Smooth Test for Testing Equality of Two Densities[1]

Anil K. Bera
University of Illinois.
E-mail:abera@uiuc.edu. Phone: 217-333-4596.

Aurobindo Ghosh
Singapore Management University.
E-mail:aurobindo@smu.edu.sg. Phone: +65-6822-0863.

Zhijie Xiao
University of Illinois.
E-mail:zxiao@uiuc.edu. Phone: 217-333-4520.

October 1, 2003.

[1]This is a preliminary draft of the paper and should not be quoted without permission. Earlier version of this paper was presented at the Midwest Econometric Group Meetings at Columbus, Ohio, 2002 and at the Tinbergen Institute, University of Amsterdam, 2003.

Abstract

The two-sample version of the celebrated Pearson (1900) goodness-of-fit problem has been a topic of extensive research and several tests like the Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling tests have been suggested. Although these tests perform fairly well as omnibus tests for comparing two samples with different probability density functions (PDF), they fail to have good power particularly against very specific alternatives like departures in location, scale, skewness and kurtosis terms. We show that a modified version of Neyman smooth test based on the empirical distribution functions (EDF) obtained from the two samples remarkably improves the detection of directions of departure. We can identify deviations in the mean, variance, skewness or kurtosis terms using what we call the ratio density function, which is the PDF of the probability integral transform of one sample based on the other sample EDF. We derived Neyman's smooth test using Rao's score principle. We also establish a bound on the relative sample sizes of the two samples that makes our test consistent. Furthermore, we suggest an "optimal" choice range of the sample size of one sample compared to the other to ensure minimal size distortion in finite samples. As an application of our procedure we compare the age distributions of employees with small employers in New York and Pennsylvania with group insurance before and after the enactment of the "Community Rating" legislation in New York. It has been a conventional wisdom that if community rating is enforced (where group health insurance premium does not depend on age or any other physical characteristics of the insured), then the insurance market will collapse since only older or less healthy patients would prefer group insurance. We find that there are significant changes in the age distribution in the population in New York owing mainly to a shift in the location and scale. Our extensive Monte Carlo studies also indicate that the suggested test has attractive finite sample properties both in terms of size and power.

1 Introduction

One of the old, celebrated problems in statistics is the *two-sample* version of Pearson (1900) goodness-of-fit problem [Lehmann (1953) and Darling (1957)]. Suppose we have two samples $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_m$ from two *unspecified* absolutely continuous distributions with cumulative distribution functions (CDF) $F(x)$ and $G(x)$, respectively. The problem is to test the hypothesis $H_0 : F = G$. Most of the tests in the literature are based on some distance measures between the two empirical distribution functions (EDF), $F_n(x)$ and $G_m(x)$. The Kolmogorov-Smirnov criterion uses [see, for instance, Darling (1957, p. 828)]

$$D_{nm} = \sqrt{\frac{nm}{n+m}} \sup_{-\infty<x<\infty} |F_n(x) - G_m(x)| . \tag{1}$$

Cramér- von Mises statistic is based on the measure [see, for example, Anderson (1962, p. 1148)]

$$W_{nm}^2 = \frac{nm}{n+m} \int_{-\infty}^{\infty} [F_n(x) - G_m(x)]^2 \, dH_{n+m}(x) , \tag{2}$$

where $H_{n+m}(x)$ is the EDF of the two samples together, i.e., $H_{n+m}(x) = [nF_n(x) + mG_m(x)]/(m+n)$. Anderson and Darling (1952) modification of (2) is given by [see Darling (1957, p. 827)]

$$A_{nm}^2 = \frac{nm}{n+m} \int_{-\infty}^{\infty} [F_n(x) - G_m(x)]^2 \, \psi(H_{n+m}(x)) \, dH_{n+m}. \tag{3}$$

Here $\psi(.)$ is some non-negative weight function chosen to accentuate the distance between $F_n(x)$ and $G_m(x)$ where the test is desired to have sensitivity. A statistically appealing weight function $\psi(u) = [u(1-u)]^{-1}$ has the effect of weighing the tails heavily since the function is large near $u = 0$ and $u = 1$ [Anderson and Darling (1954, p. 767)]. As Darling (1957, p. 824) stated the asymptotic distribution of these statistic are based on certain distribution analogues of the Glivenko-Cantelli type lemma

$$\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \to 0 \text{ with probability } 1$$

in the same way that the central limit theorem is a distribution analogue of the law of large numbers. However, the computation and implementation of these tests are not easy. Moreover, being very general, these tests do not have good finite sample power properties. In this paper, we propose a much simpler test for H_0 using Neyman (1937) smooth test principle.

Let us first consider a simpler problem, the one-sample Pearson goodness-of-fit test of $H_0^{'}: F(x) = F_0(x)$ where $F_0(x)$ is a specified CDF with $f_0(x)$ as the corresponding probability density function (PDF). Define the probability integral transform (PIT)

$$z_i = F_0(x_i) = \int_{-\infty}^{x_i} f_0(\omega)\, d\omega. \tag{4}$$

If $H_0^{'}$ is true, then $Z_1, Z_2, ..., Z_n$ are independently and identically distributed (IID) as $U(0,1)$ irrespective of F_0. And we can test H_0 by testing uniformity of Z in $(0,1)$. Therefore, in some sense, "all" testing problems can be converted into testing only *one kind of hypothesis* [see Neyman (1937, pp. 160-162) and Bera and Ghosh (2001, p. 1)]. Neyman (1937) considered the following *smooth* alternative to the uniform density

$$h(z) = c(\theta)\exp\left[\sum_{j=1}^{k}\theta_j\pi_j(z)\right],\ 0 < z < 1, \tag{5}$$

where $c(\theta)$ is the constant of integration depending on $\theta_1, \theta_2, ..., \theta_k$, $\pi_j(z)$ are orthogonal polynomials[1] of order j satisfying

$$\int_0^1 \pi_i(z)\,\pi_j(z)\,dz = \delta_{ij} \text{ where } \begin{aligned}\delta_{ij} &= 1 \text{ if } i = j \\ &= 0 \text{ if } i \neq j.\end{aligned} \tag{6}$$

We can test $H_0^{\prime}: F(x) = F_0(x)$ by testing $H_0^{''}: \theta_1 = \theta_2 = ... = \theta_k = 0$ in (5). Using the generalized Neyman-Pearson (N-P) lemma, Neyman (1937) derived the locally most powerful symmetric test for $H_0^{''}$ against the alternative H_1 :At least one $\theta_i \neq 0$, for small values of θ_i. Under H_0, asymptotically the test statistic

$$\Psi_k^2 = \sum_{j=1}^{k} u_j^2 \sim \chi_k^2 \text{ where } u_j = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\pi_j(z_i),\ j = 1, ..., k. \tag{7}$$

Neyman suggested this test to rectify some of the drawbacks of Pearson (1900) goodness-of-fit statistic [see Bera and Ghosh (2001) for more on this issue and for a

[1] Neyman (1937) used $\pi_j(y)'s$ as the orthogonal polynomials which can be obtained by using the following conditions,

$$\pi_j(y) = a_{j0} + a_{j1}y + ... + a_{jj}y^j, a_{jj} \neq 0,$$

given the restrictions of orthogonality given in (6). Solving these the first five $\pi_j(y)$ are (Neyman 1937, pp. 163-164)

$\pi_0(y) = 1,$
$\pi_1(y) = \sqrt{12}\left(y - \frac{1}{2}\right),$
$\pi_2(y) = \sqrt{5}\left(6\left(y - \frac{1}{2}\right)^2 - \frac{1}{2}\right),$
$\pi_3(y) = \sqrt{7}\left(20\left(y - \frac{1}{2}\right)^3 - 3\left(y - \frac{1}{2}\right)\right),$
$\pi_4(y) = 210\left(y - \frac{1}{2}\right)^4 - 45\left(y - \frac{1}{2}\right)^2 + \frac{9}{8}$

historical perspective], and called it a smooth test since the alternative density (5) is close to the null density $U(0,1)$ for small θs and has few intersections with the null density.

Now turning to the problem of testing $H_0 : F = G$ in the two sample case, let us start assuming that $F(.)$ is known. We construct a new random variable Z by defining $Z_j = F(Y_j)$, $j = 1, 2, ..., m$. The CDF of Z is given by

$$\begin{aligned} H(z) &= \Pr(Z \leq z) = \Pr(F(Y) \leq z) \\ &= \Pr\left(Y \leq F^{-1}(z)\right) = G\left(F^{-1}(z)\right) \\ &= G(Q(z)) \end{aligned} \tag{8}$$

where $Q(z) = F^{-1}(z)$ is the quantile function of Z. Therefore, the PDF of Z can be written as [see Neyman (1937, p. 161), Pearson (1938, p. 138) and Bera and Ghosh (2001, p. 185)]

$$h(z) = \frac{d}{dz} H(z) = \frac{g(Q(z))}{f(Q(z))}, \; 0 < z < 1. \tag{9}$$

Although this is the ratio of two PDFs, $h(z)$ is a proper density function in that sense that $h(z) \geq 0$, $z \in (0,1)$ and $\int_0^1 h(z)\, dz = 1$, if we assume that F and G are also strictly increasing functions. In the literature the PDF $h(.)$ is known under different names. Cwik and Mielniczuk (1989) termed it the *relative density,* while Parzen (1992, p. 7) and Handcock and Morris (1999, p. 22) called it the *comparison density function* and the *density ratio,* respectively. We will call it *ratio density function* (RDF) since it is both a ratio of two densities and a proper density function itself. Under $H_0 : F = G$, $h(z) = 1$, i.e., $Z \sim U(0,1)$. And, under the alternative hypothesis $H_1 : F \neq G$, $h(z)$ will differ from 1 and that provides a basis for the Neyman smooth test. Under the alternative, we take $h(z)$ as given in (5) and test $\theta_1 = \theta_2 = ... = \theta_k = 0$. Therefore, the test utilizes (9) which looks more like a "likelihood ratio". To see the exact form of $h(z)$, let us consider some particular cases. When the two distributions differ only in location; for example, $f(.) \equiv \mathcal{N}(0,1)$ and $g(.) \equiv N(\mu, 1)$, $\ln(h(z)) = \mu z - \frac{1}{2}\sigma^2$ which is *linear* in z. Similarly, if the distributions differ in scale parameter, such as, $f(.) \equiv \mathcal{N}(0,1)$ and $g(.) \equiv \mathcal{N}(0, \sigma^2)$, $\sigma^2 \neq 1$, $\ln(h(z)) = \frac{z^2}{2}\left[1 - \frac{1}{\sigma^2}\right] - \frac{1}{2}\ln \sigma^2$, a quadratic function of z. As plotted in Figure 1, the first and second-order normalized polynomials $\pi_1(z)$ and $\pi_2(z)$ can capture this type of differences in location and scale aspects of the distributions. In Figure 2, we provide plots of $h(z)$ when $f(.)$ and $g(.)$ differ respectively in skewness and kurtosis terms. These plots very closely resemble the plots of the third and the fourth normalized Legendre polynomials $\pi_3(z)$ and $\pi_4(z)$ plotted in Figure 3. Therefore, we believe that the test will not only be powerful but also will be informative on identifying particular source(s) of departure(s) from H_0.

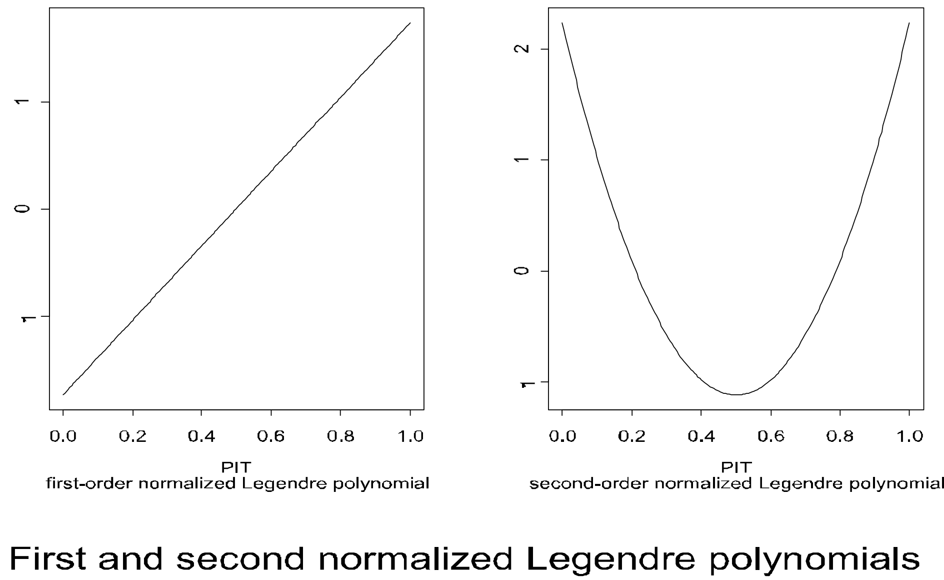


Figure 1: Plots of $\pi_1(z)$ and $\pi_2(z)$.

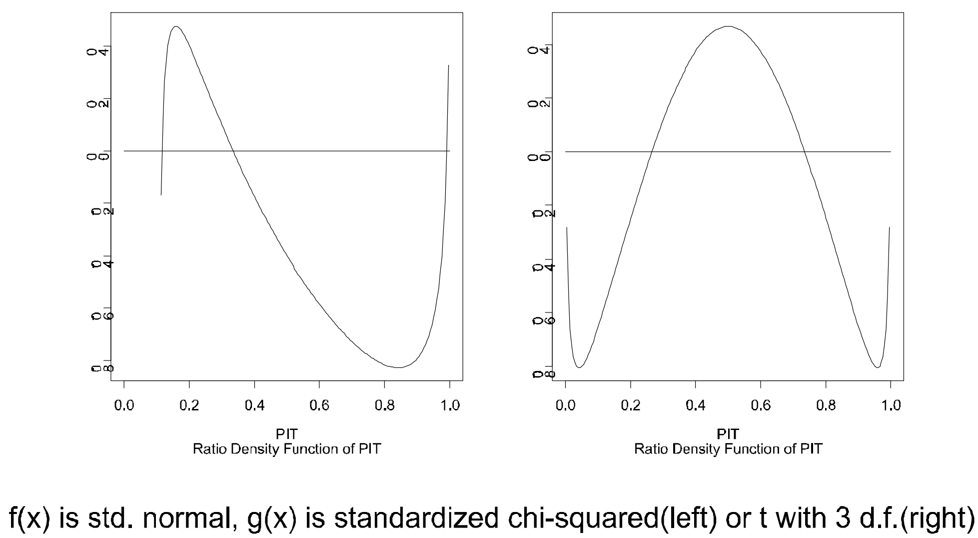


Figure 2. Plots of RDF with different skewness and kurtosis terms

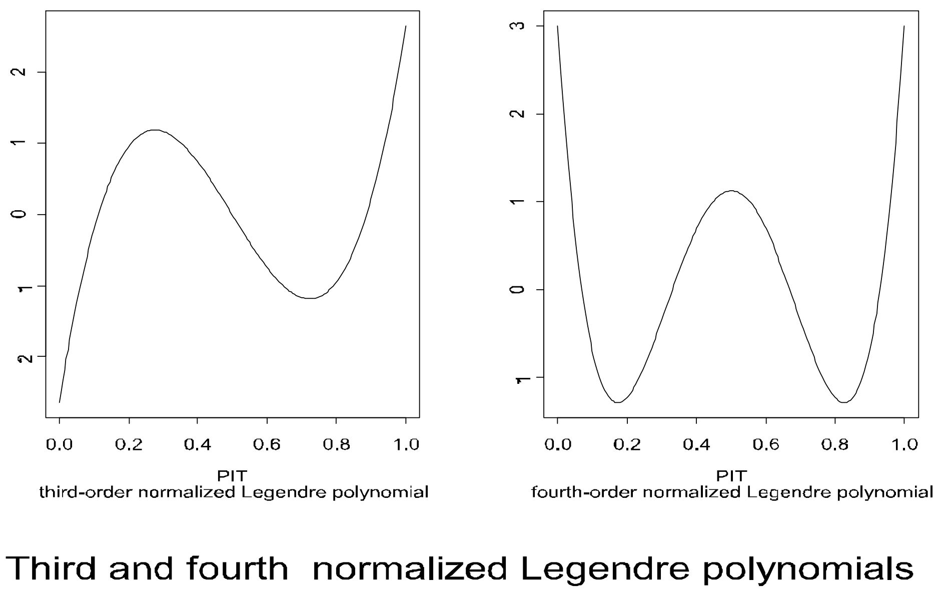


Figure 3: Plots of $\pi_3(z)$ and $\pi_4(z)$.

This approach to testing $H_0 : F = G$ is related to those based on the distance functions between F and G mentioned earlier. Under Neyman's smooth test formulation, we consider the distance between $h(z)$ and 1, *i.e.* $\frac{d}{dz}\left[G\left(F^{-1}(z)\right) - z\right]$, $0 < z < 1$. $G\left(F^{-1}(z)\right) - z$ is a familiar quantity in the literature of tests based on differences of EDFs, such as those mentioned in (1)-(3). Due to the equality [see Serfling (1980, pp. 110-111)]

$$\sup_{-\infty<x<\infty} |F(x) - G(x)| = \sup_{0<z<1} \left|G\left(F^{-1}(z)\right) - z\right|,$$

we can transform the domain from $(-\infty, \infty)$ to $(0, 1)$. Moreover, $\sqrt{m}\left[G_m\left(F^{-1}(z)\right) - z\right]$ converges to the well-known Brownian bridge process in distribution [see, for instance, Billingsley (1968, p. 104)]. Neyman's smooth test simplifies the problem even further by looking at the distance of $h(z)$ from 1 towards a particular direction specified by (5) [see Figure B for the directions using Legendre polynomials]. In the next section we derive Neyman's smooth test using Rao's (1948) score test principle for testing $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$ in (5), assuming $F(.)$ is known. In Section 3, we allow $F(.)$ to be unknown and use $\hat{z}_j = F_n(y_j)$, i.e., we get the EDF of one population and then consider the PIT of the sample from the other population. Replacing $F(.)$ by $F_n(.)$ requires careful theoretical considerations and these are also addressed in Section 3.

2 Neyman Smooth Test

For simplicity we first consider the test when $F(x)$ is known and we use of the PIT as defined by $Z_j = F(Y_j)$. Neyman (1937) derived a locally most powerful symmetric (regular) unbiased test (or critical region) for $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$ in (5) against H_1 : at least 1 $\theta_i \neq 0$, and he called it an unbiased critical region of type-C. This type-C critical region is an extension of the locally most powerful unbiased (LMPU) test (type-A region) of Neyman and Pearson (1936) from a single parameter case to a multiparameter situation. Let us denote the power function as $\beta(\theta_1, \theta_2, ..., \theta_k) = \beta(\theta) \equiv \beta$. Assuming, that the power power function $\beta(\theta)$ is twice differentiable in the neighborhood of $H_0 : \theta = 0$, Neyman (1937, pp. 166-167) formally required that an unbiased critical region of type-C of size α should satisfy the following conditions :

$$1.\ \ \beta(0, 0, ...0) = \alpha. \tag{10}$$

$$2.\ \ \beta_j = \left.\frac{\partial \beta}{\partial \theta_j}\right|_{\theta=0} = 0,\ j = 1, 2, ..., k. \tag{11}$$

$$3.\ \ \beta_{jl} = \left.\frac{\partial^2 \beta}{\partial \theta_j \partial \theta_l}\right|_{\theta=0} = 0,\ j, l = 1, 2, ..., k, j \neq l. \tag{12}$$

$$4.\ \ \beta_{jj} = \left.\frac{\partial^2 \beta}{\partial \theta_j^2}\right|_{\theta=0} = \left.\frac{\partial^2 \beta}{\partial \theta_1^2}\right|_{\theta=0},\ j = 2, 3, ...k. \tag{13}$$

5. And finally, the common value of $\left.\frac{\partial^2 \beta}{\partial \theta_j^2}\right|_{\theta=0}$ is the maximum over all such regions satisfying the conditions (10)-(13).

It is clear that conditions (10) and (11) are respectively for the size and unbiasedness. Conditions (12) and (13) ensures that equal departures from $\theta = 0$, in *all* directions should lead to the same power. Therefore, for this type-C critical region the approximate power function is $\beta(\theta) = \alpha + \frac{1}{2}\beta_{11} \sum_{j=1}^{k} \theta_j^2$. Neyman (1937) used (5) as the PDF under the alternative hypothesis. After considerable algebra [see Neyman (1937, pp. 169-180)] and using multiparameter version of the generalized Neyman-Pearson lemms, Neyman obtained his optimal test statistic. His resulting statistic, however, takes a very simple form as given below.

Proposition 1 (Neyman, 1937) *The type-C critical region is given by*

$$\Psi_k^2 = \sum_{j=1}^{k} u_j^2 \geq C_\alpha \tag{14}$$

where $u_j = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j(z_i)$, *and for large* m *the critical point* C_α *is determined from* $\Pr[\chi_k^2 \geq C_\alpha] = \alpha$.

We now show that the test statistic Ψ_k^2 can simply be obtained using Rao(1948) score (RS) test principle. Taking (5) as the PDF under the alternative hypothesis, the log-likelihood function $l(\theta)$ can be written as

$$l(\theta) = m \ln c(\theta) + \sum_{j=1}^{k} \theta_j \sum_{i=1}^{m} \pi_j(z_i). \tag{15}$$

The RS test for testing the null $H_0 : \theta = \theta_0$ is given by

$$RS = s(\theta_0)' \mathcal{I}(\theta_0)^{-1} s(\theta_0) \tag{16}$$

where $s(\theta)$ is the score vector $\partial l(\theta)/\partial\theta$ and $\mathcal{I}(\theta)$ is the information matrix $E\left[-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right]$. In our case $\theta_0 = 0$. It is easy to see that

$$\begin{aligned} s(\theta_j) &= \frac{\partial l(\theta)}{\partial\theta_j} \\ &= m\frac{\partial \ln c(\theta)}{\partial\theta_j} + \sum_{i=1}^{m} \pi_j(z_i) \\ &= m\frac{\partial \ln c(\theta)}{\partial\theta_j} + \sqrt{m}u_j, \ \ j = 1, 2, ..., k. \end{aligned} \tag{17}$$

From (5), $\int_0^1 h(z)\,dz = 1$. Differentiating this identity with respect to θ_j

$$\frac{\partial c(\theta)}{\partial\theta_j} \int_0^1 \exp\left[\sum_{j=1}^{k} \theta_j \pi_j(z)\right] dz + c(\theta) \int_0^1 \exp\left[\sum_{j=1}^{k} \theta_j \pi_j(z)\right] \pi_j(z)\,dz = 0. \tag{18}$$

Evaluating this under $\theta = 0$, we have $\left.\frac{\partial \ln c(\theta)}{\partial\theta_j}\right|_{\theta=0} = 0$ and under the null

$$s(\theta_j) = \sqrt{m}u_j. \tag{19}$$

To get the information matrix, let us first note from (17) that

$$\frac{\partial^2 l(\theta)}{\partial\theta_j\partial\theta_l} = m\frac{\partial^2 \ln c(\theta)}{\partial\theta_j\partial\theta_l}, \tag{20}$$

which is a constant. Therefore, under H_0 the (j,l)th element of the information matrix $\mathcal{I}(\theta)$ is simply $-m\partial^2 c(\theta)/\partial\theta_j\partial_l$ evaluated at $\theta = 0$. Differentiating (18) with respect to θ_l and evaluating it at $\theta = 0$, after some simplification we have

$$\left.\frac{\partial^2 c(\theta)}{\partial\theta_j \partial\theta_l}\right|_{\theta=0} + \int_0^1 \pi_j(z)\,\pi_l(z)\,dz = 0. \tag{21}$$

Using (6)

$$\left.\frac{\partial^2 C(\theta)}{\partial\theta_j \partial\theta_l}\right|_{\theta=0} = -\delta_{jl}, \tag{22}$$

and

$$\mathcal{I}(\theta_0) = mI_k, \tag{23}$$

where I_k is a $k \times k$ identity matrix. Combining (16), (20) and (23) the RS test statistic has the simple form

$$RS = \sum_{j=1}^{k} u_j^2. \tag{24}$$

The test is transformed into the problem of testing "uniform" distribution of Z_i, $i = 1, 2, ..., n$ against the smooth alternative with density function

$$h(y) = C(\theta)\exp\left[\sum_{j=1}^{k}\theta_j \pi_j(y)\right],$$

where $C(\theta)$ is a normalizing constant. If we denote $\log C(\theta) = \psi(\theta)$, we may re-write the density in the form

$$h(y) = \exp\left[\psi(\theta) + \sum_{j=1}^{k}\theta_j \pi_j(y)\right]. \tag{25}$$

This choice of the alternative can be recast into the well-known *Maximum Entropy* characterization. Suppose, we have $k-$restrictions on the random variable Z_i given by [see for example, Kagan, Linnik, Rao and Ramachandran (1973), p. 409] [2]

$$E[\pi_j(z)] = c_j,\ \ j = 1, 2, ..., k. \tag{26}$$

Then, among all possible distributions, the density function that maximizes the entropy $-E[\ln f(x)]$ has the form (25). In this sense, we are testing for the most "plausible" density function that satisfies the required conditions in (26) and maximizes the entropy measure.

The components that Neyman considered in his test is connected to an ANOVA

[2] We can interpret it as k moment conditions in a more general framework without the orthogonal polynomials π_j, this problem can be extended to a GMM framework, which is a topic of out future research.

like decomposition of the distribution. If we consider the classical Cramér- von Mises test for $H_0 : F_X(.)$ is uniformly distributed against $H_1 : F_X(.)$ is not uniformly distributed, the test statistic is

$$CvM = n \int_0^1 \left[\hat{F}_X(y) - y\right]^2 dy,$$

(or equivalently, if we test that $H_0 : F_X(.) = F_0(.)$ against $H_1 : F_X(.) \neq F_0(.)$, we consider $n \int_0^1 \left[\hat{F}_X(.) - F_0(.)\right]^2 dF_0(.)$) where $\hat{F}_X(.)$ is the empirical distribution function of X_i. Now, if we consider the following Fourier transform

$$\theta_{2j-1} = \int_0^1 \cos(2\pi jx)\, dF(x)$$
$$\theta_{2j} = \int_0^1 \sin(2\pi jx)\, dF(x),\ j = 1, 2, ...,$$

then the testing problem is equivalent to testing $H_0 : \theta_j = 0,\ j = 1, 2, ...$ against H_1 :at least one of the $\theta_j \neq 0$. If we let $\hat{\theta}_j$ be the empirical Fourier coefficients, then the CvM test can be expressed as

$$CvM = \frac{n}{2\pi^2} \sum_{j=1}^{\infty} j^{-2} \left(\hat{\theta}_{2j-1}^2 + \hat{\theta}_{2j}^2\right). \tag{27}$$

From the representation in (27), we can see that the classical CvM test (and also K-S) put increasingly smaller weights (j^{-2}) to high frequency components (i.e. put a weight j^{-2} to $e^{i2\pi jx}$ with large j). hese weights (j^{-2}) make the CvM type test effectively only use the first few components of θ_j. In Neyman's smooth test formulation, instead of putting a decreasing weight to the components, the selected components are equally weighted. This allows us to focus on the selected directions of departure from the null distribution, this makes the smooth test superior when we are trying to identify specific directions of departure and not just an overall departure [see, for example, Durbin and Knott (1972), Eubank and LaRiccia (1992) for studies on this issue]. Durbin and Knott (1972) also suggested that when using a slightly different representation for the term Z_{nj} in their decomposition (see Durbin 1972) given in (27) for testing goodness-of-fit, usually the first few $\theta_j s$ should be examined, this is essentially the same spirit as the Neyman's smooth test with $k-$components.

The basic idea of the decomposition of the Cramér- von Mises statistic is to consider a Fourier transform on $F(y)$, $y \in [0, 1]$ to the frequency domain $\theta_j, j =$

$1, 2, ...$ and thus the population version of the statistic

$$n \int_0^1 \left[F_X(y) - y\right]^2 dy$$

can be expressed as in (27). Given a finite sample $\{X_1, X_2, ..., X_n\}$, the empirical Fourier coefficients $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n$ are a *sufficient* statistic, and consequently the origibal testing problem is equivalent to the test $H_0' : \theta_j = 0,\ j = 1, 2, ..., n.$

If there is *a priori* evidence that the two samples differ mostly at low frequencies, then Neyman Smooth test with a choice of small k can indeed be a powerful test. In some applications however, it is desirable to choose the parameter "k" in the construction of Neyman's test. In other applications, we may want k to be even an increasing parameter with the sample size; say, n and choose k using some model selection criterion, such as the Schwarz's criterion (see for example, Ledwina 1994). When choosing the ideal value of k, we have to consider two aspects. One one side, a large value of k can potentially differentiate between two distribution in several directions (particularly, in high frequency data this might be useful). On the other hand, if k is too high there are two possible problems. First the effectiveness of te test in each direction would be diluted as pointed out by Neyman himself. Second, the accumulated stochastic errors discussed in Section 3 in each direction would be large and deteriorate the performance of the resulting test.

3 Testing Equality of Distributions Using the Smooth

In practice, we have to relax the assumption of known distribution function in the two sample case. We consider again two-samples of n and m observations $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$ from unknown distributions with CDFs $F(x)$ and $G(x)$, and test the hypothesis $H_0 : F = G$, *the underlying distribution being otherwise unspecified.*

Without the knowledge about the distribution function F, $Z_i = F(Y_i)$ is unknown and thus Ψ_k^2 is infeasible. Instead, we consider using the empirical distribution function (EDF)

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x), \tag{28}$$

in place of F and construct

$$\hat{Z}_i = F_n(Y_i) = \frac{1}{n} \sum_{l=1}^{n} I(X_l \leq Y_i),\ i = 1, ..., m. \tag{29}$$

where $I(.)$ is the indicator function. Substituting Z_i by $\hat{Z}_i$ in the formula of the statistic Ψ_k^2, we obtain the following generalized version of the smooth test using the EDF F_n

$$\hat{\Psi}_k^2 = \sum_{j=1}^{k} \left[\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j \left(\hat{Z}_i \right) \right]^2 . \tag{30}$$

As noted earlier, the infeasible test Ψ_k^2 converges to a χ^2-distribution with degree of freedom k and can be interpreted using Rao's score test principles. Now, we want to show that under certain regularity conditions, $\hat{\Psi}_k^2$ has the same limiting distribution. To obtain the asymptotic χ^2-distribution for $\hat{\Psi}_k^2$, we need to show that the errors coming from estimating F by the EDF F_n is asymptotically zero. As will be clear in our later analysis and in the proofs given in the Appendix, this asymptotic result does not hold true automatically even with both m and $n \to \infty$. In fact, different expansion rates for m and n are required to preserve the χ^2 limiting distribution.

We summarize the result in the following Theorem 1, whose proof is given in the Appendix.

Theorem 1 *Under the null hypothesis that F=G, if $\frac{(\log \log n)m}{n} \to 0$ as $m, n \to \infty$, $\hat{\Psi}_k^2 \Rightarrow \chi_k^2$.*

From the above Theorem 1 we can see that the size of the sample used in estimating the distribution function should be larger in magnitude than the other sample size. The reason for this requirement is the following. If we estimate the CDF of X based on observations $\{X_i\}_{i=1}^n$, F_n converges to F at the rate $\sqrt{n}$ (pointwise), i.e. the estimation error in $\hat{Z}_i$ is of the order $n^{-\frac{1}{2}}$. Thus, notice that $\pi_j(.)$ is first-order Lipschitz continuous, the accumulated estimation error in $\frac{1}{\sqrt{m}} \sum_{i=1}^m \pi_j \left(\hat{Z}_i \right)$ has the order $m^{\frac{1}{2}} n^{-\frac{1}{2}}$, which goes to zero if n increases to ∞ at a rate faster than m.

Let us try to give a heuristic justification of the results given here. To obtain consistent tests, we require that the size of the sample that is used to estimate the empirical distribution (n) should be larger than the size of the sample in the χ^2 approximation (m). Intuitively, the errors of the preliminary estimation of z is of order $n^{-\frac{1}{2}}$. These preliminary estimation errors would be accumulated in then second stage when we ake sum of m $z's$, which is necessary to obtain the assymptotic χ^2-approximation. Roughly speaking, the aggregated errors from the preliminary distributional estimation is of order $\sqrt{m}/\sqrt{n}$. To obtain consistent tests, this error term should go to zero, as the sample size increases. Consequently, the distribution function $F(.)$ should be estimated with higher accuracy than the second stage χ^2 approximation. The same problem exists in similar statistical applications. For example, the situation in our case is similar, but not exactly the, to the methods of simulation based inference (e.g. Gourieroux and Monfort 1996), where, for example, the conditional moment are estimated based on simulation. If the number of simulation increases fast enough relative to the sample size, the simulation method estimator has the same asymptotic distribution as the corresponding estimator as if the conditional moment is known.

Theorem 1 provides an upper bound for m given n so that we still have a consistent χ^2 test using the EDF F_n. Under this condition, a wide range of sample sizes can be chosen and all provie asymptotically equivalent tests, although the finite sample performance of these tests may differ substantially. A natural question to ask is: what is the optiomal rate of m relative to n?

Theorem 2 *Under the null hypothesis,*

$$\hat{\Psi}_k^2 = \chi^2 + O_p\left(\frac{1}{\sqrt{m}}\right) + O_p\left(\sqrt{\frac{m}{n}}\right),$$

and thus the optimal relative magnitude of m and n that minimizes the size distortion is $m = O\left(n^{\frac{1}{2}}\right)$.

Heuristically, we can decompose

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(\hat{Z}_i\right)$$

into

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(Z_i\right) + \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\pi_j\left(\hat{Z}_i\right) - \pi_j\left(Z_i\right)\right]. \tag{31}$$

The first part of (31), i.e. $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(Z_i\right)$, converges to a standard normal variate which contributes to the limiting χ^2 distribution.. The larger the m, the faster it goes to the limiting normal distribution. The second part of (31), i.e. $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\pi_j\left(\hat{Z}_i\right) - \pi_j\left(Z_i\right)\right]$, which is the error term coming from the estimating distribution function, converges to zero under teh conditions in Theorem 1. The larger the n relative to m, the smaller the term. To optimize the sampling properties of the test, a trade-off has to be made to balance these two components, giving an optimal relative magnitude between m and n.

Although an exact formula of the optimal sample sizes will be dependent on the optmality criterion and the exact formulation of the higher order terms are potentially not available. Theorem 2 gives the optimal relative magnitude between m and n, which substantially narrows down the range of choices for the sample sizes. Monte Carlo experiment results indicate that a simple rule of thumb choices such as $m = \sqrt{n}$

based on the guidance of Theorem 2 provides pretty good sampling performance to the propsed test.

4 Application of Neyman Smooth Test with EDF

Let us consider an insurance market where there are several insurance companies providing health insurance who are competing for clients. These clients can be grouped into different risk categories depending on their "proneness" or "propensity" of having bad health. However, the main problem that the insurance companies face is one of *adverse selection* since they cannot see beforehand what type of client they are insuring, high risk or low risk (Akerlof, 1970). Rothschild and Stiglitz (1976) claimed that in a insurance market setup where there are two types of clients, a health insurance contract based on risk categories will ensure that the high risk client chooses to pay higher premium while both high and low risk clients will fully insure. However, if it is not possible to write a health insurance contract based on the risk categories due to either legislation or other restrictions then there could be two possible scenarios. A healthier (low risk) individual will chose to buy less than complete coverage while the less healthy (high risk) individuals will buy full insurance, so the insurance market will still function although this will not the the most efficient outcome like the case where there could be risk based contracts. The second scenario happens if further regulations restrict or prohibit the selling of less than full insurance, then we have a problem that the healthy or low risk individuals will stop by coverage which means that gradually the insured population will be made up of more high risk individuals and the insurance company has to pay out more often. This will cause the premium to go up, so healthier individuals will drop their coverage even further and this cycle will finally result in a total collapse of the insurance market. This scenario is refered to as the "Adverse Seclection Death Spiral" [Buchmueller and DiNardo (2000)].

It has been a conventional wisdom that if "community rating" for setting health insurance premiums is enforced (where all the insured people have to pay exactly the same premium irrespective of their age, sex, and health conditions), more and more healthier (younger) members of a group insurance policy would drop out of the insurance plan. To find the evidence of any existence of "Adverse Selection Death Spiral", the state of New York where legislation for enforcing "community rating" was enacted in 1993 canm be compared with Pennsylvania, where there was no such legislations enacted. We would like to test for the difference between the age distributions of the adult civilian population between 18 and 64, before and after 1993 for each state to verify whether there is any difference solely due to the "death spiral" story. The data is from 1987-1996 March Current Population Survey covering questions on whether an individual have insurance coverage, and if so whether the coverage is through a small employer and other similar questions. New York and Pennsylvania were selected because these states are very similar both geographically as well as demographically.

Our objective is to use Neyman's smooth test to determine if there is a difference between the PDF of the age distributions in a state before and after 1993. The population selected was individuals who were covered by group insurance policies sponsored by their employers who had 100 or less employees. This was divided into two parts, one for the individuals before 1993 and the other for those after 1993. Fig.

4A and 4B gives the estimated PDFs using kernel density estimator.[3]

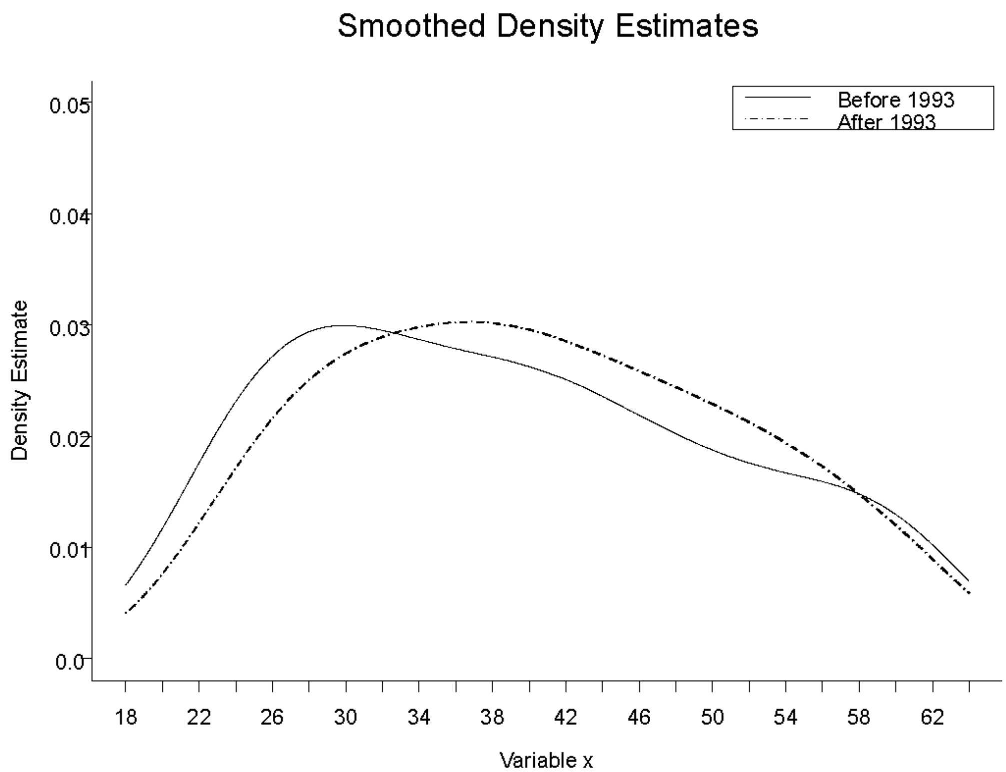


Fig 4A. Density estimates for New York.

[3]We estimated the PDF $f(x)$ of the sample $x_1, x_2, ..., x_n$ before 1993 using kernel density estimator (see Fig. 4A and Fig. 4B)

$$\hat{f}_{bef}(x) = \frac{1}{b}\sum_{j=1}^{n} K\left(\frac{x - x_j}{b}\right), \tag{32}$$

where the bandwidth is *quadruple* of $b = 1.06\min(\hat{\sigma}_x, IQR/1.34)n^{-\frac{1}{5}}$, $\hat{\sigma}_x$ is the estimated standard deviation of $x_j's$ and IQR is the interquartile range for the sample and $K(.)$ is the kernel suggested by Parzen (1962) [also see Silverman (1986) pp. 45-47].

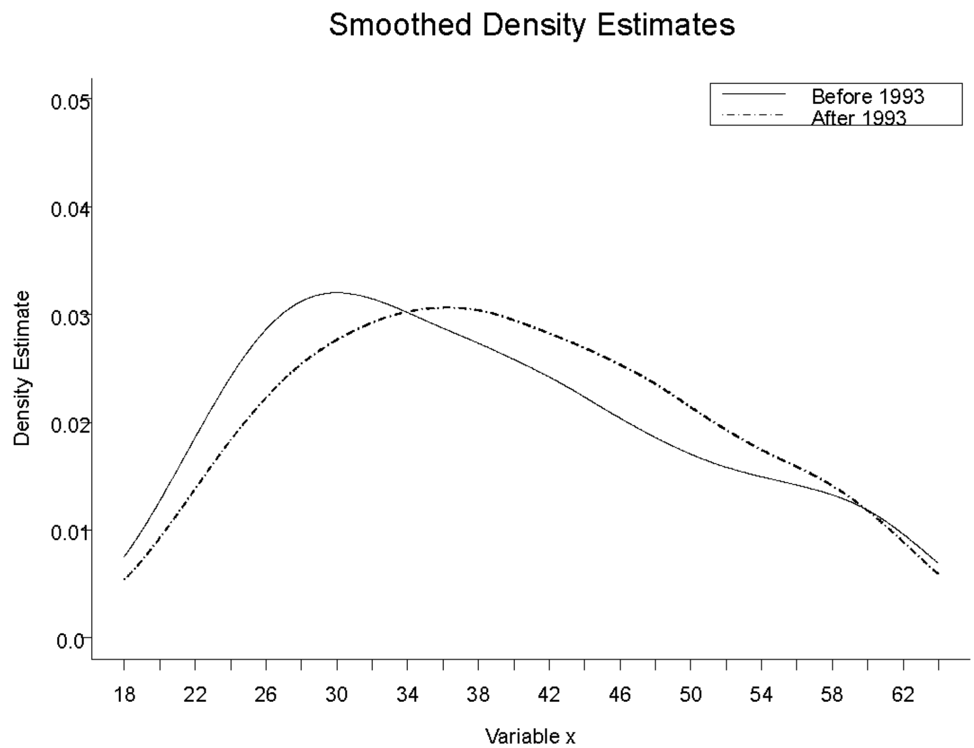
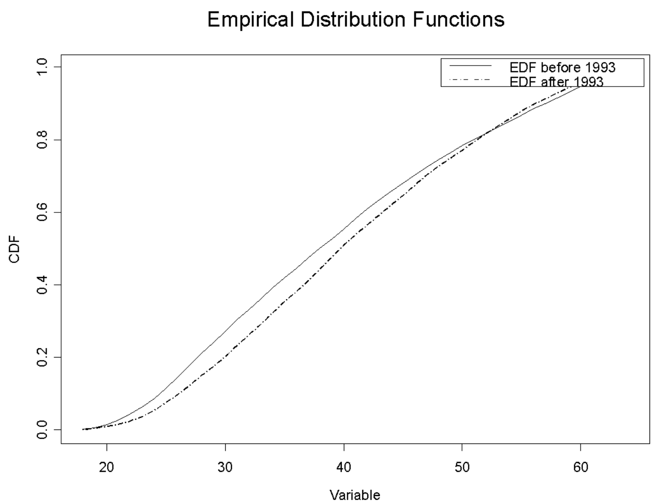


Fig 4B. Density estimates for Pennsylvania.

Suppose now, that the sample $y_1, y_2, ..., y_m$ comes from a population with PDF $g(z)$ after 1993. The probability integral transforms (PIT) of each z_i based on the density estimates are

$$z_j = F_n(y_j), \ j = 1, 2, ..., m. \tag{33}$$

We estimate the cumulative distribution function (CDF) based on the equation (33) by simply calculating the empirical distribution function (EDF) of x. If the null hypothesis of the correct specification of the model is true then $z_1, z_2, ..., z_m$ should be distributed as $U(0,1)$ (see Fig. 5B). However, if the null hypothesis is not true, then Neyman smooth test might give us the direction (s) of departure(s) from the null hypothesis.



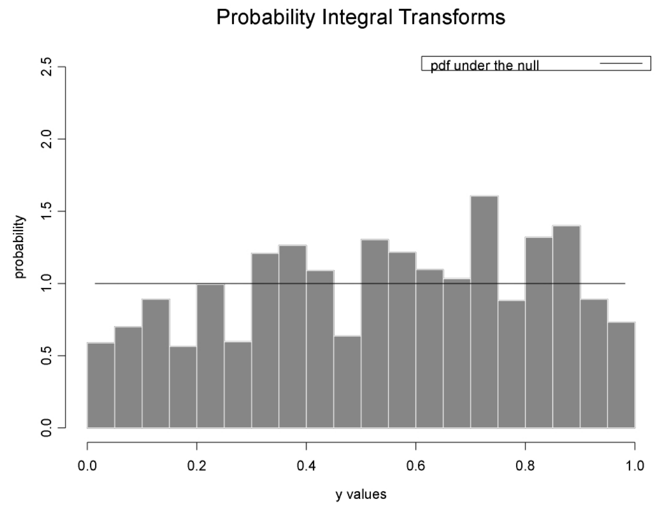


Fig 5A. Age distributions of small groups.

Fig 5B. PIT histogram for age after 1993.

Kolmogorov-Smirnov (KS) or the Cramér-von Mises (CvM) types of tests are the most commonly used tests based on the EDF for comparing two distributions. KS statistic is the maximum distance between the two EDF's in fig. 5A while CvM statistic is a weighted sum of the squares of the difference between the two CDFs. Table 1. gives us the values of all the commonly used test statistics based on the EDF of individuals covered under small employer sponsored group insurance in New York and Pennsylvania along with the .1% critical values of all the modified statistics [Stephens (1970)].

Test Statistic	New York	Pennsylvania	Critical Values Upper .1%
D^+	1.128	0.8915	1.859
D^-	4.3492	4.4053	1.859
KS	4.3492	4.4053	1.95
Kuiper	5.4809	5.3015	2.303
CvM	5.3503	5.1944	1.167
A-D	28.4875	25.2846	6.0
W	2.0858	1.9058	0.385

Table 1. Goodness-of-Fit Statistics based on EDF.

Except for the test on D^+ [which considers the positive part of (1)], all the tests clearly indicate that the two distributions are different. However, these test statistics based on the EDF fails to indicate the nature of the deviation from the null hypothesis.

In order to attempt to answer the above question we use Neyman's smooth test based on the test statistic

$$\Psi_k^2 = \sum_{j=1}^{k} u_j^2,\ u_j = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j\left(z_i\right),\ j = 1, ..., k,\ i = 1, ..., m, \tag{34}$$

where we chose $k = 4$ and $\pi_j(.)$ as the j^{th} order normalized Legendre polynomial discussed in Neyman (1937). Table 2 gives the results on the data on all the individuals under *small employer* sponsored group in New York (with n=4548 and m=2517).

Source	Ψ^2	u_1^2	u_2^2	u_3^2	u_4^2
Test Statistic	90.41531	39.15748	44.75897	3.91509	2.58377
p-value	0.0	0.0	0.0	0.04785	0.10796

Table 2. Neyman's smooth statistic and components.

In place of taking the whole population if we just use a random sample of size $m = 500$ of individuals covered under small employer sponsored group insurance in New York, the results we are given in Table 2A.

Source	Ψ^2	u_1^2	u_2^2	u_3^2	u_4^2
Test Statistic	27.6095	7.4413	13.7468	3.8773	2.544
p-value	0.000015	0.00637	0.00021	0.04894	0.11071

Table 2A. Neyman's smooth statistic and components for a sample.

Under the null, Ψ_k^2 should have a χ^2 distribution with k (here we have $k = 4$) degrees of freedom and each of the u_j^2 should be independent χ^2 with 1 degree of freedom. We can interpret the components of Ψ^2 given by u_j^2 as some measure of the j^{th} order of the departure from the null hypothesis. If we perform the same test on all the individuals who has any health insurance coverage ($\Psi^2 = 60.30782$) and those who purchased individual insurance ($\Psi^2 = 10.81446$) we see the departure from the null hypothesis with varying degrees.

Although, based on these results it appears that the population of civilian adults who had insurance before 1993 and after 1993 were distinctly different, however, as shown in the Table 3. the summary statistics of the two distribution are similar.

State	New York		Pennsylvania	
Data	Before 1993	After 1993	Before 1993	After 1993
Observations	4548	2517	3113	1875
Mean	39.2535	40.5268	38.5239	39.9712
Standard Deviation	11.85	11.0622	11.8109	11.1529
Skewness Coefficient	0.3082	0.1793	0.4106	0.216
Excess Kurtosis	-0.9381	-0.8858	-0.834	-0.8548
Minimum	18	18	18	18
1^{st} Quartile	29	32	29	31
Median	38	40	37	39
3^{rd} Quartile	48	49	47	48
Maximum	64	64	64	64

Table 3. Age distribution for small employers insured group in NY and PA.

If we perform a simple test for equality of means in the samples for New York before and after 1993 we get a p-value of 6.3×10^{-6} while the test for difference of variances gives 5.1×10^{-5}. Using the components of Neyman's smooth test we can conclude that the age distributions before and after 1993 are distinctly different and the major departure seems to be due to a difference in the first and second order terms, although third order term also plays some part in the difference in the two distributions.

5 Monte Carlo Study of Smooth test with EDF

Some preliminary size and power Monte Carlo studies have shown that the result of the smooth test have better size properties for samples where the evaluation sample size (which is used to estimate the EDF) is much bigger than the test sample size (which is used for performing the smooth test). Table 4. shows the actual size of a 5% nominal level test where the samples are drawn at random from the population of insured civilian adults who has health coverage sponsored by small employers (with less than 100 employees) before 1993 when the “community rating” legislation was enacted.

Source Null Distribution	Ψ^2 χ^2_4	u_1^2 χ^2_1	u_2^2 χ^2_1	u_3^2 χ^2_1	u_4^2 χ^2_1
actual size ($n = 2500, m = 500$)	0.0577	0.0423	0.0529	0.0497	0.083
actual size ($n = 2500, m = 50$)	0.0425	0.0473	0.0474	0.0461	0.0473

Table 4. Actual sizes for 5% nominal size smooth tests.

The results are based on evaluation sample size $n = 2500$ and the test sample size $m = 500$ and $m = 50$,, with 10000 replications.

From the results we see that if we have the test sample size n to be substantially smaller than the evaluation sample size n improves the size distortion of the resulting smooth test. We have shown in Theorem 2 that having m to be in the order of $\sqrt{n}$ is optimal from the point of view of size distortion.

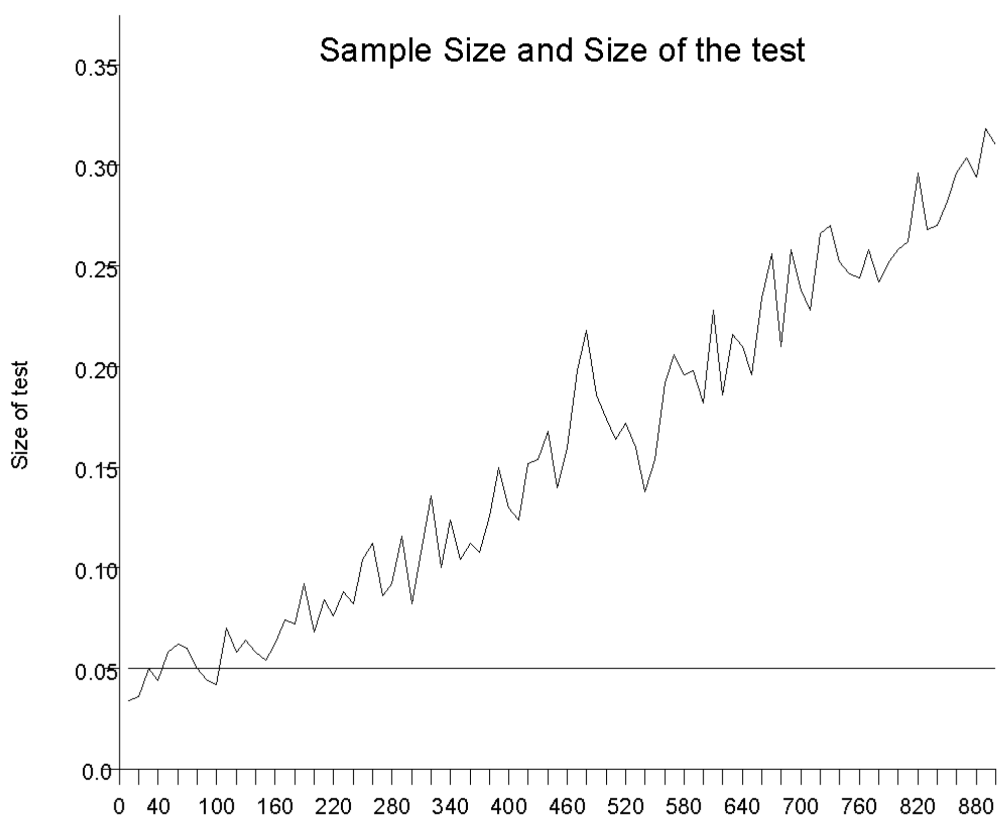


Fig 6A. Sample size and test size for $n = 900$, increasing m, no. of repl.=500.

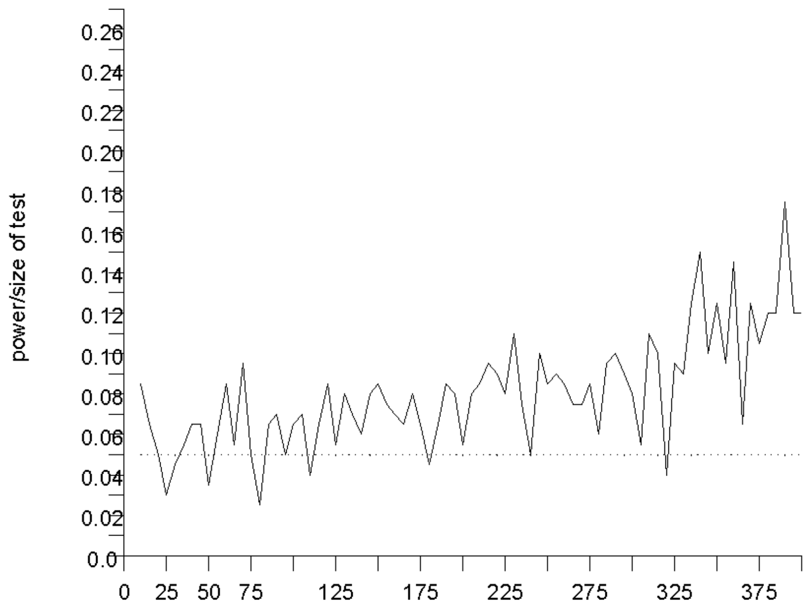


Fig 6B. Sample size and test size for $n = 2500$, increasing m, no. of repl.=200.

In Fig 9 shows a plot the size of the test with the sample size of the test sample drawn at random from the population of employees covered by small employers group insurance. The evaluation sample of size $n = 900$ is also drawn at random from the same population. For calculating the size of the smooth test we took 500 replications and evaluated the size for each value of $m = 10$ to $m = 900$ in increments of 10. The results with increments of 5 is also similar shown in figure. We observe that the test size increases more or less with the sample size m after a while. Furthermore, there could be several values of m for $n = 900$ where we have a test with a 5% actual size.

Although, subsampling from the original data of small insurance population before 1993 in New York gives us an idea about how size might be affected by the sample size keeping the evaluation sample size the same, it will not give us much idea about the power properties. Therefore, we redo the whole exercise with simulated data using a mixture of Gaussian and log-normal distributions which gives us a density close to the real data at least visually. The test size and sample size plot is given in Fig. 4. In order to get a better idea about the relative size and power properties for a data from a population of the variable X which is

$$X = B * X_1 + (1 - B) * X_2 \tag{35}$$

where $B \sim Bernoulli(0.3)$, X_1 has a $\log normal$ distribution with $\mu = -1.2$, $\sigma^2 = 4$ and X_2 is $N\left(\mu = 1.2, \sigma^2 = 1.21\right)$. Suppose, we also consider the alternative where $B \sim Bernoulli(0.5)$, $\ln X_1$ is $N\left(-0.1, 1\right)$ and finally X_2 has $N\left(1.75, 0.81\right)$. In the following charts Fig 7 and Fig 7A, we have considered $n = 625$ and gradually increased m in increments of 5 to calculate the size of the tests based on 200 replications. Fig 7A is a magnification of Fig 7 upto $m = 250$.

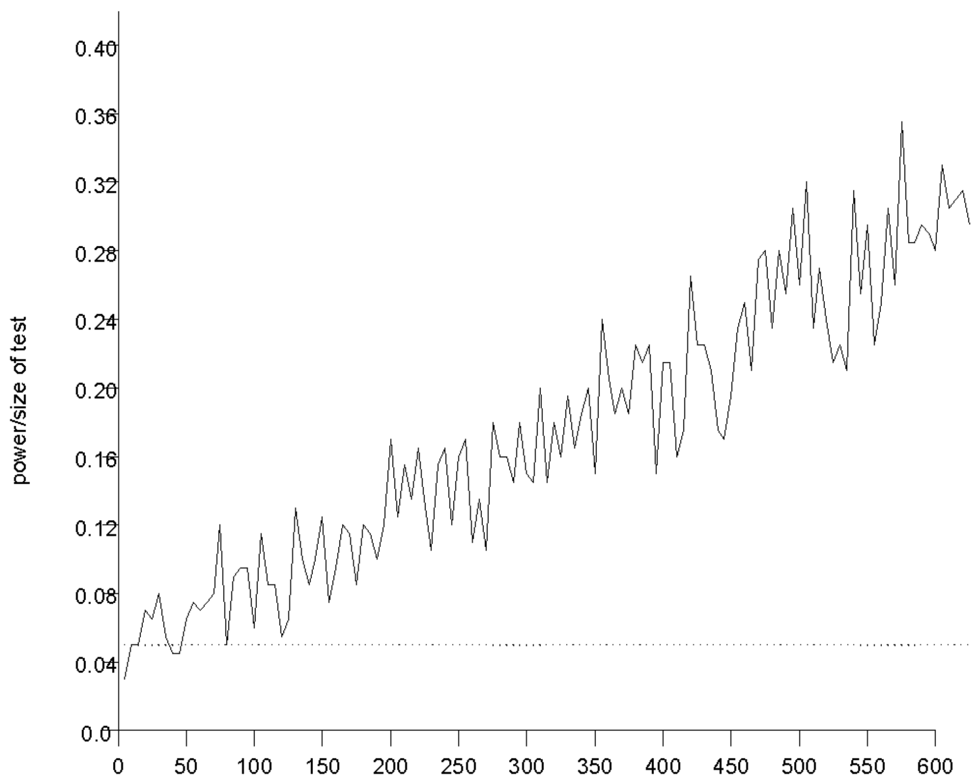


Fig 7. Sample size and the size of smooth test with simulated data using EDF.

Fig 7A. Sample size and the size of smooth test with simulated data using EDF.

Fig 8 shows the plot of the sample size m with the power of the smooth test based on the EDF of a sample of size $n = 625$.

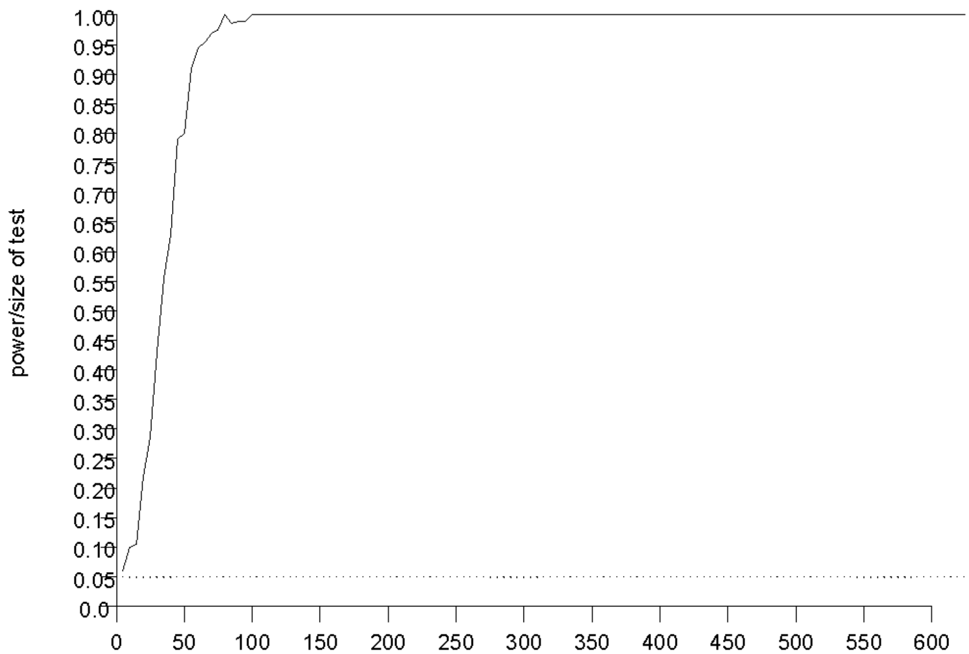


Fig 8. Sample size and power of the smooth test with EDF with simulated data.

Putting the above two charts together in Fig 9, we can get some idea about the range of the optimal sample size for the test sample m given a fixed value n of the evaluation sample (here $n = 625$). It is also worth noting that the power properties dramatically increase as we increase the size of the test sample m. For example, if $n = 2500$ then using $m = 50$, the power of the smooth test is 0.855 while the size is 0.05. If we plot the test sample size with the size and power of the test for $n = 2000$ and m between 10 and 600, we observe that the size distortion is less pronounced (see Fig. 10).

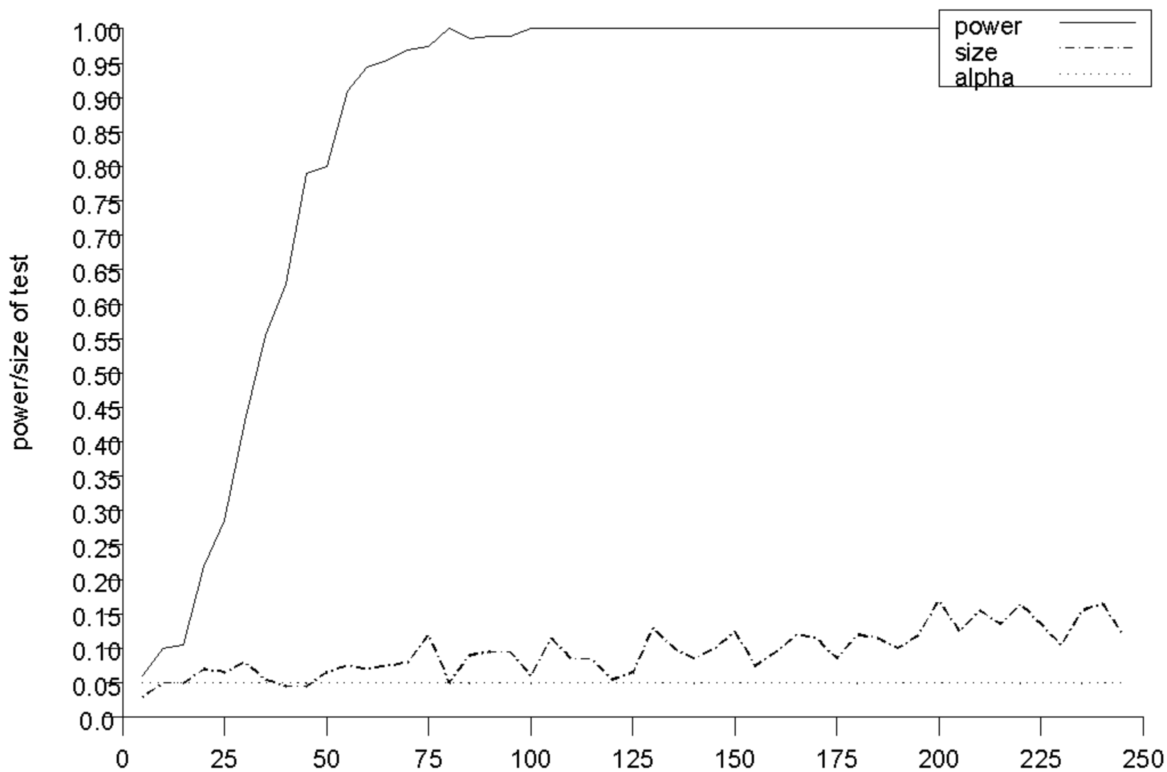


Fig 9. Size and Power of the smooth test for simulated data ($m = 625$)

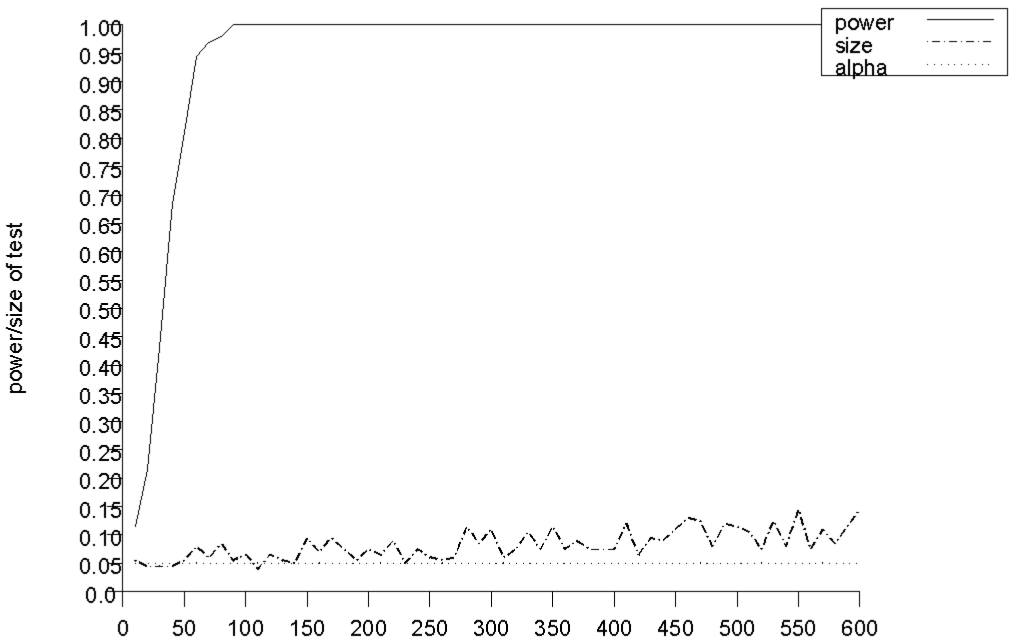


Fig 10. Size and Power of the smooth test for simulated data ($n = 2000$).

The optimal value of m relative to n and also the size and power of the smooth test can be obtained from a range of values of m with acceptable value of size. For a more analytical answer, we have already shown that the size of m should be in the order of $\sqrt{n}$ for minimizing the size distortion, however the exact relationship of m and n will depend on the values of the coefficients of the higher order terms in the Cornish-Fisher expansion discussed in the Section 3 under Theorem 2.

6 Conclusion and Future Research

We proposed a smooth test for comparing two densities in a two sample setup. Unlike traditional goodness of fit tests like the Kolmogorov-Smirnov or the Cramér-von Mises classes of tests, the smooth test for comparing two densities helps us identify the nature of the discrepancy between the two densities, and hence, the sources of departure from the null hypothesis that the two densities are identical. We have already seen that the smooth test is nothing but Rao's score test, and hence it enjoys all the optimality properties of the score test (Bera and Bilias, 2001). We have also shown that the choices of the relative sizes of the estimation and test samples are very important to get a consistent test with minimum size distortion. Unfortunately, in the current setup the exact values of the relative sizes cannot be obtained, we can only give the test sample size upto the $\sqrt{n}$ -order of the estimation sample size. Hence, we have to rely on Monte Carlo simulation to get what the relative sample sizes should be. Our Monte Carlo simulation also revealed that the smooth test has very good power properties and an optimal choice of the test sample size will also ensure that the true size is close to the nominal size of the test even in finite samples.

There are several directions of future research that we want to pursue. One of the questions that we would like to answer is how we can adjust one distribution to remove the effect of another distribution, very much in the line of what the difference-in-difference estimate tries to accomplish (Dinardo and Buchmuller, 2001). Only in that case, we can provide a viable alternative procedure for examining whether the "Adverse Selection death spiral" can be observed in the insurance market. We have looked at the global characteristics of the two distributions we are comparing, however, there could be some very local chracteristics of the density function that we have not focused on. We can extend the smooth test for comparison of two densities similar to what Fan (1996) proposed adaptively using both Neyman smooth test technique and a wavelet based test for comparing global and local departures from the null hypothesis. Last but not the least, we have to accomodate for possible dependence in the data when comparing two distributions particularly in the context of time series or panel data, we have not addressed that issue in this context.

There are several possible applications of this testing technique, we can list just a few of them. Smooth test can be used for comparing wage distributions between different genders, different ethnic groups or different residential neighborhoods like public housing projects or wealthier localites in Toronto (Oreopoulos 2002). Regional differences between unemployment rates in 50 Spanish provinces between two different regimes, 1985 and 1997 can also be investigated using the smooth test technique (López-Baro, Barrio and Artis, 2002). There has been an extensive literature in demography and public policy economics on causes adverse birth outcomes in terms of low birth weight (LBW) including socio-economic conditions and behavioral patterns like smoking habits of mothers during pregnancy [see Evans and Ringel (1999)]. Smooth test based techniques can also be used for several other studies explaining the infant mortality gap between two ethnic groups by comparing the distributions of birth weight (Miller 2002). In finance, we can look at the problem of comparing distribution of pre-tax and after-tax returns of mutual funds for taxable investors that can determine the cash inflows of these funds (Bergstresser and Poterba 2000).

7 Appendix

7.1 Proof of Theorem1

We first give two Lemmas that can be used in our proof.

Lemma 1 *If we define for $P_l(z) = \left\{(2z-1)^2 - 1\right\}^l$ for any $l \leq k$ and $l \geq 1$, then for any $r \leq 2l$*

$$\frac{d^r}{dz^r} P_l(z) \leq 2^{2l} \left[\sum_{j=0}^{\min(l,2l-r)} \frac{l!}{j!\,(l-j)!} \frac{(2l-j)!}{(2l-j-r)!} \right]. \tag{36}$$

Proof. Let us use the following simplification of the expression $\left\{(2z-1)^2 - 1\right\}^l$,

$$\begin{aligned} P_l(z) &= \left\{(2z-1)^2 - 1\right\}^l = \left\{4z^2 - 4z + 1 - 1\right\}^l \\ &= 4^l z^l (z-1)^l \\ &= 2^{2l} \sum_{j=0}^{l} \frac{l!}{j!\,(l-j)!} z^{2l-j} \end{aligned} \tag{37}$$

Since $z \in (0,1)$,

$$\begin{aligned} \frac{d}{dz} P_l(z) &= 2^{2l} \sum_{j=0}^{l} \frac{l!}{j!\,(l-j)!} (2l-j)\, z^{2l-j-1} \\ &= 2^{2l} \sum_{j=0}^{\min(l,2l-1)} \frac{l!}{j!\,(l-j)!} \frac{(2l-j)!}{(2l-j-1)!} z^{2l-j-1}. \end{aligned} \tag{38}$$

Thus (36) holds for $r = 1$. Now suppose that the equality in (38) is true for the rth derivative of $P_l(z)$ for any arbitrary $0 \leq r \leq 2l-1$, then we will show it also holds for $r+1$. We have

$$
\begin{aligned}
& \frac{d^{r+1}}{dz^{r+1}} P_l(z) \\
& = \frac{d}{dz}\left(\frac{d^r}{dz^r} P_l(z)\right) \\
& = \frac{d}{dz}\left(2^{2l}\left[\sum_{j=0}^{\min(l,2l-r)} \frac{l!}{j!\,(l-j)!} \frac{(2l-j)!}{(2l-j-r)!} z^{2l-j-r}\right]\right) \\
& = 2^{2l}\left[\sum_{j=0}^{\min(l,2l-r-1)} \frac{l!}{j!\,(l-j)!} \frac{(2l-j)!}{(2l-j-r)!}(2l-j-r)\, z^{2l-j-r-1}\right] \\
& = 2^{2l}\left[\sum_{j=0}^{\min(l,2l-r-1)} \frac{l!}{j!\,(l-j)!} \frac{(2l-j)!}{(2l-j-r-1)!} z^{2l-j-r-1}\right]. \qquad (39)
\end{aligned}
$$

Therefore, for any $r \leq 2l$,

$$
\frac{d^r}{dz^r} P_l(z) = 2^{2l}\left[\sum_{j=0}^{\min(l,2l-r)} \frac{l!}{j!\,(l-j)!} \frac{(2l-j)!}{(2l-j-r)!} z^{2l-j-r}\right]. \qquad (40)
$$

Since $z \in (0,1)$, from (40) the result follows. ■

Lemma 2 *If $\pi_l(.)$ is the normalized Legendre polynomial of degree l defined on $(0,1)$, then the first order Lipschitz condition holds, that is*

$$
|\pi_l(\hat{z}) - \pi_l(z)| \leq M\, |\hat{z} - z| \qquad (41)
$$

where M is a positive constant, and z and $\hat{z}$ are any two points between 0 *and* 1.

Proof. The Legendre polynomials are defined as [see, for instance, Kendall and Stuart (1973, p. 460)]

$$
\begin{aligned}
L_l(z) &= \left(l!2^l\right)^{-1} \frac{d^l}{dz^l}\left\{\left(z^2-1\right)^l\right\} \\
\text{where} \int_{-1}^{1} L_l(z)\, L_k(z)\, dz &= 0 \quad \text{if } l \neq k, \\
&= \tfrac{2}{2l+1} \quad \text{if } l = k.
\end{aligned} \qquad (42)
$$

The corresponding orthonormal polynomials on $(0,1)$ are

$$\begin{aligned}
\pi_l(z) &= (2l+1)^{\frac{1}{2}} L_l\left(2\left(z-\frac{1}{2}\right)\right) \\
&= (2l+1)^{\frac{1}{2}} L_l(2z-1) \\
&= (2l+1)^{\frac{1}{2}} \left(l!2^l\right)^{-1} \frac{d^l}{d(2z-1)^l}\left[(2z-1)^2-1\right]^l \\
&= (2l+1)^{\frac{1}{2}} \left(l!2^{l+1}\right)^{-1} \frac{d^l}{dz^l}\left[(2z-1)^2-1\right]^l. \qquad (43)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\pi_l(\hat{z}) - \pi_l(z) = (2l+1)^{\frac{1}{2}} \left(l!2^{l+1}\right)^{-1} \times \\
\left[\frac{d^l}{d\hat{z}^l}\left\{(2\hat{z}-1)^2-1\right\}^l - \frac{d^l}{dz^l}\left\{(2z-1)^2-1\right\}^l\right]. \qquad (44)
\end{aligned}$$

For $|z^*-z| < |\hat{z}-z|$ the mean value theorem gives

$$\begin{aligned}
\frac{d^l}{dz^l}P_l(\hat{z}) &= \frac{d^l}{dz^l}P_l(z) + (\hat{z}-z)\frac{d^{l+1}}{dz^{l+1}}P_l(z^*) \\
\text{i.e., } \frac{d^l}{dz^l}P_l(\hat{z}) - \frac{d^l}{dz^l}P_l(z) &= (\hat{z}-z)\frac{d^{l+1}}{dz^{l+1}}P_l(z^*). \qquad (45)
\end{aligned}$$

Hence, using (42) and (45) and applying Lemma 1, we have

$$\begin{aligned}
&|\pi_l(\hat{z}) - \pi_l(z)| \\
&= (2l+1)^{\frac{1}{2}} \left(l!2^{l+1}\right)^{-1} \left|\left[\frac{d^l}{dz^l}P_l(\hat{z}) - \frac{d^l}{dz^l}P_l(z)\right]\right| \\
&= (2l+1)^{\frac{1}{2}} \left(l!2^{l+1}\right)^{-1} \left|\left[(\hat{z}-z)\frac{d^{l+1}}{dz^{l+1}}P_l(z^*)\right]\right| \\
&\leq |\hat{z}-z| (2l+1)^{\frac{1}{2}} \\
&\times \left(l!2^{l+1}\right)^{-1} 2^{2l} \left[\sum_{j=0}^{\min(l,2l-1)} \frac{l!}{j!(l-j)!}\frac{(2l-j)!}{(2l-j-1)!}\right] \\
&= M|\hat{z}-z|, \qquad (46)
\end{aligned}$$

where M is a finite number. ■

Proof of Theorem 1

We only need to show that

$$\widehat{\Psi}_k^2 - \Psi_k^2 = o_p(1).$$

Notice that

$$
\begin{aligned}
\widehat{\Psi}_k^2 &= \sum_{j=1}^{k}\left[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j(\widehat{Z}_i)\right]^2 \\
&= \sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\pi_j(Z_i)+\sum_{i=1}^{m}\left[\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right]\right]^2 \\
&= \sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\pi_j(Z_i)\right]^2+\sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\left[\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right]\right]^2 \\
&+ 2\sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\pi_j(Z_i)\right]\left[\sum_{i=1}^{m}\left[\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right]\right] \\
&= \Psi_k^2 + R_{1,m,n} + R_{2,m,n}
\end{aligned}
$$

where

$$
\begin{aligned}
R_{1,m,n} &= \sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\left[\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right]\right]^2, \\
R_{2,m,n} &= 2\sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\pi_j(Z_i)\right]\left[\sum_{i=1}^{m}\left[\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right]\right].
\end{aligned}
$$

We show that, as $m,n\to\infty$, and $(\log\log n)m/n\to 0$,

$$
R_{1,m,n}=o_p(1),\ \ R_{2,m,n}=o_p(1).
$$

We first look at $R_{1,m,n}$.

$$
\begin{aligned}
R_{1,m,n} &= \sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\left[\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right]\right]^2 \\
&\le \sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\left|\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right|\right]^2.
\end{aligned}
$$

By Lemma 2, the orthonormal polynomials $\pi_j(\cdot)$ satisfy the first-order Lipschitz condition that

$$
|\pi_j(u)-\pi_j(v)|\le M\,|u-v|\,,
$$

for some finite number $M > 0$. Thus

$$\begin{aligned} R_{1,m,n} &\le \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \left| \pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right| \right]^2 \\ &\le M^2 \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \left| \widehat{Z}_i - Z_i \right| \right]^2 \\ &\le M^2 \sum_{j=1}^{k} \frac{1}{m} \left[m \max_{1\le i\le m} \left| \widehat{Z}_i - Z_i \right| \right]^2 \end{aligned}$$

Notice that for each fixed Y_i, by Donsker's Theorem we have that

$$\sqrt{n}\,[F_n(Y_i) - F(Y_i)]$$

converges in distribution to a normal variate. In addition, $\sup_u |F_n(u) - F(u)|$ is maximally of order $\sqrt{(\log\log n)/n}$ as $n \to \infty$ (see, e.g. Shorack and Wellner (1986)). Thus

$$\max_{1\le i\le m} \left| \widehat{Z}_i - Z_i \right| = O_p(\sqrt{(\log\log n)/n})$$

and

$$R_{1,m,n} \le mM^2 \sum_{j=1}^{k} \left[\max_{1\le i\le m} \left| \widehat{Z}_i - Z_i \right| \right]^2 = O_p((\log\log n)m/n),$$

which converges to 0 if $(\log\log n)m/n \to 0$ as $m, n \to \infty$.

Similarly,

$$\begin{aligned} |R_{2,m,n}| &= 2 \sum_{j=1}^{k} \frac{1}{m} \left| \sum_{i=1}^{m} \pi_j(Z_i) \right| \left| \sum_{i=1}^{m} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right| \\ &\le 2 \sum_{j=1}^{k} \frac{1}{m} \left| \sum_{i=1}^{m} \pi_j(Z_i) \right| \left[\sum_{i=1}^{m} \left| \pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right| \right]. \end{aligned}$$

Notice that $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j(Z_i)$ converges to a standard normal variable and

$$\sum_{i=1}^{m} \pi_j(Z_i) = O_p(m^{1/2}).$$

By a similar analysis to that for $R_{1,m,n}$, we have

$$\sum_{i=1}^{m} \left| \pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right| \le M \sum_{i=1}^{m} \left| \widehat{Z}_i - Z_i \right| \le Mm \max_{1\le i\le m} \left| \widehat{Z}_i - Z_i \right| = O_p(m\sqrt{(\log\log n)/n}).$$

Thus

$$
\begin{aligned}
|R_{2,m,n}| &\leq 2\sum_{j=1}^{k}\frac{1}{m}\left|\sum_{i=1}^{m}\pi_j(Z_i)\right|\left[\sum_{i=1}^{m}\left|\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right|\right] \\
&= O_p\left(\frac{1}{m}\times m^{1/2}\times m\sqrt{(\log\log n)/n}\right) \\
&= O_p(\sqrt{(\log\log n)m/n}),
\end{aligned}
$$

converging to 0 when $(\log\log n)m/n \to 0$ as $m, n \to \infty$. ■

7.2 Proof of Theorem 2

>From the proof of Theorem 1, we know that the asymptotic χ^2 test statistic $\widehat{\Psi}_k^2$ can be decomposed into

$$
\Psi_k^2 + R_{1,m,n} + R_{2,m,n}.
$$

We analyze each of these terms to show how the relative magnitude of m and n affects the testing size.

For the leading term

$$
\Psi_k^2 = \sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\pi_j(Z_i)\right]^2.
$$

By construction, conditional on X, for each j, $V_{ji} = \pi_j(Z_i)$ $(i = 1, \ldots.., m)$ are m independent and identically distributed random variables with mean zero and unit variance, thus

$$
\frac{1}{\sqrt{m}}\sum_{i=1}^{m}V_{ji} \Rightarrow N(0,1) \equiv \xi_j,
$$

where ξ_j $(j = 1, \ldots., k)$ are k independent standard normal variates. In addition, assuming that V_{ji} possesses moments up to the forth order, by standard result of Edgeworth expansion (see, e.g., Rothenberg 1984 and references therein), the probability density function of $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}V_{ji}$ has the following expansion

$$
f(x) \approx \varphi(x)\left[1+\frac{k_3H_3(x)}{6\sqrt{m}}+\frac{3k_4H_4(x)+k_3^2H_6(x)}{72m}\right]
$$

where φ is the density of standard normal, k_r is the r-th cumulant, and H_r is the Hermite polynomial of degree r defined as $H_r(x) = (-1)^r\varphi^{(r)}(x)/\varphi(x)$. In this sense, $m^{-1/2}\sum_{i=1}^{m}V_{ji}$ can be expanded into a leading term of standard normal variable ξ_j

plus a second order term, say $\frac{1}{\sqrt{m}}A_j$, of $O_p(m^{-1/2})$ and a third term, $\frac{1}{m}B_j$, of order $O_p(m^{-1})$:

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m} V_{ji} \approx \xi_j + \frac{1}{\sqrt{m}}A_j + \frac{1}{m}B_j.$$

Therefore, the term

$$\Psi_k^2 = \sum_{j=1}^{k} \frac{1}{m}\left[\sum_{i=1}^{m} \pi_j(Z_i)\right]^2$$

can be expanded as

$$\sum_{j=1}^{k} \xi_j^2 + \frac{1}{\sqrt{m}}A + \frac{1}{m}B + o_p(\frac{1}{m})$$

where the leading term $\sum_{j=1}^{k} \xi_j^2$ is the χ^2 random variable with k degree of freedom and the second term $\frac{1}{\sqrt{m}}A$ is of order $O_p(m^{-1/2})$. To obtain a good approximation of the χ^2 distribution, a large m is preferred.

Now we turn to the estimation of distribution functions.

$$\begin{aligned} R_{1,m,n} &= \sum_{j=1}^{k} \frac{1}{m}\left[\sum_{i=1}^{m}\left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i)\right]\right]^2 \\ &= \frac{m}{n}\sum_{j=1}^{k}\left[\frac{1}{m}\sum_{i=1}^{m}\sqrt{n}\left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i)\right]\right]^2 \\ &= \frac{m}{n}C. \end{aligned}$$

Notice that

$$U_{ji} = \sqrt{n}\left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i)\right] = O_p(1),$$

and $\frac{1}{m}\sum_{i=1}^{m} U_{ji} = O_p(1)$, we have

$$C = \sum_{j=1}^{k}\left[\frac{1}{m}\sum_{i=1}^{m}\sqrt{n}\left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i)\right]\right]^2 = O_p(1),$$

and thus

$$R_{1,m,n} = O_p\left(\frac{m}{n}\right).$$

Similarly,

$$
\begin{aligned}
R_{2,m,n} &= 2\sum_{j=1}^{k}\frac{1}{m}\left[\sum_{i=1}^{m}\pi_j(Z_i)\right]\left[\sum_{i=1}^{m}\left[\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right]\right] \\
&= 2\sqrt{\frac{m}{n}}\sum_{j=1}^{k}\left[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j(Z_i)\right]\left[\frac{1}{m}\sum_{i=1}^{m}\sqrt{n}\left[\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right]\right] \\
&= O_p\left(\sqrt{\frac{m}{n}}\right)
\end{aligned}
$$

since

$$
D = 2\sum_{j=1}^{k}\left[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j(Z_i)\right]\left[\frac{1}{m}\sum_{i=1}^{m}\sqrt{n}\left[\pi_j(\widehat{Z}_i)-\pi_j(Z_i)\right]\right] = O_p(1).
$$

In summary

$$
\begin{aligned}
\widehat{\Psi}_k^2 &= \Psi_k^2 + R_{1,m,n} + R_{2,m,n} \\
&= \sum_{j=1}^{k}\xi_j^2 + \frac{1}{\sqrt{m}}A + \sqrt{\frac{m}{n}}D + o_p(\frac{1}{\sqrt{m}} + \sqrt{\frac{m}{n}}),
\end{aligned}
$$

where $\frac{1}{\sqrt{m}}A$, $\sqrt{\frac{m}{n}}D$, etc., are higher order terms that brings size distortion in finite sample, but are $o_p(1)$. In particular, the leading terms are of order $O_p\left(\frac{1}{\sqrt{m}}\right)$ and $O_p\left(\sqrt{\frac{m}{n}}\right)$ respectively. To minimize the distortion coming from estimating the distribution function, we prefer n to be larger relative to m. On the other hand, to obtain fast convergence to the χ^2 limit, we want a large m. Thus, a trade-off has to be made to minimize size distortion. We balance these two terms so that they are of the same order of magnitude, giving the optimal relative magnitude as $m = O(\sqrt{n})$.
$\blacksquare\blacksquare$

References

[1] Anderson, T. W.(1962), Annals of Mathematical Statistics, 33, pp.1148-1159

[2] Anderson, T. W. and D.A. Darling (1952), Asymptotic Theory of certain "goodness of fit" criteria based on stochastic processes. Annals of Mathematical Statistics, 23, pp. 193-212

[3] Anderson, T. W. and D.A. Darling (1954), A test of goodness of fit. Journal of the American Statistical Association, 49, pp. 765-769.

[4] Andrews, D. W. K.(1994), Empirical Process methods in Econometrics. In Handbook of Econometrics, Vol IV. Eds. R. F. Engle and D. L. McFadden. Elsevier Science: Amsterdam, pp. 2247- 2294.

[5] Andrews, D. W. K. and M. Buchinsky. A three-step method for choosing the number of bootstrap replications. Econometrica, 68, pp. 23-51.

[6] Bahadur, R. R. (1966), A note on quantiles in large samples, Annals of Mathematical Statistics, 37, pp. 577-580.

[7] Bera, A. K. and A. Ghosh (2001), Neyman's smooth test and its applications in Econometrics. In: Handbook of Applied Econometrics and Statistical Inference, Eds. A. Ullah, A. Wan and A. Chaturvedi, Marcel Dekker, pp. 177-230.

[8] Billingsley, P. (1968). Convergence in Probability Measures. John Wiley: New York.

[9] Ćwik, J and J. Mielniczuk (1989), Estimating density ratio with application to discriminant analysis, Communications in Statistics: Theory and Methods, 18, pp. 3057-3069.

[10] Darling, D.A. (1955), The Cramér-Smirnov test in the parametric case. Annals of Mathematical Statistics, 26, pp. 1-20.

[11] Darling, D.A. (1957), The Kolmogorov-Smirnov, Cramér-von Mises tests. Annals of Mathematical Statistics, 28, pp. 823-838.

[12] DiNardo, J. and Buchmueller T. (2001), Did Community Rating Induce an Adverse Selection Death Spiral? Evidence from New York, Pennsylvania and Connecticut. manuscript.

[13] Durbin, J.(1973), Distribution Theory for tests based on the sample distribution function. In Regional Conference Series in Applied Mathematics, Vol. 9, SIAM: Philadelphia, PA.

[14] Durbin, J. and M. Knott. Components of Cramer-von Mises Statistics. I. Journal of the Royal Statistical Society. Series B (Methodological), 34:290-307, 1972.

[15] Eubank, R.L. and V.N. LaRiccia. Asymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. Annals of Statistics 20:1412-1425, 1992.

[16] Fan, J. (1996) Test of significance based on wavelet thresholding and Neyman's truncation. Journal of the American Statistical Association 91:674-688.

[17] Fan, J. and L-S. Huang (2001). Goodness-of-fit tests for parametric regression models. Journal of the American Statistical Association, 96, pp. 640-652.

[18] Finkelstein, H. (1971) The Law of the Iterated Logarithm for Empirical Distribution, Annals of Mathematical Statistics, 42, pp. 607-615.

[19] Gaenssler, P. and W. Stute (1979), Empirical Processes: A survey of results for independent and identically distributed random variables, Annals of Probability, 7, pp.193-243.

[20] Gourieroux, C. and A. Monfort (1991), Simulation based Inference in Models with Heterogeneity. Annales d'Econonomie et de Statistique, 20/21, pp. 69-107.

[21] Gourieroux, C. and A. Monfort (1993), Pseudo-Likelihood Methods. In Handbook of Statistics, Vol. 11, Eds, C. R. Rao, H. D. Vinod, Elsevier Science; Amsterdam, pp. 335-362.

[22] Gourieroux, C. and A. Monfort (1996), Simulation-based Econometric Methods. Oxford University Press: Oxford.

[23] Hall, P. (1992), The Bootstrap and Edgeworth Expansion. Springer-Verlag: New York.

[24] Handcock M.S. and M. Morris (1999), Relative Distribution Methods in Social Sciences. Springer: New York.

[25] Kendall, M.G. and A. Stuart (1969), The Advanced Theory of Statistics, 3rd Edition, Vol. 2. Hafner: New York.

[26] Kagan, A.M., Yu. V. Linnik and C. R. Rao (Translated from Russian text by B. Ramachandran), (1973) Characterization problems in mathematical statistics. Wiley: New York.

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

[28] Lehmann, E. L. (1953), The Power of Rank Tests, Annals of Mathematical Statistics, 24, pp. 23-43.

[29] Neyman J. "Smooth test" for goodness of fit. Skandinaviske Aktuarietidskrift 20:150-199, 1937.

[30] Neyman, J. and E.S. Pearson (1936). Contributions to the theory of testing statistical hypothesis I: Unbiased critical regions of Type A and A_1. Statistical Research Memoirs 1:1-37.

[31] Parzen E. (1992), Comparison change analysis. In Nonparametric Statistics and Related Topics, ed. A. Saleh. Elsevier: Holland, pp. 3-15.

[32] Pettitt, A. N. (1977), A Cramér-von Mises type goodness of fit statistic related to $\sqrt{b_1}$ and b_2. Journal of the Royal Statistical Society. Series B (Methodological), 39, 364-370.

[33] Pyke, R. (1984). Asymptotic results for empirical and partial-sum processes: A review. The Canadian Journal of Statistics, 12, pp.241-264.

[34] Rao, C. R. (1948). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. Proceedings of the Cambridge Philosophical Society 44:50-57.

[35] Seillier-Moiseiwitsch, F. (1993). Sequential Probability Forecasts and the Probability Integral Transform, International Statistical Review, 61, pp. 395-408.

[36] Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley & Sons: New York.

[37] Shorack, G. R. and J. Wellner (1986) Empirical Processes with Applications to Statistics. John Wiley & Sons: New York.

[38] Strassen, V. (1964). An invariance principle for the Law of the Iterated Logarithm. Z. Wahrscheinlivhkeitstheorie und verw. Gebiete 3 211-226.