Return to Contents PDF Version

Analysis Method: kolsmir

Brian T. Luke (lukeb@ncifcrf.gov)

The Kolmogorov-Smirnov test (K-S test) simply measures the maximum difference in the cumulative fraction plots of the two States.

$F_1(x)$ is the fraction of samples in the first State that have an intensity less than *x*, while $F_2(x)$ is the fraction of samples in the second State. The metric is simply given by the following.

$D = \max|F_1(x) – F_2(x)|$ over all *x*.

The peaks should be ranked from highest to lowest value of *D*.

In general, the intensities are ranked from lowest to highest and x assumes each value of the intensity (meaning that the point with this intensity does not count towards the total). Increment the total in each state starting with the sample with the lowest intensity and stopping with the sample with the next-to-highest intensity and find the largest difference after each increment.

This method is only valid for a two-State problem.

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	17	0	0	3	5	7	2	15	11	66	11	206	31	563	44
45	20	6	4	9	8	59	16	175	41	572	101	1675	178	3545	337
60	26	0	0	1	1	12	3	75	20	401	30	1451	76	3687	159
90	29	11	9	47	22	293	61	1545	171	4469	421	7811	969	9572	1941
150	39	12	4	142	30	1232	140	5365	485	9130	1253	9942	2697	9999	5052
300	53	155	62	2829	358	9137	1491	9994	3968	10000	7598	10000	9592	10000	9983

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 *D* 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 *D* 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 better than those for Feature-b. For datasets with 300 cases and 300 controls, approximately 91.4% of the features with Za=20 produced a higher *D* value than any observed feature with no information. In contrast, if 2Zb=20, only 14.9% of the features had higher *D* 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 *D* value greater than 17 if Za=60, meaning that the range of intensities for one State is only 40% 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 75% of the range of the larger intensity State. If the features have the form of Feature-b, the region of overlap increases from 52.5% to 85% as the number of Cases and Controls increases from 30 to 300.

(Last updated 4/29/07)