

Tests for Evaluating Rank Histograms from Ensemble Forecasts

Ian Jolliffe
University of Exeter
i.t.jolliffe@ex.ac.uk

Cristina Primo
ECMWF

Outline

- Rank histograms
 - Definition
 - Testing for Uniformity
- Elmore's artificial example
- Chi-squared goodness-of-fit test
- Alternatives to the chi-squared test
 - Cramér-von Mises tests
 - Decomposition of the chi-squared statistic
- Examples
 - Elmore again
 - 500hPa heights
- Final Remarks

Rank Histograms (aka Talagrand diagrams)

- Consider an ensemble of $(k-1)$ members forecasting a variable X . These are used to divide the overall range of values for X into k bins.
- A verifying observation will then fall into one of these k bins.
- Suppose now we have n such ensemble forecasts and corresponding observations. These are used to construct a histogram with k bins.

Rank Histograms – Testing for Uniformity

- If the ensemble members and the verifying observation all come from the same probability distribution (desirable), then the probability of the verifying observation falling into a particular bin is the same for all bins.
- Thus the rank histogram should be roughly ‘flat’ or uniform.
- Because of sampling variation it won’t be exactly flat – we wish to test whether any deviations from ‘flatness’ are large enough that they unlikely to have arisen by chance.

Elmore's artificial example

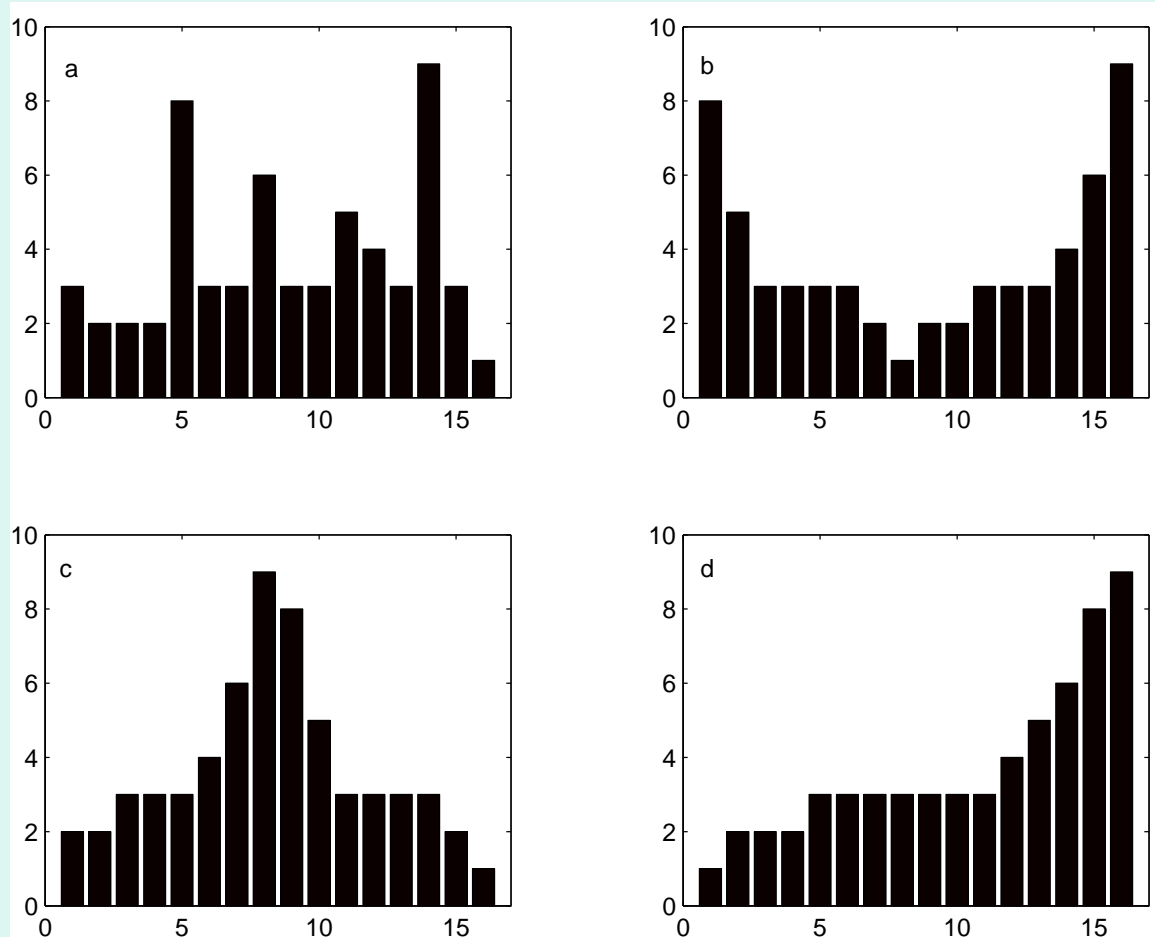
16 bins
60 observations

Bottom right – bias/trend

Top right – under-dispersion

Bottom left – over-dispersion

Top left – roughly flat??

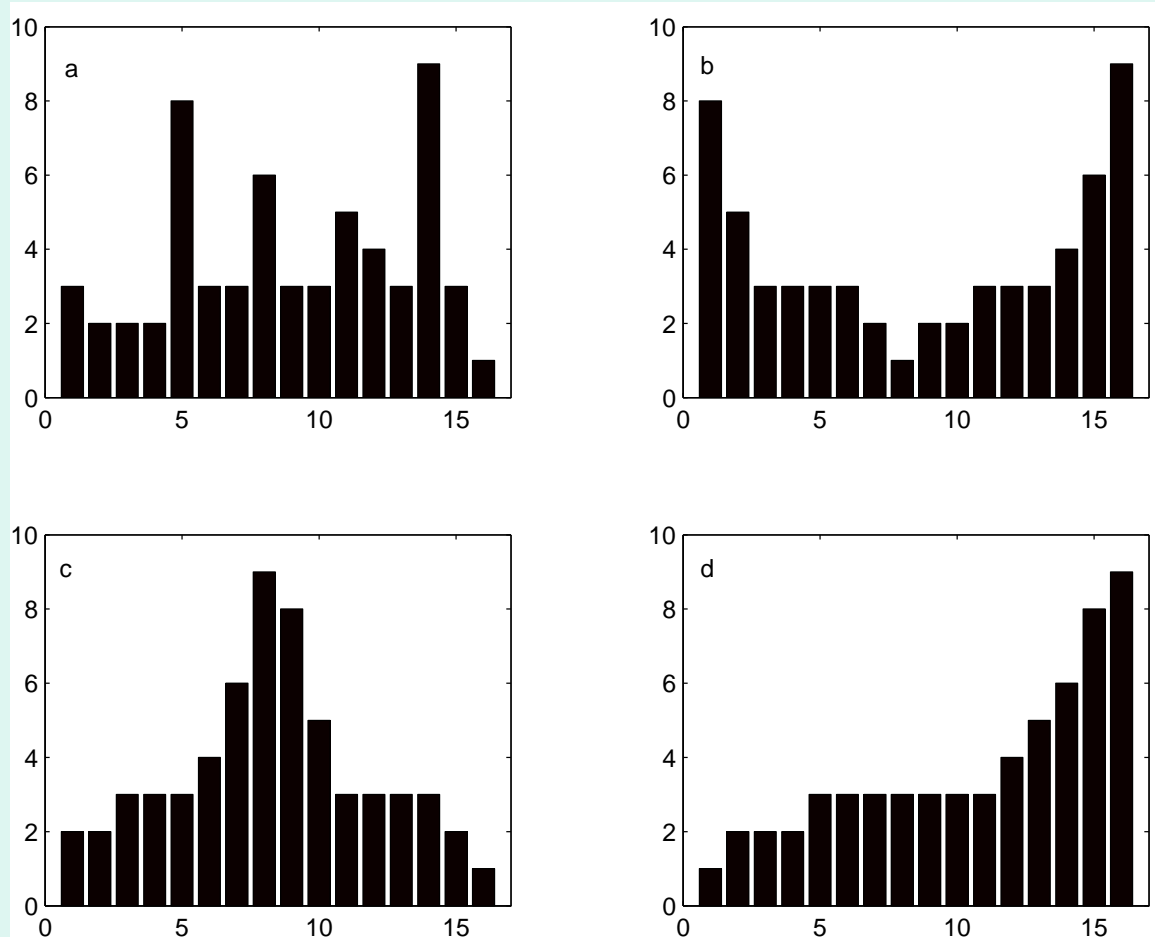


Elmore K L (2005) Weather & Forecasting, 20, 789-795

10IMSC August 2007

Elmore's artificial example II

In fact the data in the top-left panel were generated randomly so that the ensemble members and verifying observations are from the same distribution. Hence deviations from flatness are due to chance. The other three panels have the same bin frequencies, but rearranged in a way that deviations from flatness appear unlikely to have arisen by chance.



The χ^2 goodness-of-fit test

- The best-known general test that data come from a particular distribution has test statistic

$$T = \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}$$

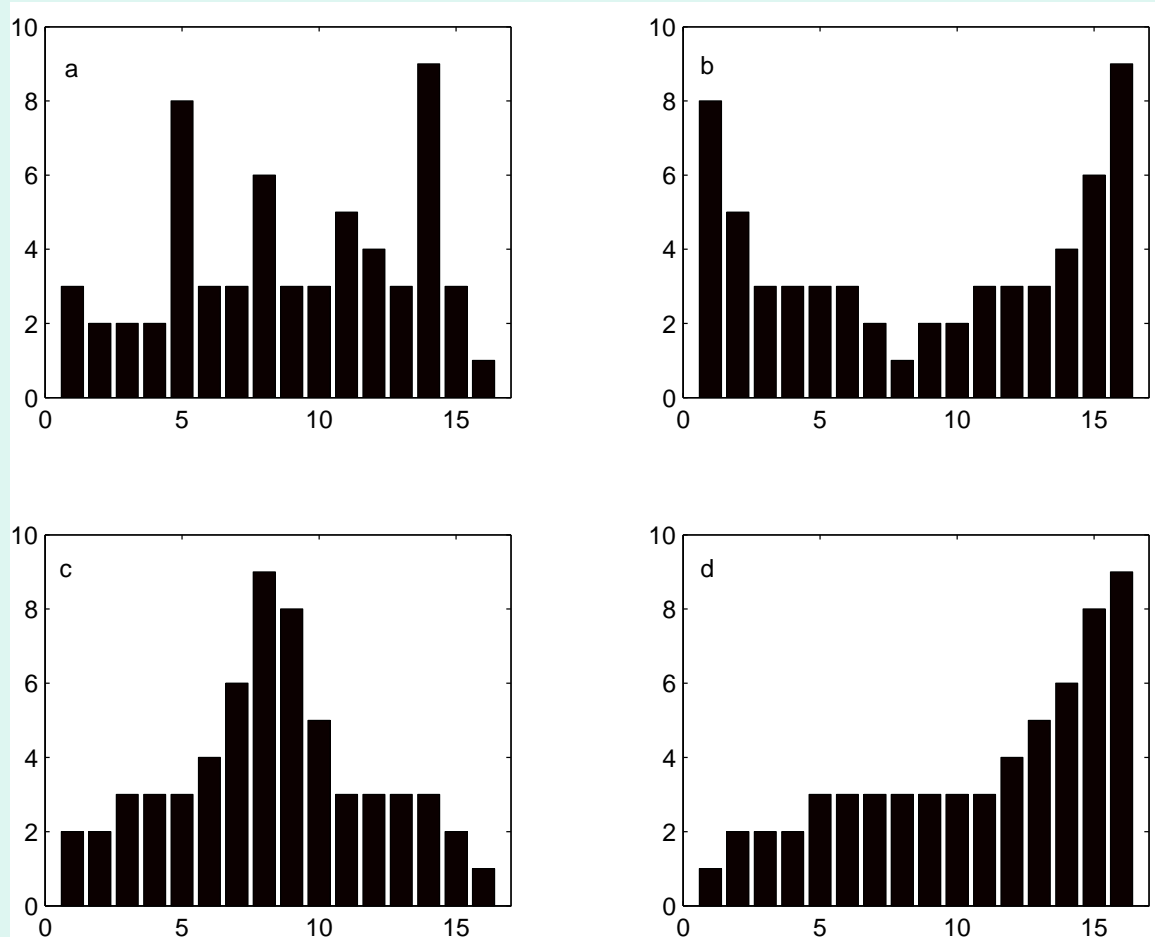
- Here n_i , e_i are the observed and expected (given the hypothesised distribution) number of observations in the i^{th} bin.
- For the uniform distribution, $e_i = n/k$
- Under the null hypothesis that the hypothesised distribution is correct, T has a χ^2 distribution with $(k-1)$ degrees of freedom.

The χ^2 goodness-of-fit test II

- The χ^2 is a good general test – it has some power to detect *all* types of deviation from the null hypothesis (NH).
- However because it spreads its power widely (and thus thinly) it is not very good at detecting specific alternatives to the NH.

Elmore's artificial example III

For the top-left panel, $T=19.467$. Comparing this to χ^2_{15} gives a p-value of 0.193 – the deviations from ‘flatness’ could have easily arisen by chance (as indeed they did). This seems plausible, but T has the same value for the other three panels, leading to the same conclusion, which seems a lot less plausible.



Alternatives to χ^2 (Cramér-von Mises)

- Elmore suggests using members of a family of Cramér–von Mises tests. These are based on comparing the *cumulative* distribution of the observation with that of the hypothesised distribution.
- They have the advantage that they take the order of the bins into account, and are more powerful than χ^2 at detecting alternatives such as those in Elmore's example (numbers later).
- One disadvantage is that they need special tables to assess their 'significance'

Alternatives (decomposing χ^2)

- The overall χ^2 statistic T can be decomposed into the sum of $(k-1)$ terms, each term having (approximately) independent χ^2_1 distributions under the null hypothesis of uniformity.
- There are restrictions on the way the decomposition is done, but by suitable choices we can isolate terms corresponding to bias/trend, over/under-dispersion etc.

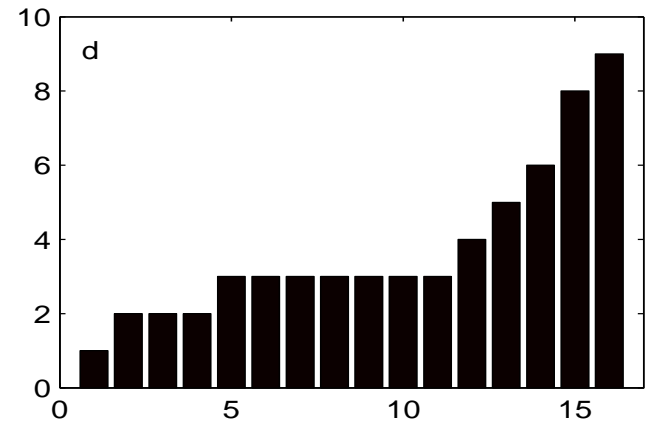
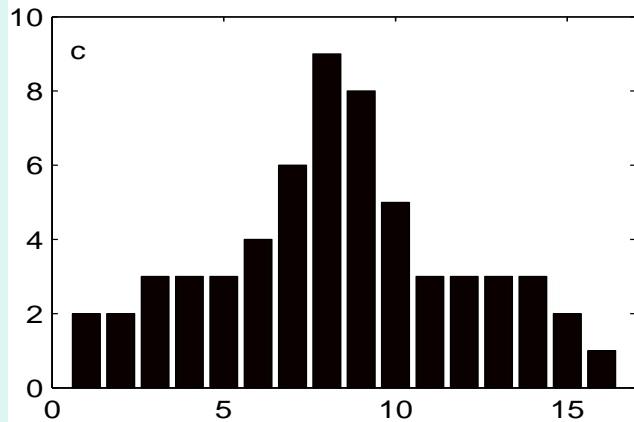
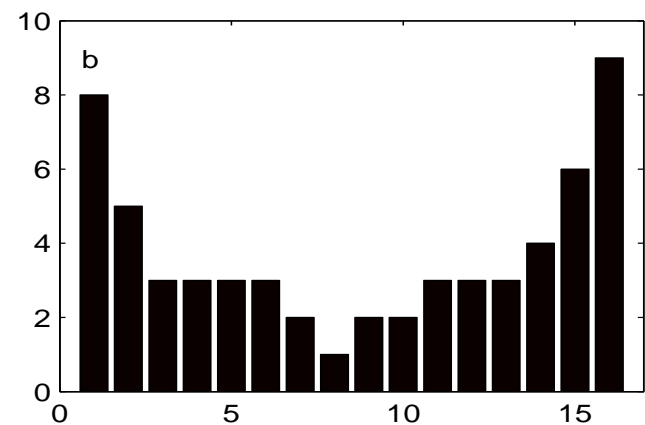
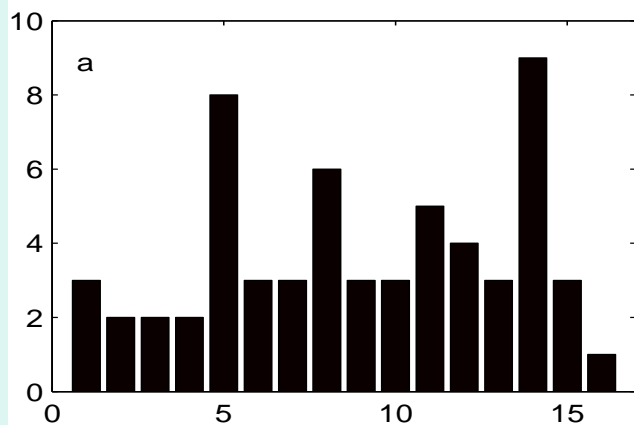
Alternatives (decomposing χ^2 II)

- We need not find $(k-1)$ individual terms. For example if trend/bias and over/under dispersion are the only deviations from uniformity of interest, we can decompose T into 3 components, one each for the deviations of interest, each with 1 degree of freedom and a third component representing all other deviations, with 3 degrees of freedom.

Alternatives (decomposing χ^2 III)

- In the results that follow, we label various 1 degree of freedom components as:
 - Linear (= bias/trend)
 - Ends (contrasts end categories with all others)
 - V-shape (represents a different sort of over/under dispersion to Ends)
- Resid_1, Resid_2, represent the $(k-1)$ degree of freedom terms after removing Linear + Ends and Linear + V-shape respectively.
- Cramér-von Mises gives the smallest p-value found by Elmore.

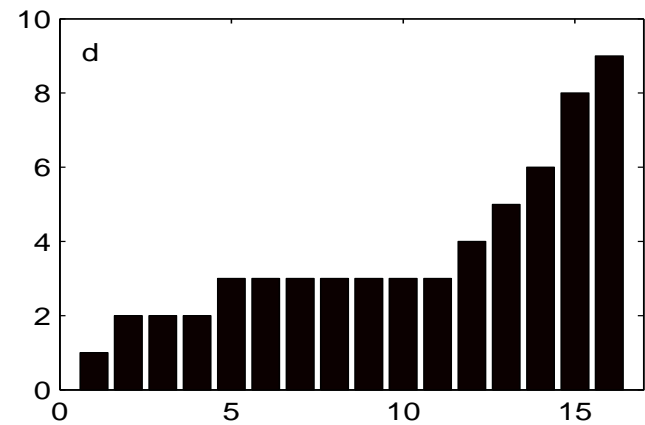
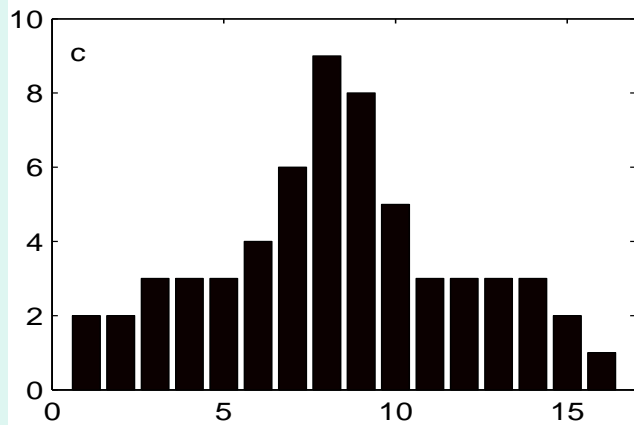
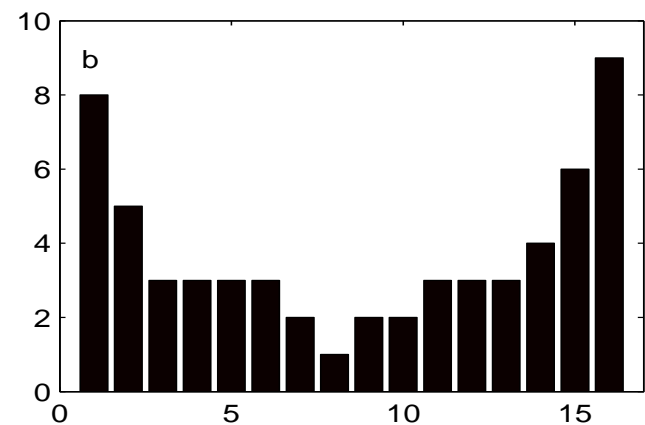
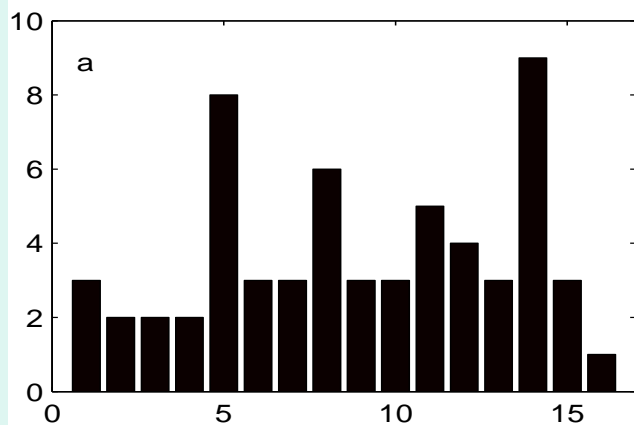
Elmore example again



Elmore example – p-values

	Panel a	Panel b	Panel c	Panel d
T	0.193	0.193	0.193	0.193
C-von M	0.379	0.0013	0.0013	0.0001
Linear	0.474	0.575	0.737	0.0001
Ends	0.172	0.0002	0.079	0.321
V-shape	0.284	0.0001	0.0001	0.128
Resid_1	0.195	0.965	0.235	0.995
Resid_2	0.165	0.985	0.985	0.9996

Elmore example again



Elmore example - comments

- Decomposition is at least as powerful as Cramér-von Mises.
- V-shape does better than Ends, but note lower Ends p-value for b than for c .
- Very large p-values for Residuals reflect the artificial nature of the example, where deviations from uniformity are designed to be of a specific form.

Northern hemisphere 500hPa geopotential height

Although not strictly correct, we treat these data as consisting of 420 independent observations of 14-member ensemble forecasts.

500hPa heights – p-values

T	Linear	Ends	V-shape	Resid_1	Resid_2
0.00001	0.0065	0.00003	0.011	0.024	0.0006

- Highly significant value for T.
- Clear underdispersion, and Ends has much smaller p-value than V-shape.
- Linear also indicates evidence of bias.
- Resid_2 clearly shows deviation other than Linear and V-shape is present.

Caveats

- P-values are approximate.
- Some restrictions (orthogonality) on decomposition.
- Assumes independence of forecasts.

Virtues

- More powerful than T – may identify deviations from uniformity when T does not.
- If T does identify deviations, decomposition can tell you the nature of these deviations.
- Easier to use and more flexible than Cramér-von Mises tests, and at least as powerful in the examples examined.

Questions?



500hPa Geopotential height data – more details

- Daily 24-hr forecasts of NH 500hPa heights for 2006-2007 winter season, created from the NCEP GEFS system.
- 14 ensemble members – hence 15 bins.
- Total number of forecasts = $290304 = 84 \text{ days} \times 3456 \text{ gridpoints}$.
- Strong temporal and spatial dependence – take 25 spatial degrees of freedom and divide number of time points by 5 (Toth), and **for illustration** treat the data as 420 independent observations.

Some background mathematics

- Let L be a $(k \times k)$ matrix whose rows are orthonormal, with elements l_{rj} , whose last row's elements are all $1/\sqrt{k}$.

- Let
$$x_i = \frac{(n_i - e_i)}{\sqrt{e_i}}, \text{ so } \sum_{i=1}^k x_i^2 = T$$

- Finally let
$$u_r = \sum_{i=1}^k l_{ri} x_i$$

Background mathematics II

- Then $u_1^2, u_2^2, \dots, u_{(k-1)}^2$ are asymptotically independent χ^2 random variables each with one degree of freedom.
- By choosing the first few rows of L appropriately, it is possible to isolate parts of T which are sensitive to particular types of deviation from uniformity such as bias/trend and over/under dispersion.