Analysis Method: student (previously named *fimp* [Hab-05])

Brian T. Luke (lukeb@ncifcrf.gov)

This procedure calculates the z-score for two independent sets of samples under the null hypothesis that their intensities are part of the same distribution. Therefore, this method is only valid for a two-State examination. If there are n_1 Cases and n_2 Controls, and the average of the intensities in each State are I_1 and I_2, respectively, the unsigned z-score is determined from the following equation.

$$z_l = | I_1 - I_2 | \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

In this equation, σ_1^2 and σ_2^2 are the variances of the intensities in the Cases and Controls, respectively.

$$\sigma_1^2 \approx \frac{\sum^{Cases} (I_n - I_1)^2}{(n_1 - 1)}$$

$$\sigma_2^2 \approx \frac{\sum^{Controls} (I_n - I_2)^2}{(n_2 - 1)}$$

The equations on the right represent the unbiased estimate of the variance for a sample of intensities in each State. The features are then ordered from highest to lowest values of z_l.

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	4.569	3	1	11	5	27	16	67	46	184	95	457	217	1055	387
45	3.869	28	12	79	58	232	180	656	401	1611	893	3237	1681	5390	2849
60	3.851	33	24	128	85	450	290	1251	774	2934	1587	5306	2956	7738	4807
90	3.870	50	45	237	176	1003	662	2783	1742	5666	3506	8286	5690	9606	7835
150	4.300	38	33	322	191	1560	990	4732	2829	8206	5818	9702	8289	9984	9587
300	4.096	293	237	2518	1855	7259	5600	9719	8817	9994	9862	10000	9998	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 maximum value of the z-

score 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 the z-score 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 recognizes putative biomarkers represented by Feature-a slightly better than those for Feature-b. For datasets with 300 cases and 300 controls, approximately 72.6% of the features with Za=20 produced a higher z-score than any observed feature with no information. In contrast, if 2Zb=20, only 56% of the features had higher z-scores. 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 z-score greater than 4.569 if Za=55, meaning that the range of intensities for one State is only 45% 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 60% to 70% of the range of the larger intensity State. If the features have the form of Feature-b, the region of overlap increases from 65% to 90% as the number of Cases and Controls increases from 30 to 300.

(Last updated 4/4/07)