Statistics & Decisions [to appear], 1–22 (2008)

Goodness Of Fit Testing Using A Specific Density Estimate

C.J. Albers, W. Schaafsma

Received: September 2007; Accepted: April 2008

Summary: To test the hypothesis $H_0 : f = \psi$ that an unknown density f is equal to a specified one, ψ, an estimate $\hat{f}$ of f is compared with ψ. The total variation distance $||\hat{f} - \psi||_1$ is used as test statistic.

The density estimate $\hat{f}$ considered is a peculiar one. A table of critical values is provided, this table is applicable for arbitrary ψ.

Relations with other methods, Neyman's smooth tests in particular, are discussed and power comparisons are performed. In certain situations, our test is recommendable. An example from practice is provided.

AMS 2000 subject classification: 62G10, 62H15
Key words and phrases: Goodness of fit, Neyman smooth tests, χ^2 test, power analysis

1 Introduction

After Karl Pearson's breakthrough paper (1900) about his χ^2 test, many improvements were suggested. Neyman (1937), for example, considered continuous analogues of Pearson's problem. We concentrate the attention on such analogue.

Problem. Given are the ordered outcomes $x_{[1]} < x_{[2]} < \ldots < x_{[n]}$ of an independent random sample $X_1, \ldots, X_n$ from a probability distribution on $\mathbb{R}$ with a 'smooth' density f, not unlike a given density $\psi = \Psi'$. Required is a statement about the truth or falsity of the hypothesis H_0: $f = \psi$ of equality.

The statistician who has to solve this problem may be appalled by the abundance of proposals. Pearson's test depends on a classification of the data. Neyman's smooth test (1937) (see Section 6) requires that one specifies an orthonormal basis for an L_2 space and restricts the attention to the first $k+1$ basis vectors. The Kolmogorov test (Kolmogorov, 1933) is yet another possibility. In the past decade, pre-test procedures (cf. Albers et al., 2000, 2001) and data-driven tests (cf. Ledwina, 1994, Kallenberg and Ledwina, 1995, Inglot and Ledwina, 1996) were developed.

We start from the idea that it is natural to choose some estimate $\hat{f}$ of f and to compare this estimate with the postulated density ψ by rejecting H_0 if $\hat{f}$ and ψ are 'too different'. This idea, dating back to Bickel and Rosenblatt (1973), is commonly used in goodness-of-fit theory (see Hart (1997) for a summary). In our construction, H_0 will be rejected if the area $||\hat{f} - \psi||_1 = \int |\hat{f}(x) - \psi(x)| \, \mathrm{d}x$ between the graph of ψ and that of $\hat{f}$ is sufficiently large. The density estimate $\hat{f}$ (see Albers and Schaafsma, 2003) we recommend will be constructed in Section 2. It is not a kernel estimate in the usual sense. The null distribution of the test statistic is studied to determine P-values and to construct a table of critical values. This table will be reported in Section 5 which, together with Section 2, contains the essence of this paper. (Sections 3 and 4 provide elaborations for special cases useful in making interpretations.)

Our estimate $\hat{f}$ depends on the sample size n and on the degree m of a specific polynomial. That is why the notation $\hat{f} = f_n^{(m)}$ is used, together with $t_n^{(m)} = ||f_n^{(m)} - \psi||_1$ for the outcome of the test statistic $T_n^{(m)}$. In Section 8 we shall recommend choice of $m = \lfloor n^{1/3} \rfloor$. The P-value $\mathrm{P}_0(T_n^{(m)} \geq t_n^{(m)}) = \alpha(x)$ will be used as degree of belief in H_0. Here P_0 refers to the distribution of $T_n^{(m)}$ under H_0. If H_0 is rejected for $\alpha(x)$ smaller than some nominal level, then one is acting according to the general Neyman-Pearson theory. In practice, this is often fairly natural.

If H_0 is maintained then one will usually proceed under the assumption that $f = \psi$. If H_0 is rejected then one will sometimes proceed on the basis of an estimate of f. We do *not* recommend to use the density f_n^m with $m = \lfloor n^{1/3} \rfloor$ which we use in testing H_0, but the density $f_n^{(m)}$ with $m = \lfloor n^{1/2} \rfloor$ (as outlined in Albers and Schaafsma (2003)). (The use of the P-value as 'degree of belief' is considerably questionable from a foundational point of view. See, e.g. Salomé et al. (1999)).

Applying the probability transform $x_i \to u_i = \Psi(x_i)$ we obtain

$$u_{[0]} = 0, \qquad u_{[i]} = \Psi(x_{[i]}) \quad (i = 1, \ldots, n), \qquad u_{[n+1]} = 1.$$

Note that $\Psi(X_i)$ has distribution function $G = F \circ \Psi^{-1}$, quantile function $B = G^{-1} = \Psi \circ F^{-1}$, density function $g(u) = f(\Psi^{-1}(u))/\psi(u)$, quantile density $b(p) = B'(p)$,

3.89	7.44	8.65	9.40	10.00	11.27	11.52	14.23	15.52	15.63
16.39	17.33	18.37	21.12	21.76	22.54	23.29	23.36	24.17	24.57

Table 1.1 Data of the example considered in Section 1

etcetera. The hypothesis H_0: $f = \psi$ is equivalent to H_0: $g \equiv 1$ and to H_0: $b \equiv 1$. It is interesting to note that the distribution of the test statistic $||f_n^{(m)} - \psi||_1$ does not depend on the density ψ to be tested. If applications are made then the density estimate $f_n^{(m)}$ is displayed together with the null density ψ.

Example. Throughout this manuscript, we shall work with the following theoretical example (a concrete application is considered in Section 10). Consider the data $x_{[1]}, \ldots, x_{[20]}$ given in Table 1.1. The information is provided that the underlying density f is such that the support $\{x; f(x) > 0\} = (0, 25)$. We want to test H_0: $f = \psi$ where ψ is the density of the uniform distribution of $(0, 25)$. Figure 1.1 presents graphs of the density estimates $f_{20}^{(m)}$ to be specified in Section 2 ($m = 1, 2, 3, 4$). To test H_0: $f = \psi$, we consider either one of the shaded L_1 areas $||f_{20}^{(m)} - \psi||_1$ ($m = 1, 2, 3, 4$) which are .141, .212, .252, and .277. The data in Table 1.1 have actually been obtained by sampling from the distribution on $(0, 25)$ with density $f(x) = x/625 + 1/50$. The density estimate $f_{20}^{(2)}$ is closer to f than $f_{20}^{(1)}$, and $f_{20}^{(4)}$ is even closer, whilst in this example, $f_{20}^{(3)}$ is the 'best estimate' of f. Table 5.1 (properly extended) provides the P-values .029, .028, .032, and .034 if one uses the shaded areas $||f_{20}^{(m)} - \psi||_1$ ($m = 1, 2, 3, 4$) to test H_0. These P-values are less different than one might expect. The reason is that the underlying test statistics $T_{20}^{(m)}$ are strongly correlated (see Section 4).

Note that Karl Pearson's test requires the specification of the number $k + 1$ of cells such that the χ_k^2 distribution applies. If we take $k = 1$, then we arrive at the two-sided sign test which, for our data, provides $P = .263$. If we take $k = 2$, then we have to work with the exact null distribution of Pearson's statistic. Computations provided $P = .14$.

An elementary discussion. Confronted by the differences between these P-values, the reader will, hopefully, appreciate the following preparation to more sophisticated discussions Sections 7, 8, and 9 (the quick reader might continue with the last sentence of this section). The data of Table 1.1 were obtained by sampling from the distribution indicated because this allows computation of powers using formulas from elementary analysis. The first step is to apply the probability transform where x_i is replaced by $u_i = \Psi(x_i) = x_i/25$. The true distribution of $U_i = \Psi(X_i)$ has distribution function $G = F \circ \Psi^{-1}$ where $G(u) = F(25u)$ and $g(u) = f(\Psi^{-1}(u))/\psi(u) = 25f(25u) = \frac{1}{2} + u$. We concentrate the attention on the formulation $H_0 : g \equiv 1$ or, equivalently, $H_0 : b \equiv 1$, where b is the quantile density.

If a simple alternative is considered, e.g. H_1: $g(u) = 2u$, then we can apply the Neyman-Pearson Fundamental Lemma. For this special alternative H_1 we reject H_0 if $\prod_{i=1}^n u_i$ is sufficiently large or, equivalently, if $-2\sum \log(u_i)$ is sufficiently small. It is well known that the distribution of $-2\sum \log(U_i)$ is χ_{2n}^2 if H_0 is true. The P-value $P\left(\chi_{2n}^2 \leq -2\sum \log(u_i)\right) = P\left(\chi_{40}^2 \leq 21.62\right)$ thus obtained, for the example, is equal to

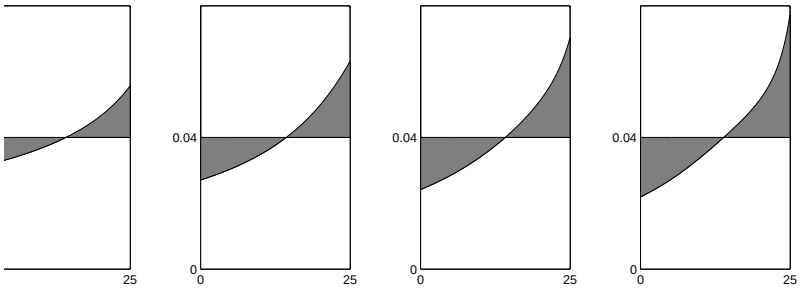

Figure 1.1 Density functions for the data of Table 1.1, for $m = 1$ (left) to $m = 4$ (right). Shaded areas are the test statistics $t_{20}^{(m)}$. (In practice, we recommend to use $m = \lfloor 20^{1/3} \rfloor = 2$ to test H_0 and $m = \lfloor 20^{1/3} \rfloor = 4$ to estimate f.)

.0078. Hence H_0 is rejected at all levels of significance $\alpha \geq .0078$.

In practice we do not know which simple alternative to choose. That is why we study the omnibus test based on some test statistic $T_n^{(m)}$ with outcome $t_n^{(m)} = ||f_n^{(m)} - \psi||_1$. In the present context, $||f_{20}^{(1)} - \psi||_1$ happens to coincide with $|\bar{u} - 1/2| = .141$ because $\bar{u} = 20^{-1} \sum_{i=1}^{20} u_i = .641$. It is of interest for later interpretations to note that, due to chance fluctuations, this outcome is considerably larger than the value $\int_0^1 u(u + \frac{1}{2})\, du = .583$ to be expected if one samples from the true density f.

Elementary power computations (for the true density f, and the corresponding density g) were made for the tests based on the test statistics with outcomes $\prod u_i$, $\sum u_i$, $\prod(u_i + \frac{1}{2})$, and $\sum \text{sign}(u_i - \frac{1}{2})$ or, equivalently, for those with outcomes $\sum h(u_i)$ with $h : (0, 1) \to \mathbb{R}$ defined by $h_1(u) = \log u$, $h_2(u) = u$, $h_3(u) = \log(u + \frac{1}{2})$, and $h_4(u) = \text{sign}(u - \frac{1}{2})$, respectively. If one rejects H_0: $f = \psi$ or, equivalently, H_0: $g \equiv 1$ if $\sum h(u_i)$ is sufficiently large, then one is using a level-α test which is Uniformly Most Powerful (UMP) level-α for testing H_0 against all alternatives of the form $g(\theta) = c(\theta)\exp(\theta h(u))$ with $\theta > 0$. The maximum power in the true density $g_\theta(u) = \frac{1}{2} + u$ is obviously achieved if $h_3(u) = \log(\frac{1}{2} + u)$ is used. Using the asymptotic normality of $\sum h(U_i)$, both under H_0: $g \equiv 1$ and under H_1: $g(u) = \frac{1}{2} + u$, approximate powers can be computed analytically. Using $\mu = \text{E}_0(h(U))$ and $\sigma^2 = \text{Var}_0(h(U))$ to denote mean and variance of $h(U)$ under H_0, and $\mu' = \text{E}_1(h(U))$ to denote the mean under H_1, the power of the one-sided level-α test is approximately given by $1 - \Phi(z_\alpha - \delta)$ where Φ is the distribution function of the standard-normal distribution, $z_\alpha = \Phi^{-1}(1 - \alpha)$, and $\delta = n^{1/2}(\mu' - \mu)/\sigma$. For $h = h_i$ and $n = 20$ as in Table 1.1, we obtain $\delta_i \approx 1.12$, 1.29, 1.28, and 1.12 respectively. Powers $1 - \Phi(z_\alpha - \delta_i)$ in the true distribution are approximately equal to $1 - \Phi(1.645 - \delta_i) = .30$, .36, .36, and .30 if $\alpha = .05$ and the one-sided level-α tests are used. They are about $1 - \Phi(1.960 - \delta_i) = .20$, .25, .25, and .20 if the two-sides size-α tests are used (with equal tail probabilities). That H_0 was rejected at all levels of significance $\alpha > .0078$ if h_1 is used and at all levels $\alpha > .029$ if h_2 or h_3 is used is more surprising than the non-occurrence of statistical significance if

i	j	ρ_{ij}	
1	2	$\frac{1}{2}\sqrt{3}$	= .866
1	3	$\frac{-\pi^2+12-3(\log 2)^2-6\mathrm{Li}_2(-\frac{1}{2})}{6\sqrt{4-3(\log 3)^2}}$	= .915
1	4	$\log 2$	= .693
2	3	$\frac{4-3\log(3)}{2}\sqrt{\frac{3}{4-3(\log 3)^2}}$	= .990
2	4	$\frac{1}{2}\sqrt{3}$	= .866
3	4	$\frac{\frac{3}{2}\log 3-2\log 2}{\sqrt{1-\frac{3}{4}(\log 3)^2}}$	= .850

Table 1.2 Correlations $\rho_{i,j}$ for the four types of test statistic. (The computation ρ_{13} uses $\int_0^1 \log(u)\log(u+\frac{1}{2})\,\mathrm{d}u = -\frac{1}{2}\mathrm{Li}_2(-\frac{1}{2}) + 2 - \frac{1}{12}\pi^2 - (\log\sqrt{2})^2 - \log\sqrt{27} + \log 2$ (cf. Lewin, 1991) where $\mathrm{Li}_2(z) = \int_z^0 t^{-1}\log(1-t)\,\mathrm{d}t$ is the second polylogarithmic function).

Pearson's χ^2-test is used. (Due to chance fluctuations, the sample reported in Table 1.1 is such that, as already observed, $\bar{u} = .641$ is considerably, but not significantly, larger than $\mathbf{E}\ h_2(U) = .583$.)

A peculiar drawback of the two-sided tests based on h_1, h_2, h_3, and h_4 (either with equal tail probabilities under H_0 or with adapted values such that unbiasedness is achieved for all alternatives of the form $g_\theta(u) = c(\theta)\exp(\theta h(u))$ with $\theta \neq 0$) is that these level-α tests are *not* unbiased size-α for testing H_0 against the omnibus alternative A: densities $f \neq \psi$ exist beyond the exponential family such that the probability of rejecting H_0 is less than α if this density is the true one. (This drawback is not restricted to tests of the form indicated, see the end of Section 8.) Finally, we note that the correlations computed under H_0 and presented in Table 1.2 indicate that the tests based on h_2 and h_3 are almost equivalent whereas, in spite of $\delta_1 \approx \delta_4$, the tests based on h_1 and h_4 are considerably different.

The intuitions following from these computations are in line with the discussions to be presented in Sections 7,8, and 9.

2 Specification of the proposed test statistic

To test H_0: $f = \psi$, consider the area

$$||\hat{f} - \psi||_1 = ||\hat{g} - 1||_1 = ||\hat{b} - 1||_1$$

between the graph of ψ and that of $\hat{f}(= f_n^{(m)})$. Note that the first equality follows from the fact that the L_1 norm corresponds to the total variation norm which is invariant under bimeasurable bijections. (This invariance is the main reason why we consider the L_1 norm as more 'natural' than, e.g., the L_2 norm which is behind the smooth tests of Neyman, that of Pearson included, see Section 8.) The second equality can be established

by noting that $||b-1||_1$ is equal to

$$\begin{aligned}\int_0^1 |B'(p)-1| \ \mathrm{d}p &= \int_0^1 \left|(G^{-1})'(p)-1\right| \ \mathrm{d}p \\ &= \int_0^1 \left|\frac{1}{g(G^{-1}(p))}-1\right| \ \mathrm{d}p \\ &= \int_0^1 \left|\tfrac{1}{g(u)}-1\right| \ \mathrm{d}G(u) \\ &= ||g-1||_1.\end{aligned}$$

To define the special estimate $f_n^{(m)}$, we start from the Bernstein polynomial approximation

$$B_n(p) = \sum_{i=0}^{n+1} u_{[i]} \binom{n+1}{i} p^i (1-p)^{n+1-i}$$

of degree $n+1$ to the empirical quantile function (see Muñoz Perez and Fernández Palacín, 1987, De Bruin et al., 1999). This special estimate $B_n(p)$ of $B(p)$ is attractive because the derivative

$$b_n(p) = \sum_{i=0}^{n} (u_{[i+1]} - u_{[i]}) \binom{n}{i} (n+1) p^i (1-p)^{n-i}$$

is a true probability density function: it is positive and integrates up to one. By numerical transformation (via $F_n = B_n^{-1} \circ \Psi$), an estimate f_n of f is obtained. To increase performance, Albers and Schaafsma (2003) replaced b_n by a smoothed version $b_n^{(m)}$ (the degree of B_n is lowered from $n+1$ to $m+1$, and, hence, that of b_n from n to m). In the density estimation case it was suggested to take $m = \lfloor n^{1/2} \rfloor$. In the present context of testing H_0: $b = 1$ some further smoothing is indicated. We recommend a choice of $m = \lfloor n^{1/3} \rfloor$ if an omnibus test is required. For motivation behind our recommendation, see Sections 8 and 9.

The idea to use some quantile-function estimate in hypothesis testing is not new, and dates back to Parzen (1979). LaRiccia (1991), for example, gives an approach using such quantile function to test H: $F \in \mathcal{F}$ where $\mathcal{F}$ is some class of distribution functions. We, however, are fascinated by the crucial problem of testing the simple (i.e. not composite) hypothesis H_0: $f = \psi$. (Our test, with $m = \lfloor n^{1/3} \rfloor$, is not recommendable if ψ is obtained by using the data to specify some particular $\Psi \in \mathcal{F}$; the rationale behind our fascination is primary 'philosophical': we are interested in 'the limits of reason', see Section 9 and (Albers, 2003).)

The definition of $B_n^{(m)}$ (and, hence, of $b_n^{(m)}$, $g_n^{(m)}$, $f_n^{(m)}$, etcetera) is as follows. Let $B_m(p|u_1,\ldots,u_m)$ correspond to $B_n(p)$ if $n = m$ and the outcomes $u_1,\ldots,u_m$ (unordered) have to be evaluated. Define the U-statistic

$$B_n^{(m)}(p) = \binom{n}{m}^{-1} \sum_{1 \le \alpha_1 < \ldots < \alpha_m \le n} B_m(p|u_{\alpha_1},\ldots,u_{\alpha_m}).$$

This can be rewritten as the L-statistic

$$B_n^{(m)}(p) = p^{m+1} + \sum_{j=1}^{m} \binom{m+1}{j} p^j (1-p)^{m+1-j} \sum_{i=j}^{n-m+j} \frac{\binom{i-1}{j-1}\binom{n-i}{m-j}}{\binom{n}{m}} u_{[i]}.$$

Differentiation provides $b_n^{(m)}$ as a convex combination of densities of Beta$(i+1, m+1-i)$ distributions $(i = 0, \ldots, m)$. (See the beginning of Section 6 for explicit expressions.) Note that $B_n^{(m)}$ is the distribution function of a probability distribution on $(0,1)$ (with a density) and that, hence, $F_n^{(m)} = (B_n^{(m)})^{-1} \circ \Psi$ is such that its derivative $\hat{f} = f_n^{(m)}$ is a genuine probability density function: it is nonnegative everywhere and integrates up to one. In practice, computations of $b_n^{(m)}$ and of $f_n^{(m)}$ are performed via numerical differentiation of $B_n^{(m)}$ and of $F_n^{(m)}$. (In Albers (2003, Chapter 4) results can be found about the asymptotic distribution of $b_n^{(m)}$ and $f_n^{(m)}$ if $m = n$; for other values of m, suggestions are made.)

Though we are primarily interested in using $T_n^{(m)}$ with outcome

$$t_n^{(m)} = ||f_n^{(m)} - \psi||_1 = ||b_n^{(m)} - 1||_1$$

as test statistic (and with $m = \lfloor n^{1/3} \rfloor$)), some other test statistics could be discussed as well, e.g. that based on the Kolmogorov distance with outcome

$$\tilde{t}_n^{(m)} = ||F_n^{(m)} - \Psi||_\infty = ||B_n^{(m)} - I||_\infty$$

where $I(p) = p$ (see the end of Section 3 for an elaboration in the case $m = 1$). Note that $\tilde{t}_n^{(m)}$ is an analogue of the test statistic of Kolmogorov's test (see Section 7). The quick reader is invited to continue with Section 5. The Sections 3 and 4 are about the special cases $m = 1$ and $m = 2$. Although of limited practical interest, they do provide a useful basis for interpretations, both for $m = 1, 2$ as for larger m.

3 The case $m = 1$

Ignoring the degenerate case $m = 0$ where the smoothing is so strong that $B_n^{(0)}(p) = p$ and, hence, $f_n^{(0)}$ equals ψ and does not depend on the data, we start with $m = 1$ where

$$B_n^{(1)}(p) = (1-\bar{u})p^2 + \bar{u}(2p - p^2)$$

is a convex combination of the quantile function $2p - p^2$ of the Beta$(1, 1/2)$ distribution and the quantile function p^2 of the Beta$(1/2, 1)$ distribution. (Note that this does not imply that the inverse $G_n^{(1)}$ of $B_n^{(1)}$ is a convex combination of Beta distributions.) For $\bar{u} = 1/2$ the uniform distribution appears.

Theoretical intermezzo. It is of some theoretical interest to consider the quantile functions $B_\theta(p) = (1-\theta)p^2 + \theta(2p - p^2)$ for arbitrary $\theta \in [0,1]$. Here $B_n^{(1)}(p)$ corresponds to $B_\theta(p)$ if $\theta = \bar{u}$. An elementary analysis provides

$$G_\theta(u) = \begin{cases} (2\theta-1)^{-1}(\theta - \sqrt{\theta^2 - (2\theta-1)u}) & \text{if } \theta > 1/2 \\ u & \text{if } \theta = 1/2 \\ (1-2\theta)^{-1}(-\theta + \sqrt{\theta^2 + (1-2\theta)u}) & \text{if } \theta < 1/2 \end{cases}$$

with density

$$g_\theta(u) = \frac{1}{2\sqrt{\theta^2 + (1-2\theta)u}} \qquad (0 < u < 1)$$

(for $\theta = 0$ the distribution function of Beta$(1, 1/2)$ is obtained, for $\theta = 1$ that of Beta$(1/2, 1)$). It is possible to extend this family $\{g_\theta | \theta \in [0,1]\}$ of densities by allowing arbitrary $\theta \in \mathbb{R}$. This extension, however, serves no practical purpose because we are interested in the testing of H_0: $g \equiv 1$ and, hence, in obtaining good power properties for densities 'not too far from $g_{1/2}$'. If X_θ is a random variable with density function g_θ, then (for arbitrary $\theta \in \mathbb{R}$) $\mathbf{E}\, X_\theta = \int_0^1 u g_\theta(u)\, \mathrm{d}u = \frac{1}{3}\theta + \frac{1}{3}$. In a parametric approach to the testing of $\mathrm{H}_0 : g \equiv 1$, the attention might be concentrated on level-α tests which are 'optimal' if g belongs to the parametric family $\{g_\theta | \theta \in \Theta\}$ of densities just considered. The locally most powerful unbiased size-α test rejects H_0: $\theta = \frac{1}{2}$ if $\bar{u}$ is sufficiently far from $\frac{1}{2}$. The most stringent size-α test may be obtained by rejecting for large values of $\prod_{i=1}^n (g_\theta(u_i) + g_{1-\theta}(u_i))$ with θ chosen such that the shortcoming is maximum. This will correspond to the most stringent somewhere most powerful unbiased size-α test. Elaborations are not presented because, just like the tests studied at the end of Section 1 (for exponential subalternatives), these 'optimal' tests (for alternatives of the form g_θ) will fail to be unbiased size-α for testing H_0: $g \equiv 1$ against the omnibus alternative A: $g \neq 1$. Alternatives g exist (beyond the one-parameter subalternatives), where the power is substantially smaller than the nominal level of significance α (we return to this at the end of Section 8). (**End of intermezzo.**)

In Section 2 the test statistics $T_n^{(m)}$ and $\tilde{T}_n^{(m)}$ were defined. For $m = 1$ we have

$$\begin{aligned} t_n^{(1)} &= ||b_n^{(1)} - 1||_1 \\ &= |2\bar{u} - 1| \int_0^1 |1 - 2p|\, \mathrm{d}p \\ &= |\bar{u} - \tfrac{1}{2}| \end{aligned}$$

and

$$\begin{aligned} \tilde{t}_n^{(1)} &= \sup_p \left| B_n^{(1)}(p) - p \right| \\ &= \sup_p |2\bar{u} - 1| p(1-p) \\ &= \tfrac{1}{2}|\bar{u} - \tfrac{1}{2}|. \end{aligned}$$

Conclusion. If one chooses $m = 1$, then both $T_n^{(m)}$ and $\tilde{T}_n^{(m)}$ lead to using the deviation of $\bar{u}$ from $1/2$ as test statistic. The corresponding P-value is, approximately, given by $\mathrm{P}(\chi_1^2 \leq 12(\bar{u} - \frac{1}{2})^2) = 2\Phi(-\sqrt{12n}|\bar{u} - \frac{1}{2}|)$. This test corresponds to that of Neyman (1937) if a polynomial of degree 1 is used. A drawback is that the test is not unbiased size-α for testing H_0: $f = \psi$ against the omnibus alternative A: $f \neq \psi$.

4 The case $m = 2$

The exact equivalence with a Neyman smooth test vanishes if $m = 2$ because then we have that

$$\begin{aligned} B_n^{(2)}(p) &= p^3 + \frac{3p(1-p)}{\binom{n}{2}} \sum_{i=1}^{n} (n - i + p(2i - n - 1)) u_{[i]} \\ &= p + 3p(1-p)\varepsilon + 3p(1-p)(p - \tfrac{1}{2})\delta \end{aligned}$$

where $\varepsilon = \bar{u} - \frac{1}{2}$ is based on the sample mean $\bar{u}$ and

$$\delta = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |u_i - u_j| - \tfrac{1}{3}$$

is based on Gini's mean difference

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |u_i - u_j| = \frac{2}{n(n-1)} \sum_{i=1}^{n} (2i - n - 1) u_{[i]}.$$

Note that both the sample mean $\bar{u}$ and Gini's mean difference are U-statistics as well as L-statistics. We introduced ε and δ because, under H_0,

$$\mathcal{L}\, n^{1/2} \begin{pmatrix} \varepsilon \\ \delta \end{pmatrix} \longrightarrow \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{bmatrix} \right),$$

with $\sigma^2 = 1/12$ and $\tau^2 = 1/45$, we exactly have $\mathrm{Var}\,(\varepsilon) = n^{-1}\sigma^2$ and $\mathrm{Cov}\,(\varepsilon, \delta) = 0$ (Nair, 1936). Locke and Spurrier (1978) suggests that instead of $\bar{u}$ and g other statistics (e.g. $\sum (u_i - \frac{1}{2})^2/n$, and $-\sum \log(u_i(1-u_i))$) could equally well be used to provide goodness-of-fit tests for uniformity. See Section 8 for further discussion (and note that the examples just considered are of the same form as those already considered at the end of Section 1, namely with $h(u) = (u - \frac{1}{2})^2$ and $h(u) = \log(u - \frac{1}{2}) - \log u$, respectively).

It follows from the limit theorem just established that, under H_0,

$$\mathcal{L}\, n \left(12\varepsilon^2 + 45\delta^2 \right) \to \chi^2_2 = \mathrm{Gamma}(1, \tfrac{1}{2}),$$

and that, hence, using any positive multiple of $12\varepsilon^2 + 45\delta^2$ as test statistic, the approximate P-value

$$\mathrm{P}_2^{(1)} = \mathrm{P}(\chi^2_2 \geq n(12\varepsilon^2 + 45\delta^2)) = \exp(-n(3\varepsilon^2 + 11.25\delta^2))$$

is obtained. We, however, prefer an 'exact' approach based on the test statistic $T_n^{(2)}$ with outcome

$$t_n^{(2)} = ||b_n^{(2)} - 1||_1 = 3 \int_0^1 \left| -3\delta p^2 + (3\delta - 2\varepsilon)p + (\varepsilon - \tfrac{1}{2}\delta) \right| \, \mathrm{d}p$$

In practice ε and δ are known, and the numerical computation of this integral is straightforward. Deriving distributional properties of $T_n^{(2)}$ for given ε and δ is straightforward as well.

$\alpha = 10\%$

n	m								
	2	3	4	5	6	7	8	9	10
10	.227	.278	.314	.344	.371	.397	.421	.443	.466
20	.161	.196	.221	.240	.258	.272	.286	.299	.311
50	.102	.123	.138	.151	.161	.170	.178	.185	.191
100	.072	.087	.098	.106	.113	.119	.125	.130	.135

$\alpha = 5\%$

n	m								
	2	3	4	5	6	7	8	9	10
10	.269	.328	.368	.401	.427	.456	.479	.501	.525
20	.191	.231	.260	.281	.300	.315	.329	.342	.355
50	.121	.146	.164	.177	.188	.197	.205	.213	.220
100	.085	.103	.115	.125	.132	.139	.145	.150	.155

$\alpha = 1\%$

n	m								
	2	3	4	5	6	7	8	9	10
10	.347	.423	.473	.512	.548	.575	.600	.624	.643
20	.249	.302	.338	.364	.387	.405	.418	.433	.445
50	.158	.191	.212	.231	.242	.254	.263	.268	.278
100	.112	.134	.150	.162	.170	.178	.184	.191	.196

Table 5.1 Some critical values for $m = 2, \ldots, 10$. For a more extensive table, see `http://-mcs.open.ac.uk/cja235` .

The exact distribution of $T_n^{(2)}$, under H_0 has been studied using simulation experiments. Table 5.1 provides critical values $t_{n,\alpha}^{(2)}$, for $\alpha = .10$, $.05$, and $.01$.

Conclusion. With respect to the example of Section 1 we have $\varepsilon = .141$, $\delta = -.038$. The χ_2^2-test discussed in this section provides the approximate P-value $\mathrm{P}_2^{(1)} = .023$. Using (an extension of) Table 5.1, it follows from $t_n^{(2)} = .212$ that $\mathrm{P}_2 = .029$.

5 The general case

The results of the previous two sections can be generalized to arbitrary $m \leq n$. In Section 4 exact representations were given in terms of the sample mean and Gini's mean difference. For $m \geq 3$ theoretical results can still be derived but they are too complicated to be of interest. In practice, the numerical computation of $t_n^{(m)} = ||f_n^{(m)} - \psi||_1 = ||b_n^{(m)} - 1||_1$ and obtaining distributional properties of $T_n^{(m)}$, for a given sample, are straightforward. To test H_0: $f = \psi$, one can use the simulation-based critical values reported in Table 5.1. (We recommend the choice $m = \lfloor n^{1/3} \rfloor$.)

The figures in Table 5.1 were obtained as follows. Given some choice (m, n), a sample of size n was drawn from the standard uniform distribution providing an outcome

$t_n^{(m)}$ of the test statistic $T_n^{(m)}$. This process was repeated 100 000 times. Percentiles taken from the empirical distribution of $T_n^{(m)}$ were reported.

6 Relation with Neyman's smooth tests

As indicated at the end of Section 2, the quantile density estimate $b_n^{(m)} = (B_n^{(m)})'$ is a *positive* polynomial function on $[0,1]$; it is a *convex* combination of the densities of Beta$(i+1, m+1-i)$ distributions $(i = 0, \ldots, m)$. This representation is very fortunate, because it implies that the $b_n^{(m)}$ and, hence, the density estimates $g_n^{(m)}$ and $f_n^{(m)}$ are genuine probability densities. Note that the weight of the density $(m+1)\binom{m}{i}p^i(1-p)^{m-1}$ of Beta$(i+1, m+1-i)$ can be obtained by elaborating on

$$b_n^{(m)} = \binom{n}{m}^{-1} \sum_{1 \le \alpha_1 < \ldots < \alpha_m \le n} b_m^{(m)}(p \mid u_{\alpha_1}, \ldots, u_{\alpha_m})$$

or, equivalently, by differentiating $B_n^{(m)}(p)$. The first approach provides the weight

$$\binom{n}{m}^{-1} \sum_{1 \le \alpha_1 < \ldots < \alpha_m \le n} \left(u_{[\alpha_i+1]} - u_{[\alpha_i]}\right)$$

which, obviously, is positive. It is a matter of elementary combinatorics to write

$$\sum_{1 \le \alpha_1 < \ldots < \alpha_m \le n} u_{[\alpha_i+1]} = \sum_{h=i+1}^{n+1-m+i} \binom{h-1}{i}\binom{n+1-h}{m-i} u_{[h]}$$

and to establish that the weights thus obtained correspond to those obtained by differentiating $B_n^{(m)}(p)$.

The mathematician might discuss an alternative basis of the linear space of functions on $[0,1]$, e.g. that of orthogonal (ordinary, or trigonometric) polynomials. This can be done with respect to the estimation of b but is of particular interest if we are discussing the estimation of the density $g = G'$ of $U_1 = \Psi(X_1)$ (with $G = F \circ \Psi^{-1} = B^{-1}$).

Let $\varphi_0 \equiv 1, \varphi_1, \varphi_2, \ldots$ be any system of linearly independent functions on $L_2[0,1]$. (Note that $L_2[0,1] \subset L_1[0,1]$. We do not regard it as a severe restriction if the density g to be estimated is supposed to be in $L_2[0,1]$.) The Gram-Schmidt orthogonalization process provides the orthonormal basis $\psi_0, \psi_1, \psi_2, \ldots$ of (a subspace of) $L_2[0,1]$. Note that

$$\begin{aligned} \psi_0 &\equiv \varphi_0 \equiv 1, \\ \psi_{r+1} &= \left(\varphi_{r+1} - \sum_{i=0}^{r} \frac{(\varphi_{r+1}, \psi_i)\psi_i}{(\psi_i, \psi_i)}\right) \Big/ \left|\left|\varphi_{r+1} - \sum_{i=0}^{r} \frac{(\varphi_{r+1}, \psi_i)\psi_i}{(\psi_i, \psi_i)}\right|\right|_2, \end{aligned}$$

$(r = 1, 2, \ldots,)$. If a function $h \in L_2[0,1]$ (a quantile density or a probability density) can be written as a linear combination of $\varphi_0, \ldots \varphi_k$ then it can equally well be written as

a linear combination of $\psi_0, \ldots, \psi_k$. A useful orthonormal basis is that of the normalized shifted Legendre polynomials

$$\psi_r(u) = (-1)^r \sqrt{2r+1} \sum_{k=0}^{r} \binom{r}{k} \binom{r+k}{k} (-u)^k, \quad r = 0, 1, \ldots$$

which is obtained by applying the Gram-Schmidt process to $\varphi_h(u) = u^h, \quad (h = 0, 1, \ldots)$. We elaborate on two lines of thought.

(1) Focussing on the quantile densities, and starting from the estimate $b_n^{(m)}$, we can consider $\varphi_h(p) = p^h \quad (h = 0, 1, \ldots)$ and determine the weights $w_{n,h}$ such that $b_n^{(m)}(p) = \sum w_{n,h} p^h$. The deviations from the 'ideal' weights $w_{n,0} = 1$, $w_{n,1} = \ldots = 0$, corresponding to H_0: $b \equiv 1$, are

$$(2\bar{u} - 1) = 2\varepsilon, \qquad 2(1 - 2\bar{u}) = 4\varepsilon$$

in case $m = 1$ (see Section 3),

$$3(\varepsilon - \tfrac{1}{2}\delta), \qquad -6\varepsilon + \tfrac{9}{4}\delta, \qquad -9\delta$$

in case $m = 2$ (see Section 4), etc. They can be used as the basis of a χ^2_m test (see the title of Pearson's original paper). It is obvious, however, that in practice more weight should be attached to earlier standardized deviations than to later ones. This is done in a (more or less) 'natural' way if we use $T_n^{(m)}$ as test statistic. (Motivation is primarily mathematical; the discussion in Section 4 shows that the weight assigned to the first squared standard deviation is much, perhaps too much, larger than that assigned to the second.)

(2) Focussing on probability densities in $L_2(0,1)$, Neyman (1937) provides a general approach to the problem of testing H_0: $g \equiv 1$ on the basis of the outcome $u_1, \ldots, u_n$ of an independent random sample $U_1, \ldots, U_n$ from a distribution with density g. The structure of $L_2(0,1)$ was used by choosing a number k and some system $\varphi_0 \equiv 1, \varphi_1, \ldots, \varphi_k$ of linearly independent functions on $(0,1)$ or, preferably,the system $\psi_0, \ldots, \psi_k$ obtained from $\varphi_0, \ldots, \varphi_k$ via orthonormalization. Assuming that $g \in L_2(0,1)$, one can think about the projection $1 + \sum_{j=1}^{k} (g, \psi_j)\psi_j$ of g on the $k+1$ dimensional subspace spanned by $\varphi_0, \ldots, \varphi_k$ or, equivalently, by $\psi_0, \ldots, \psi_k$. Here the inner-products (Fourier-coefficients) (g, ψ_j) correspond to the expectations $\theta_j = \mathbf{E}\, \psi_j(U_i) = \int \psi_j(u) g(u)\, \mathrm{d}u$ which can nicely be estimated by using the sample means $\hat{\theta}_j = n^{-1} \sum_{i=1}^{n} \psi_j(u_i)$, providing the estimate $\hat{g} = 1 + \sum_{j=1}^{k} \hat{\theta}_j \psi_j$ of the true density g. In this L_2-approach it is convenient to use $n||\hat{g} - 1||_2^2$, i.e.

$$n \sum_{j=1}^{k} \hat{\theta}_j^2 = \sum_{j=1}^{k} \left\{ n^{-1/2} \sum_{i=1}^{n} \psi_j(u_i) \right\}^2$$

as test statistic because its distribution under H_0 is approximately that of χ^2_k. (Note that $\mathrm{E}_0 \psi_j(U) = (\psi_j, 1) = 0$, etc.) This suggests to use the P-value

$$\mathrm{P}\left(\chi^2_k \geq n^{-1} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} \psi_j(u_i) \right)^2 \right)$$

as degree of belief in H_0. The choice

$$\varphi_0 \equiv 1, \quad \varphi_1 = \mathbf{1}_{[p_0, p_0+p_1)}, \quad \varphi_2 = \mathbf{1}_{[p_0+p_1, p_0+p_1+p_2)}, \quad \ldots, \quad \varphi_k = \mathbf{1}_{(1-p_k, 1]}$$

provides Karl Pearson's P-value

$$\mathrm{P}\left(\chi_k^2 \geq \sum_{j=0}^{k} \frac{(n_j - np_j)^2}{np_j} \right)$$

where n_j is the number of observations in cell j ($j = 0, \ldots, k$).

Many authors have discussed the choice of the number k. Karl Pearson himself stated 'Thus, if we take a very great number of groups our test becomes illusory. We must confine our attention in calculating P to a finite number of groups, and this is undoubtedly what happens in actual statistics. The number k of degrees of freedom will rarely exceed 30, often not greater than 12', (see Pearson, 1900). Later generations of statisticians, dealing with Neyman's smooth tests, have made other recommendations about k. Kallenberg et al. (1985) states with respect to Pearson's test: 'In a classical paper by Mann and Wald (1942), a rule is given to let k increase with n roughly at the rate $n^{2/5}$ when using intervals with equal probability under H_0. More recent numerical work, however, has shown that for particular alternatives, a small fixed value of k often gives much better power (cf. Best and Rayner, 1981)'. Regarding the choice of the number of components k in Neyman's test, Rayner and Best (1989) states that '$k \leq 4$ will usually suffice'. (See Inglot et al. (1990, 1994), Kallenberg et al. (1985) for extensive analyses in this respect.)

All χ_k^2 tests considered have in common that the k underlying test statistics (in Section 4 the sample mean and Gini's mean difference) are used as the basis of the consideration: all other possibilities are ignored. With respect to Neyman's smooth test this implies that the *first* basis vectors $\varphi_0, \ldots, \varphi_k$ (or, equivalently, $\psi_0, \ldots, \psi_k$) are incorporated and, hence, an intuitive idea exists that the earlier basis vectors (lower degree polynomials) are more important than later ones. This suggests that it may be advantageous to replace the unweighted combination of the χ_1^2-statistics $\left\{ n^{-1/2} \sum_{i=1}^{n} \psi_j(u_i) \right\}^2$ by a weighted sum providing the P-value

$$\mathrm{P}\left(w_1 Z_1^2 + \ldots + w_k Z_k^2 \geq \sum_{j=1}^{k} w_j \left\{ n^{-1/2} \sum_{i=1}^{n} \psi_j(u_i) \right\}^2 \right)$$

where $Z_1^2, \ldots, Z_k^2$ are independent χ_1^2 variables. With respect to an idealized context, the choice $w_j = j^{-1/2}$ is discussed in Schaafsma and Steerneman (1981) as one of the possibilities to obtain a substantial improvement of power properties in the subalternative defined by $\delta_1^2 \geq \delta_2^2 \geq \ldots \geq \delta_k^2$ where $\delta_j = (\psi_j, g)$. It follows from Section 4 that using $T_n^{(m)}$ as test statistic is in line with this idea of using decreasing weights. (The fact that $T_n^{(m)}$ is an L_1-norm difference whereas $n||\hat{g} - 1||_2^2$ is an L_2-norm difference is of minor interest.)

Remark. The estimate $\hat{g}$ of the unknown true density g is usually not a probability density itself: it is true that $\int_0^1 \hat{g}(u)\, \mathrm{d}u = 1$ but usually not true that $\hat{g}(u) \geq 0$ $(0 < u <$

1). There are many ways to adapt $\hat{g}$ such that a probability density is obtained. Using approach **(1)** is one of the possibilities. Another one is the maximum-entropy approach described in Jaynes (2003): suppose we have estimates $\hat{\theta}_j = n^{-1}\sum_{i=1}^{n}\psi_j(u_i)$ of the expectations $\theta_j = (\psi_j, g)$ and are interested in the true density g of U_i $(i = 1, \ldots, n)$. Our estimate $\hat{g}$ of g 'should' satisfy the restrictions $\int_0^1 \psi_j(u)g(u)\,\mathrm{d}u = \hat{\theta}_j,\ (j = 1, \ldots, k)$ and be such that the (Shannon) entropy

$$-\int_0^1 g(u)\log(g(u))\mathrm{d}u$$

is maximum. The solution to this optimalization problem is, somewhat surprisingly, that $\hat{g} = g_{\hat{\theta}}$ where

$$g_\theta(u) = \exp\left(\theta_1\psi_1(u) + \ldots + \theta_k\psi_k(u) - c(\theta)\right)$$

defines an exponential family and $\hat{\theta}$ is the maximum likelihood estimate of θ. If one imposes the model that $g \in \{g_\theta;\ \theta \in \mathbb{R}^k\}$ and tests H_0: $\theta = 0_k$ versus A: $\theta \neq 0_k$ by applying the Wilks-Wald asymptotics to the Neyman-Pearson likelihood-ratio principle, then one arrives at the χ^2_k test based on $n||\hat{\theta}||_2^2$ described.

7 Relations with other goodness of fit tests

We are fascinated by the total-variation (or L_1) distance $||f - \psi||_1$ and the Kolmogorov distance $||F - \Psi||_\infty$. The underlying motivation is largely mathematical: the total-variation distance is invariant under bijective mappings while the Kolmogorov distance is invariant under monotonous transformations. Under certain additional assumptions we have that $||f - \psi||_1 = 2||F - \Psi||_\infty$. We always have $||f - \psi||_1 \leq 2||F - \Psi||_\infty$ (see, e.g., Loève, 1955). Both distances are such that they do not change if distribution functions $G = F \circ \Psi^{-1}$ are replaced by corresponding quantile functions. The test statistic $\tilde{T}_n^{(m)} = ||F_n^{(m)} - \Psi||_\infty$ is obtained by replacing the unknown true quantile function B in $||F - \Psi||_\infty = ||B - 1||_\infty$ by the corresponding estimate $B_n^{(m)}$ which is a continuous and increasing analogue of the empirical quantile function. Kolmogorov's test (1933) is based on $||\hat{B} - 1||_\infty$ where $\hat{B}$ is the empirical quantile function. As the true quantile function is smooth, the estimates $B_n^{(m)}$ will be closer to the truth, on the average, than the discontinuous functions $\hat{B}$ on which they are based. That is why it is reasonable to expect that the power properties of the tests based on $||f_n^{(m)} - \psi||_1$ and $||F_n^{(m)} - \Psi||_\infty$ are somewhat better than those based on Kolmogorov's test. Much will depend, however, on the alternative hypotheses to be considered and on the choice of m to be made.

A delicate issue is as follows. If one accepts that the context asks for a test statistic of the form $||\hat{f} - \psi||_1$ then the question arises which nonparametric density estimate $\hat{f}$ one should use. In De Bruin et al. (1999) it was made very clear that the estimator $f_n = f_n^{(n)}$ studied there is 'not unreasonable though some further improvement is possible'. Such improvement can be achieved by using $f_n^{(m)}$ instead of f_n, or by using a kernel estimator k_n, preferably with the bandwidth determined such that the method is optimal for estimating ψ itself (note that ψ is given). The comparison between the tests based on

the specific statistic $||f_n^{(m)} - \psi||_1$, with $m = \lfloor n^{1/3} \rfloor$ recommended, and $||k_n - \psi||_1$ will depend on a large number of specifications with respect to k_n, e.g. of the basic kernel and its bandwidth. The comparison will also depend on the alternative hypotheses for which power comparisons are made, etc. Arguments in favor of $T_n^{(m)} = ||f_n^{(m)} - \psi||_1$ (and $\tilde{T}_n^{(m)} = ||F_n^{(m)} - \Psi||_\infty$) include that *the distribution of the test statistic under H_0 does not depend on ψ*. Critical values of the distribution of $T_n^{(m)}$ can be found in Table 5.1. ($\tilde{T}_n^{(m)}$ has not yet been considered.) For the test statistics $||k_n - \psi||_1$ additional simulation studies would be needed for any k_n and ψ of interest.

Conclusion. A plethora of methods exists to test H_0: $f = \psi$. One class of methods is that of χ_k^2 tests. These tests have in common that they are based on k 'deviations from the probable' (see the title of Pearson, 1900). These deviations $t_j - \mu_j$ have their origin in test statistics T_j with expectations μ_j under H_0. If these T_j constitute a 'correlated system' (see, again, the title of Pearson, 1900), as is the case in general, then they can be combined by using $(T - \mu)'\Sigma^{-1}(T - \mu)$ as omnibus statistic. Here Σ is the covariance matrix of T under H_0 and the (asymptotic) distribution under H_0 is that of χ_k^2. Even for fixed value k, many χ_k^2 tests exist because the attention can be restricted to different $(k + 1)$ dimensional subspaces of $L_2([0, 1])$ (see Section 6 and note that the χ_k^2 tests corresponding to different bases $(\varphi_0, \varphi_1, \ldots, \varphi_k)$ of such $(k+1)$-dimensional subspace are not equivalent). Section 4 shows that χ_k^2 tests may also appear in a different manner.

Other tests have their origin in the mathematical argument that $||\hat{f} - \psi||_1$, or $||\hat{F} - \Psi||_\infty$, or $\int(\hat{F} - \Psi)^2 \, \mathrm{d}\Psi$, etc., 'should' be chosen as test statistic. Note that $||\hat{f} - \psi||_1$ is invariant under bimeasurable bijections and that $||\hat{F} - \Psi||_\infty$ and $\int(\hat{F} - \Psi)^2 \, \mathrm{d}\Psi$ $(= \int_0^1 (G(u) - u)^2 \, \mathrm{d}u)$ are invariant under monotonous transformations.

The practical statistician has to choose one specific testing method from this plethora. Followers of the Neyman-Pearson theory will argue that the choice of test statistic should depend on the alternatives to ψ which have to be taken into account. At the beginning of Section 1 we deliberately did not specify any alternative because we hoped that a test statistic $||f_n^{(m)} - \psi||_1$ with specific value of m, e.g. $m = \lfloor n^{1/3} \rfloor$, is 'universally recommendable' if H_0: $f = \psi$ has to be tested in the case of sufficient smoothness and regularity of f and ψ. We shall see in Section 8 that such 'universally recommendable' test does not exist. For alternatives with density g (after the probability transform) symmetric around $1/2$, our test is less 'usually' powerful than Neyman's smooth test based on $\varphi_h(u) = u^h$ $(h = 0, \ldots, k)$. Our test, however, has very good power properties if H_0: $f = \psi$ has to be tested against alternatives where g is a monotonous function of u or, equivalently, where the likelihood ratio f/ψ is monotonous. This conclusion, however, affects the idea that $T_n^{(m)}$ with $m = \sqrt[3]{n}$ is 'universally recommendable'. Other test statistics of the form $||\hat{F} - \Psi||_\infty$ or $\int(\hat{F} - \Psi)^2 \mathrm{d}\Psi$, etc., either with $\hat{F} = F_n^{(m)}$ or with $\hat{F}$ the empirical distribution function, will also not be 'universally recommendable'.

8 Power Comparisons

In Miller and Quesenberry (1979) and Inglot et al. (1994), power properties were determined for χ_k^2 tests in order to study the choice of k that is most appropriate. It is in this

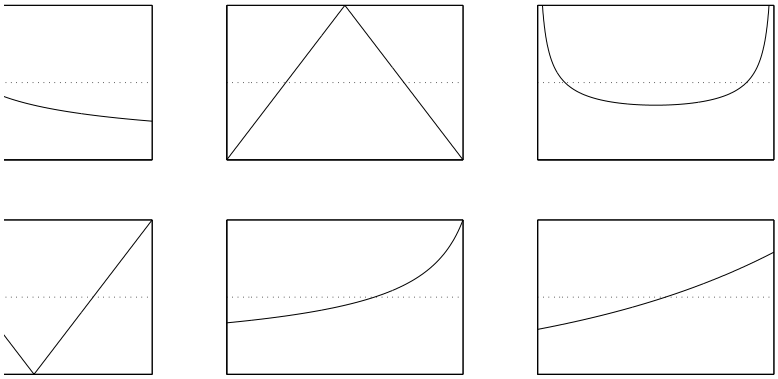

Figure 8.1 Top row, from left to right g_1, g_2 and g_3. Bottom row, from left to right g_4, g_5 and g_6. All horizontal axes go from 0 to 1, the vertical axes from 0 to 2

respect that the attention is concentrated on the alternatives

- $g_1(u) = 1/(2\sqrt{u})$
- $g_2(u) = 2 - 4|u - 1/2|$
- $g_3(u) = (1/\sqrt{u} + 1/\sqrt{1-u})/4$
- $g_4(u) = 4|u - 1/2|$

discussed in Miller and Quesenberry (1979) and the alternatives

- $g_5 = 2/\sqrt{9 - 8u}$, which is g_θ with $\theta = \frac{3}{4}$, and
- $g_6 = \mathrm{e}^u/(\mathrm{e} - 1)$, which is $\tilde{g}_\theta$ with $\theta = 1$,

which appeared in the theoretical intermezzo of Section 3 (and at the end of Section 1) (see Figure 8.1). Note that g_1 and g_3 are not in $L_2(0, 1)$. Powers for various choices of k and m and various sample sizes are reported in Table 8.1. As Neyman's test for $k = 1$ (and $\varphi_0 \equiv 1, \varphi_1(u) = u$) is in exact agreement with our test for $m = 1$ (see Section 3), the differences between the columns under $k = 1$ and under $m = 1$ are caused by randomization and approximation errors, respectively. The column under $m = 1$ is obtained as follows. For the *monotonous* alternatives g_1, g_5, and g_6 we computed the noncentrality parameters δ_i as in Section 1 providing $\delta_1 = 3^{-1/2}n^{1/2}$, $\delta_5 = 12^{-1/2}n^{1/2}$ and $\delta_6 = |(\mathrm{e} - 1)^{-1} - \frac{1}{2}|12^{1/2}n^{1/2} = .081^{1/2}n^{1/2}$ for the test based on $|\bar{u} - \frac{1}{2}|$ (see the end of Section 1 and Section 3). From these δ's the powers under $m = 1$ were obtained by using the formula $\Phi(-1.960 + \delta)$.

The results for g_1, g_5 and g_6 reported in Table 8.1 are in line with what one should expect: the alternatives g_5 and g_6 were chosen (see Section 3) such that it is 'optimal' to choose $k = m = 1$. For increasing k, Neyman's test looses power faster than our test

altern.	n	Neyman				Using $t_n^{(m)}$							
		$k=1$	2	3	4	$m=1$	2	3	4	5	6	7	8
g_1	10	47	51	52	53	57	49	50	49	48	47	47	48
	20	74	77	78	78	73	73	74	74	74	74	74	73
	50	98	99	99	99	98	98	98	98	98	98	99	99
g_2	10	0	21	11	11	1	1	1	2	4	8	11	13
	20	0	62	48	39	1	1	2	8	16	26	37	40
	50	0	99	97	95	1	0	19	57	77	86	91	93
g_3	10	10	30	30	35	10	11	12	13	13	14	16	19
	20	10	45	44	52	10	11	12	13	14	18	21	24
	50	10	79	76	84	10	12	14	19	30	39	49	55
g_4	10	11	26	23	19	11	12	13	15	16	17	21	25
	20	11	63	58	59	11	12	13	16	18	23	31	38
	50	11	96	94	96	11	11	15	25	44	59	70	78
g_5	10	15	13	13	12	15	16	16	16	16	16	15	15
	20	25	21	19	18	25	26	26	26	26	25	25	25
	50	54	47	43	39	53	56	56	56	56	55	55	55
g_6	10	14	10	9	9	14	15	15	15	15	15	14	14
	20	24	18	15	14	25	25	25	25	25	25	24	24
	50	52	42	36	54	52	54	53	53	53	52	52	52

Table 8.1 Rejection percentages (at $\alpha = 5\%$) for Neyman's smooth tests (with $\varphi_j(u) = u^j$, $j = 0, \ldots, k$) with $k = 1, \ldots, 4$ and the tests based on $t_n^{(m)}$ with $m = 1, \ldots, 6$. The Neyman data for $g_1, \ldots, g_4$ are obtained from Miller and Quesenberry (1979). The numbers in column $m = 1$ are obtained using the method described in Section 8. All other percentages are based 10000 Monte Carlo-replications. The correspondence between columns $k = 1$ and $m = 1$ suggests that the simulations and the asymptotic results are sufficiently reliable (except for the result for g_1 and $m = 10$ where the asymptotics is unreliable).)

does for increasing m. The reason is obvious: our test stays closer to the test studied in Section 3 (see Section 4). The alternative g_1 is such that Neyman's test is a bit better because it is faster in picking up additional information.

For the *symmetric* alternatives g_2, g_3 and g_4 we computed the variances σ^2 of $(\bar{U} - \frac{1}{2})n^{1/2}$ and compared these with the variance $\sigma_0^2 = 12^{-1}$ under H_0. The powers in the column under $m = 1$ are next computed by using the formula $2\Phi(-1.960\sigma_0/\sigma)$. For g_2 we have $\sigma^2 = 24^{-1}$ and, hence, $2\Phi(-1.960\sqrt{2}) = .006$. For g_3 and g_4 we have $\sigma^2 = 7/60$ and $\sigma^2 = 8^{-1}$ with corresponding powers approximately $.098$ and $.110$, respectively.

The results for g_2, g_3, and g_4 reported in Table 8.1 are in line with what one should expect: the lack of dispersion of g_2, compared with the uniform density, has the effect that the power is less than 5% if the choice $k = m = 1$ is made. This shows that the test based on $|\bar{u} - \frac{1}{2}|$ is *not* unbiased size-α. For $k = m \geq 2$, Neyman's test is preferable for these symmetric alternatives because our test puts relatively more weight on the deviation $|\bar{u} - \frac{1}{2}|$. It is not true, however, that, e.g., Neyman's test for $k = 2$ is unbiased size-α. To establish this, we considered the case where U has the discrete distribution $\frac{1}{2}\mathbf{1}_{1/2-1/\sqrt{12}} + \frac{1}{2}\mathbf{1}_{1/2+1/\sqrt{12}}$. We do not suggest that our test is unbiased size-α.

9 General conclusions

The problem of testing H_0: $f = \psi$ against A: $f \neq \psi$, or A: $||f - \psi||_1 > 0$, is too 'ill-posed' to be settled satisfactorily. Classical χ_k^2 tests like those of Pearson or of Neyman (and those studied in Section 4) are asymptotically of size-α, but they are not 'optimal' in an overall sense.

The choice of the number of degrees of freedom k in these χ_k^2 tests is difficult to make. In Section 1 we cited Kallenberg et al. (1985) which claims that a small fixed choice of the number of cells in a χ^2 test gives best power. Rayner and Best (1989) made a similar statement. Ledwina (1994) stated that 'recommendations in statistical literature are sometimes confusing'. Schaafsma and Steerneman (1981) considered an idealized context where 'decreasing weights' are assigned to the χ_1^2 distributed components of χ^2. Recent papers (Ledwina, 1994, Inglot and Ledwina, 1996, Kallenberg and Ledwina, 1995, Inglot et al., 1994) on Neyman's test prescribe the use of data-driven methods, where the choice of k depends on the data set. One of the suggestions is to use Schwarz's Bayesian Information Criterion to choose the dimension for the appropriate exponential model for the data, and to use this dimension as k.

Fascinated by the mathematical formulation A: $||f - \psi||_1 > 0$ of the alternative hypothesis we started our investigations in the hope that a satisfactory compromise would be achieved by rejecting H_0 for sufficiently large outcomes of

$$t_n^{(m)} = ||f_n^{(m)} - \psi||_1$$

and a specific choice of m, e.g. $m = \lfloor n^{1/3} \rfloor$. The power computations in Section 8 indicate that (1) the choice of m is much less crucial than the choice of k in χ_k^2 tests, (2) for $m = k \geq 2$ the χ_k^2 test is definitely preferable if alternatives $f \neq \psi$ are considered

193	195	205	213	219	224	241	245	246	247	248	250	252
252	253	254	256	257	258	265	266	267	267	268	269	269
270	270	272	272	276	276	276	280	280	283	283	284	285
288	289	290	291	293	297	299	299	300	305	318	335	347

Table 10.1 Azimuth measurements by Bom (1978)

such that the corresponding density is symmetric around $\frac{1}{2}$ as is the case with g_2, g_3, and g_4 in Table 8.1, (3) for alternatives f with f/g monotonously increasing or monotonously decreasing (see g_1, g_5, and g_6 in Table 8.1) rejecting H_0 for large outcomes of $t_n^{(m)}$ with $m = \lfloor n^{1/3} \rfloor$ seems to provide the 'satisfactory comprimise' we are looking for. However, Table 8.1 suggests that a data-dependent approach for finding m might yield a more satisfactory compromise.

Conclusion. Testing H_0: $f = \psi$ versus A: $f \neq \psi$ is a Pandora's box. Consensus about a testing method cannot easily be attained. Note that in the approach of Section 6 a specific choice of basis functions $\varphi_0, \ldots, \varphi_k$ is needed. Our test, with $m = \lfloor n^{1/3} \rfloor$, provides a 'very reasonable' approach if H_0 has to be tested against the subalternative A' of A defined by monotonicity of f/ψ. We suggest that it is also a reasonable approach if H_0 has to be tested against the wider subalternative A'' defined by stochastic inequality, i.e. by $F \geq \Psi$. If the alternatives of interest are different, e.g. because ψ has been adapted to location/scale characteristics of the sample, then one should not proceed with our test (at least not with the choice $m = \lfloor n^{1/3} \rfloor$ indicated). It will then be difficult to compromise between the plethora of tests available.

10 An example from archaeology

Starting with Van Giffen (1925, 1926), many scientists made statements about the preference direction of Dutch passage mounds or, more precisely, the chamber in the interior of such dolmen. An east-west preference direction was documented. Various definitions of the main direction of (the chamber of) passage mounds are proposed and corresponding 'azimuth measurements' are reported in literature. The azimuth of an (undirected) line segment is obtained by measuring the number of degrees, from south via west and north, to provide a value between 180° and 360°. In some protocols it was mentioned that the actual azimuth measurement reported is the average of two azimuth measurements, one derived from the eastern end of the mound and one from the western end.

Table 10.1 reports $n = 52$ ordered azimuth measurements, collected by Bom (1978). We regard these values $x_{[1]}, \ldots, x_{[52]}$ as the outcomes of the order statistics corresponding to an independent random sample from a distribution with density f on $[180, 360]$ (such that $\lim_{x \searrow 180} f(x) = \lim_{x \nearrow 360} f(x)$; we shall ignore this additional information). We shall test the null hypothesis

$$\mathrm{H}_0^{(1)}: \quad f(x) = \tfrac{1}{180}, \quad 180 < x < 360,$$

Test	$H_0^{(1)}$	$H_0^{(2)}$
$T_{52}^{(3)}$	2	40
$T_{52}^{(4)}$	0	29
Neyman ($k = 3$)	0	8
Neyman ($k = 4$)	0	3
χ^2 ($k = 3$)	0	14
χ^2 ($k = 4$)	0	2

Table 10.2 P-values (in %) for the testing of $H_0^{(1)} or H_0^{(2)}$ on the basis of the data in Table 10.1. The Neyman tests are applied with $\varphi_j(u) = u^j$, $j = 0, \ldots, k$, and $k = 3$ and 4. Pearson's χ_2-test results are based on $(k + 1) = 4$ and 5 equiprobable classes

of uniformly distributed azimuth values, as well as the null hypothesis

$$H_0^{(2)}: \quad f(x) = \tfrac{1}{90}\left(1 - \tfrac{1}{90}|x - 270|\right), \quad 180 < x < 360,$$

that f is the density of the mean $\frac{1}{2}(X_1 + X_2)$ of two independent random variables, both uniformly distributed on $[180, 360]$. The motivation for formulating $H_0^{(2)}$ originates from the remark that azimuth values were sometimes obtained by taking the average of two values, one from the eastern end and one from the western end. (The testing of $H_0^{(2)}$ should be regarded as a mathematical exercise rather than as something of genuine archaeological interest.)

Table 10.2 provides results in the form of P-values. Our test is used with both $m = 3$ and $m = 4$ because $3 < \sqrt[3]{52} \approx 3.73 < 4$. We compared this with other tests discussed in this paper. All tests considered for $H_0^{(1)}$ have P-values below 2%. Neyman's test (with $k = 4$) and Pearson's χ^2 test (with 5 equiprobable classes and, thus, $k = 4$ degrees of freedom) reject $H_0^{(2)}$ at $\alpha = 5\%$. The other tests considered, do not reject this hypothesis, and our test (both with $m = 3$ as $m = 4$) has considerably larger P-values than the other ones. This illustrates the conclusion of Section 9.

Acknowledgements. We are grateful to referees of this article and, especially, to the reader Wilbert Kallenberg, for their helpful comments. We thank the archaeologist Jan Talen for his assistance w.r.t. the example of Section 10.

References

C.J. Albers. *Distributional inference: the limits of reason*. PhD thesis, University of Groningen, 2003. Also available from `http://irs.ub.rug.nl/ppn/243136900`.

C.J. Albers and W. Schaafsma. Estimating a density by adapting an initial guess. *Computational Statistics & Data Analysis*, 42(1-2):27–36, 2003.

W. Albers, P.C. Boon, and W.C.M. Kallenberg. Size and power of pretest procedures. *Annals of Statistics*, 28(1):195–214, 2000.

W. Albers, P.C. Boon, and W.C.M. Kallenberg. Power gain by pre-testing? *Statistics & Decisions*, 19(3):253–276, 2001.

P. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimators. *Annals of Statistics*, 1:1071-1095, 1973.

F. Bom. *Eerste Nederlandse hunebeddengids*. Ankh-Hermes, Deventer, 1978.

R. De Bruin, D. Salomé, and W. Schaafsma. A semi-Bayesian method for nonparametric density estimation. *Computational Statistics & Data Analysis*, 30:19–30, 1999.

A.E. van Giffen. *De hunebedden in Nederland*, volume 1. Oosthoek, Utrecht, 1925.

A.E. van Giffen. *Atlas van de hunebedden in Nederland*. Oosthoek, Utrecht, 1926.

J.D. Hart. *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer-Verlag, New York, 1997.

T. Inglot and T. Ledwina. Asymptotic optimality of data-driven Neyman's tests for uniformity. *Annals of Statistics*, 24(5):1982–2019, 1996.

T. Inglot, Y. Jurlewitz, and T. Ledwina. On Neyman-type smooth tests of fit. *Statistics*, 21:549–568, 1990.

T. Inglot, W.C.M. Kallenberg, and T. Ledwina. Power approximations to and power comparison of smooth goodness-of-fit tests. *Scandinavian Journal of Statistics*, 21(2): 131–145, 1994.

E.T. Jaynes. *Probability Theory: The Logic Of Science*. Cambridge University Press, 2003.

W.C.M. Kallenberg and T. Ledwina. Consistency and monte carlo simulation of a data driven version of smooth goodness-of-fit tests. *Annals of Statistics*, 23(5):1594–1608, 1995.

W.C.M. Kallenberg, J. Oosterhoff, and B.F. Schriever. The number of classes in chi-squared goodness-of-fit tests. *Journal of the American Statistical Association*, 80 (392):959–968, 1985.

A. Kolmogorov. Sulla determinazione emprica di una legge di distribuzione. *Giornalle dell Instituto Italiano degli Attuari*, 4:1–11, 1933.

V.N. LaRiccia. Smooth goodness-of-fit tests: a quantile function approach. *Journal of the American Statistical Association*, 86(414):427–431, 1991.

T. Ledwina. Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, 89(427):1000–1005, 1994.

L. Lewin, editor. *Structural properties of polylogarithms*, volume 37 of *Mathematical Surveys and Monographs*. American Statistical Society, Providence RI, 1991.

C. Locke and J.D. Spurrier. On tests of uniformity. *Communications in statistics A. Theory and methods*, 7(3):241–258, 1978.

M. Loève. *Probability theory*. Princeton University Press, 1955.

H.B. Mann and A. Wald. On the choice of the number of class intervals in the application of the chi square test. *Annals of Mathematical Statistics*, 13:306–317, 1942.

F.L. Miller and C.P. Quesenberry. Power studies of tests for uniformity II. *Communications in statistics B. Simulation and computation*, 8:271–290, 1979.

J. Muñoz Perez and A. Fernández Palacín. Estimating the quantile function by Bernstein polynomials. *Computational Statistics & Data Analysis*, 5:391–397, 1987.

U.S. Nair. The standard error of Gini's mean difference. *Biometrika*, 28:428–436, 1936.

J. Neyman. 'Smooth' test for goodness of fit. *Skandinavisk Aktuarietidskrift*, 20:149–199, 1937.

E. Parzen. Nonparametric statistical data modelling. *Journal of the American statistical association*, 74(365):105, 1979.

K. Pearson. 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 be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

J.C.W. Rayner and D.J. Best. *Smooth tests of goodness of fit*. Oxford University Press, 1989.

D. Salomé, R. de Bruin, and W. Schaafsma. Q-values for χ^2 problems. *Statistics & Decisions*, 17, 1999.

W. Schaafsma and A.G.M. Steerneman. Discriminant analysis when the number of features is unbound. *IEEE transactions on systems, man, and cybernetics*, SMC-11(2): 144–151, 1981.

Casper J. Albers
Department of Mathematics & Statistics
The Open University
Walton Hall
Milton Keynes, MK7 6AA
United Kingdom
c.j.albers@open.ac.uk

Willem Schaafsma
Department of Mathematics
University of Groningen
PO Box 800
9700AV Groningen
The Netherlands