% chance of having a *V* value lower than 1.5699 if Za=50, meaning that the range of intensities for one State is only 50% that of the other. As the number of Cases and Controls increases from 45 to 150, the range of the smaller intensity State increases from 55% to 85% of the range of the larger intensity State. If the features have the form of Feature-b, the region of overlap increases from 70% to 85% as the number of Cases and Controls increases from 30 to 150.

(Last updated 4/29/07)