STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

SP 29

D. van Dantzig and J. Hemelrijk

Comment on the paper: "Distibutionfree tests in time-series based on the breaking of records", of F.G.Foster and A.Stuart.

Overdruk uit:
Journal of the Royal Statistical Society,B 16
(1954) p.19-21.

The following contributions were received in writing after the meeting:

Messrs. D. VAN DANTZIG and J. HEMELRIJK: The subject of records in time-series is a very important one and the contribution to this theory made in the paper by F. G. Foster and A. Stuart seems to us to be very interesting and helpful in its further development.

There are fields of research where records are of central interest, e.g., the investigation of floods, which at present is carried out rather intensively in the Netherlands for obvious reasons. The usual way of treating subjects of this kind is to apply the theory of extreme values as developed by M. Fréchet, L. H. C. Tippett, R. A. Fisher and E. J. Gumbel. It is, however, likely that investigations based on records may lead to additional results. Anyway the use of records seems to be highly relevant in this context.

We have therefore applied the method of Foster and Stuart to the highest seawater levels attained during each of the years 1864–1953 in Hoek van Holland (these being the years for which data about this quantity are available) and to some parts of this period. For tied observations the mean of the number of records for all permutations of the observations of each tie has been recorded. We return to the question of the treatment of ties below. Table 1 gives the results.

TABLE 1

Application of the d, d′ *and* D-*test to the Highest Water-levels in Hoek van Holland in a number of Years*

Years	*Forwards*		*Backwards*		d	d′	D	n
	Upper	*Lower*	*Upper*	*Lower*				
1864–1953	6	0	0	$7\frac{5}{8}$	$5\frac{1}{8}$	$-7\frac{5}{8}$*	13*	90
1894–1953	1	4	0	6	−3	−6*	3	60
1864–1952	5	$\frac{5}{8}$	6	$6\frac{5}{8}$	$4\frac{1}{8}$	$-\frac{5}{8}$	5	89
1865–1952	4	$4\frac{1}{2}$	6	$6\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	0	88
1894–1952	0	4	6	5	−4	1	−5	59

* Indicates significance at the 0·05 level.

To understand the large differences in the results for the different periods it should be pointed out that the highest level reached in 1864 was exceptionally low (in fact this was the lowest value recorded in the whole period) and that the values of 1894 and 1953 were the two highest ones, 1953 exceeding 1894. Using the normal approximation for the probability distributions of d, d' and D, the standard errors from Tables 3 and 4 of the paper and a level of significance 0·05, the d' gives a significant result, indicating an upward trend, for the first two periods of the table, and D does the same for the first period. On the other hand, omission of the year 1953 changes the "backwards" results considerably, as might be expected, 1953 being the highest value recorded. In this case the only significant results are the ones mentioned above and no contradictory results emerge from these data. Nevertheless the example shows that the first and the last observations of the series have a very large influence on the results and the same holds for a small number of observations at the beginning and the end, even if the first and the last observation are not exceptionally small or large. This seems to us to be a drawback, restricting the applicability of the method; it is a consequence of the fact that only absolute records are counted. The fact that in our example there happens to be a small *real* upward trend, caused by the perennial rise of the sea level, does not influence our argument. An unlucky choice of the beginning or the end of the period—and psychological factors are apt to have a large influence here—may either generate a spurious significance or obliterate a real difference. In order to do full justice to the authors of the present interesting paper it must be remarked that the Gumbel-method suffers from a similar drawback.

This difficulty might perhaps be met with if an extension of the theory to observations exceeding (or falling short of) m directly preceding observations could be made, for fixed values of m or possibly for different values of m simultaneously. The derivation of the probability distributions of statistics based on such notions is likely to be difficult but not unfeasible and the influence of the first and last observations of the series would be brought in better proportion to the influence of the middle part of the series. The possibility of drawing a wrong conclusion because of one or two outlying observations—caused by error or by investigating a series especially after a record has been found, as is the case in the example of the high-water-levels—would be greatly reduced. Such an extension might have another beneficial effect by making the test sensitive to fluctuations in the time series and also to alternative hypotheses of a more complicated character, viz., hypotheses stating that the time series contains a number of observations interspersed at unknown moments between the other ones, which originate from another population containing higher (or lower) values than the population giving the ordinary observations (viz. "dangerous" as opposed to "ordinary" years). Of course these remarks are not meant as a criticism on the important work done by the authors of the paper, but rather as a stimulus to continue their studies on the subject. In fact we hope to profit by the results obtained already and we would be grateful for further developments.

Another remark refers to section 10 of the paper: the treatment of ties. The authors propose to randomize the ties after the observations have been made and they refer to a paper of Lehmann, stating that he has proved the preservation of optimal properties for two-sample tests when this procedure is applied. Lehmann's result, however, only refers to the property of unbiasedness of a test and not to its power. It seems very likely, that this randomization procedure diminishes the power of the tests. In fact this can be proved to be true for a large class of tests containing, e.g., the sign test, where omitting the ties (zero-differences in that case) gives a larger power than distributing them at random over the positive and negative differences. The sign test may be regarded as a special case of Friedman's method of m rankings (viz. when there are only two columns) and omitting the ties coincides in this case with the mid-rank method which is the usual procedure for Friedman's method. For Wilcoxon's test also it may be proved that for a sufficiently large number of observations the power of the mid-rank method is larger than when the ties are randomized. This result may be made plausible by the following argument. The difference between the mid-rank method and the randomization of ties consists in the addition, in the last case, of a random variable to the test statistic; this random variable, however, has the same probability distribution if the hypothesis tested is true as when this is not so. Thus the probability distribution of the resulting test statistic, under an alternative hypothesis, is partly remodelled after the model of the distribution under the hypothesis tested. One might also explain the loss of power intuitively by the loss of information inherent in the randomization, the information, contained in the knowledge that the tied observations *are* equal, being lost.*

* Of course, if randomization is introduced, when using a test statistic with a discrete probability distribution, in order to raise the true level of significance to the nominal value of the level of significance, the power will increase at least in a neighbourhood of the hypothesis tested. This, however, is a different type of application from the one considered here.

As an analogue of the mid-rank method we have, in computing the above Table 1, recorded the number of records as the average of this number over all permutations of tied observations. This does not change the mean of the statistics d, d' and D under the hypothesis, but it does affect their variance. Our conjecture is that the variance will be reduced a little, as is the case for Wilcoxon's two sample test. This has not been taken into account in the above calculations. Another procedure might be to count the equalling of a record as a record too. The number of records would then always remain an integer. It does not seem unfeasible to work out the standard errors and limiting distributions for these modified statistics—under some restrictions for the number of ties—and this would make the theory applicable to discrete probability distributions as well as to continuous ones.

Dr. R. M. SUNDRUM: At the conclusion of their interesting paper, the authors mention the possibility of extending their test to sequential form. To apply the Wald likelihood-ratio form of the sequential test would require a much greater knowledge of the sampling distributions in the non-null case than we have at present. The following suggestion, based on a simple sequential procedure suggested by Chia Kuei Tsao (1953, *Annals of Math. Stat.*, **24**, p. 141, Abstract) may be found interesting.

Consider the three hypotheses: $H_0 =$ no trend; $H_1 =$ an upward trend; and $H_2 =$ a downward trend. Let n_1 be the number of upper record values and n_2 the number of lower record values, and k, l, two positive integers, $k > l$. Sampling continues till one of the following relations holds and the corresponding decision is made.

Accept H_1 if $n_1 = k$ before $n_2 = l$.

Accept H_2 if $n_2 = k$ before $n_1 = l$.

Accept H_0 if $n_1 = l$ before $n_2 = k$ $(n_2 > l)$
and if $n_2 = l$ before $n_1 = k$ $(n_1 > l)$.

The choice of k, l has to be made so as to meet specified requirements regarding the Type I and II errors. It should be simple to obtain one relation between k and l providing control of Type I error. Ideally, the other relation between k and l should be such as to meet a specified requirement as to power. However, this form of the sequential test is such that we may alternatively fix the size of the expected sampling number and then subject to this choose k and l by intuitive considerations.

The authors replied in writing as follows :

As Dr. Armitage and Mr. Ehrenberg pointed out, there are not many distribution-free methods of estimation, but this may simply be because the appropriate methods have not been investigated. In the present case, for example, some function of the numbers of upper and lower records will estimate Δ, the regression coefficient, and Dr. Chandler gives some relevant information here. However, the function will differ according to the distributional form of the alternative hypothesis considered: for estimation, as for power calculations, distribution-free statistics lose their simplicity.

Mr. Ehrenberg's criticism of our application to the rainfall figures at Oxford applies equally to a wide class of significance tests. We were, as he supposed, using the standard procedure of regarding the time series as a realization of a stochastic process, and we set out to test the hypothesis that the process is a sequence of independent identical variates.

The procedure recommended for the resolution of tied observations was the only one we had seen mentioned in the literature, and we are grateful to Professors van Dantzig and Hemelrijk for their discussion of this point. A recent abstract in the *Annals of Mathematical Statistics* outlines further results along the lines they indicate. We agree that record-values analysis seems to be more appropriate to many situations, such as the water levels example they give, to which extreme-values analysis has in the past been applied.

The continuous approximation suggested by Dr. Armitage is very interesting and should prove of value in investigating the asymptotic properties of record tests.

Several speakers, including Dr. Vajda and Dr. Good, made related remarks on the use of "local" records, for which each observation is compared with a fixed number of immediately preceding observations, and on what might be called "sectional" records, for which a series is divided into separate sections, records being counted only within sections. We agree with Mr. Beale that the use of local records is likely to have less power when testing against a linear trend;

indeed, to take the extreme case, the difference-sign test (discussed in the paper) may be regarded as a local record test of order 2, and this test has very low power. Such, or similar, procedures, however, would be required to deal with the type of alternative hypothesis discussed by Mr. Page and Dr. Good.

The value of increasing the weight attached to later records is quite clear from Mr. Beale's calculations, and this is another subject which would be well worth exploring.

As Professor Kendall said, problems concerning the content of the orthants of a multi-normal distribution may be put into the form of problems concerning a sequence of independent normal variates. In the particular case when these variates are identical, it follows immediately from equation (1) of our paper that if $X_1, \ldots X_n$ are multinormally distributed with a correlation matrix whose off-diagonal elements are all $+\frac{1}{2}$, the probability that the X_i are all positive is exactly $\frac{1}{n+1}$. This is true for any n, odd or even, and we are reluctant to believe that no further progress can be made with the exact solution of the important general problem. The method of approximation outlined by Professor Kendall seems likely to require algebraic heroics.

Dr. Cox's very interesting binomial tests have remarkably high powers, and it appears that even these can be improved upon slightly by weighting procedures. It seems worth pointing out that records still remain slightly easier to count. This is the case in particular if sequential procedures are envisaged.

The ingenious stopping rule given by Dr. Sundrum seems to be equally applicable to a large number of tests of randomness and we look forward to hearing more about this in the future.

We are grateful to the contributors to the discussion, from which we have learnt a great deal.