JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 6(4), 241–254

Components of the Pearson-Fisher Chi-squared Statistic

G.D. RAYNER†
National Australia Bank and Fellow of the University of Wollongong
Institute of Mathematical Modelling and Computational Systems
University of Wollongong, Wollongong, NSW 2522, Australia

Abstract. The Pearson-Fisher chi-squared test can be used to evaluate the goodness-of-fit of categorized continuous data with known bin endpoints compared to a continuous distribution, in the presence of unknown (nuisance) distribution parameters. Rayner and McAlevey [11] and Rayner and Best [9],[10] demonstrate that in this case, component tests of the Pearson-Fisher chi-squared test statistic can be obtained by equating it to the Neyman smooth score test for a categorized composite null hypothesis under certain restrictions. However, only Rayner and McAlevey [11] provide even brief details as to how these restrictions can be used to obtain any kind of decomposition. More importantly, the relationship between the range of possible decompositions and the interpretation of the corresponding test statistic components has not previously been investigated. This paper provides the necessary details, as well as an overview of the decomposition options available, and revisits two published examples.

Keywords: categorized composite null hypothesis, chi-squared statistic decomposition, Neyman smooth test, Pearson-Fisher.

1. Introduction

The chi-squared goodness-of-fit test is essentially an *omnibus* test, but in many situations it decomposes into asymptotically independent *component* tests that can provide useful and illuminating interpretations of the way in which the data fit (or do not fit) the hypothesized distribution. The nature of the particular decomposition can be related to a particular orthogonal scheme: different schemes correspond to the component tests optimally detecting various kinds of departures from the null hypothesis distribution. In the absence of a compelling reason otherwise, for interpretable tests, Rayner and Best [9] recommend using orthogonal polynomials to produce component tests that detect moment departures from the null hypothesis (that is, the first component detects a mean shift, the second a change in scale, the third a change in skewness, etc.).

† Requests for reprints should be sent to G.D. Rayner,Institute of Mathematical Modelling and Computational Systems University of Wollongong, Wollongong, NSW 2522, Australia.

In their book, Rayner and Best [9] decompose the chi-squared goodness-of-fit test statistic for both simple and composite goodness-of-fit hypotheses about uncategorized continuous data. Here, the terms simple and composite refer, respectively, to the absence or presence of (nuisance) distributional parameters that must be estimated. However, for categorized continuous data where the bin endpoints are known (see for example the datasets in Tables 1 and 2), only the simple hypothesis case (of no nuisance parameters) has been thoroughly explored. In all of the above three cases this is done by equating the chi-squared statistic to an appropriate decomposed Neyman smooth score test statistic based on any chosen orthonormal scheme.

For categorized continuous data where the bin endpoints are known and where nuisance parameters are present (referred to in Rayner and Best's book as the categorized composite case), only *restrictions* on the decomposed Neyman smooth score test statistic are provided to enable the decomposition to be performed. No method of constructing the test statistic components satisfying these restrictions is provided other than a comment that such a decomposition can be "based on a Helmert matrix".

In Rayner and McAlevey [11] and Rayner and Best [10] some examples are provided that use the categorized composite constructions outlined in Rayner and Best [9]. Even here however, only a set of restrictions are provided, and although the component test statistics have evidently been calculated in these examples, the method used to do so is briefly presented as a linear programming problem, and not discussed in any detail. In fact, the method used to construct the component test statistics here results in the interpretation of the r-th component as some kind of unknown contrast in the first $r + constant$ data cells.

The basic problem is that for the categorized composite case (categorized continuous data where the bin endpoints are known), the relationship between a particular decomposition of the Pearson-Fisher chi-squared statistic and the corresponding orthogonal scheme has not yet been made clear. This stands in contrast to Rayner and Best's [9] informative decompositions of the chi-squared goodness-of-fit test statistic for uncategorized simple, categorized simple, and uncategorized composite null hypotheses.

This paper addresses each of the above deficiencies. First, section 2 introduces the problem along with the current state of the literature. Section 3 describes my method for constructing components of the Pearson-Fisher chi-squared test according to any chosen orthonormal scheme. The examples of Rayner and McAlevey [11] and Rayner and Best [10] are revisited in section 5, which also discusses the difficulties involved in obtaining the relevant MLE's.

2. Notation

Consider n data points gathered from a continuous distribution that are grouped into m categories specified by the $m+1$ bin endpoints $c_0 < c_1 < \ldots < c_{m-1} < c_m$. These data are probably better described as *grouped continuous* rather than strictly *categorical*. See for example the datasets in Tables 1 and 2. Because the category bin endpoints are available we can express the m null hypothesis cell probabilities $p = (p_1, \ldots, p_m)^T$ in terms of $\beta = (\beta_1, \ldots, \beta_q)^T$, the q unspecified (nuisance) parameters of the null hypothesis distribution.

To calculate the chi-squared test statistic, the m null hypothesis cell probabilities $p(\beta) = (p_1(\beta), \ldots, p_m(\beta))^T$ must first be estimated. Using maximum likelihood estimation (MLE) methods (see section 5.1) to do so will result in known asymptotic test statistic distributions, and the resulting chi-squared statistic will be the Pearson-Fisher chi-squared statistic X^2_{PF}. It also simplifies the problem considerably (if, as is usually the case, $q < m$), requiring only MLE's for the q nuisance parameters β to be found, as the MLE's for the cell probabilities are then given by $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m)^T = p(\hat{\beta})$.

Following the notation of Rayner and Best ([9], chapter 7), define $\hat{D} = \text{diag}(\hat{p}_1, \ldots, \hat{p}_m)$ and W to be the q by m matrix with components the derivatives $W_{u,j} = (\partial p_j / \partial \beta_u)$ evaluated at $p = \hat{p}$ (where $u = 1, \ldots, q$ and $j = 1, \ldots, m$). Let

$$F = \hat{D}^{-1} - \hat{D}^{-1}\hat{p}\hat{p}^T\hat{D}^{-1} - \hat{D}^{-1}\hat{W}^T\left(\hat{W}\hat{D}^{-1}\hat{W}^T\right)^{-1}\hat{W}\hat{D}^{-1}. \tag{1}$$

Now define $\hat{H}$ to be the $(m-q-1) \times m$ matrix that satisfies

$$\hat{H}^T\hat{H} = F \tag{2}$$

subject to the three restrictions

$$\hat{H}\hat{p} = 0,\ \hat{H}\hat{W}^T = 0 \text{ and } \hat{H}\hat{D}\hat{H}^T = I_{m-q-1}. \tag{3}$$

Let $N = (N_1, \ldots, N_m)^T$ be the number of observations in each of the m categories and $n = \sum N$. Then, with $\hat{H}$ as specified and $\hat{V} = \hat{H}N/\sqrt{n}$, Rayner and Best's ([9], p.116) Neyman smooth score test statistic becomes the same as the Pearson-Fisher chi-squared statistic $X^2_{PF} = \hat{V}^T\hat{V}$, and can be decomposed into $m-q-1$ asymptotically independent χ^2_1 distributed component test statistics given by squared elements of the vector $\hat{V}$. See Rayner and Best ([9], p.116) for details.

Clearly, the difficulty here is constructing $\hat{H}$ so that equations (1), (2), and (3) are satisfied, and the resulting components are usefully interpretable. The literature is rather uninformative on this matter. Rayner

and Best ([9], p.116) mention that $\hat{H}$ can be "...based on a Helmert matrix" in some way, though how is not clear. Other references (Rayner and Best, [9]) refer to Rayner and McAlevey's [11] approach, though they acknowledge that it is not unique. Their construction method uses the fact that the r-th row of $\hat{H}$ is subject to $q+r+1$ constraints due to the requirements of equation (3) and an additional introduced restriction of orthonormality. There are m elements to be solved for in each of the $m-q-1$ rows of $\hat{H}$, and elements of the r-th row after the $(q+r+1)$-th are taken to be zero. The problem then reduces to a linear programming task. The resulting interpretation of the vector of component test statistics $\hat{V}$ due to this construction method is that "...$\hat{V}_r$ is a *contrast in the first* $q+r+1$ *cells*".

3. Constructing $\hat{H}$

This section uses the restrictions in equations (2) and (3) to develop a new method for obtaining $\hat{H}$ from any given square orthonormal matrix of the appropriate size. This allows all possible $\hat{H}$'s to be considered as corresponding to a particular choice of orthonormal matrix. The desired choice of $\hat{H}$ can then be made, by first selecting the appropriate orthonormal scheme.

Rayner and Best ([9], proof of Corollary 7.1.1, p.114) prove that for

$$K = \hat{D}^{-1/2}\hat{p}\hat{p}^T\hat{D}^{-1/2} + \hat{D}^{-1/2}\hat{W}^T(\hat{W}\hat{D}^{-1}\hat{W}^T)^{-1}\hat{W}\hat{D}^{-1/2}, \qquad (4)$$

then $I_m - K$ has rank $m-q-1$. Since $I_m - K$ has rank $m-q-1$ then $F = \hat{D}^{-1/2}(I_m - K)\hat{D}^{-1/2}$ also has this rank, and therefore possesses $m-q-1$ non-zero eigenvalues.

Obtain the $m-q-1$ non-zero eigenvalues $\lambda_1, \ldots, \lambda_{m-q-1}$ (arranged in non-decreasing order) and normalized eigenvectors $f_1, \ldots, f_{m-q-1}$ of F. Define the $m \times m$ matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{m-q-1}, 0, \ldots, 0)$. Also let $U_1 = (f_1, \ldots, f_{m-q-1})$ and $U = (U_1, U_2)$ where U_2 is an arbitrary $m \times (q+1)$ matrix of normalized column vectors chosen to be orthogonal to $f_1, \ldots, f_{m-q-1}$. One possible choice for U_2 is a Gram-Schmidt orthnormalization of the columns of $(\hat{p}, \hat{W}^T)$, since by equation (8) these are orthogonal to F, although any choice of U_2 is equivalent as long as U is orthonormal. With U and Λ defined in this way, we have $FU = U\Lambda$ and $UU^T = U^TU = I_m$ so that

$$U^TFU = \Lambda \text{ and } F = U\Lambda U^T. \qquad (5)$$

Note that when actually computing the decomposition in equation (5), it is better to replace $\hat{D}^{-1}\hat{p}\hat{p}^T\hat{D}^{-1}$ (the second term in F) with its equivalent,

an $m \times m$ matrix of ones. In addition, this construction can sometimes be more efficiently computed using the singular-value decomposition of F (see Datta, [3]).

Define the $m \times m$ matrix $\Lambda^* = \text{diag}(\lambda_1^{-1}, \ldots, \lambda_{m-q-1}^{-1}, 0, \ldots, 0)$. Let J_i $(i = 1, \ldots, m)$ be the $m \times m$ matrix that is zero everywhere except for the first i diagonal elements, which are unity. Then $\Lambda^*\Lambda = J_{m-q-1}$ and defining the $(m-q-1) \times m$ matrix $G = \hat{H}U\Lambda^{*1/2}$ means that equations (2) and (5) give

$$G^T G = \Lambda^{*1/2} U^T \hat{H}^T \hat{H} U \Lambda^{*1/2} = J_{m-q-1}. \tag{6}$$

Because G is an $(m - q - 1) \times m$ matrix, this equation tells us that the first $m - q - 1$ columns of G must be orthonormal vectors, which are the only non-zero elements of G. Then for any given G satisfying equation (6) a corresponding

$$\hat{H} = G\Lambda^{1/2} U^T \tag{7}$$

is defined. Note that $\hat{H}$ does not depend on the particular U_2 chosen.

For distinct eigenvalues $\lambda_1, \ldots, \lambda_{m-q-1}$ and a given G, U_1 (and therefore $\hat{H}$) is uniquely defined up the signs of the eigenvectors $f_1, \ldots, f_{m-q-1}$. Without loss of generality, assume the sign of each eigenvector's leading element is positive. Because a change in the sign of the i-th eigenvector f_i is equivalent to changing the sign of the i-th column of G $(i = 1, \ldots, m-q-1)$, each $\hat{H}$ corresponds uniquely to a given G, and therefore uniquely to the given orthonormal scheme selected.

This $\hat{H}$ satisfies the restrictions in equation (3). From Rayner and Best ([9], p.116) $\hat{p}^T \hat{D}^{-1} \hat{p} = 1$ and $W\hat{D}^{-1}\hat{p} = 0$ so that $\hat{p}^T \hat{D}^{-1} W^T = 0$. Using these expressions it is easy to show that

$$\hat{p}^T F = 0 \text{ and } \hat{W}F = 0. \tag{8}$$

This means that both $\hat{p}$ and the q rows of $\hat{W}$ are orthogonal to all columns of F, so are orthogonal to each of $f_1, \ldots, f_{m-q-1}$ and we have $U_1^T \hat{p} = 0$ and $U_1^T \hat{W}^T = 0$. Therefore, since only the first $m - q - 1$ columns of $\Lambda^{1/2}$ are nonzero, we have

$$\hat{H}\hat{p} = G\Lambda^{1/2} U^T \hat{p} = G\Lambda^{1/2} \begin{pmatrix} U_1^T \hat{p} \\ U_2^T \hat{p} \end{pmatrix} = 0$$

and

$$\hat{H}\hat{W}^T = G\Lambda^{1/2} U^T \hat{W}^T = G\Lambda^{1/2} \begin{pmatrix} U_1^T \hat{W}^T \\ U_2^T \hat{W}^T \end{pmatrix} = 0.$$

Also, from equation (1) and the transpose of the expressions in (8),

$$F^T DF = F^T \left(I_m - \hat{p}\hat{p}^T \hat{D}^{-1} - \hat{W}^T \left(\hat{W} \hat{D}^{-1} \hat{W}^T \right)^{-1} \hat{W} \hat{D}^{-1} \right) = F^T.$$

This expression, along with equations (5), (7) and the fact that $U^T U = I$, shows that $\hat{H}$ satisfies the final restriction since

$$\begin{aligned} & \hat{H}\hat{D}\hat{H}^T = G\Lambda^{1/2} U^T D U \Lambda^{1/2} G \\ = \; & G\Lambda^{*1/2} U^T (U\Lambda U^T D U \Lambda U^T) U \Lambda^{*1/2} G^T \\ = \; & G\Lambda^{*1/2} U^T (F^T D F) U \Lambda^{*1/2} G^T = G\Lambda^{*1/2} U^T (F^T) U \Lambda^{*1/2} G^T \\ = \; & G\Lambda^{*1/2} U^T (U\Lambda U^T) U \Lambda^{*1/2} G^T = GG^T = I_{m-q-1}. \end{aligned}$$

The definition of G (see equation (6)) and equation (7) together describe how to obtain $\hat{H}$: the first $m-q-1$ columns of G are chosen to be any square orthonormal matrix, the remaining elements of G are zero. Then $\hat{H}$ is formed from this G using equation (7). But what orthonormal matrix should be used?

4. Interpreting the Component Statistics

Clearly the composite hypothesis case (when nuisance parameters are present) should generalize the simple hypothesis case (where the null hypothesis distribution is completely defined). In the simple hypothesis case there are no nuisance parameters so $q = 0$, $\hat{W}$ does not exist, and F only consists of the first two terms in equation (1). In the simple case, Rayner and Best ([9], section 5.3, p.63; and appendix 3, p.147) assign the r-th row of $\hat{H}$ (corresponding to the component test statistic V_r, $r = 1, \ldots, m-q-1$) to be values of the orthogonal polynomials $h_r(x_i)$ evaluated at $x_i = 0, \ldots, m-1$ and defined by the equations

$$\sum_{i=1}^{m} h_r(x_i) p_i = \left\{ \begin{array}{l} 1, \; r = 0 \\ 0, \; r \neq 0 \end{array} \right\} \text{ and } \sum_{i=1}^{m} h_r(x_i) h_s(x_i) p_i = \left\{ \begin{array}{l} 1, \; r = s \\ 0, \; r \neq s \end{array} \right\} \quad (9)$$

where $h_0(x) = 1$ and $h_r(x)$ is a polynomial of degree r. The component test statistic V_r arising from this choice can be written as $V_r = (HN)_r / \sqrt{n} = \sum_{i=1}^{m} h_r(x_i) N_i / \sqrt{n}$. This V_r is said to correspond to departures of the r-th moment of the categorized data from the r-th moment of the hypothesized distribution.

It is clear that for $r = 1, \ldots, m-q-1$ the r-th row of G corresponds to the same row of H and thus to the component test statistic V_r. It is desirable to keep the moment interpretations for V_r, but this is difficult.

This question governs which orthonormal scheme for G is selected, and warrants further investigation.

I recommend selecting the r-th row of G to be values of a (linear combination of the p_j weighted orthogonal) polynomial. The particular linear combination is chosen to ensure compatibility with the simple hypothesis case (where no nuisance parameters are present).

For H_0 an $(m-q-1) \times m$ matrix, choose its rows to be values of the orthogonal polynomials $h_r(x_i)$ $(r \neq 0)$ evaluated at $x_i = 0, \ldots, m-1$ and defined by the equation (9) above (for details see Emerson, [4]). Define

$$F_0 = H_0^T H_0 \tag{10}$$

Obtain U_0, Λ_0 and Λ_0^* from this F_0 in the same way as U, Λ and Λ^* are obtained from F (see equation (5)). Then let

$$M = U_0 \Lambda_0^{*1/2} \Lambda^{1/2} U^T. \tag{11}$$

Since we know what H should be in the simple case (that is, H_0), we can obtain F in the simple case (F_0) and calculate the corresponding G ($G_0 = H_0 U_0 \Lambda_0^{*1/2}$). Then, using this G to find H in the composite case gives $H = G_0 \Lambda^{1/2} U^T = H_0 U_0 \Lambda_0^{*1/2} \Lambda^{1/2} U^T = H_0 M$ and $V = HN/\sqrt{n} = H_0 M N/\sqrt{n}$ provides the desired component test statistics.

For the simple case (no nuisance parameters present) $q = 0$ and $F = F_0$ so that $H = H_0 M = H_0$ (though M is not the identity matrix) and this method gives the same results as the simple case method of Rayner and Best ([9], chapter 5). Note that when elements of p are very close or the same, obtaining U or U_0 is a numerically sensitive operation. For this reason I recommend ensuring that the elements of p differ by a small amount (I use 10^{-5}).

Interestingly, since H_0 has only $m-q-1$ rows and there are $m-1$ orthogonal polynomials $h_r(x)$ (for $r = 1, \ldots, m-1$) to choose from, for $q \neq 0$ there is some freedom in the order of the moment departures that can be examined.

For distributions where the parameters fitted represent moments (eg Normal) then one is unlikely to be interested in examining moment departures of the order of the parameters fitted, so $r = q+1, \ldots, m-1$ would be chosen. For other distributions (eg Beta and possibly Poisson) we may still be interested in moment departures of low order despite obtaining the parameters from the data, so one could then choose $r = 1, \ldots, m-q-1$. Note that whichever group of moment departures are examined, the resulting component test statistics will always be asymptotically chi-squared by definition.

5. Examples

In this section I first discuss issues relating to the MLE of parameters using grouped data, then revisit Example 5.2 (see Table 1) from Rayner and Best [9] and the example in section 3 (see Table 2) from Rayner and McAlevey [11]. In both these examples, a goodness-of-fit test is performed to judge how well *grouped continuous* data with known bin endpoints fit a normal distribution (with unspecified nuisance parameters μ and σ). Programs to perform the Pearson-Fisher chi-squared decomposition for the normal distribution, along with output for the examples considered in this paper, are available from the author on request.

5.1. MLE's for grouped data

To find $\hat{H}$ we first need $\hat{p} = p(\hat{\beta})$, the MLE's of the m null hypothesis cell probabilities $p(\beta)$.

Rayner and McAlevey [11] and Rayner and Best ([9],[10]) all indicate that categorised MLE's are used with the Pearson-Fisher chi-squared statistic, but do not emphasize the vital information that the category endpoints $c_0, \ldots, c_m$ are available, so that the data are in fact *grouped continuous* instead of *categorical*. For truly categorical data, only the vector of observed counts N is available, and the usual multinomial MLE's $\hat{p} = N/n$ are obtained. More information (such as the category endpoints) must be available in order to introduce the underlying distribution parameters β into the likelihood function. Note that for the elements of $\hat{p}$ to sum to unity the support of the distribution being fit to must be (c_0, c_m). For example, when fitting the normal distribution (which has infinite support) the first and last category endpoints must be $-\infty = c_0 < \ldots < c_m = \infty$.

The Sheppard corrected grouped mean and standard deviation (Kendall and Stuart, [6], Vol 1, sections 2.20 and 3.18, p.47, pp. 77-80) are sometimes used in place of MLE's when estimating parameters of the underlying distribution (D'Agostino and Stephens, [2], p.548). The resulting estimates are generally closer to the correct MLE's, although Kendall and Stuart ([7], Vol 2, Exercise 18.24-18.25, p.74-75) note that the grouped MLE correction is not generally the same as the Sheppard correction.

In the following examples we will require MLE's for grouped normal data with unknown mean and variance. Here there are $q = 2$ nuisance parameters, $\beta = (\beta_1, \beta_2) = (\mu, \sigma)$. Let $\Phi(x)$ be the distribution function for the standard normal distribution $N(0, 1)$. The log-likelihood is then

$$\ell(\mu, \sigma) = constant + \sum_{j=1}^{m} N_j \log p_j(\mu, \sigma) \tag{12}$$

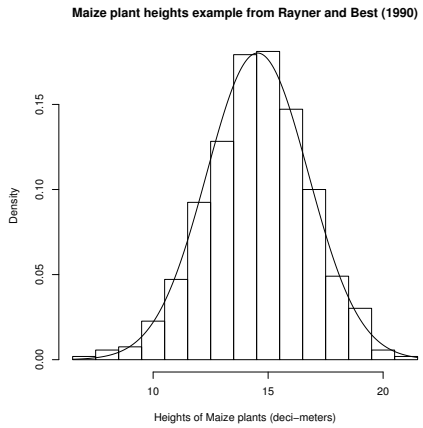

Figure 1. Density histogram of Maize plant data in Table 1 from Rayner and Best [10] superimposed on the fitted normal distribution with parameters $(\hat{\mu}, \hat{\sigma}) = (14.539603, 2.213820)$.

where $p_j(\mu, \sigma) = \Phi\left(\frac{c_j - \mu}{\sigma}\right) - \Phi\left(\frac{c_{j-1} - \mu}{\sigma}\right)$ for $j = 1, \ldots, m$.

It is difficult to maximize this log-likelihood analytically, so numerical methods are used. In a similar fashion to the approach used in Rayner and Rayner [8], I use the Nelder-Mead simplex minimization algorithm (as implemented in the statistical package R, see Ihaka and Gentleman [5]) on the negative log-likelihood in equation (12), using the Sheppard corrected grouped mean and standard deviation as the starting value.

5.2. Example from Rayner and Best [10]

Table 1. Distribution of the heights of Maize plants (in decimeters).

Class center	7	8	9	10	11	12	13	14
Frequency	1	3	4	12	25	49	68	95
Class center	15	16	17	18	19	20	21	
Frequency	96	78	53	26	16	3	1	

This example uses the EMEA data set in Table 1 from D'Agostino and Stephens ([2], p.548) and assesses if these data set are normally distributed (with the unspecified mean and variance as the nuisance parameters). The

Table 2. Distribution of the heights of mothers (in inches).

Upper limit	55	57	59	61	63	65	67	69	∞
Frequency	3	8.5	52.5	215	346	277.5	119.5	23.5	6.5

data set provides the observed number of maize plants in each of 15 height categories specified by the category class center.

Before analysing these data as if they are *grouped continuous* we must decide what we can assume about data with heights $(-\infty, 6.5)$ and $(21.5, \infty)$, since fitting these data to the normal distribution assumes heights in these ranges have positive probability. Because only the "class centers" of the categories are provided, there are two ways of approaching the data.

1. Either these categories were included in the data (and there were zero observations in these categories), so that the number of categories is $m = 17$, and we obtain MLE's agreeing *exactly* with Rayner and Best [10] of $(\hat{\mu}, \hat{\sigma}) = (14.539603, 2.213820)$ along with a not dissimilar $X^2_{PF} = 7.051491$ (Rayner and Best, [10], find $X^2_{PF} = 6.54$).

2. Alternatively, the first and last categories were actually $(-\infty, 7.5)$ and $(20.5, \infty)$, so there are $m = 15$ categories, we obtain different MLE's of $(\hat{\mu}, \hat{\sigma}) = (14.539722, 2.217189)$ and $X^2_{PF} = 6.226699$.

Rayner and Best [10] present $m - q - 1 = 12$ component statistics, which (since there are $q = 2$ nuisance parameters) implies they are considering $m = 15$ categories. Confusingly however, they use the MLE's obtained for $m = 17$ categories.

Taking the first approach (where $m = 17$) and examining moment departures of order $r = 3, \ldots, 16$, we obtain the following $\hat{V}_r$'s and corresponding p-values (since asymptotically $\hat{V}_r^2 \sim \chi^2_1$):

$$\begin{array}{lll}
\hat{V}_3 = -1.53\ (0.13), & \hat{V}_4 = 0.47\ (0.64), & \hat{V}_5 = -0.52\ (0.60), \\
\hat{V}_6 = -0.61\ (0.54), & \hat{V}_7 = 0.52\ (0.60), & \hat{V}_8 = -0.51\ (0.61), \\
\hat{V}_9 = 1.09\ (0.27), & \hat{V}_{10} = -0.44\ (0.66), & \hat{V}_{11} = -0.45\ (0.65), \\
\hat{V}_{12} = -0.16\ (0.87), & \hat{V}_{13} = -0.83\ (0.40), & \hat{V}_{14} = -0.48\ (0.63), \\
\hat{V}_{15} = -0.79\ (0.43), & \hat{V}_{16} = 0.39\ (0.70). &
\end{array}$$

Here, as in Rayner and Best's [10] analysis, the p-values are all relatively large, so there do not seem to be moment departures of any order considered.

Taking the second approach (where $m = 15$) and examining moment departures of order $r = 3, \ldots, 14$, we obtain:

$$\begin{array}{lll} \hat{V}_3 = -1.52\ (0.13), & \hat{V}_4 = 0.69\ (0.49), & \hat{V}_5 = -0.81\ (0.42), \\ \hat{V}_6 = -0.15\ (0.88), & \hat{V}_7 = -0.19\ (0.85), & \hat{V}_8 = -0.03\ (0.98), \\ \hat{V}_9 = 1.13\ (0.26), & \hat{V}_{10} = 0.03\ (0.97), & \hat{V}_{11} = 0.80\ (0.43), \\ \hat{V}_{12} = 0.41\ (0.68), & \hat{V}_{13} = 0.69\ (0.49), & \hat{V}_{14} = -0.41\ (0.68). \end{array}$$

Here, all p-values are also relatively large.

Interestingly, the first few $\hat{V}_r$'s are similar for each approach, though the higher order components are progressively more affected by the "edge effect" differences between the two approaches. This is to be hoped for, since both interpretations of the data category bin values are reasonable. It may be a better idea to fit some kind of truncated normal distribution to this dataset (possibly with the truncation endpoints included as nuisance parameters).

Of course, it is certainly not good statistics to apply 12 or 14 significance tests to a data set and focus on the most critical of these. Rayner and Best [10] recommend that when *testing* a distributional hypothesis (rather than investigating it in an EDA manner) only the initial components (say, $\hat{V}_3$ and $\hat{V}_4$), along with a residual test formed from the remaining components (that is, $X^2_{PF} - \hat{V}_3^2 - \hat{V}_4^2$) be used. Since each $V_r^2 \sim \chi^2_1$ is (asymptotically) independent, the null distribution of such residual tests is easily obtained.

For the example data set, both approaches give non-significant $\hat{V}_3$ and $\hat{V}_4$; while for $m = 17$, $X^2_{PF} - \hat{V}_3^2 - \hat{V}_4^2 = 4.499918$ (with p-value 0.97 from χ^2_{12}), and for $m = 15$, $X^2_{PF} - \hat{V}_3^2 - \hat{V}_4^2 = 3.432863$ (also with p-value 0.97 from χ^2_{10}). So if we were actually *testing* for normality, in practice the same conclusion would be made using either approach. Of course, it is important to be consistent and use the appropriate MLE's corresponding to the class structure chosen!

Note that while Rayner and Best [10] come to the same conclusion, the $\hat{V}_r$'s they obtained are not similar to either of the two possible approaches shown above, and should be interpreted differently.

5.3. Example from Rayner and McAlevey ([11])

Here a normal distribution is fitted to the heights of 1052 mothers grouped into $m = 9$ classes (see Table 2), taken from Snedecor and Cochran ([12], Example 5.12.5, p.78). No explanation is given for the "half mothers" that are observed - perhaps these are observations falling on the class boundaries?

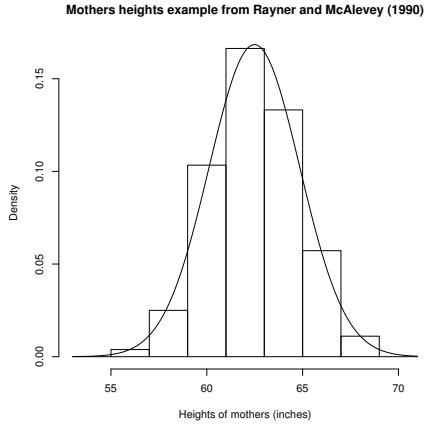

Figure 2. Density histogram of Mother's heights data in Table 2 from Rayner and McAlevey [11] superimposed on the fitted normal distribution with parameters $(\hat{\mu}, \hat{\sigma}) = (62.486285, 2.368791)$. Note that since the extreme classes are half-infinite, the histogram height corresponding to these classes is zero despite the fact that they are not empty.

This time the class limits are quite clear, and we can obtain the MLE's as $(\hat{\mu}, \hat{\sigma}) = (62.486285, 2.368791)$. These estimates have log-likelihood of $1.672713 * 10^{-4}$ larger than Rayner and McAlevey's [11] values of $(\hat{\mu}, \hat{\sigma}) = (62.4865, 2.3678)$.

These MLE's lead to a chi-squared test statistic of $X^2_{PF} = 12.69994$ instead of Rayner and McAlevey's [11] value of $X^2_{PF} = 13.16$, and examining moment departures of order $r = 3, \ldots, 8$, we obtain the following $\hat{V}_r$ and corresponding p-values (since asymptotically $\hat{V}_r^2 \sim \chi^2_1$)

$$\hat{V}_3 = 0.61\ (0.54),\ \hat{V}_4 = 2.59\ (0.01),\ \hat{V}_5 = -1.06\ (0.29),$$
$$\hat{V}_6 = 2.08\ (0.04),\ \hat{V}_7 = 0.44\ (0.66),\ \hat{V}_8 = 0.08\ (0.94)$$

From an EDA point of view, there seems to be a discrepancy in terms of the 4th and 6th order moments, but nowhere else. On the other hand, if we were *testing* for normality, then $\hat{V}_3$ is non-significant but $\hat{V}_4$ is quite significant, and $X^2_{PF} - \hat{V}_3^2 - \hat{V}_4^2 = 5.642718$ (with p-value 0.23 from χ^2_4) is non-significant, so we would probably reject normality as a model for these data because of their tail weight.

Interestingly, Rayner and McAlevey [11] also reject normality here because of their $\hat{V}_2 = -2.2904$ with p-value 0.02 and their $\hat{V}_6 = 2.0650$ with p-value 0.04. Their component statistics are different to those obtained above, and must be interpreted differently. Their component statistics lead

to the conclusion that normality should be rejected because the "...fifth and ninth cells are less normal-like than their predecessors".

It is worth considering that these results might be due the somewhat unrealistic edge effect assumptions we have incorporated about the extreme data classes. Tail weight could be quite heavily influenced by the extreme classes, and in Rayner and McAlevey's analysis, one of the culprit cells (the ninth) is itself an extreme class. As with the Rayner and Best [10] example, it may be better to consider fitting some kind of truncated distribution to these data.

6. Conclusion

This paper provides, for the first time, a complete understanding of the options available for Rayner and Best's [9] decomposition of the Pearson-Fisher chi-squared statistic. In addition, unlike previous analyses using this method (see for example Rayner and Best [9] [10]; Rayner and McAlevey [11]), comprehensive details are provided to enable researchers to perform these tests. The example analyses of Rayner and Best [10] and Rayner and McAlevey [11] are revisited and re-analysed using a far more interpretable decomposition than the original analysis.

The approach outlined in a conference paper by Best and Rayner [1] has recently produced very good approximations to the component values given for the examples in section 5.2. In a private communication, Best and Rayner indicate they produce 3rd and 4th order components of -1.535 and 0.682 (compared to -1.52 and 0.69, if we assume $m = 15$ classes) for the maize plants example in section 5.2; and 0.603 and 2.588 (compared to 0.61 and 2.59) for the mothers heights example in section 5.3. This agreement is probably a result of the fact that fitting data by MLE and moment matching produces fairly similar results for the normal distribution.

Note that for the approach provided in my paper, whichever group of moment departures is examined, the resulting Pearson-Fisher component test statistics will always be asymptotically chi-squared by definition. In contrast, the Pearson chi-squared components obtained by Best and Rayner [1] may not have a known distribution if parameter estimates of comparable order to moments used for fitting the distribution. However, the Best and Rayner [1] method is clearly of interest for future work given the empirical agreement obtained for the examples included despite the different nature of their approach.

Expressing Pearson's test in terms of its components explains why this test often has weak power. This test assesses deviations from the null hypothesis distribution with *equal weight* for each of its $m-q-1$ components. For the examples included here, $m - q - 1$ was as large as 14. Examining

all of these dimensions reduces the effectiveness of the test for detecting departures in terms of the (usually more important) earlier moments. Using these component tests, it is unnecessary to dilute the test power: we can test using the first few component test statistics along with the sum of the remaining components.

In addition, each component corresponds to a specified moment difference between the data and the hypothesized distribution. This allows a more EDA approach to investigating departures from the null hypothesis distribution in terms of interpretable components.

Acknowledgements

This paper is based on research conducted while lecturing at Deakin University, Geelong. Thanks to Dr. John Rayner and Dr. John Best for helpful discussions.

References

1. D. J. Best and J. C. W. Rayner. Chisquared components as tests of fit for discrete distributions. *Presented at the International Conference for Statistics, Combinatorics, and Related Areas*. University of Wollongong, Australia, 2002.
2. R. B. D'Agostino and M. A. Stephens. *Goodness-of-fit Techniques*. New York: Marcel Dekker, 1986
3. B. N. Datta. *Numerical Linear Algebra and Applications*. Brooks/Cole Publishing Company, 1995.
4. P. L. Emerson. Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics* 24:695–701, 1968.
5. R. Ihaka and R. Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5:299–314, 1996.
6. M. Kendall and A. Stuart. *The Advanced Theory of Statistics* Vol 1, 4th Edition. Charles Griffin Ltd, 1977.
7. M. Kendall and A. Stuart. *The Advanced Theory of Statistics* Vol 2, 4th Edition. Charles Griffin Ltd, 1977.
8. G. D. Rayner and J. C. W. Rayner. Categorised regression, *Submitted*, 2002.
9. J. C. W. Rayner and D. J. Best. *Smooth Tests of Goodness of Fit*. New York: Oxford University Press, 1989.
10. J. C. W. Rayner and D. J. Best. Smooth Tests of Goodness of Fit: An Overview. *International Statistical Review* 58, 9–17, 1990.
11. J. C. W. Rayner and L. G. McAlevey. Smooth goodness of fit tests for categorised composite null hypotheses. *Statistics & Probability Letters* 9:423–429, 1990.
12. G.W. Snedecor and W. G. Cochran. *Statistical Methods*. 7th Edition. Ames, IA: Iowa State University Press, 1982.