Written by: Lars Holden.
Date: 25.10.2009

UP , uniform positioning of points within the bin and between bins

Tracks

1. Track 1: unmarked points

Questions

Q-1 Where in the genome are the points in track 1 positioned more towards the borders of the bin than expected by chance?

Q-2 Where in the genome are the points in track 1 positioned more towards the middle of the bin than expected by chance?

Q-3 Where in the genome are the points in track 1 positioned more towards the left part of the bin than expected by chance?

Q-4 Where in the genome are the points in track 1 positioned more towards the right part of the bin than expected by chance?

Q-5 Where in the genome are the points in track 1 positioned more non-uniformly inside the bin than expected by chance?

Q-6 Is the number of point in the bin larger/smaller/different than expected from the total number of points in the genome?

Comments:

- In question 1-5 we assume the points as independently, uniformly distributed in the bin and in question 6 independently, uniformly distributed in the genome.
- Significance is determined by means of p-values. For Q-1, small p-values identify regions where the points in track 1 are closer to the borders of the bin than expected. Similar for Q-2, Q-3, Q-4 and Q-5 and for Q-6 if the number of points in the bin are too many/few than expected.

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

H_0: Points have a uniform distribution in the bin.
For question 5, the hypothesis is

H_0: Points have a uniform distribution in the genome.

Remarks

- Question 1-5 is similar to a UP, US question if the entire bin is inside a segment.

Alternative hypotheses

$\mathbf{H}_1$: Points tend to be positioned towards the borders of the bin.

$\mathbf{H}_2$: Points tend to be positioned towards the middle of the bin.

$\mathbf{H}_3$: Points tend to be positioned towards the left part of the bin.

$\mathbf{H}_4$: Points tend to be positioned towards the right part of the bin.

$\mathbf{H}_5$: Points are unequally distributed within bin.

$\mathbf{H}_5$: Points are unequally distributed between the bins.

Testing against $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$ and $\mathbf{H}_4$

When testing against H_1 and H_2, let d_i, $1 = 1, \ldots n$, be the relative position, but now scaled such that the value is -1 at both borders and 1 in the middle of the bin (and thus 0 halfway between the middle and the border).

When testing against H_3 and H_4, let d_i, $1 = 1 \ldots n$, be the relative position of points within the bin scaled such that the value is -1 at the left end and 1 at the right end.

To test the first four hypotheses above, we may use the Wilcoxon sign-rank test. For n larger than 20-30, we may also use the t-test, which is markedly less time-consuming.

The Wilcoxon test is done in the following way: Rank the d_i without regard to sign; with 1 assigned to the observation closest to 0 (any zeros are neglected). Then compute $W+$ and $W-$ as the sums of the value of the ranks of the originally positive and negative observations, respectively. Significance levels are based on the fact that if H_0 is true, then there are $2n$ equally likely ways for the n ranks to receive signs. As test statistic, we use $W = MIN(W-, W+)$. For small samples ($N \leq 30$), the critical regions must be found from some table. For $N > 30$, the test statistic W approaches a normal distribution with a mean of $n(n+1)/4$ and a variance of $n(n+1)(2n+1)/24$. However, to increase speed, we should consider using the t-test when $n > 20$. The t-test to use is the standard one-sample test.

Testing H_5

To test against the alternative H_5, one may use the Kolmogorov test.

Testing H_6

To test against the alternative H_6, one may use a binomial test where n is the total number of points in the bin and $p = n * B/L$ where B in number of base pairs in the bin and T is number of based pairs in the genome.

Remark

The alternatives are formulated such that a one-sided test may appear most appropriate, except for H_5 and H_6.