Analysis Method: chisq

Brian T. Luke (lukeb@ncifcrf.gov)

This procedure can be used for any number of states. The full intensity range over all N samples is divided into R regions. $O_{r,s}$ is the number of observed samples in State *s* having an intensity in region *r*, and is one cell in an *R*x*S* tableau. The expected value for this cell, $E_{r,s}$, is the number of samples in the row times the number of samples in the column divided by the total number of samples.

$$E_{r,s} = \left(\sum_{s=1}^{S} O_{r,s}\right)\left(\sum_{r=1}^{R} O_{r,s}\right)\Big/ N$$

An approximate X^2 value measuring the difference between observed and expected is

$$X_l^2 = \sum_{s=1}^{S}\sum_{r=1}^{R}\left\{\left(O_{r,s} - E_{r,s}\right)^2 \Big/ E_{r,s}\right\}$$

This is for (*R*-1)(*S*-1) degrees of freedom, but since all features will have the same values for *R* and *S*, they can be ordered from highest to lowest X_l^2.

In general, the expected number of samples must be five or more, and this can be used to determine the number of regions *R*. If State *s* contains N_s samples, this always represents the sum of this column in determining the expected number of samples. If the State with the smallest number of samples has N_s^m samples, we have the following.

$$\left(\sum_{s=1}^{S} O_{r,s}\right) N_s^m / N \geq 5$$

Therefore a region is determined once the total number of samples from any State in this intensity region is at least $5N/N_s^m$. The number of regions, *R*, is then the integer that is at most $N_s^m/5$.

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	27.2	0	0	2	0	2	0	13	2	27	8	58	19	167	31
45	28.8	6	3	29	10	77	36	208	78	675	219	1753	507	3481	867
60	34.4	14	4	46	16	148	90	308	237	1472	522	3044	959	5060	1944
90	44.0	36	22	163	94	620	277	1932	850	4503	1918	7449	3925	9237	6147

150	68.0	16	18	185	66	1073	408	3708	1658	7312	4034	9413	6912	9951	8929
300	104.8	516	300	3618	2221	8302	6438	9924	9333	9998	9962	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 maximum value of X_l^2 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 X_l^2 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).

For larger datasets, the results are quite similar if the feature has type Feature-a or Feature-b. This is caused by the larger number of regions and a comparable disparity between the number of Cases and Controls in the upper regions for Feature-a and the upper and lower regions for Feature-b. As the number of samples decreases, the number of regions decreases, and therefore the range of intensity within each region increases. There still would be a reasonable disparity between the Cases and Controls in the upper region(s) for Feature-a, but the disparity is reduced for the upper and lower region(s) of Feature-b. Therefore, the ability to generate a sufficiently large X_l^2 value is easier for features of type Feature-a than Feature-b.

For the largest dataset (300 Cases and 300 Controls), X_l^2 exceeds the maximum value obtained for features with no information at least 50% of the time until Za=20 for type Feature-a (one State has an intensity range that is 80% of the other), or until 2Zb=20 for Feature-b (90% overlap of the intensity ranges). In either case 10% of all samples have intensity in a range that is not covered by samples in the other State. For the smallest dataset (30 Cases and 30 Controls), a sufficiently large X_l^2 is not obtained for type Feature-a until Za=65 (the range of one State is only 35% of the range of the other), or 2Zb=100 for type Feature-b (there is only a 50% overlap of the intensity ranges). For type Feature-a, 32.5% of all samples have intensity in a range that is not covered by samples in the other State, while for type Feature-b this increases to 50% of all samples.

(Last updated 4/4/07)