For the Statistics Editor

In the first edition, to distinguish between the two contributions on goodness of fit, after we had written and submitted it, the title of the Rayner & Best article was changed to *Goodness of fit Tests and Diagnostics*. This wasn't the brief we were given and didn't well reflect the content. A different title for this edition is certainly necessary. We suggest Assessing Statistical Distribution Models: Goodness of fit Tests.

We have two prejudices we hope will be indulged. We never use

- the hyphenated 'goodness-of-fit' and
- the Pearson *chi-squared* test.

In regard to the latter it seems to us that a test cannot be uniquely identified by its distribution, so we prefer simply the Pearson test, in line with the Pearson-Fisher and other tests. Exceptions to this rule occur in references, where if an article includes 'goodness-of-fit' or *chi-squared* test in the title, these unfortunate choices are unavoidable.

The word count is a bit long at 6976 words.

There are 50 references, not 15 to 30.

Author details

J.C.W. Rayner
National Institute for Applied Statistics Research Australia, University of Wollongong, NSW 2522, Australia
and School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia
e-mail: John.Rayner@newcastle.edu.au
fax: 61 2 49216898
O. Thas
Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium and
National Institute for Applied Statistics Research Australia, University of Wollongong, NSW 2522, Australia
e-mail: Olivier.Thas@UGent.be
D.J. Best
School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia
e-mail: John.Best@newcastle.edu.au

Contents (in place of an abstract)

Keywords:
Cholesky components, data driven testing, diagnostic components, empirical distribution function tests of fit, model selection, multinomial models, Neyman-Barton smooth tests, orthonormal functions.

Assessing Statistical Distribution Models: Goodness of fit Tests

1. Introduction

The purpose of a one-sample test of fit is to give an objective measure of how well a probability model agrees with observed data. Otherwise, as Pearson (1900, p. 171) said: 'the comparison of observation and theory in general amounts to a remark - based on no quantitative criterion.' Reviews of Pearson's paper may be found in Plackett (1983) and Rayner and Best (2009, Sect. 2.2). The multiple sample goodness of fit problem assesses whether data from a number of populations are consistent with having come from the same population. If the answer is affirmative, then the one sample question may be appropriate. We will not consider the multiple sample problem here.

In many situations it will be valuable to find out if data for a particular event are consistent with an established pattern or model. Specifically, (a) can extreme events such as flood heights, insurance losses and equity risks be modelled by the extreme value distribution, (b) are right skewed data, that might arise, for example, in life testing and reliability contexts, consistent with the Rayleigh distribution, and (c) are data, usually binomial in similar scenarios but too over-dispersed to be so in the current context, consistent with the beta-binomial? This sometimes occurs with taste-test and market research data.

Rayner and Best (2009) identify three benefits of having applied a goodness of fit test:

- a compact description of the data;
- illumination of the data generation mechanisms; and
- validation or not of distributional assumptions necessary for the application of powerful parametric procedures.

The best of modern methods add another benefit: selection of a more valid model than that originally hypothesized.

Here we discuss tests for multinomial models, tests based on the empirical distribution function (EDF), and the construction of the Neyman-Barton smooth tests. In the final section we then address some modern developments in smooth testing: diagnostics, Cholesky components and model selection. Other tests of fit have been suggested, but we shall not consider these here, except to briefly mention two. Correlation tests are sometimes suggested in conjunction with quantile-quantile or probability plots of the data. The Shapiro-Wilk test of normality, for example, may be considered a correlation test. See D'Agostino and Stephens (1986). Tests based on the empirical Laplace transform are a general class of tests and often have competitive power. See, for example, Henze and Meintanis (2002, 2012).

2. Tests for Multinomial Models

For Pearson's test, sometimes called the Pearson chi-squared test, observations O_1, O_2, …, O_m are assumed to come from m non-overlapping classes that are expected to contain E_1, E_2, …, E_m observations. The test statistic is

$$X_P^2 = \sum_{i=1}^{m} (O_i - E_i)^2 / E_i .$$

Pearson (1900) assumed that parameters of the probability model giving rise to the cell expectations were known, and showed that the asymptotic distribution of X_P^2 is the χ_{m-1}^2 distribution.

There is a considerable literature on how, when there is choice, the cells may best be constructed. The recommendation that for the χ^2 approximation to be adequate the cell expectation be at least five, goes back to Sir Ronald Fisher writing in the 1920s. This advice is now seen as extremely conservative.

Most interesting applications involve estimating parameters. Pearson incorrectly asserted that estimating parameters makes no difference to the asymptotic null distribution of the test statistic. If the cell expectations are calculated from parameters estimated efficiently from the grouped data, then Sir Ronald Fisher, again writing in the 1920s, showed that if *q* parameters are estimated from the grouped data using maximum likelihood the asymptotic distribution is χ_{m-q-1}^2. However if estimation is via maximum likelihood using the ungrouped data Chernoff and Lehmann (1954) showed that Pearson's statistic no longer had an asymptotic chi-squared distribution. See Rayner et al. (2009) for a fuller discussion of cell construction and multinomial testing when parameters need to be estimated.

There are many possible generalizations of Pearson's test. Neyman's version is based on the test statistic

$$X_N^2 = \sum_{i=1}^{m} (O_i - E_i)^2 / O_i .$$

One interesting possibility is the Cressie-Read family, based on the test statistic

$$I_\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{m} O_i \left\{ \left(\frac{O_i}{E_i}\right)^\lambda - 1 \right\}.$$

Taking $\lambda = 1$ leads to X_P^2, $\lambda = -1$ leads to a modified likelihood ratio statistic, $\lambda = 0$ is the usual likelihood ratio statistic, $\lambda = -0.5$ is the Freeman-Tukey statistic, and $\lambda = -2$ leads to X_N^2. Other choices can lead to tests with excellent properties; see Read and Cressie (1988).

3. EDF Tests

Much data is essentially continuous, and grouping it to permit the use of X_P^2 and other statistics based on the multinomial loses information and often power. A different approach that does not group the data is to compare the empirical distribution function (EDF) with the distribution function specified by the hypothesized probability distribution. The Russian mathematician Kolmogorov (1933) proposed the first such test. EDF tests of fit, especially for continuous distributions, are reviewed in D'Agostino and Stephens (1986). Tests of fit for discrete distributions using the EDF approach are considered in Choulakian et al. (1994). See also Spinelli and Stephens (1997).

If we have *n* data points, the EDF $F_n(x)$ is given by

$$F_n(x) = (\text{number of observations} \leq x)/n.$$

Tests of fit for a hypothesized distribution function, $F(x)$, can be made by looking at the difference between $F_n(x)$ and $F(x)$ for each data point. Kolmogorov (1933) suggested the statistic

$$KS = \max_x |F_n(x) - F(x)|$$

while a generally more powerful test is based on the Anderson-Darling (1954) statistic. Its calculation is illustrated in Example 1 below.

It can be shown that the EDF test statistics are linear combinations of components with decreasing weights. The components are similar to those defined in Section 4 for smooth tests. See, for example, Stephens (1974). If the null hypothesis specifies a discrete distribution where the number of classes is fixed and not too large (say five or less) we suggest calculating X_P^2 to assess goodness of fit. Otherwise, the Anderson-Darling test is a good choice, as it weights the important early components heavily and gives a powerful omnibus assessment.

3.1 Example 1: Testing Normality

Consider the following set of examination marks from 20 students:

53, 15, 70, 73, 79, 48, 91, 20, 24, 91, 87, 15, 3, 78, 78, 62, 16, 15, 20, 32.

Are these data normally distributed?

First we apply the Pearson-Fisher test with class boundaries – 12.5, 12.5, 37.5, 62.5, 87.5, 112.5 and midpoints 0, 25, 50, 75, 100. The cell counts are 1, 8, 3, 6, 2. We find X_{PF}^2 = 5.22 with p-value 0.07 using the χ_2^2 distribution. The normality assumption would seem to be marginal.

We now apply the Anderson-Darling test. We calculate $\bar{x}$ = 48.5 and s = 30.8 and then find $y_i = (x_i - \bar{x})/s$ for i = 1, …, 20, where the x_i are the 20 exam marks. Next use a computer routine or tables of the standard normal distribution to find

$$z_i = F(y_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_i} \exp(-x^2/2)dx$$

and order the z_i as $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$. Then the Anderson-Darling statistic, AD say, is

$$AD = -n - n^{-1} \sum_{i=1}^{n} (2i-1)\{\log z_{(i)} + \log z_{(n+1-i)}\},$$

which, for the above data, takes the value 0.931. From tables of critical values such as D'Agostino and Stephens (1986, Table 4.9), or using MINITAB's Normplot routine, we find a p-value of 0.014. This implies the normality hypothesis is doubtful and the marks are not well summarized by $\bar{x} \pm s$. We revisit this example later.

3.2 Example 2: Testing the Poisson Assumption

Consider the number of people in a shopping mall contributing to a charity collector during five-minute intervals:

Number of contributors	0	1	2	3	4	5	6	7	8
Number of intervals	9	27	26	17	10	8	4	3	1

The sample size is $n = 105$. Let the observed and expected counts assuming a Poisson distribution be o_j and $e_j = np_j$, respectively, for $j = 0, 1, \ldots, 8$. Pooling cells from 7 onwards to get all cell expectations greater than one, we find $X^2_{PF} = 8.62$ with p-value 0.20 using the χ^2_7 distribution. At all reasonable levels the Poisson assumption is acceptable.

Now let $z_j = \sum_{i=0}^{j} (o_i - e_i)$, in which case

$$AD = \frac{1}{n}\sum_{j=0}^{7} \frac{\hat{z}_j^2 \hat{p}_j}{\hat{h}_j(1-\hat{h}_j)}$$

where the '^' indicates we have estimated the Poisson parameter by the sample mean. We find that $AD = 1.178$, and, using Spinelli and Stephens (1997, Table 1), the *p*-value is between 0.025 and 0.050. Thus the distribution is not well fitted by a Poisson model. It seems plausible that contributors come in clusters rather than at random, as a Poisson model would imply. The Pearson-Fisher test is far less critical of the Poisson assumption than the Anderson-Darling test.

4. Neyman Smooth Tests

One problem with Pearson's chi-squared test is that it is an omnibus test, with some power against general alternative distributions to the hypothesized distribution, but with relatively poor power against important directional alternatives. Alternatives which are often of great practical importance are differences between the data and the mean, variance and, to a lesser extent, the skewness and kurtosis of the specified distribution. A test that does have good power at detecting these moment alternatives is Neyman's (1937) 'smooth' test. It has the useful property that it can readily be partitioned into *components* that give a powerful and detailed scrutiny of the data. See Rayner and Best (2009, Sect. 4.1) for a review of Neyman (1937).

The Neyman smooth tests may be applied to both categorized and uncategorized distributions and distributions that may or may not involve nuisance parameters. See Rayner and Best (1990) and Rayner et al. (2011) for overviews. Most modern authors use orthonormal functions, as did Neyman (1937), and permit nuisance parameters, contrary to Neyman (1937). The smooth tests may be used to test for very general distributions, provided orthonormal functions on the specified distribution may be constructed. Recurrence formulae for orthonormal polynomials are given for univariate distributions in Rayner et al. (2008) and for bivariate distributions in Rayner et al. (2013).

The test is based on the score statistic, and hence is weakly optimal, and the test statistic has a convenient asymptotic chi-squared distribution. In our preferred formulation components are readily available, and are asymptotically independent, each asymptotically having the standard normal distribution. The components are often the basis for well-known tests of fit in their own right, but more importantly,

may permit a detailed, convenient and informative scrutiny of the data. The order of the test (see below) is at the user's discretion, and it may be chosen to give an omnibus or directional test, or something between.

Suppose we wish to test for a probability density function $f_X(x; \beta)$, where $\beta = (\beta_1, \ldots, \beta_q)^T$ is a $q \times 1$ vector of parameters (such as μ and σ in the normal case). This null probability density function is nested in the order k probability density function

$$g_k^N(x;\theta,\beta) = C(\theta,\beta)\exp\{\sum_{i=1}^{k}\theta_i\pi_i(x;\beta)\}f_X(x;\beta)$$

(N is for Neyman) where $\theta = (\theta_1, \ldots, \theta_k)^T$ is a vector of k real parameters, $C(\theta, \beta)$ is a normalizing constant that is assumed to exist, and $\{\pi_i(x; \beta)\}$ is a set of functions orthonormal on $f_X(x; \beta)$ with $\pi_0(x; \beta) = 1$ for all x. Thus for $f_X(x; \beta)$ continuous,

$$\int_{-\infty}^{\infty}\pi_i(x;\beta)\pi_i(x;\beta)f_X(x;\beta)\mathrm{d}x = 1 \text{ for } i = j \text{ and zero otherwise.}$$

Sometimes $C(\theta, \beta)$ may not exist but the statistic we are about to define will. Using a random sample $X_1, \ldots, X_n$ we test for $f_X(x; \beta)$ parametrically, by testing H_0: $\theta = 0$ against K: $\theta \neq 0$, with β a vector of nuisance parameters. For small k the alternatives vary 'smoothly' from the null, in the sense of Neyman (1937). As an example of this, in testing for the standard normal distribution, the order one alternative is a normal distribution with a mean shift, the order two alternative is a normal distribution with both a mean shift and a variance shift, and the third order alternative is more involved, with mean, variance and skewness differences from the hypothesized normal distribution.

Neyman (1937) developed this approach for the case of no nuisance parameters. In this situation the probability integral transformation converts the problem of testing for a completely specified distribution to testing for the uniform (0, 1) distribution. Thomas and Pierce (1979) and Kopecky and Pierce (1979) considered smooth tests for general distributions involving nuisance parameters. Their tests are based on a family of alternatives specified using powers of the cumulative distribution function rather than orthogonal polynomials as Neyman (1937) used, and as a consequence they require tables of coefficients for their implementation. Our preferred formulation permits nuisance parameters and uses orthonormal functions, thereby avoiding the need for tables of coefficients. For a more detailed account of the history of the smooth tests, see Rayner and Best (2009, Sect. 1.2).

When no nuisance parameters are present, the smooth test statistic based on the model $g_k^N(x;\theta,\beta)$ is

$$S_k = V_1^2 + \ldots + V_k^2$$

in which the components V_r are given by

$$V_r = \sum_{j=1}^{n}\pi_i(X_j;\beta)/\sqrt{n}\,.$$

When testing for a distribution from an exponential family of distributions, such as the binomial, Poisson, geometric or normal, the likelihood equations for $\hat{\beta}$, the maximum likelihood estimator of β, are $\hat{V}_r = V_r(\hat{\beta}) = 0$ for $r = 1, \ldots, q$. The model $g_k^{\mathrm{N}}(x;\theta,\beta)$ results in a singular asymptotic covariance matrix for $(\hat{V}_r)$. This can be resolved by removing $\theta_1, \ldots, \theta_q$ from the model, that is, by modifying $g_k^{\mathrm{N}}(x;\theta,\beta)$ to $C(\theta,\beta)\exp\{\sum_{i=q+1}^{q+k}\theta_i\pi_i(x;\beta)\}f_X(x;\beta)$. What is happening here is that $\theta_1, \ldots, \theta_q$ are playing the same role as the elements of β, so the presence of all these parameters in the model leads to a redundancy. The test statistic is

$$\hat{S}_k = \hat{V}_{q+1}^2 + \ldots + \hat{V}_{q+k}^2$$

in which $\hat{V}_r = \sum_{j=1}^{n}\pi_i(X_j;\hat{\beta})/\sqrt{n}$.

When testing for a distribution outside of an exponential family the score test statistic is not a simple sum of squares: it is a quadratic form. This will be considered in the next section.

It may occur that in $g_k^{\mathrm{N}}(x;\theta,\beta)$ a normalizing constant $C(\theta, \beta)$ cannot be found. An alternative smooth model is due to Barton (1953, 55, 56):

$$g_k^{\mathrm{B}}(x;\theta,\beta) = \{1 + \sum_{i=1}^{k}\theta_i h_i(x;\beta)\}f(x;\beta).$$

(B is for Barton.) These densities don't require the existence of the normalizing constant, but may be negative. This negativity may be handled using adjustments such as those proposed in Gajek (1986) and in Glad et al. (2003), but this is only a problem for nonparametric density estimation and not for the smooth tests discussed here. For the Barton model the relevant previous results for the Neyman alternatives remain true.

Example 2 Revisited:
For the Poisson, the appropriate orthonormal polynomials are known as Poisson-Charlier polynomials. We have

$$\pi_r(y) = \sqrt{(\lambda^r/r!)}\sum_{j=0}^{r}(-1)^{r-j}({}^r\mathrm{C}_j)(j!)(\lambda^{-j})({}^y\mathrm{C}_j),$$

where $y = 0, 1, 2, \ldots$, from which the orthonormal polynomials to second order are

$$\pi_0(y) = 1 \text{ for all } y,\ \pi_1(y) = (y/\lambda - 1)\sqrt{\lambda},\ \pi_2(y) = \{(y(y-1)/\lambda^2 - 2y/\lambda + 1\}\lambda/\sqrt{2}.$$

Due to the estimation of λ, $\hat{V}_1 = 0$. For the data here $\hat{V}_2 = 2.12$, suggesting a model with greater dispersion than the Poisson. Note that $\hat{V}_2$ is a standardized version of Fisher's Poisson Index of Dispersion.

For count data the objective is to test for a multinomial distribution with specified cell probabilities $p_1, \ldots, p_m$. When there are no nuisance parameters this is

done by imbedding these probabilities in the smooth alternative cell probabilities

$$\pi_j = C(\theta)\exp\{\sum_{i=1}^{k}\theta_i h_{ij}\}p_j, j = 1, ..., m$$

and testing H_0: $\theta = 0$ against K: $\theta \neq 0$. In this formulation the θ_i, $i = 1, ..., k$ are real parameters, $\theta = (\theta_1, ..., \theta_k)^T$, $C(\theta)$ is a normalising constant so that $\pi_1 + \ldots + \pi_m = 1$, and the h_{rs} satisfy, for $r, s = 1, ..., m$,

$$\sum_{j=1}^{m} h_{rj}h_{sj}p_j = \delta_{rs}, \text{ with } h_{mj} = 1, j = 1, ..., m.$$

Given cell counts $N_1, \ldots, N_m$, the score test statistic when $k = m - 1$ is Pearson's $X_P^2 = \sum_{j=1}^{m}(N_j - np_j)^2/(np_j)$; more familiarly $X_P^2 = \sum_{i=1}^{m}(O_i - E_i)^2/E_i$.

A generalisation of this formulation to allow for nuisance parameters leads to the Pearson-Fisher test when the cell probabilities are estimated by maximum likelihood from the multinomial, and to the Rao-Robson (1974) test when the cell probabilities are estimated by maximum likelihood from the ungrouped data. These are thus identified as smooth tests.

It should be emphasized that using different orthonormal functions in the smooth model results in different tests that detect different alternatives. Which orthonormal functions should be chosen depends on which alternatives one hopes to most powerfully detect. If the hypothesized distribution were uniform, we could choose the Legrendre polynomials, as did Neyman (1937). If it were desirable to test for periodic alternatives we could consider using the series $\{\sin(i\pi x)\ \sqrt{2}\}$, or the Walsh functions as did Hamdan (1964). The aim is to find an orthonormal series that represents the alternatives of interest in as few terms as possible. Greater power results from doing this. A helpful R package is available at http://www.biomath.ugent.be/~othas/smooth2/Home.html.

Each smooth test for a particular distribution requires an independent study. The small sample distributions of the test statistic and its components should be investigated. Generally the approach to the asymptotic chi-squared distribution is so slow that we recommend finding p-values using the parametric bootstrap. Moreover the powers of the smooth tests should be compared with appropriate competitor tests. In general the most powerful smooth test of fit test is likely to be the sum of the squares of the first few non-zero components; $\hat{S}_{q+k}$ with $k = 2$, 3 or 4. We recommend augmenting this by the use of the components $\hat{V}_{q+1}^2, ..., \hat{V}_{q+k}^2$ in an exploratory data analysis fashion. However, in particular cases there are variations to this advice.

Sometimes the data are only available in grouped form. See Best and Rayner (2007) for testing for the grouped exponential and Best et al. (2008) for testing for the grouped normal.

Example 1 Revisited:
For the exam mark data in Section 2.1, calculate $y_i = (x_i - \bar{x})/s$. The orthonormal polynomials appropriate for testing normality are often called the Hermite-Chebyshev polynomials. We have

$$\pi_0(y) = 1 \text{ for all } y,\ \pi_1(y) = y,\ \pi_2(y) = (y^2 - 1)/\sqrt{2},\ \pi_3(y) = (y^3 - 3y)/\sqrt{6},$$
$$\pi_4(y) = (y^4 - 6y^2 + 3)/\sqrt{24},\ \pi_5(y) = (y^5 - 10y^3 + 15y)/\sqrt{120}, \text{ and}$$
$$\pi_6(y) = (y^6 - 15y^4 + 45y^2 - 15)/\sqrt{720}.$$

Given these we calculate, for the above data, $\hat{V}_1 = \hat{V}_2 = 0$, due to the estimation of μ and σ, and $\hat{V}_3 = -0.0361$, $\hat{V}_4 = -1.4656$, $\hat{V}_5 = -0.0462$, $\hat{V}_6 = 1.8907$. Note that $\hat{V}_3$ and $\hat{V}_4$ are standardized versions of the usual skewness and kurtosis coefficients. As $\hat{V}_3, \ldots, \hat{V}_6$ have asymptotic standard normal distributions, we see that $\hat{V}_6$ is near significance at the 5 percent level and that $\hat{V}_4$ and $\hat{V}_6$ are much larger than the odd-order moment statistics $\hat{V}_3$ and $\hat{V}_5$. These observations suggest an alternative probability model that is symmetric and shorter tailed than the normal. Thus a uniform distribution is a possibility so that we could summarize the data as 'evenly' spread over the range (0, 100) rather than as 48.5 ± 30.7. If the odd order $\hat{V}_r$ had dominated the even order values then a skewed alternative model is suggested, while if both odd and even $\hat{V}_r$ are large, $|\hat{V}_r| > 1.65$ say, then a symmetric alternative with longer tails then the normal, such as the Laplace, is indicated.

5. Modern Developments in smooth testing

5.1 Are components diagnostic?

When testing for a distribution from an exponential family the smooth test statistic is the sum of the squares of specified components. In particular a single component can be used to test for the distribution. There is a considerable literature in recent years on whether a test based on the rth component may be interpreted as diagnosing the failure or not of the data to be consistent with the rth moment of the specified distribution. Simulation studies demonstrate that for virtually any null distribution component tests are not diagnostic. It is now known that if V_r is significantly large, then any or all of the non-null moments up to the $2r$th could be the cause.

This issue would all be resolved if, in the score test statistics, the asymptotic covariance matrix was replaced by one that estimated the component variances and covariances consistently, under both the null and alternative hypotheses. Henze and Klar (1996), Henze (1997) and Klar (2000) worked in this vein. Unfortunately the simulation studies in Klar (2000) and Thas et al. (2009) show that convergence of these 'rescaled' components to their asymptotic limits is extremely slow, with samples as large as 10,000 required to achieve satisfactory results. Thus rescaling does not create diagnostic components and inference, for example to obtain p-values, should involve resampling methods rather than the asymptotic distributions of the statistics involved. Nevertheless, although in the finite samples that occur in practice the rescaled components may be somewhat 'tainted' by higher order moments, it is reasonable to say they are more diagnostic than the raw, unscaled components.

In the power studies we have sighted, powers based on the rescaled components are often less than those based on the unscaled components, and where power gains are achieved it is often at the expense of power loss for alternatives elsewhere in the parameter space.

Our personal preference is to not use rescaled components. When testing for

distributions within exponential families, for formal testing we prefer to use $\hat{V}_{q+1}^2 + ... + \hat{V}_{q+k}^2$ with k = 2, 3 or 4, depending on the hypothesized distribution, and use the individual components in a data analytic fashion, as in Example 1 revisited. The first significant component beyond the qth at best *suggests* that the corresponding moment is the cause of the model failure. Outside of exponential families the situation is a little different.

5.2 Generalised smooth tests and Cholesky components

The score test statistic for a smooth model is a quadratic form in the components. If the null hypothesis specifies a distribution from an exponential family of distributions then the score test statistic $\hat{S}_k$ has the appealing form of being a sum of squares of components that under the null hypothesis are asymptotically independent and asymptotically standard normal. Outside of exponential families this convenient form cannot be expected. This includes when testing for the zero-inflated Poisson, extreme-value, negative binomial, generalized Pareto (see Rayner et al. 2009, Chapter 11), gamma (De Boeck et al. 2011) and inverse Gaussian distributions (Best et al. 2012). The difficulty for these distributions is that the components are not even asymptotically uncorrelated, so the significance of one component may be associated with the significance of others.

Outside of exponential families an alternative approach is to use the generalized score test. Information about generalised score tests is given, for example, in Rippon and Rayner (2010) and Thas (2010). The generalised smooth tests use the generalised score tests, with the $q \times 1$ nuisance parameter β being estimated by solving $V_r = 0$, for $r = 1, \ldots, q$. The solution, $\widetilde{\beta}_0$, is a method of moments estimator. These estimators are not usually fully efficient, and their use often means estimating efficiency is sacrificed to gain interpretable components.

The generalised score test statistic for the Neyman and Barton models is of the form $\widetilde{V}^T\widetilde{\Sigma}^{-1}\widetilde{V}$, where $\widetilde{\Sigma}$ is a consistent estimator of the asymptotic covariance matrix of the score $\widetilde{V} = (\tilde{V}_{q+1}, \ ... \ + \tilde{V}_{q+k})^{\mathrm{T}}$. A *Cholesky decomposition* of $\widetilde{\Sigma}^{-1}$ gives $\widetilde{\Sigma}^{-1} = \mathrm{M}\mathrm{M}^T$, where M is upper triangular. Putting $\widetilde{V}^* = \mathrm{M}^T\widetilde{V} = (\widetilde{V}_{q+1}^*, \ ... \ + \widetilde{V}_{q+k}^*)^{\mathrm{T}}$ gives $\widetilde{V}^T\widetilde{\Sigma}^{-1}\widetilde{V} = \widetilde{V}^{*T}\widetilde{V}^*$. Thus the generalised score test statistic is of the form $(\widetilde{V}_{q+1}^*)^2 + ... + (\widetilde{V}_{q+k}^*)^2$. Since for most distributions of interest a multivariate central limit theorem applies and $\widetilde{V}^*$ has asymptotic covariance matrix the identity, the elements $\widetilde{V}_r^*$ of $\widetilde{V}^*$ are asymptotically independent and asymptotically standard normal.

Since M^T is lower triangular, $\widetilde{V}_r^*$ is the sum of the first r elements of $\widetilde{V}$. It follows that the previous discussion about whether or not the components $\hat{V}_r$ in the smooth test are diagnostic applies equally to the components $\widetilde{V}_r^*$ in the generalised smooth test.

The outcome of this discussion is that when testing for *any* distribution, the generalized score test and Cholesky decomposition together yield components that are equally as convenient as those resulting from the score test when testing for distributions from exponential families. Moreover, because of their construction using the Cholesky decomposition, the data and the model agree in moments up to the qth. If the rth ($r > q$) is first significant Cholesky component, this suggests the data and the

model do not agree in moments up to the rth, although the significance may be due to moments up to the $2r$th.

The only caveat on this approach is that the orthonormal polynomial of order r requires the existence of the first $2r$ moments of the null distribution. Thus, for example, using this approach it is not possible to directly test for the Cauchy, which has no moments of any order. One option would be to use the probability integral transformation to essentially test for the continuous uniform distribution on (0, 1). However the resulting components are then difficult to interpret in terms of the original null hypothesis.

5.3 Data driven testing

One of the difficulties associated with the definitions of $g_k^{\mathrm{N}}(x;\theta,\beta)$ and $g_k^{\mathrm{B}}(x;\theta,\beta)$ is the determination of the order k. In finite samples mis-specifying the order k can seriously affect the power of the test. If k is chosen to be large, the test will give good protection against a wide range of alternatives. However, the power to detect any particular alternative in the parameter space will decrease as k increases. Conversely, if k is chosen to be small, the test will be quite directional, giving good protection against a small set of alternatives, and none against all others.

The idea behind data-driven smooth tests is to let the data make the decision about the order. See Ledwina (1994), Kallenberg and Ledwina (1995, 1997a, 1997b) and Inglot et al. (1997). The order is chosen to optimize a specified criterion. If L_k is the likelihood of a random sample of size n from a distribution that is an alternative of order k, it would be natural to maximize L_k by choice of k in some set, say $\{1, 2, \ldots, d\}$, where d is specified before sighting the data. Here d is the maximum order one is prepared to accept. However this procedure would simply choose $k = d$. A penalty term is needed to discourage complexity. The Bayesian Information Criterion (BIC)

$$\mathrm{BIC}_k = -2\log L_k + k \log n$$

is appropriate, but the computationally easier modified BIC, $\hat{S}_k + k \log n$ is often preferred. The optimal order, K say, is taken to be the smallest order that maximizes BIC_k. The test statistic is then chosen to be the score or generalised score statistic involving the first K components. This reduces to the sum of squares of the first K components or Cholesky components, as previously defined. An alternative option is to use Aikaike's information criterion (AIC)

$$\mathrm{AIC}_k = -2\log L_k + 2k,$$

or its modified form $\hat{S}_k + 2k$. BIC penalizes complex models more heavily than AIC. The point is that many different model selection rules are possible. In different circumstances different rules will be appropriate.

Since the order is no longer a predetermined constant but a random variable, the test statistic $\hat{S}_K$ is no longer asymptotically χ^2 distributed, even when testing for distributions from exponential families. Although asymptotic distribution theory has been developed for some of the data-driven tests, critical values and p-values are best determined using resampling methods.

These data-driven smooth tests have competitive power. Inasmuch as they give protection against mis-specifying the order, it would be unreasonable to expect

them to have greater power than all smooth tests of specified order. If there is any vagueness in the class of alternatives one hopes to best detect, data-driven smooth tests are an excellent choice.

A more flexible method and one of wider applicability, but one which does not result in omnibus consistent tests, is given in Claeskens and Hjort (2004). They fix the maximal order, and they also consider subset selection. This means that for a given maximal order, say d, the model selection criterion can select any subset of indexes from $\{1, \ldots, d\}$ and the data-driven test statistic is then built from the corresponding components V_j with j in the selected index set.

5.4 Model selection

Suppose now that the order k of a smooth alternative has been determined, and a smooth or generalized smooth test of this order applied. If this test statistic or indeed, any of its components is significant, what can be said about the true model? The moment interpretation of the components may not provide helpful insight. The insignificant θ_r in the order k alternative could be replaced by zero, significant θ_r by their method of moments estimators $\tilde{\theta}_r$, and β by its method of moments estimator under the null hypothesis, $\tilde{\beta}_0$. This approach seems intuitively reasonable, but are there better options?

In Rayner et al. (2009, Chapter 10) two approaches are described: model selection through hypothesis testing and model selection using model selection criteria. Here we focus on the first of these. First suppose $S_h = \{1, 2, \ldots, d\}$ is an index set called the *horizon*. We may use either Neyman or Barton models, but in iterative work it is far more convenient to work with the Barton smooth models

$$g_S^B(x;\theta_S,\beta) = \{1 + \sum_{i\in S} \theta_i h_i(x;\beta)\} f(x;\beta)$$

in which $S \subset S_h$, $\theta_S = \{\theta_i: i \in S\}$ and $\{h_i(x; \beta)\}$ is a set of functions orthonormal on $f(x; \beta)$. The nonpositivity of $g_S^B(x;\theta_S,\beta)$ is not an issue for score tests. Clearly $g_{S_h}^B(x;\theta_{S_h},\beta)$ is the most complex model we are prepared to accept. A horizon of four clearly limits the maximal order more than one of 44 does.

We describe model selection through hypothesis testing, which is similar to the familiar forward selection and backward elimination techniques used in regression analysis. In forward selection at the uth step the model is $g_S^B(x;\theta_S,\beta)$ and we consider whether or not to add a single θ_i term to the model, where $\theta_i \in \theta_{S_u}$, for every possible θ_i not already in the model. A slightly modified score test is derived to test each of these hypotheses. If any are significant at a predetermined level, then the θ_i corresponding to the most significant test is added to the model. Backward elimination is similar, with the least significant θ_i being eliminated from successively reduced models until only significant terms remain in the model. The score test statistics change at each iteration.

It has previously been argued that when a data-driven smooth test rejects the null hypothesis it may be informative to investigate the selected components, even though the diagnostic property may not be guaranteed. Another way forward, however, is to plot the density estimate that corresponds to the selected model. The Barton model is particularly convenient for this purpose. Suppose that $\hat{S}$ contains the

indexes of the selected components, and the appropriate nonparametric density estimate is given by

$$g^B_{\hat{S}}(x;\tilde{\theta}_{\hat{S}},\tilde{\beta}_0) = \{1 + \sum_{i\in\hat{S}} \tilde{\theta}_i h_i(x;\tilde{\beta}_0)\} f(x;\tilde{\beta}_0).$$

This density estimate is referred to as the *improved density estimate*, but it can only be considered as a genuine density estimate after correcting for the non-positivity, using the methods proposed in Gajek (1986) and Glad et al. (2003). In Rayner et al. (2009, Chapter 10) we suggest using the plot of the improved density as the basis for the formulation of the conclusions. Since the improved density is merely a graphical representation of the information contained in the data-driven smooth test statistic, the conclusions will be consistent.

6. Examples and Conclusion

Example 3: Angus Data

Angus (1982) gives 20 operational lifetimes in hours:

6278, 3113, 5236, 11584, 12628, 7725, 8604, 14266, 6125, 9350,
3212, 9003, 3523, 12888, 9460, 13431, 17809, 2812, 11825, 2398.

In testing the null hypothesis that the data follow the exponential distribution, the Chernoff-Lehmann test with four classes, the Anderson-Darling, Cramer-von Mises and Shapiro-Wilk tests give bootstrap p-values 0.01, 0.01, 0.01 and 0.00 respectively. It would appear that the data are not consistent with exponentiality, as Angus (1982) also found. See Figure 1.

We also find $\hat{S}_4 = \hat{V}_2^2 + \ldots + \hat{V}_5^2 = 8.8$ with χ^2_4 p-value is 0.07 and bootstrap p-value based on 1,000 bootstrap runs of 0.02. In addition $\hat{V}_2 = -1.7$, $\hat{V}_3 = -1.9$, $\hat{V}_4 = -1.4$, $\hat{V}_5$ –0.8. The corresponding asymptotic two-sided p-values are 0.10, 0.06, 0.16 and 0.44, with bootstrap p-values based on 1,000 bootstrap runs 0.00, 0.00, 0.04 and 0.22 respectively. Here and elsewhere use of the asymptotic null distribution of the components and the test statistic itself can be misleading. In general it is more desirable to use bootstrap p-values

The first two components contribute most to $\hat{S}_4$, suggesting at least dispersion and skewness departures from what would be expected under the exponential model. However the cause of the significance of the order two component could be due to the third and fourth moments of the true distribution differing from those of the exponential; both corresponding components have p-values less than 5%. Similarly the significance of the order three component could be due to moments of the true distribution up to the sixth, and at this point the order six component hasn't been assessed.

Now suppose an horizon of {2, …, 6} is taken; order one is omitted because the rate is a nuisance parameter and fills the same role as θ_1. In forward selection at the first step the test statistics corresponding to θ_2 to θ_5 are all significant at the 5% level, but that corresponding to θ_3 is most significant, so that term is included in the model, and this new model assessed. This next assessment finds the term corresponding to θ_4 should be included in the model. A new assessment shows that no further terms are significant, and assessment stops.

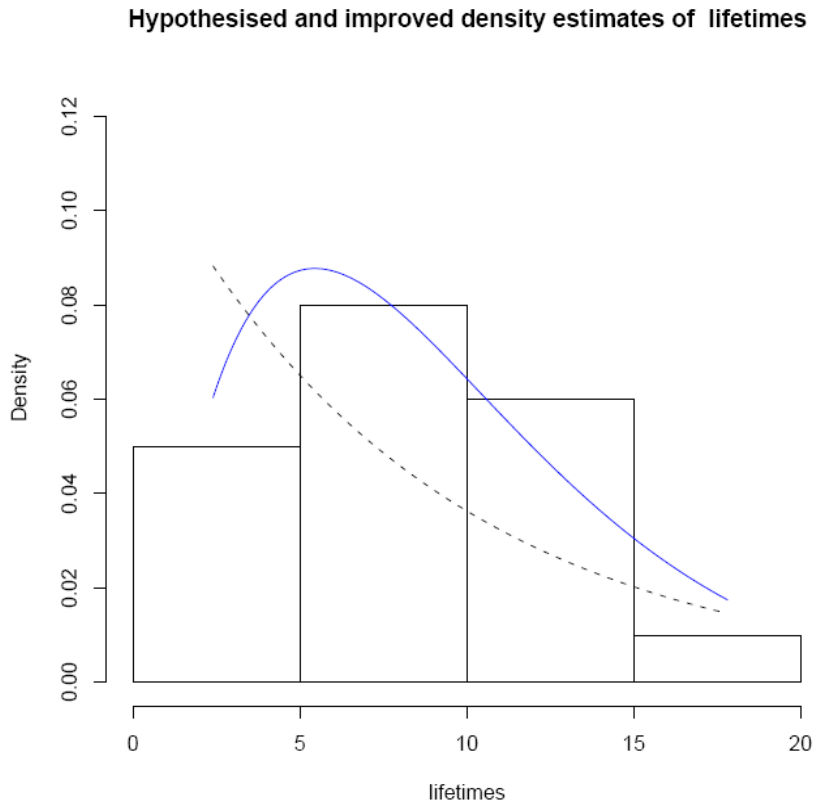

Figure 1: Histogram, density estimates of the hypothesized exponential distribution (dashed line) and the improved density (full line) for the Angus lifetimes data. Lifetimes are given in thousands of hours. The counts in the four classes are 5, 8 6 and 1.

In backward elimination the initial model includes θ_2 to θ_6 and successively eliminates first θ_6, then θ_5, then θ_2, and then finds that no further terms warrant removal. Although it need not be the case in general, both forward selection and backward elimination agree here.

Figure 1 shows the hypothesized and the improved density estimates. From this it may be concluded that the true density is probably not a monotonically decreasing density such as the hypothesised exponential density. The improved density suggests that there is a mode close to 5 and that the distribution is skewed with a tail toward the larger lifetimes.

Statistical model assessment is at the heart of good statistical practice, and is the genesis of modern statistics. In general we suggest using the Anderson-Darling test as an omnibus test. For a more detailed scrutiny of the data we recommend testing using the smooth or generalized smooth tests with uncorrelated components. These tests generally have competitive power and the uncorrelated components facilitate exploratory data analysis. In exploratory data analysis a significant rth component suggests non-zero θ_r and hence the data and the null model differ in moments up to the rth. However with arbitrary alternatives moments up to the $2r$th may be the cause of the model failure. Model selection techniques may be used to construct smooth models consistent with the data. It is more informative and accurate to base exploratory data analysis on these models and their graphical representations than on

the components themselves. However these techniques may use many criteria, so there are many possible density plots. Each selected model has implications about which moments may be the cause of model failure, but it is important to remember that this is exploratory data analysis, and perhaps it is the envelope of possible densities that gives insights rather than any particular choice.

Bibliography

Angus J E 1982 Goodness-of-fit tests for exponentiality based on a loss-of-memory type functional equation. *Journal of Statistical Planning & Inference* **6**: 241-251

Anderson T W, Darling D A 1954 A test of goodness of fit. *Annals Mathematical Statistics* **23**: 193-212

Barton D E 1953 On Neyman's smooth test of goodness of fit and its power with respect to a particular system of alternatives. *Skandinavisk Aktuarietidskr* **36**: 24-63

Barton D E 1955 A form of Neyman's Ψ^2 test of goodness of fit applicable to grouped and discrete data. *Skandinavisk Aktuarietidskr* **38**: 1-16

Barton D E 1956 Neyman's Ψ^2 test of goodness of fit when the null hypothesis is composite. *Skandinavisk Aktuarietidskr* **39**: 216-245

Best D J, Rayner J C W 2007 Chi-squared components as tests of fit for the grouped exponential distribution. *Computational Statistics & Data Analysis*: **51**, 3946-3954

Best D J, Rayner J C W, Thas O 2008 X^2 and its components as tests of normality for grouped data. *Journal of Applied Statistics* **35**(5): 481-492

Best D J, Rayner J C W, Thas O 2012. Comparison of some tests of fit for the inverse Gaussian distribution. *Advances in Decision Sciences*, Article ID 150303, 9 pages, DOI:10.1155/2012/150303

Boulerice B, Ducharme G R 1995 A note on smooth tests of goodness of fit for location-scale families. *Biometrika* **82**: 437-8

Carolan A M, Rayner J C W 2001 One sample score tests for the location of modes of nonnormal data. *Journal of Applied Mathematics and Decision Sciences* **5**(1): 1-19

Chernoff H, Lehmann E L 1954 The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals Mathematical Statistics* **25**: 579-86

Claeskens G, Hjort N 2004 Goodness of fit via non-parametric likelihood ratios. *Scand J Statist* **31**: 487–513

Choulakian V, Lockhart R A, StephensM A 1994 Cramer-Von Mises statistics for discrete distributions. *Canadian Journal of Statistics* **22**: 125-37

D'Agostino R B, Stephens M A 1986 Goodness of fit *Techniques*. Marcel Dekker, New York

De Boeck B, Thas O, Rayner J C W, Best, D J 2011 Smooth tests for the gamma distribution. *Journal of Statistical Computation & Simulation* **81**, 843-855

Gajek G 1986 On improving density estimators which are not bona fide functions. *Annals of Stat*istics **14**: 1612–18

Glad I, Hjort N, Ushakov N 2003 Correction of density estimators that are not densities. *Scandinavian Journal of Statist*istics **30**: 415–27

Hamdan MA 1964 A smooth test of goodness of fit based on the Walsh functions. *Australian Journal of Statistics* **6**: 130-6

Henze N 1997 Do components of smooth tests of fit have diagnostic properties? *Metrika* **45**: 121-130

Henze N, Klar B. 1996 Properly rescaled components of smooth tests of fit are diagnostic. *Australian & New Zealand Journal of Statistics* **38**: 61-74

Henze N, Meintanis S 2002 Tests for exponentiality based on the empirical Laplace transform. *Statistics* **36** 147-161

Henze N, Meintanis S 2012 Goodness-of-fit tests for the gamma distribution based on the empirical Laplace transform. *Communications in Statistics-Theory and Methods* 41: 1543-1556

Klar B 2000 Diagnostic smooth tests of fit. *Metrika* **52**: 237-252

Inglot T, Kallenberg W, Ledwina T 1997 Data driven smooth tests for composite hypotheses. *Annals of Statistics* **25**: 1222-1250

Kallenberg W, Ledwina T 1995 Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. *Annals of Statistics*: **23**: 1594-1608

Kallenberg W, Ledwina T 1997a Data driven smooth tests for composite hypotheses: Comparison of powers. *J Statist Comput Simul*, **59**: 101-121

Kallenberg W, Ledwina T 1997b Data driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association* **92**: 1094-1104.

Ledwina T 1994 Data driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, **89**: 1000-1005

Kolmogorov A N 1933 Sulla determinazione empirica di una legge di sitribuzione. *Giornale dell' Instituto Italiano desli Attuari* **4**: 83-91

Kopecky K J, Pierce D A 1979 Efficiency of smooth goodness-of-fit tests. *Journal of the American Statistical Association* **74**: 393-7

LaRiccia V N 1991 Smooth goodness of fit tests: a quantile function approach. *Journal of the American Statistical Association* **86**: 427-31

Neyman J 1937 'Smooth' test for goodness of fit. *Skandinavisk Aktuarietidskr* **20**: 150-99

Pearson K 1900 On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophy Magazine*, 5th series **50**: 157-75

Pena E A 1998 Smooth goodness of fit tests for the baseline hazard in Cox's proportional hazards model. *Journal of the American Statistical Association* **93**: 673-92

Plackett R L 1983 Pearson Karl and the chi-squared test. *International Statistical Review* **51**: 59-72

Rao K C, Robson D S 1974 A chi-square statistic for goodness-of-fit tests with the exponential family. *Communication in Statistics*. **3**: 1139-53

Rayner, J.C.W. and Best, D.J. (1990). Smooth tests of goodness of fit: an overview. *International Statistical Review*, **58**, 1, 9-17

Rayner J C W, Best D J, Mathews K L 1995 Interpreting the skewness coefficient. *Communications in Statistics-Theory and Methods* **24**(3): 593-600

Rayner J C W, Rayner G D 1988 *S*-sample smooth goodness of fit tests: Rederivation and Monte Carlo assessment. *Biometrical Journal* **40**: 651-63

Rayner, J C W, Thas O. and Best D J 2009 *Smooth Tests of Goodness of Fit: Using R* (2nd ed.). Wiley, Singapore

Rayner, J.C.W., Thas, O. and Best, D J 2011 *Smooth tests of goodness of fit*. *Wiley Interdisciplinary Reviews: Computational Statistics*. **3**(5): 97–406,

September/October 2011. DOI: 10.1002/wics.171
Rayner, J.C.W., Thas, O. and De Boeck, B. (2008). A generalised Emerson recurrence relation. *Australian and NZ Journal of Statistics* **50**(3): 235-240
Rayner, J C W, Thas, O, Pipelers, Peter and Beh, Eric J (2013) Calculating bivariate orthonormal polynomials by recurrence *Australian & NZ Journal of Statistics* **55**(1): 15-24
Rippon Paul, Rayner, J C W 2010 Generalised score and Wald tests. *Advances in Decision Sciences* Article ID 292013, 8 pages,. doi:10.1155/2010/292013.
Read T R C, Cressie N A C 1988 *Statistics for Discrete Multivariate Data*. Springer-Verlag, New York
Scholz F W, Stephens M A 1987 K-sample Anderson-Darling tests. *Journal of the American Statistical Association*. **82**: 918-24
Spinelli J J, Stephens M A 1997 Cramer-Von Mises tests of fit for the Poisson distribution. *Canadian Journal of Statistics* **25**: 257-68
Stephens M A 1974 Components of goodness of fit statistics. *Annals of the Institute of Henri Poincare* **10**: 37-54
Thas O (2010) *Comparing distributions*. Springer, New-York
Thas O, Rayner, J C W, Best, D J, De Boeck B. 2009 Informative statistical analyses using smooth goodness of fit tests. *Journal of Statistical Theory and Practice*, **3**: 705-733
Thomas D R, Pierce D A 1979 Neyman's smooth goodness-of-fit test when the hypothesis is composite. *Journal of the American Statistical Association* **74**: 441-5

J. C. W. Rayner, O. Thas and D. J. Best

6976 words