Cramér-von Mises statistics for discrete distributions

by V. Choulakian, R. A. Lockhart and M. A. Stephens*

Summary

Cramér-von Mises statistics are developed for use in testing for discrete distributions, and tables are given for tests for the discrete uniform distribution.
Some key words: EDF statistics; goodness-of-fit; components.

1 Introduction

The Cramér-von Mises family of goodness-of-fit statistics is a well-known group of statistics used to test fit to a continuous distribution. In this article we extend the family to provide tests for discrete distributions. The statistics examined are the analogues of those associated with the names of Cramér-von Mises, Watson and Anderson-Darling, called W^2, U^2 and A^2 respectively, and their components. We provide formulas for the test statistics, and asymptotic percentage points for the test for a uniform distribution with k cells. The tests are based on the empirical distribution function (EDF) of the sample. They are closely related to Pearson's X^2 test, and to Neyman-Barton smooth tests; in particular, all the tests can be broken down into components, as has been observed by many authors. It is suggested that A^2 be used to test the overall null hypothesis in general, and U^2 for the particular case where observations are counts around a circle. Their components can be used to test for particular types of departure from the null.

In section 2, we define the test statistics and give the general distribution theory. In section 3 the solution of the uniform case is given, together with

*A shortened version of this paper was presented at the Wilks Conference, Princeton, February 1992, in honour of the work of Professor G. S. Watson.

two examples; in section 4 modified versions of the statistics are discussed. In section 5 power studies are given which show that A^2 is a good omnibus test statistic. Finally, in section 6 we discuss the use of components as individual test statistics and demonstrate the use of a graphical procedure called the Z-plot to determine, when a statistic is found to be significant, the type of departure from the null.

2 EDF tests for discrete data

The Kolmogorov-Smirnov statistic appears to be the only EDF statistic which has been studied extensively for testing goodness-of-fit with discrete data; see, for example, Schmid (1958), Conover (1972), Horn (1977), Pettitt and Stephens (1977), Wood and Altavela (1978) and Stephens (1986). Schmid (1958) derived the asymptotic null distribution of the Kolmogorov-Smirnov statistic when the hypothesized cumulative distribution function possessed a finite number of discontinuities and was increasing between the discontinuities. Wood and Altavela (1978) extended Schmid's results to cumulative distribution functions having a countable number of discontinuities, and found that the Kolmogorov-Smirnov statistic can be calculated using the multivariate normal distribution. Hirotsu (1986) and Nair (1987) have discussed problems involving two samples with results closely related to those given below. Freedman (1981) has discussed U^2 for testing uniformity of counts around a circle.

The W^2, U^2 and A^2 statistics for discrete data are defined as follows. Consider a discrete distribution with k cells labelled $1, 2, \ldots, k$, and with probability p_i of falling into cell i. Suppose N independent observations are given; let o_i be the observed number of outcomes in cell i, and let $Np_i = e_i$ be the expected number in cell i. Let $S_j = \sum_{i=1}^{j} o_i$ and $T_j = \sum_{i=1}^{j} e_i$. Then S_j/N and $H_j = T_j/N$ correspond to the empirical distribution function $F_N(x)$ and the cdf $F(x)$ in the continuous case. Suppose $Z_j = S_j - T_j, j = 1, 2, \ldots, k$. The Cramér-von Mises statistics W^2, U^2 and A^2 for a discrete distribution are then defined by

$$W^2 = N^{-1} \sum_{j=1}^{k} Z_j^2 p_j; \tag{1}$$

$$U^2 = N^{-1} \sum_{j=1}^{k} (Z_j - \bar{Z})^2 p_j; \tag{2}$$

$$A^2 \;=\; N^{-1}\sum_{j=1}^{k} Z_j^2 p_j/\{H_j(1-H_j)\} \tag{3}$$

where $\bar{Z} = \sum_{j=1}^{k} Z_j p_j$. Note that $Z_k = 0$ in these summations, so that the last term in W^2 is zero. The last term in A^2 is of the form $0/0$, and is set equal to zero.

It is convenient to put these expressions into matrix notation. Let a prime, e. g. $\mathbf{Z}'$, denote the transpose of a vector or matrix. Let $\mathbf{I}$ be the $k \times k$ identity matrix, and let $\mathbf{p}'$ be the $1 \times k$ vector $(p_1, p_2, \ldots, p_k)$. Suppose $\mathbf{D}$ is the $k \times k$ diagonal matrix whose j-th diagonal entry is $p_j, j = 1, \ldots, k$ and let $\mathbf{G}$ be the diagonal matrix whose (j,j)-th element is $H_j(1-H_j), j = 1, \ldots, k$. The S_j and T_j may be defined in terms of the o_i and e_i. Arrange these quantities into column vectors $\mathbf{S}$, $\mathbf{T}$, $\mathbf{o}$, $\mathbf{e}$ (so that, for example, the j-th component of $\mathbf{S}$ is $S_j, j = 1, \ldots, k$). Then $\mathbf{Z} = \mathbf{Ad}$, where $\mathbf{d} = \mathbf{o} - \mathbf{e}$ and $\mathbf{A}$ is the $k \times k$ partial-sum matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix}.$$

The definitions become

$$W^2 \;=\; \mathbf{Z}'\mathbf{D}\mathbf{Z}/N; \tag{4}$$

$$U^2 \;=\; \mathbf{Z}'(\mathbf{I} - \mathbf{D}\mathbf{1}\mathbf{1}')\mathbf{D}(\mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D})\mathbf{Z}/N; \tag{5}$$

$$A^2 \;=\; \mathbf{Z}'\mathbf{D}\mathbf{G}^{-1}\mathbf{Z}/N. \tag{6}$$

The well-known Pearson X^2 statistic, defined by $X^2 = \sum_{i=1}^{k}(o_i - e_i)^2/e_i$, is

$$X^2 = (\mathbf{d}'\mathbf{D}^{-1}\mathbf{d})/N = \mathbf{Z}'\mathbf{A}^{-1'}\mathbf{D}^{-1}\mathbf{A}^{-1}\mathbf{Z}/N.$$

Distribution theory. The covariance matrix of $\mathbf{o}$ is $\mathbf{\Sigma}_0 = N(\mathbf{D} - \mathbf{p}\mathbf{p}')$ and that of $\mathbf{Z}$ is $\mathbf{\Sigma} = \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}'$, with entries $\Sigma_{ij} = N(\min(H_i, H_j) - H_i H_j)$. All four statistics above are of the general form $\mathbf{Z}'\mathbf{M}\mathbf{Z}$, where $\mathbf{M}$ is symmetric. We may write a typical statistic S as

$$S = \mathbf{Z}'\mathbf{M}\mathbf{Z} = \sum_{i=1}^{k-1} \lambda_i (\mathbf{w}_i{}'\mathbf{Z})^2, \tag{7}$$

where λ_i are the $k-1$ non-zero eigenvalues of $\mathbf{M\Sigma}$ and $\mathbf{w}_i$ are the corresponding eigenvectors, normalized so that $\mathbf{w}_i{}'\mathbf{\Sigma}\mathbf{w}_j = \delta_{ij}$. In (7), the term $s_i^2 = (\mathbf{w}_i{}'\mathbf{Z})^2$ is called the i-th **component** of the statistic. As $N \to \infty$, the distribution of $\mathbf{Z}/\sqrt{N}$ tends to the multivariate normal with mean 0 and variance Σ/N. The distribution of a typical s_i tends to univariate normal, mean 0, variance 1 and asymptotically the s_i are independent. A typical statistic, S, has an asymptotic distribution

$$S \sim \sum_{i=1}^{k-1} \lambda_i s_i^2 \tag{8}$$

which is a sum of independent weighted χ_1^2 variables. To calculate percentage points for this distribution it is necessary only to find the weights λ_i for the appropriate statistic and then use the method of Imhof (1961). This method has been used to find the asymptotic distribution of W^2, U^2, and A^2 when the tested distribution is discrete uniform with k cells, and percentage points for these statistics are given in Table 1.

Calculation of λ_i and components. To find the λ_i and $\mathbf{w}_i$ in (7), it is convenient to work with $\mathbf{X} = \mathbf{M}^{1/2}\mathbf{Z}$; the covariance of $\mathbf{X}$ is then $\mathbf{\Sigma}_X = \mathbf{M}^{1/2}\mathbf{\Sigma}\mathbf{M}^{1/2}$. The statistic is $S = \mathbf{X}'\mathbf{X}$ and can be written in a form similar to (7) using the eigenvalues and eigenvectors of $\mathbf{\Sigma}_X$. It may be shown that the eigenvalues λ_i of $\mathbf{\Sigma}_X$ are the same as those of $\mathbf{M\Sigma}$, and corresponding eigenvectors $\mathbf{v}_i$ are related to $\mathbf{w}_i$ above by $\mathbf{w}_i = \mathbf{M}^{1/2}\mathbf{v}_i$. The required normalization is now $\mathbf{v}_i{}'\mathbf{\Sigma}_X\mathbf{v}_j = \delta_{ij}$ which follows from the normalization of $\mathbf{w}_i$ given above. Since $\mathbf{\Sigma}_X\mathbf{v}_j = \lambda_j\mathbf{v}_j$, it follows that the $\mathbf{v}_i$ are jointly orthogonal, and of length $\mathbf{v}_i{}'\mathbf{v}_i = 1/\lambda_i$. The statistic $S = \mathbf{Z}'\mathbf{M}\mathbf{Z}$ may be written

$$S = \mathbf{X}'\mathbf{X} = \sum \lambda_i(\mathbf{v}_i{}'\mathbf{X})^2 = \sum(\mathbf{u}_i{}'\mathbf{X})^2, \tag{9}$$

where $\mathbf{u}_i = \sqrt{\lambda_i}\mathbf{v}_i$ is an eigenvector of $\mathbf{\Sigma}_X$ with length 1. Finally, the component $(\mathbf{w}_i{}'\mathbf{Z})^2$ is found from $(\mathbf{u}_i{}'\mathbf{X})^2$ by dividing by λ_i.

An advantage to the above form (9) is that $\mathbf{\Sigma}_X$ is symmetric and $\mathbf{M\Sigma}$ is not. It is often easier to find the λ_i and $\mathbf{v}_i$ algebraically for a symmetric matrix; this is the case for the discrete uniform distribution considered below. Also, when eigenvalues and eigenvectors must be found numerically, computer packages are more readily available if the matrix is symmetric, and the packages usually give eigenvectors of unit length.

The components s_i^2 take different values for different statistics, since they depend on $\mathbf{M}$ through the eigenvectors $\mathbf{w}_i$; however, $\sum s_i^2$ is the same for all the statistics and equals the well-known Pearson X^2. This is illustrated

in section 6. For different statistics, the different components s_i^2, when compared with the χ_1^2 distribution, may be used to test for different kinds of departure from the tested distribution. We return to these applications later.

3 The discrete uniform distribution

3.1 Statistic W^2

We now consider tests for the discrete uniform distribution with k cells and with $p_j = 1/k$ for all j.

For W^2, we have $\mathbf{M} = \mathbf{D}/N$; thus $\mathbf{X} = \mathbf{D}^{1/2}\mathbf{Z}/\sqrt{N} = \mathbf{Z}/\sqrt{kN}$, and $\mathbf{\Sigma}_X$ has entries

$$(\mathbf{\Sigma}_X)_{ij} = \{k \min(i,j) - ij\}/k^3. \tag{10}$$

The last row and last column of $\mathbf{\Sigma}_X$ are all zeros (this occurs because the final term is included in the sum (1)), so one eigenvalue, say λ_k, is zero. For the others, suppose $\mathbf{Q}$ is the matrix $\mathbf{\Sigma}_X$ with the last row and column omitted. The inverse of $\mathbf{Q}$ is k^2 times $\mathbf{P}^*$, where $\mathbf{P}^*$ is the $(k-1) \times (k-1)$ tridiagonal matrix with 2 along the main diagonal and -1 on the subdiagonal and superdiagonal. The eigenvalues and eigenvectors of this matrix are well known. The eigenvalues of $\mathbf{Q}$ can then be found. They are

$$\lambda_i = \frac{1}{2k^2\{1 - \cos(i\pi/k)\}}, \qquad i = 1, 2, \ldots, k-1.$$

The orthonormal eigenvector $\mathbf{u}_i$ corresponding to λ_i has j-th component $u_{ij} = (2/k)^{1/2} \sin(\pi ij/k), j = 1, 2, \ldots, k$. The i-th component s_i^2 is then $(\mathbf{u}_i{}'\mathbf{X})^2/\lambda_i$.

Asymptotic percentage points of W^2, given by (8), are in Table 1. Monte Carlo studies show that the points for k cells and N observations converge quickly, as $N \to \infty$, to the asymptotic points for k cells. This rapid convergence is similar to the continuous case. Note also that, as $k \to \infty$, the values $\lambda_i \to 1/(i^2\pi^2)$, the λ-values for the continuous uniform test, and the percentage points converge, as expected, to the asymptotic points for W^2 for this test.

3.2 Statistic U^2

It is convenient to write U^2 in the form $U^2 = \frac{1}{kN}\sum_{i=1}^k (Z_i - \bar{Z})^2$, where $\bar{Z} = \sum_{i=1}^k Z_i/k$. Let $\mathbf{Y}$ be the vector with components $Y_i = (Z_i - \bar{Z})/(kN)^{1/2}$,

so that $U^2 = \mathbf{Y}'\mathbf{Y}$. We now need the eigenvalues and eigenvectors of the covariance matrix $\mathbf{\Sigma}_Y$ of $\mathbf{Y}$. Write $\mathbf{B} = \mathbf{I} - \mathbf{1}\mathbf{1}'/k$; then it may be shown that $\mathbf{\Sigma}_Y = \mathbf{B}\mathbf{\Sigma}_X\mathbf{B}$. Since $\mathbf{B}$ is idempotent, the eigenvalues of $\mathbf{\Sigma}_Y$ are the same as those of $\mathbf{B}\mathbf{\Sigma}_X$. From W^2, we have $\mathbf{\Sigma}_X\mathbf{w}_i = \lambda_i\mathbf{w}_i$; then $\mathbf{B}\mathbf{\Sigma}_X\mathbf{w}_i = \lambda_i\mathbf{B}\mathbf{w}_i = \lambda_i(\mathbf{w}_i - \bar{w}_i\mathbf{1})$, where $\bar{w}_i = \sum_{j=1}^{k} w_{ij}/k$. Thus the eigenvalues of $\mathbf{B}\mathbf{\Sigma}_X$ are the same as those of $\mathbf{\Sigma}_X$, provided the eigenvector $\mathbf{w}_i$ has mean $\bar{w}_i = 0$. For k odd, this occurs when i is even, and the eigenvalues of $\mathbf{\Sigma}_Y$ then occur *in pairs*; the values, each occurring twice, are

$$\lambda_i = \frac{1}{2k^2\{1 - \cos(i\pi/k)\}}, \qquad i = 2, 4, \ldots, k-1. \tag{11}$$

The corresponding orthonormal eigenvectors are the $\mathbf{u}_i$ of W^2 above, now called $\mathbf{r}_i$, and another set $\mathbf{r}_i^*$ whose j-th components are $(2/k)^{1/2}\cos(\pi ij/k)$. For k even, λ_i is again given by (11), for $i = 2, 4, \ldots, k-2$, and each occurs twice; the eigenvectors are again $\mathbf{r}_i$ and $\mathbf{r}_i^*$, $i = 2, 4, \ldots, k-2$. There is a further eigenvalue $\lambda_{k-1} = 1/4k^2$ with corresponding eigenvector $\mathbf{r}_{k-1} = (1/k)^{1/2}(-1, 1, -1, 1, \ldots, 1)$. Let λ_i^* denote the i-th eigenvalue when the complete set of $k-1$ eigenvalues has been arranged in descending order, and let $\mathbf{u}_i{}^*$ denote the corresponding eigenvector. The i-th component s_i^2 of U^2 is then $(\mathbf{u}_i{}^{*\prime}\mathbf{X})^2/\lambda_i^*$.

For k odd, the asymptotic distribution (8) may, for U^2, be written as a sum of weighted exponentials. The distribution function of such a sum may be put in closed form. For k even there is a similar but more complicated expression. However, it is easy to obtain percentage points by Imhof's method, and these are given in Table 1. U^2 has already been discussed by Freedman (1981), who gave λ's for certain values of k, and suggested fitting Pearson curves to obtain asymptotic points, but did not give a general formula for the λ-values or their eigenvectors.

3.3 Statistic A^2

The analysis for A^2, given by (3), can follow the same lines as for W^2, with matrix $\mathbf{M} = \mathbf{D}\mathbf{G}^{-1}/N$ (note that $\mathbf{D}$ and $\mathbf{G}$ are both diagonal). The algebra to find λ_i and $\mathbf{v}_i$ is now more complicated, and we omit the details. The interesting result is that, for k cells, the eigenvalues are $\lambda_i = 1/\{i(i+1)\}, i = 1, 2, \ldots, k-1$, and $\lambda_k = 0$. The non-zero values are *exactly the same* as for the continuous case, up to $k-1$, whereas for W^2 and U^2, the λ_i only tend to the continuous lambdas as $k \to \infty$.

To obtain components s_i^2 of A^2, define vector $\mathbf{w}_i$, with components

$w_{ij}, j = 1, 2, \ldots, k$, given by

$$w_{ij} = \{t_i(j) - t_i(j-1)\}/c_i, \tag{12}$$

where $c_i^2 = (k+i)!/\{(2i+1)(k-i-1)!\}$, and $t_i(j)$ is the i-th Chebyshev polynomial for discrete values (Erdelyi, 1953, page 223, with $N = k$). The normalising constant ensures that $\mathbf{w}_i'\mathbf{\Sigma}\mathbf{w}_j = \delta_{ij}$. Then component $s_i^2 = k(\mathbf{w}_i'\mathbf{Z})^2/N$. The $t_i(j)$ can be found from the recurrence relation

$$(i+1)t_{i+1}(j) = (2i+1)(2j-k+1)t_i(j) - k^2(k^2-i^2)t_{i-1}(j), \tag{13}$$

together with $t_0(j) = 1$ and $t_1(j) = 2j - k + 1$. Eigenvalues λ_i and eigenvectors $\mathbf{w}_i$ have already been given by Hirotsu (1986) and by Nair (1987) in connection with similar problems, arising from the analysis of ordered contingency tables, and involving the matrix here called $\mathbf{\Sigma}_X$.

3.4 Test for the discrete uniform distribution

The test of H_0 : observations $x_r, r = 1, 2, \ldots, N$ are from a discrete uniform distribution with k cells is as follows:

1. Calculate the statistics from formulas (1)-(3) above, with $p_j = 1/k$.
2. Refer the statistic to the appropriate part of Table 1, for a distribution with k cells, and reject H_0 at level α if the statistic exceeds the given point for level α. Although the points given are asymptotic, Monte Carlo studies show that they provide good accuracy for N as low as 10.

Statistic U^2 should be used for cells which occur around a circle because its value does not depend on which cell is chosen to be the first. All three statistics may be used for cells along a line. We give power studies below which suggest that overall A^2 is the recommended statistic for such cells, particularly as it is effective in detecting changes in the tail of the distribution.

3.5 Example 1

Pettitt and Stephens (1977) give an example, from Siegel (1956), of data with $k = 5$ cells. $N = 10$ subjects were asked to rank photographs according to preference; the same photograph was presented in five tones, so that there was a natural ordering of the cells. The observed preferences

for tone, or cell, i were, for $i = 1, 2, \ldots, 5$, the values o_i =0, 1, 0, 5, 4. It was desired to test for no preference in tone, that is for equal probabilities for the cells, so that $e_i = 2$ for $i = 1, \ldots, 5$. Then $\mathbf{Z}' = (-2, -3, -5, -2, 0)$ and $(\mathbf{Z} - \bar{\mathbf{Z}})' = \frac{1}{5}(2, -3, -13, 2, 12)$. These give values $W^2 = 0.84$, $U^2 = 0.264$ and $A^2 = 3.83$, with p-values 0.007, 0.019 and 0.009 respectively. The p-values, taken from Table 1, must be treated with caution since the sample size is small, but, nevertheless, those for W^2 and A^2 are clearly significant, near the 0.008 level. Pettitt and Stephens found that the Kolmogorov-Smirnov statistic was also significant at this level, while the Pearson X^2 statistic (whose value is 11) had exact p-value 0.04 and an approximate p-value, from the asymptotic χ^2 distribution, of 0.024.

3.6 Example 2

In this example we illustrate the use of U^2. The data are taken from Edwards (1961), and consist of counts of births of children with anencephalitis, for the years 1940–1947, in Birmingham, England. The counts for January to December are 10, 19, 18, 15, 11, 13, 7, 10, 13, 23, 15, 22. It is desired to test H_o: the counts are uniform over the months, against the possibility of a seasonal effect. The total is 176, so $e_i = 14\frac{2}{3}$ ignoring the slight variability in the lengths of the months. Such data is often displayed as counts around a circle divided into 12 monthly sectors. For these data a goodness-of-fit statistic for testing H_o should not depend on which month of the year is regarded as the first, so U^2 is the statistic of choice.

The value of U^2 is 0.214 and, from table 1, with $k = 12$ cells, the p-value is 0.031. The Pearson X^2 statistic is 18.727, with $p = 0.066$ when compared with χ^2_{11}. Rayner and Best (1989) have also tested these data for uniformity, but used the Kolmogorov statistic with tables given by Pettitt and Stephens (1977), and also V_2, the second component of X^2, partitioned using Chebyshev polynomials. Both these statistics depend on the month that is regarded as origin.

4 Modified versions of W^2 and A^2

The definitions of W^2, U^2 and A^2 given in Section 2 have been chosen to be analogous to the corresponding statistics for testing specified continuous distributions. However, they can be modified in various ways (as can the continuous statistics; see, for example, de Wet and Venter, 1973), to give greater prominence to certain parts of the tested distribution. One such modification is to omit the p_j in definitions (1)–(3); then, for long-tailed

distributions such as the Poisson, the modified statistic will give more weight to accuracy in the long tail. We denote the statistics modified in this way by W_m^2 and A_m^2. For the discrete uniform test, of course, all $p_j = 1/k$; the new statistics, as well as the percentage points, are k times their old values, and therefore the test is unchanged.

We illustrate the modified statistics by finding A^2 and A_m^2 for the following example.

4.1 Example 3

Best and Rayner (1987, Section 4) give an example of data with $k = 5$ and $N = 20$ observations. The $p_i, i = 1, 2, \ldots, 5$ are 0.05, 0.3, 0.3, 0.3, 0.05, and the observed values in the cells are 1, 4, 11, 4, 0. These lead to values $\mathrm{Z}_i = 0, -2, 3, 1, 0$. Best and Rayner use their statistics V_1^2 and V_2^2, which, asymptotically, have χ_1^2 distributions and are independent, and also the statistic $V_1^2 + V_2^2$, with asymptotic χ_2^2 distribution. They obtain values $V_1^2 = 0.2, V_2^2 = 2.666, V_1^2 + V_2^2 = 2.866$, and give p-values 0.605, 0.103, 0.077. The first and last of these are incorrect; they should be 0.655 and 0.239. They also calculate Pearson's X^2 statistic = 6.5 (correct) and give the incorrect p-value 0.0034 and reject H_0, inviting us to "better" their results. We start by giving the correct p-value for $X^2 = 6.5$ on 4 d. f., which is 0.165 — then H_0 cannot be rejected even at the 10% level. If A^2 is used as the test statistic, the eigenvalues of $\mathbf{\Sigma}_X$ are $\lambda_i = 0.515, 0.258, 0.133$ and 0.044, and the value of A^2 is 1.173; this has p-value 0.280 when compared with distribution (8). If A_m^2 is used, the λ_i are 1.834, 0.987, 0.757, 0.422, and $A_m^2 = 3.910$, with p-value 0.396. Clearly the null hypothesis cannot be rejected.

5 Power studies

We now consider the power of the Cramér-von Mises statistics for testing uniform p_i, especially against trend alternatives. The natural statistic for comparison is Pearson's X^2, which is usually used for discrete distributions, but we include also the Kolmogorov-Smirnov statistics S^+, S^-, and S, which were studied by Pettitt and Stephens (1977). These are defined by $S^+ = \max_j Z_j; S^- = \max_j(-Z_j); S = \max(S^+, S^-)$. Pettitt and Stephens compared S with X^2, using certain families indexed by a constant δ, called $A_1(\delta)$ and $A_2(\delta)$, as alternatives to equal values of the p_j. For family $A_1(\delta)$, $p_i = \{i^\delta - (i-1)^\delta\}/\delta, i = 1, \ldots, k$. If $\delta = 2$, $p_i = (2i-1)/n^2$, and the p_i increase linearly. For family $A_2(\delta)$, $p_i = 1/n - \delta$, for $1 \leq i \leq n/2$, and

$p_i = 1/n + \delta$ for $n/2 < i \leq n$. For family A_1, there is a steady trend, and for family A_2 there are two blocks of equal values of the p_i. Two other alternatives have been added, called A_3 and A_4; these have cell probabilities which are U-shaped — symmetric around the centre cells, and with higher probabilities at the ends.

Table 2 gives power results for Monte Carlo samples of size 20, for the test of H_0: all p_i equal (to $1/n$), against the alternatives shown. It is clear that EDF statistics are much better than the familiar X^2 when the alternative is a trend in the values of the p_i. For the U-shaped p_i, U^2 is the best statistic, a result corresponding to continuous data, where U^2 detects a change in variance rather than mean, but A^2 still holds its own against X^2. Since trend alternatives are often likely to be the case, A^2 is recommended as an omnibus test statistic. Of course, if the direction of trend is known, either S^+ or S^- can do better, as is seen from the Table.

6 Use of components

We have seen that each statistic may be decomposed into components, such that the entire statistic, for example A^2, is a weighted sum of the components. The individual components can be expected to describe certain features of the data, and the importance of each component is assessed by the weight it is given. Several authors have discussed components — for example, for W^2, Durbin and Knott (1972) and Durbin, Knott and Taylor (1975); for W^2, U^2 and A^2, Stephens (1974); for X^2, and for Neyman–Barton tests, Lancaster(1969), Rayner and Best (1989). It has been suggested that the components be examined in order, and used as test statistics by comparing them to their (asymptotic) χ^2_1 distributions. In addition, differences of the type $A^2 - \lambda_1 s_1^2$, $A^2 - \lambda_1 s_1^2 - \lambda_2 s_2^2$, etc., can be compared to their asymptotic distributions.

We shall concentrate our discussion on the use of components with tests for discrete distributions and, in particular, for the discrete uniform distribution. The statistic X^2 is an interesting special case; because the λ_i are all equal to 1 it may be decomposed in many different ways. For the uniform discrete test, Rayner and Best (1989) use Chebyshev polynomials to decompose X^2. These are successively constant, linear, quadratic, etc., and are equivalent to using the mean, variance, third moment, etc. of the cell numbers. Thus a significant low-order component can be interpreted in terms of fairly simple departures from the tested distribution. In section 3 we have shown that the components of A^2 and A^2_m are also based on Cheby-

shev polynomials and lead to similar components. Neyman-Barton tests for discrete distributions again use the same polynomials. The difference in the various statistics is that the X^2 and Neyman-Barton tests give the successive components equal weights while A^2 and A^2_m give them decreasing weights.

When distributions are tested where the cells have unequal probabilities, as in Example 3 above, it is still possible to define polynomials in cell numbers of increasing order for use in decomposing X^2. Rayner and Best (1989, Appendix 3) give formulas for the first two such components; (but note two errors: their V_r should be divided by $n^{1/2}$, and in the constant C the coefficient of S_3 should be $-2A^2Y$.) However, the components of A^2 or A^2_m are no longer simple polynomials which are interpretable in terms of successive moments of cell numbers. We illustrate with Example 3. The four components of A^2 are 0.369, 1.506, 4.397 and 0.229; the third of these is significant at the 5% level, with $p = 0.036$. The components of A^2_m are 0.239, 1.313, 0.261, and 4.687, and here it is the fourth component which is significant at the 5% level, with $p = 0.030$. Notice that both sets of components add to 6.5 which is the value of X^2.

Even if components are directly interpretable in terms of sample moments of cell numbers, one may question the use of a goodness of fit statistic to test, say, that the mean is correct. For a parametric alternative, for example, there might be a better test available from likelihood ratio test theory.

An important objection to testing components in order, even though components of the Chebyshev type are attractive for the reasons given above, is that in general, it is difficult to decide where to stop — and, of course, if the data is used to make the decision, it is impossible to assess the correct α-level of the overall procedure. It is also difficult to establish a stopping rule independent of the data, because the "best" number of components to use depends on the tested distribution and also the type of alternative it is desired to detect. The question is especially acute for X^2 and Neyman-Barton tests, both of which give components equal weights. Authors who have discussed how many components of these test statistics to use include Miller and Quesenberry (1979), Solomon and Stephens (1983), and Rayner and Best (1989). There is general agreement that low order components give most power and after a certain order, the addition of further components may dilute the power. Nevertheless, for some alternatives, the departure from the null may be detected only by a quite high component. In the case of the EDF statistics, W^2, U^2 and A^2 we recommend use of the entire statistic for testing and components, when the statistic is significant, to examine where the departure lies when this can be easily interpreted. This

recommendation is based on three properties of these statistics. They are easily calculated directly from the definitions, they are based on natural measures of difference between the tested and observed distributions, and the components are given decreasing weights – thus dilution is taken care of in a natural way without totally omitting higher components. The power studies in section 6 bear out the overall effectiveness of the statistics used in this way.

The statistic U^2 is an interesting special case. This statistic can be used for cells along a line but as we have shown it is the natural statistic for observations around a circle because it does not depend on which cell is considered the first. An important alternative to uniformity, for example with monthly data, would be a periodic distribution of cell counts. The individual components of U^2 depend on the origin; however, the sum of the two components for the same eigenvalue will be independent of the origin. Such "double" components are asymptotically distributed χ^2_2. There is one extra component which is itself cyclically invariant and is asymptotically χ^2_1. The first components have a natural interpretation in terms of frequencies of a cyclic departure from the null. For example the first of these "double" components, if significant, shows a departure from uniformity with 1 peak and 1 trough 6 months apart; the second double component shows a departure with 2 peaks and 2 troughs separated by 3 months, etc. This continues until the final component represents peaks and troughs occurring every other month. In example 2, the first 5 double components are 6.636, 3.489, 1.136, 1.375, 1.637 while the last single component is 4.455. The first and last are significant with P-values 0.036 and 0.035. The significant first double component indicates a cyclic effect with period 1 year, but the last is more difficult to interpret.

6.1 The Z-plot

One method of examining the data is to plot the values of Z_j against j. It is useful also to plot the values of w_{ij} (or $-w_{ij}$) against j, when the $\mathbf{w}$ pattern can be interpreted usefully in relation to the tested distribution. For example, consider the components of A^2 for the discrete uniform test; the components w_{ij} of $\mathbf{w}_i$, as i increases are successively constant, linear in j, quadratic in j, etc. If s_1^2 is large the Z_j take roughly the same pattern as the w_{1j}. This indicates a shift in mean. This is so because s_1 is proportional to $\sum_0^k Z_j = -\sum j(o_j - e_j)$ which is the sample mean minus its expected value. Similarly if s_2^2 is large the Z_j are approximately linear, indicating a shift in variance, and so on. A plot, carefully interpreted, can give more

information than a test based on a single component much as a plot of residuals after fitting a regression is more informative than a simple test for homoscedasticity.

We illustrate this procedure by plotting the Z-plot and the plots of $-\mathbf{w}_1$ and $-\mathbf{w}_2$ for Example 1 in Figure 1. The Z values are all negative indicating clearly that the observed distribution is stochastically larger that the hypothesized. If the values were closer than they are to $-\mathbf{w}_1$ this would indicate a direct shift in mean; the merit of the plot lies in showing the more subtle effect of stochastic ordering.

If the cells do not have equal probabilities the $\mathbf{w}_i$ arising form either A^2 or A_m^2 are more difficult to interpret. This is demonstrated by Figure 2 in which we plot $\mathbf{w}_3$ for A^2 and $\mathbf{w}_4$ for A_m^2 for Example 3. These are the eigenvectors which yield the largest components in the respective statistics.

Consider the statistic A^2. The Z-plot is reasonably close to $\mathbf{w}_3$, which is roughly cubic. This can be interpreted as a difference in the length of the tails. For statistic A_m^2 the Z-plot is close to the plot of $\mathbf{w}_4$. This is even closer to a cubic polynomial and is much easier to interpret. (The fact that the tested distribution is symmetric makes the interpretation of the $\mathbf{w}$'s easier. In general, corresponding eigenvectors become increasingly more difficult to interpret as the cell probabilities depart from uniformity.) With reference to the example we can make no strong conclusions since neither the overall statistic A^2 nor A_m^2 was close to being significant. We give the Z-plots to demonstrate the possibilities but also the difficulties of interpreting a single significant component when the tested distribution is not uniform.

Similarly if the Z-plot crosses the axis once a difference in variability is indicated.

References

Best, D. J. and Rayner, J. C. W. (1987) Goodness-of-fit for grouped data using components of Pearson's X^2. *Computational Statistics and Data Analysis*, **5**, 53-57.

Conover, W. J. (1972) A Kolmogorov goodness-of-fit test of continuous distributions. *Journal of the American Statistical Association*, **67**, 591-596.

de Wet, T. and Venter, J. H. (1973) Asymptotic distributions for quadratic forms with applications to tests of fit. *Annals of Statistics* **2**, 380-387.

Durbin, J. and Knott, M. (1972) Components of Cramér-von Mises statistics, I. *Journal of the Royal Statist. Soc. B*, **34**, 290-306.

Durbin, J., Knott, M. and Taylor, C. C. (1975) Components of Cramér-von Mises statistics, II. *Journal of the Royal Statist. Soc. B*, **37**, 290-307.

Edwards, J. H. (1961) The recognition and estimation of cyclic trends. *Ann. Hum. Gen.* **25**, 83-86.

Erdelyi, A. (1953) *Higher transcendental functions.* Bateman Project, 2. New York: McGraw Hill.

Freedman, L. S. (1981) Watson's U_N^2 statistic for a discrete distribution. *Biometrika*, **68**, 708-711.

Hirotsu, C. (1986) Cumulative chi-squared statistic as a tool for testing goodness of fit. *Biometrika*, **73**, 165-174.

Horn, S. D. (1977) Goodness-of-fit tests for discrete data: a review and an application to a health impairment scale. *Biometrics*, **33**, 237–248.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419–426.

Lancaster, H. O. (1969) *The Chi-Squared Distribution.* New York: Wiley.

Miller, R. L. and Quesenberry, C. P. (1979) Power studies of tests for uniformity, II. *Commun. Statist. B - Simulation and Computation*, **8**, 271-290.

Nair, V. N. (1987) Chi-squared-type tests for ordered alternatives in contingency tables. *Journal of the American Statistical Association*, **87**, 283-291.

Pettitt, A. N. and Stephens, M. A. (1977) The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, **19**, 205-210.

Rayner, J. C. W. and Best, D. J. (1989) *Smooth Tests of Goodness of Fit.* New York: Oxford University Press.

Schmid, P. (1958) On the Kolmogorov and Smirnov limit theorems for discontinuous distribution functions *Ann. Math. Statist.* **29**, 1011-1027.

Siegel, S. (1956) *Non-parametric statistics for the behavioural Sciences*. New York: McGraw Hill.

Solomon, H. and Stephens, M. A. (1983) On Neyman's statistic for testing uniformity. *Commun. Statist. B - Simulation and Computation*, **12**, 127-134.

Stephens, M. A. (1974) Components of goodness-of-fit statistics. *Annals Inst. Henri Poincare, B*, **10**, 37-54.

Stephens, M. A. (1986) Tests based on EDF statistics. Chapter 4 in *Goodness-of-Fit Techniques* (R. B. d'Agostino and M. A. Stephens, eds.) New York: Marcel Dekker.

Wood, C. L. and Altavela, M. M. (1978) Large-sample results for Kolmogorov-Smirnov statistics for discrete distributions. *Biometrika* **65**, 235–239.