UP - MS, Located in highly marked segments

Tracks

- Track 1: Unmarked points
- Track 2: Segments with an attached variable/mark

Remark: Mark of Track 2 assumed to be real numbers or an ordered categorical variable (including the binomial case).

Questions

Is there within the considered bin a correlation between mark values of track 2 and the number of points in the segments of track 1?

Simple model: Segments either of equal length or the segment length is unimportant

We then have a set of pairs (number of points, mark value), and we may use correlation tests. Specific assumptions on the relations between the number of points and the mark value will normally be difficult to establish, and we thus use the well-known non-parametric test based on Kendall's tau. Note that the null hypothesis of no correlation may be rejected either due to some relation between the marks and points or due to factors affecting both, e.g. both number of points and mark values being systematically high in certain areas within the bin.

Alternative model: For fixed mark value, the number of points is approximately proportional to segment length.

Assume that the values of the mark are real numbers, and define:

X_i: Number of points in segment i.

Y_i: Value of variable in segment i.

L_i: Length of segment i.

Choose model (e.g. based on graphical displays):

Model 1: $X_i/L_i = \alpha + \beta Y_i$

Model 2: $\ln(X_i + \epsilon)/L_i = \alpha + \beta Y_i$

Model 3: $\ln(X_i + \epsilon)/L_i = \alpha + \beta \ln Y_i + \epsilon$

The selected model is tested by ordinary regression. Note that extension to more than one mark/variable is simple, as is the use of nominal variables (using general linear models).