From: Lars Holden
Date: 21.10.2009

MP - US, same distribution of marks in the segments in US

Tracks

1. Track 1: marked points
2. Track 2: unmarked segments

We assume the marks are ordered categorical or not ordered categorical.

Question

Is the distribution of the marks inside each segment in track 2 the same?

Comment:

- We assume the position of the points in track 1 and segments in track 2 are fixed. We permute only the marks of the points in track 1.
- Significance is determined by means of p-values. Small p-values identify bins where the marks of the segments in track 1 are not independent of the marks of the segments of track 2.
- We may extend the question to also include the distribution of marks among the points that are outside the segments.
- The p-values are found by an analytic calculation or MC simulation.

Bins

The genome (or the areas of the genome under study) are divided into small regions, called bins. The tests are performed in each bin.

Hypothesis tested

For each bin i we have the null hypothesis
$\mathbf{H}_0$: *The mark of the points in track 1 has the same distribution for all segments in track 2.*

The alternative hypothesis is:
$\mathbf{H}_1$: *The distribution of the mark of the points in track 1 depends on the segment in track 2.*

Statistics and rejection of the null hypothesis, categorical variables

Let r be the number of categories for marks of the points in track 1 and let c be the number of segments in track 2 in the bin. Furthermore, let $O_{i,j}$ be the number of observations of points in track 1 with mark equal i in segment j in track 2 in the bin. The table with the $O_{i,j}$ values is a contingency table with r rows and c columns.

Let N be the total number of points inside segments, i.e. $N = \sum_{i=1}^{r} \sum_{j=1}^{c} O_{i,j}$. If the marks of the pairs are independent, we expect $O_{i,j} \approx E_{i,j}$ where

$$E_{i,j} = \frac{1}{N} \sum_{k=1}^{r} O_{k,j} \sum_{k=1}^{c} O_{i,k}.$$

Let

$$X = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

Under the null hypothesis X is χ^2-distributed with $(r-1)(c-1)$ degrees of freedom. This is an approximation that is considered accurate if all $O_{i,j} > 10$. (ref. Wikipedia/Pearson's chi-square test). We find the p-value from this distribution.

The combinations of i and j that give the largest contribution to the double sum in X, are the cells where the contribution for rejecting the hypothesis is largest. The combinations of i and j where $\frac{|O_{i,j} - E_{i,j}|}{E_{i,j}}$ is largest, is an estimate for where the deviation from same probability is largest.