**Daphna Joel, Amir Averbuch and Isaac Meilijson, Tel Aviv University, May 2018**

EXPLANATION of the Matlab code that estimates gender disparity as a function of cluster size.

This program has been written by us for our own use, with no attempts to optimize anything not make user-friendly. The following description make help to understand what it does.

Rows 8 to 23 construct data pdt and stdp, the raw estimated mean and standard deviation of the proportion p (a random variable), as a function of cluster size. The proportion p is defined as the maximal between proportion of male and female in the cluster.

Rows 25 to 50 are a regularized smooth version of the estimators.

The input XX to analysiscluster is a matrix with three columns: cluster size, number male, number female in cluster.

XX2 is a 4-column matrix, where the first three are XX and the fourth is the maximum of the two entries in the second and third columns of XX.

datain is a 4-column matrix, where the first three are XX and the fourth is the statistic with which work is performed, a normalized version of the female-male disparity.

datain(:,4)=(1/4)\*((datain(:,3)-datain(:,2)).^2./datain(:,1)-1)./(datain(:,1)-1);

The data [datain(:,1), datain(:,4)] is sorted in increasing order of the cluster size datain(:,1) and submitted to equal-weight symmetric moving average wddt of the second column and its empirical standard deviation sddt in these windows.

Implemented theoretical calculations lead to estimated mean of p as pdt and estimated standard deviation of p as stdp, obtained in rows 22 and 23, displayed in Figures 1 and 2. Their regularized smooth version dd of pdt is displayed in Figure 3, together with plus and minus one (smooth) standard deviation cc.

Figure 3 is the main output of the program.