Henry Braun

A Simple Method for Testing Goodness of Fit in the Presence of Nuisance Parameters

Preprint
May
1978
8

Institute of Mathematical Statistics
University of Copenhagen

Henry Braun*

A SIMPLE METHOD FOR TESTING GOODNESS OF FIT IN THE PRESENCE OF NUISANCE PARAMETERS.

Preprint 1978 No. 8

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

February 1978

Summary.

This paper presents a method, based on the empirical distribution function, for testing goodness of fit (gf) under composite null hypotheses. After the unknown parameters are estimated from the entire data set, the procedure calls for the transformed sample to be randomly partitioned into a large number of groups, and a gf statistic calculated for each group. These statistics are used to construct a test which can attain, asymptotically, any desired level α, and which requires for its implementation only standard tables of critical values. The procedure is particularly recommended when an a priori grouping of the sample can be employed and, hence, heterogeneous alternatives are quite plausible. It is shown that, under these alternatives, the power of the procedure compares favourably with that of other methods.

0. Introduction.

Historically, the primary concern of the classical theory of goodness-of-fit (gf) has been the testing of simple hypotheses. With the exception of chi-square tests, the problem of composite null hypotheses, until fairly recently, has received little analytical attention. Undoubtedly, a major obstacle is that the presence of nuisance parameters severely complicates the distribution theory, not only for statistics based on the sample distribution function but also for other gf methods. The implementation of Barton's [1956] extension to composite hypotheses of Neyman's [1937] "Smooth goodness-of-fit test", for example, requires extensive specialized tables. Similarly, the distribution of the Shapiro and Wilks [1965] statistic for testing normality has proved intractable and their tables are based on Monte Carlo studies. Though, subsequently, a test for exponentiality was proposed (Shapiro and Wilk [1972]), these methods do not appear to be broadly applicable.

A more intuitive approach involves first estimating the unknown parameters, and then carrying out a probability integral transform of the observations employing the estimated distribution function. Durbin [1973] has shown that under regularity conditions, as the sample size tends to infinity, the resulting sequence of empirical processes converges weakly to a Gaussian process. This result provides a rigorous foundation for computing gf statistics for the estimated empirical process. Unfortunately, the distribution of the limit process depends on the underlying distribution of the observations, so that different tables are required for each particular application. An extensive survey of recent work in this area can be found in Neuhaus [1977].

To obviate the need for new tables, preparation of which requires considerable numerical work, various methods which permit the use of standard tables have been suggested. Durbin [1976] discusses two such techniques involving the use of randomization, and points out that their convenience may be outweighed by their undoubtedly poor power characteristics.

In this paper another randomization device is presented and its properties investigated. The procedure consists of firstly using the entire sample to estimate the nuisance parameters and, secondly, randomly dividing the data into a fairly large number of groups, no group ordinarily containing more than about 10 to 15 percent of the sample. Employing the (same) estimated distribution function, each group of observations is mapped into [0,1] and for each, a gf statistic is computed. Given that there are m groups and a test of approximate level α is desired, then each statistic should be composed with the upper α/m percent point of the distribution appropriate to testing a simple hypothesis with the same number of observations as in the group. The null hypothesis is rejected if any of the m statistics exceed their critical values.

The intuitive justification for using the standard distribution is based on the assumption that no one group has a disproportionate influence on the value of the estimator. If ,indeed, that is the case, then from the standpoint of a single group, the parameter estimates appear to be superefficient (compare Darling [1955] p. 2-3). Hence, the effect of the estimation should become negligible as the total sample size tends to infinity, notwithstanding the fact that m such significance tests have been carried out. A rigorous presentation of the argument will be given in Section 1.

The above procedure has been motivated by a practical application (Braun [1977]) to data which consisted of closed birth interval lengths, hypothesized to have a gamma distribution involving unknown nuisance parameters. The adequacy of the model was of some interest, but no tables were available to carry out a proper test of fit. However, a natural grouping of the data by the parity of the interval led to the notion of separately testing each group of interval lengths of the same parity. The heuristics of the proposed procedure, outlined above, suggested the plausibility of using standard tables.

Although in this case the grouping was not done randomly, it seemed sufficient that the values of the parameter estimates played no role. In fact, it might be expected that the procedure would be most readily applied to those problems in which a natural a priori grouping is available, particularly if an alternative hypothesis of heterogeneity is being entertained. For example, in a test of normality carried out on data collected over several days, it might be suspected that one day's work differs from the others' and, thus, testing each day's data separately is intuitively attractive.

The power of the proposed procedure under such alternatives is studied in Section 2, where it is shown that, under certain conditions, the power approaches 1 as the sample size gets very large, while the power of the method carrying out a single significance test on the entire sample has power which is asymptotically only the level of the test. Heuristic arguments are subsequently advanced which suggest that the proposed procedure does well even for moderate sample sizes.

1. MAIN RESULT

A. Asymptotic Validity

Let $X_1, X_2, \ldots$ be iid with distribution function $F(\cdot,\theta)$ where θ is a vector of parameters that is partitioned as $\theta = \binom{\theta_1}{\theta_2}$. Suppose the true (unknown) value of θ is $\theta_0 = \binom{\theta_{10}}{\theta_{20}}$ and that the hypothesis to be tested is H_0: $\theta_1 = \theta_{10}$. Thus, θ_2 represents a vector of nuisance parameters.

For a given sample size N, let $\hat{\theta}_{2N} = \hat{\theta}_{2N}(X_1, \ldots, X_N)$ be an estimator of θ_2 and let $\hat{\theta}_N = \begin{pmatrix}\theta_{10} \\ \hat{\theta}_{2N}\end{pmatrix}$. Let ν denote the closure of a given neighbourhood of θ_0. Following Durbin [1973], we make the following assumptions:

Assumption 1:

$$N^{\frac{1}{2}} (\hat{\theta}_{2N} - \theta_{20}) = N^{-\frac{1}{2}} \sum_{i=1}^{N} \ell(x_i, \theta_0) + \epsilon_N$$

where

(0) $x_1, \ldots, x_N$ are iid observations from $F(\cdot, \theta)$,

(i) ℓ is measurable and for a random observation x $E[\ell(x, \theta_0) | \theta = \theta_0] = 0$,

(ii) $E[\ell(x, \theta_0)\ell(x, \theta_0)' | \theta = \theta_0] = L(\theta_0)$, a finite nonnegative - definite matrix,

(iii) $\epsilon_N \overset{P}{\to} 0$.

Assumption 2:

(i) $F(x,\theta)$ is continuous in x for all $\theta \varepsilon \nu$.

(ii) Let $x(t,\theta) = \inf \{x: F(x,\theta) = t\}$ be the inverse transformation of $t = F(x,\theta)$. Then the vector-valued function $g(t,\theta_1,\theta_2)$ defined by

$$g(t,\theta_1,\theta_2) = \left. \frac{\partial F(x,\theta)}{\partial \theta} \right|_{\substack{x = x(t,\theta_1) \\ \theta = \theta_2}} ,$$

exists and is continuous in (t,θ_1,θ_2) for all $\theta_1 \times \theta_2 \,\varepsilon\, \nu \times \nu$ and all $0 \leqq t \leqq 1$.

Remark: Assumption 1 is slightly simpler than assumption A1 (Durbin [1973], p.281) because we are not concerned here with a sequence of alternatives. Assumption 2 corrects a small error in Durbin's assumption A2.

Once $\hat{\theta}_{2N}$ has been calculated, the procedure calls for partitioning the sample into groups of possibly different sizes. For convenience of exposition, it is assumed in the sequel that the groups are of equal size $n(N)$ so that the number of groups is $m(N) = N/n(N)$. The dependence of n and m on N is usually suppressed in the notation. A gf statistic is then calculated for each group. For fixed $\alpha \varepsilon (0,\frac{1}{2})$, let $z_n(\alpha)$ denote the upper α percent point of the standard distribution of the statistic for sample size n and let $\alpha_m = 1 - (1-\alpha)^{1/m}$. The null hypothesis is rejected if and only if the largest of the calculated statistics exceeds $z_n(\alpha_m)$.

<u>Theorem 1</u>: Suppose the procedure described above employs either a Kolmogorov-Smirnov (KS) statistic or the Cramer-von Mises (CM) statistic. Then under Assumptions 1 and 2, the level of the procedure $\to \alpha$ as $N \to \infty$, provided that $n(N) = o(N^p)$ for some $0 < p < \frac{1}{2}$.

Theorem 1 is a direct consequence of the following two propositions which are established under its hypotheses, following the introduction of some notation. Define

$$\hat{F}^{(i,N)}(t) = n^{-1}(\#\text{ observations } x \text{ in group } i \text{ such that } F(x,\hat{\theta}_N) \leqq t)$$

and

$$\hat{y}^{(i,N)}(t) = n^{\frac{1}{2}}[\hat{F}^{(i,N)}(t) - t], \quad i = 1,2,\ldots,m.$$

$\hat{T}^{(i,N)}$ denotes either one of the KS statistics:

$$\max\{0, \sup_{0\leqq t\leqq 1} \hat{y}^{(i,N)}(t)\},\ \max\{0, \sup_{0\leqq t\leqq 1} -\hat{y}^{(i,N)}(t)\}, \sup_{0\leqq t\leqq 1} |\hat{y}^{(i,N)}(t)|,$$

or the CM statistic:

$$\int_0^1 [\hat{y}^{(i,N)}(t)]^2 dt.$$

The corresponding symbols without the carets are employed when $\hat{\theta}_N$ is replaced by θ_0.

Proposition 1: $\max_{1 \leq i \leq m} |\hat{T}^{(i,N)} - T^{(i,N)}| < \gamma_N = o_p(N^{-q})$
for some $q > 0$.

Proof: The proof depends on a basic relation derived in Durbin [1973]. Letting $\hat{t}_N = \hat{t}_N(t) = F(x(t,\hat{\theta}_N),\theta_0)$, we have

$$\hat{F}^{(i,N)}(t) = n^{-1}(\text{\# observations in group } i \text{ such that } x \leq x(t,\hat{\theta}_N))$$

and

$$F^{(i,N)}(\hat{t}_N(t)) = n^{-1}(\text{\# observations } x \text{ in group } i \text{ such that } F(x,\theta_0) \leq \hat{t}_N(t)).$$

Thus,

$$\hat{F}^{(i,N)}(t) = F^{(i,N)}(\hat{t}_N(t)), \tag{2.1}$$

Showing that $\hat{F}^{(i,N)}(\cdot)$ and $F^{(i,N)}(\cdot)$ are related by a random time transformation. The proof now differs slightly according to which class of statistics is considered.

(a) KS statistics. As a consequence of (2.1)

$$\begin{aligned} \hat{y}^{(i,N)}(t) &= n^{\frac{1}{2}}[\hat{F}^{(i,N)}(t) - t] \\ &= n^{\frac{1}{2}}[F^{(i,N)}(\hat{t}_N) - \hat{t}_N] + n^{\frac{1}{2}}[\hat{t}_N - t] \\ &= y^{(i,N)}(\hat{t}_N) + n^{\frac{1}{2}}[\hat{t}_N - t]. \end{aligned}$$

Since

$$\sup_{0\leq t\leq 1} | y^{(i,N)}(\hat{t}_N(t)) | = \sup_{0\leq t\leq 1} | y^{(i,N)}(t) | ,$$

$\hat{T}^{(i,N)}$ and $T^{(i,N)}$ can differ by at most

$$\gamma_N \equiv n^{\frac{1}{2}}\delta_N \equiv n^{\frac{1}{2}} \sup_{0\leq t\leq 1} | \hat{t}_N(t) - t | .$$

The proof of Lemma 1 in Durbin [1973] shows that

$$\delta_N = \sup_{0\leq t\leq 1} |\hat{t}_N - t| = o_p(N^{-r}), \ r < \tfrac{1}{2},$$

so that $\gamma_N = o_p(N^{-q})$ some $q > 0$ as long as $n = o(N)$.

(b) CM statistic. Again from (2.1),

$$\hat{T}^{(i,N)} = n \int_0^1 [\hat{F}^{(i,N)}(t) - t]^2 \, dt$$

$$= n \int_0^1 [F^{(i,N)}(\hat{t}_N(t)) - t]^2 \, dt.$$

Suppose the observations in group i are denoted by $x_1,\ldots,x_n$. Then define

$$s_j = F(x_j,\theta_o) \qquad j = 1,\ldots,n,$$

and let $t_1,\ldots,t_n$ be determined implicitly by the relations

$$s_j = \hat{t}_N(t_j) \qquad j = 1,\ldots,n.$$

Writing G(t) for $F^{(i,N)}(\hat{t}_N(t))$, it is clear that G(t) is the empirical df of the fictitious sample $t_1,\ldots t_n$, and $\hat{T}^{(i,N)}$ is the corresponding CM statistic. On the other hand, $T^{(i,N)}$ is the CM statistic of the sample $s_1,\ldots,s_n$. Now, using the formula

$$T^{(i,N)} = \sum_{j=1}^{n} (s_j - \frac{j-\frac{1}{2}}{n})^2 + 1/(12n),$$

together with the fact that

$$|t_j - s_j| < \delta_N \qquad j = 1,\ldots,n,$$

we obtain

$$|\hat{T}^{(i,N)} - T^{(i,N)}| < n\delta_N + \delta_N^2 \equiv \gamma_N.$$

If $n = o(N^p)$ some $p < \frac{1}{2}$, then $\gamma_N = o_p(N^{-q})$ some $q > 0$. □

Proposition 2:

$$|P\{\max_{1\leq j\leq m} \hat{T}^{(j,N)} \geq z_n(\alpha_m)\} - P\{\max_{1\leq j\leq m} T^{(j,N)} \geq z_n(\alpha_m)\}|$$

$$\to 0 \text{ as } N \to \infty.$$

Proof: Since

$$\max_{1\leq j\leq m} | \hat{T}^{(j,N)} - T^{(j,N)} | < \gamma_N,$$

it follows that

$$P\{\max_{1\leq j\leq m} T^{(j,N)} > z_n(\alpha_m) + \gamma_N\} < P\{\max_{1\leq j\leq m} \hat{T}^{(j,N)} > z_n(\alpha_m)\}$$

$$< P\{\max_{1\leq j\leq m} T^{(j,N)} > z_n(\alpha_m) - \gamma_N\} \qquad (2.2)$$

The proof consists of showing that the change in the overall significance level caused by perturbing $z_n(\alpha_m)$ by γ_N is asymptotically negligible as $N \to \infty$. This requires fairly precise knowledge of the functional relationship between critical values and their corresponding significance levels, particularly when both the sample size and the critical values tend to infinity.

For the sake of brevity, only the case of Smirnov's statistic $\sqrt{n} \sup_{0 \leq t \leq 1} y^{(i,N)}(t)$ will be considered. If $\Phi_n^+(\cdot)$ denotes the df of the statistic for sample size n, then Smirnov [1941] showed that

$$\rho_n(z) = 1 - \Phi_n^+(z) = e^{-2z^2}\{1 - \frac{2z}{3\sqrt{n}} + O(\frac{1}{n})\} \qquad (2.3)$$

as $n \to \infty$ for $z = O(n^{1/6})$.

Since $z_n(\alpha_m) = O(\log N)$, it is legitimate to use (2.3) to evaluate $\rho_n(z_n(\alpha_m) + \gamma_N)$. In fact,

$$\rho_n(z_n(\alpha_m) + \gamma_N) = e^{-2z_n^2(\alpha_m)}[e^{-4z_n(\alpha_m)\gamma_N - 2\gamma_N^2}] \times$$

$$[1 - \frac{2z_n(\alpha_m) - 2\gamma_N}{3\sqrt{n}} + O(\frac{1}{n})]$$

$$= \rho_n(z_n(\alpha_m))[1 + O_P(1/n)]$$

$$= \alpha_m[1 + O_P(\frac{1}{n})],$$

with a similar result for $\rho_n(z_n(\alpha_m) - \gamma_N)$. Thus, the first and last expressions in (2.2) both tend to α as $N \to \infty$.

An asymptotic expansion for the distribution of Kolmogorov's statistic can be found in Korolyuk [1955], while similar results for the CM statistic can be found in Anderson and Darling [1952] or Mogul'skii [1977].

Practical Considerations

An important question in the application of this procedure is the choice of the group size(s). As Proposition 1 makes clear, if the CM statistic is to be applied then the maximal group size n should be no more than $N^{\frac{1}{2}}$. In the case of KS statistics more latitude is permitted, but caution suggests that n be about .1N to .15N.

If the group structure is intrinsic to the problem, then another form of the question often arises. It may happen that some or all of the group sizes are rather small, say of the order of ten. In this situation, are the asymptotic results presented above relevant? The answer seems to be yes, provided that the total sample size is large and $\hat{\theta}_{2N}$ is a good estimator of θ_2.

When N is large, Proposition 1 shows that γ_N (which bounds $|T^{(i,N)} - \hat{T}^{(i,N)}|$) tends to be small, particularly if n is small. Proposition 2 is more problematic since the proof requires n to tend to infinity, though at an admittedly slow rate. However, the work of Stephens [1970] provides some empirical evidence that even for small values of n, perturbing $z_n(\alpha/m)$ by γ_N changes the corresponding critical value by a negligible amount.

Stephens has shown how the common gf statistics can be modified so that the critical values for $n = \infty$ can be used for all sample sizes. For example, if D_n denotes Kolmogorov's statistic for sample size n, then Stephens suggests calculating

$$\tilde{D} = D_n(n^{\frac{1}{2}} + .12 + .11n^{-\frac{1}{2}}),$$

and then treating $\tilde{D}$ as if it were distributed as $\lim_{n\to\infty} \sqrt{n}D_n$. Between the significance levels α and the corresponding critical values $z(\alpha)$, the following relation holds:

$$\alpha = 2 \exp(-2z^2(\alpha)).$$

Furthermore, Stephens states that the approximation is quite good even for n very small, and that its accuracy increases as the significance levels become more extreme. Such modifications of $T^{(i,N)}$ and $\hat{T}^{(i,N)}$ would differ by no more than $O(\gamma_N)$, and changing $z(\alpha/m)$ by $O(\gamma_N)$ has little effect on the corresponding significance level.

Because they both involve grouping, it would be of interest to compare the method suggested here with Durbin's half-sample method (Durbin [1976]).

The latter procedure carries out a single significance test on the N transformed observations, but the probability integral transform employs an estimate of θ_2 based on a randomly chosen half-sample. Durbin showed that as $N \to \infty$, the effect of the estimation on the distribution of the empirical process becomes negligible. The two methods use grouping in different phases: one when carrying out the significance testing, the other when constructing the estimator of the nuisance parameters.

Certainly these methods should be considered when the non-randomized procedure cannot be implemented for lack of tables or because the distribution of the transformed observations depends on the values of the nuisance parameters. But choosing between them is difficult without supporting numerical evidence. However, the present method might be preferred when there is a natural grouping so that less arbitrariness is involved, or, when the testing problem is only a component of a larger study and the values of the nuisance parameters are of interest. In such cases the method which uses the final estimates (provided H_0 is accepted) in carrying out the test might be easier to justify.

Finally when heterogeneous alternatives are a distinct possibility, the power of the method presumably can be enhanced by employing jackknife methods. That is, when the groups are balanced, the probability integral transform of the i^{th} group uses $\hat{\theta}_{2N}^{(i)}$ where $\hat{\theta}_{2N}^{(i)}$ denotes the estimator of θ_2 constructed from all the observations except those in the i^{th} group. If H_0 is accepted, the usual jackknife estimator of θ_2 based on the pseudo values constructed from $\{\hat{\theta}_{2N}^{(i)}\}$ can be used for further investigations.

2. POWER

A. Preliminaries

It is widely assumed that procedures involving extraneous randomization have poor power characteristics. Unfortunately, in the area of gf testing, very little seems to be known about the magnitude of the power loss even for different homogeneous alternatives. In the case of the present procedure, such calculations could be carried out exactly, if rather laboriously, using the formulas in Suzuki [1968]. This is postponed to a later investigation. For the moment, it must be assumed that the present method is generally not as powerful as the one based on the full (estimated) empirical process, and it's use can be recommended only when the latter's implementation is impractical or impossible. In any application, the set of P-values generated by the procedure can give only an indication of the adequacy of the fit.

As stated previously, the apparent arbitrariness of the procedure is diminished when a natural grouping of the data can be employed. In such a case it is often quite plausible to fear "heterogeneous alternatives". This phrase refers to the situation in which the observations in a large majority of groups conform to the null hypothesis, while those in the remaining groups differ in one or more distributional characteristics. The k-sample slippage hypothesis is a classical parametric example. Since the present procedure tests each group separately, it should prove particularly sensitive to heterogeneous alternatives and some asymptotic results in this direction are presented in the following subsection. The remainder of the section develops heuristic arguments which suggest that against these alternatives, the power of the procedure compares favourably with that of other methods, even with only moderate sample sizes. At present, the discussion is limited to KS statistics only.

B. Asymptotics

Let the sample consist of m groups each containing n independent observations for which the null hypothesis, H_0, specifies a common distribution $F(\cdot,\theta)$ (cf. Section 1.A). Suppose that, in fact, only $m-\ell$ groups conform to H_0, while the remaining observations follow a different common distribution G. It will prove convenient to phrase the argument in terms of the procedures followed by three statisticians S_1, S_2, and S_3.

S_1 observes only the $m-\ell$ groups conforming to H_0 and his data (after transformation) is denoted by $0 \leqq t_1 \leqq t_2 \leqq \cdots \leqq t_{(m-\ell)n} \leqq 1$, while S_2 and S_3 both observe the entire sample; their observations (after transformation) are denoted by $0 \leqq v_1 \leqq v_2 \leqq \cdots \leqq v_{mn} \leqq 1$. S_1 and S_2 compute

$$\hat{d}_1^{\,+} = \max_{1 \leq j \leq (m-\ell)n} \left[\frac{j}{(m-\ell)n} - t_j\right]$$

and

$$\hat{d}_2^{\,+} = \max_{1 \leq i \leq mn} \left[\frac{i}{mn} - v_i\right]$$

respectively. On the other hand, S_3 computes Smirnov statistics $\hat{e}_1^{\,+}, \ldots, \hat{e}_m^{\,+}$ for each group separately. H_0 is rejected by S_1 if $\hat{D}_1^{\,+} = [(m-\ell)n]^{\frac{1}{2}} \hat{d}_1^{\,+}$ exceeds $\hat{z}_{(m-\ell)n}(\alpha)$, by S_2 if $\hat{D}_2^{\,+} = (mn)^{\frac{1}{2}} \hat{d}_2^{\,+}$ exceeds $\hat{z}_{mn}(\alpha)$, and by S_3 if $\max_{1 \leqslant k \leqslant m} \sqrt{n}\hat{e}_k^{\,+}$ exceeds $z_n(\alpha_m)$. The caret signifies percentage points modified for parameter estimation. Interest centers on comparing the performance of S_2 and S_3. The following theorem shows that in certain circumstances as the sample size becomes large, S_3's power tends to one, but S_2's tends to α, the level of the test. This last result is proved by showing that, asymptotically, S_2 can do no better than S_1.

<u>Theorem 2</u>: In the notation of Theorem 1 and the preceding paragraphs, suppose that in addition to Assumptions 1 and 2, the following conditions hold:

(i) $\ell(n/m)^{\frac{1}{2}} \to 0$ as $N \to \infty$,

(ii) Let $h_N = \hat{\theta}(X_1, \ldots, X_{mn}) - \hat{\theta}(X_1, \ldots, X_{(m-\ell)n})$, where X_i $(1 \leqq i \leqq (m-\ell)n)$ are iid as F and X_i $((m-\ell)n + 1 \leqq i \leqq mn)$ are iid as G.
$h_N = o_P((mn)^{-\frac{1}{2}})$,

(iii) $G \neq F$.

Then, as $N \to \infty$,

$$P\{S_2 \text{ rejects } H_0\} \to \alpha$$

and

$$P\{S_3 \text{ rejects } H_0\} \to 1.$$

Proof: We first consider the case of H_0 simple, so that no estimation is required, and develop a relation between d_1^+ and d_2^+.

Suppose that $t_j = v_{j+k}$ for some $1 \leqq j \leqq (m-\ell)n$ and some $1 \leqq k \leqq \ell n$. Then

$$\left| \left(\frac{j+k}{mn} - v_{j+k}\right) - \left(\frac{j}{(m-\ell)n} - t_j\right) \right|$$

$$\leqq \left| \frac{j}{mn} - \frac{j}{(m-\ell)n} \right| + \frac{k}{mn} \leqq \frac{2\ell}{m}. \tag{2.1}$$

That is, the discrepancies assigned by S_1 and S_2 to the same observation differ by, at most, $2\ell/m$. Of course, S_2 must also compute the discrepancies at the ℓn observations not available to S_1. For fixed $t_1,\ldots t_{(m-\ell)n}$, the most extreme situation occurs when these ℓn observations are all less than t_1 and, in this case, no discrepancy can exceed ℓ/m. Consequently,

$$d_2^+ \leqq d_1^+ + 3\ell/m,$$

so that

$$D_2^+ \leqq \sqrt{\frac{m}{m-\ell}}\, D_1^+ + 3\ell\sqrt{\frac{n}{m}}\,. \tag{2.2}$$

Thus, if $\ell\sqrt{n/m} \to 0$ as the total sample size tends to infinity, then the power of S_2's test tends to α and this remains the case no matter how distant the alternative.

On the other hand, under the conditions of the theorem, $\sqrt{n}\, e_i^+ = O_P(\sqrt{n})$ for each of the discordant groups. Since $z_n(\alpha_m) = O(\sqrt{\log m})$, S_3 must have asymptotic power 1. This result holds even if one considers a sequence of alternatives converging to the null distribution at an appropriate rate.

The presence of nuisance parameters does not alter these conclusions for, corresponding to each observation x among the $(m-\ell)$ groups conforming to H_0, we have for some $1 \leqq j \leqq (m-\ell)n$ and some $1 \leqq k \leqq \ell n$,

$$t_j = F(x, \hat{\theta}^{(1)}_{(m-\ell)n})$$

and

$$v_{j+k} = F(x, \hat{\theta}^{(2)}_{mn}),$$

where $\hat{\theta}^{(1)}_{(m-\ell)n}$ and $\hat{\theta}^{(2)}_{mn}$ are the estimates of θ employed by S_1 and S_2, respectively. Now,

$$|t_j - v_{j+k}| = |F(x, \hat{\theta}^{(1)}_{(m-\ell)n}) - F(x, \hat{\theta}^{(2)}_{mn})|$$

$$\leqq |\hat{\theta}^{(1)}_{(m-\ell)n} - \hat{\theta}^{(2)}_{mn}| \left| \frac{\partial F(x,\theta)}{\partial \theta} \right|_{\theta = \theta^*},$$

where θ^* lies between $\hat{\theta}^{(1)}_{(m-\ell)n}$ and $\hat{\theta}^{(2)}_{mn}$. Hence, $|t_j - v_{j+k}| = o_p((mn)^{-\frac{1}{2}})$ and (2.1) becomes

$$\left|[(j+k)/mn - v_{j+k}] - [j/(m-\ell)n - t_j]\right| \leqq 2\ell/m + o_p((mn)^{-\frac{1}{2}}). \tag{2.3}$$

Thus,

$$\hat{D}_2^+ \leqq (m/(m-\ell))^{\frac{1}{2}} \hat{D}_1^+ + 2\ell(n/m)^{\frac{1}{2}} + o_p(1), \tag{2.4}$$

and, consequently, the power of S_2's procedure must tend to α as the sample size increases.

S_3's estimate of θ coincides with that of S_2's and, hence, his estimate of the underlying distribution converges uniformly in probability to the null distribution. In fact,

$$|F(x,\hat{\theta}^{(2)}_{mn}) - F(x,\theta_0)| = \max\{0_p(\ell/m), 0_p([(m-\ell)n]^{-\frac{1}{2}})\}.$$

It is readily apparent that S_3's power tends to 1, as the sample size increases, inasmuch as $\sqrt{n}\,\hat{e}_i^+$ is still $0_p(\sqrt{n})$ for the discordant groups. □

Remark: Condition (ii) of the theorem must be verified in each particular application. It is certainly satisfied by the sample mean if F and G have finite expectations, since

$$\hat{\theta}_{mn} = \hat{\theta}_{(m-\ell)n} - (\ell/m)\,\hat{\theta}_{(m-\ell)n} + (\ell/m)[\sum_{j=(m-\ell)n+1}^{mn} x_j/(\ell n)].$$

However, estimators satisfying Assumption 1 should, in general, be quite well behaved in this respect.

C. Finite Sample Results

One may ask whether the asymptotic results of Theorem 2 are at all relevant to sample sizes common in practice. Since purely analytical methods are unlikely to prove tractable, a large-scale Monte Carlo study is needed to give an informative answer. However, the heuristic argument presented below does indicate that the suggested procedure should have good power, even in only moderately large samples. The argument is an indirect one, in that it does not involve the calculation of any rejection probabilities. Rather, it consists of studying a class of sample configurations which should be fairly typical under the alternative hypothesis. For such samples, it is possible to develop an approximate relation between the statistics computed by S_2 and S_3 from which we can derive some idea of the relative performance of the two procedures.

Suppose that $\ell = 1$; that is, of the m groups, exactly one does not conform to H_0, and that its distribution is stochastically smaller than that specified by H_0. After estimating the nuisance parameters, the two statisticians compute the two-sided Kolmogorov statistic on the transformed data. Let d_3 denote the (unnormalized) statistic calculated by S_3 for the discordant group, d_2 denote the value of the statistic computed by S_2 for the whole sample, and d_2^* the statistic based on the (m-1)n observations corresponding to the groups conforming to H_0. We propose to show that, roughly speaking,

$$d_2 \leqq d_2^* + d_3 m^{-1}. \tag{2.5}$$

The argument is somewhat similar to the one in Theorem 2. Suppose that $0 \leqq t_1 \leqq t_2 \leqq \cdots \leqq t_{(m-1)n}$ denote the observations for which d_2^* is calculated. Assuming that d_2^* is fairly typical, we construct samples which make d_2 as large as possible consistent with the value of d_3. In reality, the different configurations of the discordant group affect the final sample configuration through their contribution to the parameter estimates. One aspect of the heuristic nature of the discussion is that this effect is ignored.

Assuming the t_i's to be fixed, d_2 becomes larger the smaller the observations in the discordant group. But if $d_3 = n^{-\gamma}$ (say) for some $\gamma \varepsilon (0,\frac{1}{2})$, then no more than about $n^{1-\gamma}$ of these can be smaller than t_1, without violating the constraint. The change in the discrepancy at t_1 is then no more than

$$\frac{n^{1-\gamma}+1}{mn} - \frac{1}{n(m-1)} \sim n^{-\gamma} m^{-1}.$$

If w_n, the largest of these observations, falls between t_{j-1} and

t_j, then the change in the discrepancy at t_j is (letting $j = \alpha(m-1)n$)

$$\frac{\alpha(m-1)n + n}{mn} - \alpha = \frac{1-\alpha}{m} .$$

But, since $d_3 = n^{-\gamma}$, we must have $w_n \geqq 1 - n^{-\gamma}$. Thus, $1-\alpha$ should not, in general, exceed $n^{-\gamma}$ so that $(1-\alpha)/m \leqq n^{-\gamma} m^{-1}$. Intuitively, these are the two extreme cases. If nothing unusual occours between t_1 and t_α, it should be true that

$$d_2 \leqq d_2^* + n^{-\gamma} m^{-1},$$

for the discrepancies at the new observations can not exceed the RHS of the expression.

The "inequality" (2.5) is sharper than (2.1) and more useful for our present purposes because the bound for d_2 involves the value of d_3. It is convenient to carry out the remainder of the discussion in terms of the modified versions of d_2 and d_3 (see Section 1.C). We therefore construct

$$\begin{aligned} \tilde{D}_3 &= d_3(n^{\frac{1}{2}} + .12 + .11\ n^{-\frac{1}{2}}) \\ &= n^{\frac{1}{2}-\gamma} + e_3(n). \qquad (2.6) \end{aligned}$$

The modification of d_2 is more problematic inasmuch as the appropriate formula depends on the particular null hypothesis being tested. For the sake of convenience, we suppose that the null distribution is normal with unspecified mean and variance. (It should be emphasized that the general validity of the conclusions do not depend on this particular choice .) Employing the formula given in Pearson and Hartley [1972], page 359, we construct

$$\tilde{D}_2 = d_2((nm)^{\frac{1}{2}} - 0.01 + 0.85\ (nm)^{-\frac{1}{2}}).$$

In the sequel, we suppose that d_2 assumes the upper bound given in (2.5) and, hence,

$$\tilde{D}_2 = (d_2^* + n^{-\gamma} m^{-1}) (nm)^{\frac{1}{2}} + (d_2^* + n^{-\gamma} m^{-1}) (-.01 + 0.85 (nm)^{-\frac{1}{2}})$$

$$= [(nm)^{\frac{1}{2}} d_2^* + n^{\frac{1}{2}-\gamma} m^{-\frac{1}{2}}] + e_2 (n,m). \qquad (2.7)$$

Consider a sequence of problems, indexed by n, in which $(nm)^{\frac{1}{2}} d_2^*$ remains roughly constant. If γ is held fixed as n increases, then it follows from (2.6) and (2.7) that $\tilde{D}_3$ increases, but $\tilde{D}_2$ decreases. That is P_3, the P-value of $\tilde{D}_3$, becomes more extreme, while P_2, the P-value of $\tilde{D}_2$, becomes less extreme. Table I presents some numerical examples which show that in fairly typical situations $P_2 > mP_3$, indicating that the procedure based on $\tilde{D}_3$ provides a more sensitive test of H_0.

In view of Theorem 2, this last result is not very surprising, involving, as it does, increasingly extreme values of $\tilde{D}_3$. However, a more informative comparison can be constructed in the following manner. Suppose α is held fixed, but $\gamma = \gamma(n)$ is allowed to vary with n, so that $d_3 = n^{-\gamma(n)}$ is just significant at level α_m. This is equivalent to considering a sequence of tests of fixed overall level α. How then do the P-values of $\tilde{D}_2$ behave in this sequence of problems ?

Interestingly, the P-values increase (i.e. become less extreme), just as before. This follows from the fact, which we shall now prove, that $\gamma(n)$ increases with n. If Z is a random variable distributed as $\tilde{D}_3$, then the significance level corresponding to Z = z is

$$\alpha = \alpha(z) = 2 \exp(-2z^2). \tag{2.8}$$

In our problem, approximating α_m by α/m, (2.8) becomes $\alpha/m(n) = 2 \exp[-2(n^{\frac{1}{2}-\gamma(n)} + e_3(n))^2]$, whence

$$\gamma(n) = \tfrac{1}{2} - \log[(-\tfrac{1}{2}\log \alpha/(2m(n)))^{\frac{1}{2}} - e_3(n)] / \log n,$$

which is an increasing function of n. Returning to (2.7), the crucial term in the expression for $\tilde{D}_2$ is $n^{-\gamma(n)} (n/m(n))^{\frac{1}{2}}$. Under the hypotheses of our theorem, $n/m(n)$ decreases in n. Since $n^{-\gamma(n)}$ also decreases in n, $\tilde{D}_2$ must decrease in n, as well. A numerical illustration can be found in Table II.

TABLE I

Comparison of P-values, γ fixes.

n	m	N	q_n	$\tilde{D}_3$	$\tilde{D}_2$	P_3	mP_3	P_2
8	8	64	.415	1.776	1.238	3.6×10^{-3}	2.9×10^{-2}	7.0×10^{-4}
12	18	216	.368	1.943	1.055	1.05×10^{-3}	1.9×10^{-2}	8.0×10^{-3}
16	32	512	.330	2.074	.954	3.68×10^{-4}	1.2×10^{-2}	2.6×10^{-2}
20	50	1000	.295	2.183	.899	1.45×10^{-4}	7.25×10^{-3}	4.5×10^{-2}

TABLE II

Comparison of P-values, overall level fixed.

n	$\gamma(n)$	$\tilde{D}_2$	P_2
8	.250	1.238	7.0×10^{-4}
12	.262	1.041	9.2×10^{-3}
16	.271	.934	3.2×10^{-2}
20	.277	.876	5.8×10^{-2}

In carrying out the computations for Table I, n and m were chosen to satisfy the relations

$$n = 2N^{1/3} \text{ and } m = N^{2/3}/2 = n^2/8.$$

In addition, γ was fixed to be .25 and $(mn)^{\frac{1}{2}} d_2^*$ to be .6. That this choice of γ is a reasonable one for the sample sizes employed, can be inferred from the q_n column in Table I. The q_n are defined by

$$q_n = P\{\max_{1 \leq i \leq n} \Phi(X_i) \leq 1 - n^{-.25}\},$$

where Φ is the df of a standard normal variate and $X_1, \ldots X_n$ are iid $N(-1.5,1)$. The values of q_n indicate that $d_3 \geq n^{-.25}$ is not an uncommon occurance even for moderately distant alternatives. Finally, fixing $d_2^* = .6/(mn)^{\frac{1}{2}}$ is roughly equivalent to locating d_2^* at the median of the distribution.

The column labelled mP_3 gives the overall level of the procedure so that the computed value of $\tilde{D}_3$ is just significant at level α/m, leading to a rejection of H_0 by S_3. This should be compared with the P_2 column which contains the P-values of $\tilde{D}_2$. The rather extreme values of P_2 for $n = 8$ and 12 are probably more due to the crudeness of the bound (2.5) for small sample sizes than to any particular merit of the procedure. Note that in contrast to mP_3, P_2 rapidly increases with n.

In Table II, the values of $\gamma(n)$ were chosen to keep mP_3 fixed at $2.9\text{x}10^{-2}$, its value for $n = 8$ in Table I. Although P_2 decreases with n, the changes, in comparison with the corresponding values of Table I, are not as dramatic as those in mP_3.

Taking $n = 20$ as an example, mP_3 changes from $7.25\text{x}10^{-3}$ to $2.9\text{x}10^{-2}$, an increase by a factor of 4. On the other hand, P_2

changes from 4.5×10^{-2} to 5.8×10^{-2}, an increase by a factor of only 1.5.

Acknowledgments:

The author would like to thank M.L. Eaton, P. Gaenssler, and S. Lauritzen for useful conversations.

References

[1] Anderson, T.W. and Darling, D.A. (1952): "Asymptotic theory of certain goodness of fit criteria based on stochastic processes." Ann. Math. Statist. 23, 193-212.

[2] Barton, D.E. (1956): "Neyman's ψ_k^2 test of goodness of fit when the null hypothesis is composite." Skand. Aktuarietidskr. 39, 216-245.

[3] Durbin, J. (1973a): "Weak convergence of the sample distribution function when parameters are estimated." Ann. Statist. 1, 279-290.

[4] Durbin, J. (1973b): "Distribution theory for tests based on the sample distribution function." Regional Conference Series in Applied Mathematics 9. SIAM, Philadelphia.

[5] Durbin, J. (1976): "Kolmogorov-Smirnov tests when parameters are estimated" in "Empirical Distributions and Processes" (ed. P. Gaenssler and P. Révész), Lecture Notes in Mathematics 566. Springer-Verlag, Berlin.

[6] Braun, H. (1977): "A stochastic model for birth histories with applications to the study of fertility patterns." T.R. 124, Series 2, Department of Statistics, Princeton University.

[7] Darling, D.A. (1955). "Cramér-Smirnov test in the parametric case." Ann. Math. Statist. 26, 1-20.

[8] Korolyuk, V.S. (1955): "Asymptotic expansions for Kolmogorov-Smirnov's tests" (in Russian). Izv. Akad. Nauk. S.S.S.R., 19, p. 103.

[9] Mogul'skii, A.A. (1977): "Remarks on Large Deviations for the ω^2 statistic." Theory Prob. Applications, 22, 166-171.

[10] Neuhaus, G. (1977): "Asymptotic theory of goodness of fit tests when parameters are present: A survey." Lecture held at 10^{th} European meeting of Statisticians, Leuven, Belgium.

[11] Neyman, J. (1937): "'Smooth test' for goodness of fit." Skand. Aktuarietidskr. 20, 149-199.

[12] Pearson, E.S. and Hartley, H.O. (1972): "Biometrika tables for statisticians" Cambridge University Press.

[13] Shapiro, S.S. and Wilk, M.B. (1965): "An analysis of variance test for normality (complete samples)." Biometrika 52, 591-611.

[14] Shapiro, S.S. and Wilk, M.B. (1972): "An analysis of variance test for the exponential distribution (complete samples)." Technometrics 14, 355-70.

[15] Smirnov, N.V. (1944): "Approximate laws of distribution of random variables from empirical data." Uspekhi Mat. Nauk. 10, 179-206. (in Russian).

[16] Stephens, M.A. (1970): "Use of the Kolmogorov-Smirnov, Cramer - von Mises and related statistics without extensive tables." J. Roy. Statist. Soc. Ser. B, 32, 115-122.

[17] Suzuki, G. (1968): "Kolmogorov-Smirnov tests of fit based on some general bounds." JASA, 63, 919-924.

PREPRINTS 1977

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Asmussen, Søren & Keiding, Niels: Martingale Central Limit Theorems and Asymptotic Estimation Theory for Multitype Branching Processes.

No. 2 Jacobsen, Martin: Stochastic Processes with Stationary Increments in Time and Space.

No. 3 Johansen, Søren: Product Integrals and Markov Processes.

No. 4 Keiding, Niels & Lauritzen, Steffen L. : Maximum likelihood estimation of the offspring mean in a simple branching process.

No. 5 Hering, Heinrich: Multitype Branching Diffusions.

No. 6 Aalen, Odd & Johansen, Søren: An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations.

No. 7 Johansen, Søren: The Product Limit Estimator as Maximum Likelihood Estimator.

No. 8 Aalen, Odd & Keiding, Niels & Thormann, Jens: Interaction Between Life History Events.

No. 9 Asmussen, Søren & Kurtz, Thomas G.: Necessary and Sufficient Conditions for Complete Convergence in the Law of Large Numbers.

No. 10 Dion, Jean-Pierre & Keiding, Niels: Statistical Inference in Branching Processes.

PREPRINTS 1978

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Tjur, Tue: Statistical Inference under the Likelihood Principle.

No. 2 Hering, Heinrich: The Non-Degenerate Limit for Supercritical Branching Diffusions.

No. 3 Henningsen, Inge: Estimation in M/G/1-Queues.

No. 4 Braun, Henry: Stochastic Stable Population Theory in Continuous Time.

No. 5 Asmussen, Søren: On some two-sex population models.

No. 6 Andersen, Per Kragh: Filtered Renewal Processes with a Two-Sided Impact Function.

No. 7 Johansen, Søren & Ramsey, Fred L.: A Bang-Bang Representation for 3x3 Embeddable Stochastic Matrix.

No. 8 Braun, Henry: A Simple Method for Testing Goodness of Fit in the Presence of Nuisance Parameters.