How to Approximate a Histogram by a Normal Density

Lawrence D. BROWN and J. T. Gene HWANG (formerly Jiunn T. HWANG)*

Which normal density curve best approximates the sample histogram? The answer suggested here is the normal curve that minimizes the integrated squared deviation between the histogram and the normal curve. A simple computational procedure is described to produce this best-fitting normal density. A few examples are presented to illustrate that this normal curve does indeed provide a visually satisfying fit, one that is better than the traditional $\bar{x}$, s answer. Some further aspects of this procedure are investigated. In particular it is shown that there is a satisfactory answer that is independent of the bar width of the histogram. It is also noted that this graphical procedure provides highly robust estimates of the sample mean and standard deviation. We demonstrate our technique by using data including Newcomb's data of passage time of light and Fisher's iris data.

KEY WORDS: Graphical approximation; Least squares approximation; Robust estimation.

1. INTRODUCTION

The composition of a numerical random sample is conveniently pictured by its histogram. For many classes of data one expects the underlying population to be approximately normal, and hence the histogram of the sample also to be approximately normal. If so, it may be further convenient to smooth the histogram by approximating it by a suitable normal density curve.

Figures 1 and 2 illustrate this process with two historic sets of statistical data. The data in Figure 1 are Newcomb's classical measurements of the passage time of light. (See Stigler 1977.) The data in Figure 2 are measurements of iris sepal width for 150 plants of three species, as presented in Fisher (1936). (See also Andrews and Herzberg 1985, pp. 5–8.) In each case the approximating normal density curve is chosen in the commonsense fashion—its mean, μ is $\bar{x}$, the sample mean, and its standard deviation σ, is s, the sample standard deviation as defined by $s^2 = (n-1)^{-1}\Sigma(x_i - \bar{x})^2$.

In each figure the approximating normal density provides a reasonable visual fit to the underlying histogram; however, in both cases (and particularly in the first) the visual fit can be significantly improved by using a different value of μ and σ. This fact is displayed in Figures 3 and 4.

The intent of this article is to describe a method of choosing μ and σ so as to provide the best fit to a given histogram. The "fit" will be measured in a least square sense. This is mathematically convenient; it enables mathematical precision in our answer, and it appears in examples to provide a normal density that does indeed provide a visually satisfying fit.

Finding the best μ, σ in the above sense requires the solution of a pair of simultaneous transcendental equations. These can easily be solved numerically. We used Gauss on an IBM PC, but any other standard programming language will suffice.

The bar width of the histogram has an influence on the choice of the approximating normal density. However, for small to moderate bar widths (those below, say, $\sigma/3$), this influence is very minor. For a given set of data the best approximating normal density converges to a limiting answer as the bar width converges to zero. This limiting answer therefore provides a normal density that is a good fit for any histogram drawn from the data, so long as the bar width is not too large. The simultaneous equations needed to calculate this limiting answer are somewhat easier to compute, to manipulate, and to solve than are the corresponding equations that take into account the bar width.

One other feature of this limiting answer may also be of interest. The approximating normal density is of course determined by values $\tilde{\mu}$, $\tilde{\sigma}$ computed from the data, as described in Theorems 3.1 or 4.1 or Corollary 3.1. These values $\tilde{\mu}$ and $\tilde{\sigma}$ are highly robust estimators of the corresponding population values μ and σ. This desirable robustness property is not shared by $\bar{x}$ and s.

2. NORMALIZING THE HISTOGRAM

A normal density curve encloses an area of one. For this reason it is appropriate before fitting to a histogram that the histogram itself should be rescaled so that it encloses an area of one. To do this, let

ξ_O = left endpoint of first histogram bar,
ξ_M = right endpoint of last histogram bar,
b = bar width, $\xi_j = \xi_O + bj$,
$M = (\xi_M - \xi_O)/b$ = number of histogram bars,
n = sample size, and
n_j = number of observations in the jth bar interval.

A scaled histogram is given by

$$h(t) = C \cdot n_j \quad \text{if} \quad \xi_{j-1} < t \leq \xi_j, j = 1, \ldots, M, \qquad (2.1)$$

with $C = (bn)^{-1}$ to give area one.

3. FITTING A HISTOGRAM WITH A NORMAL DENSITY

The basic mathematical results to be derived are actually valid for fitting any nonnegative function by a normal density. Thus we let $g(\cdot)$ denote a nonnegative function on the line. In our applications g will be an area one histogram, but that is not required in Theorem 3.1. The objective as mentioned in the introduction, is

*Lawrence D. Brown and J. T. Gene Hwang are Professors, Department of Mathematics, Cornell University, Ithaca, NY. 14853 and are supported in part by NSF-DMS 9107842.

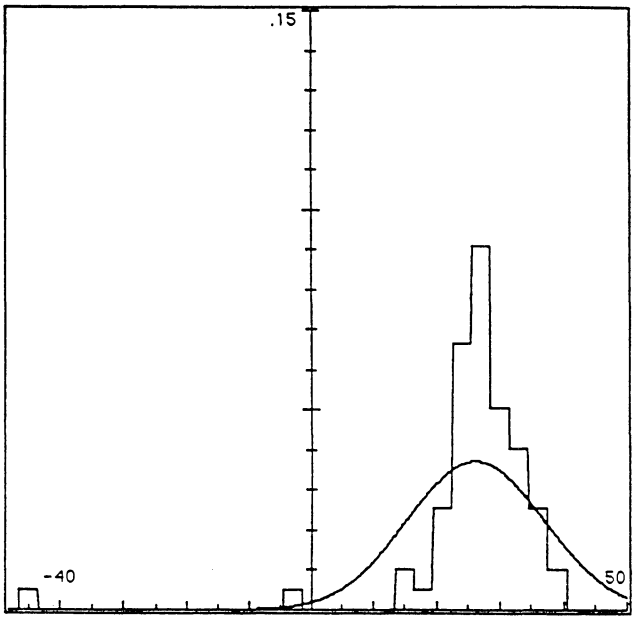

Figure 1. Histogram of Newcomb's Data and a Normal Density With the Same Mean and Standard Deviation.

to find μ, σ to minimize

$$D(\mu, \sigma) = \int (\varphi_{\mu,\sigma}(t) - g(t))^2 \, dt, \qquad (3.1)$$

where $\varphi_{\mu,\sigma}$ denotes the normal density with mean μ and standard deviation σ. In a slightly different nonparametric approach Rudemo (1982) proposed a similar criterion. Here is the first result.

Theorem 3.1. Values (μ, σ) minimizing (3.1) must exist. Any such values satisfy

$$\int (t - \mu)\varphi_{\mu,\sigma}(t)g(t) \, dt = 0, \qquad (3.2)$$

and

$$4\sigma\sqrt{\pi} \int \left(1 - \frac{(t-\mu)^2}{\sigma^2}\right)\varphi_{\mu,\sigma}(t)g(t) \, dt = 1. \qquad (3.3)$$

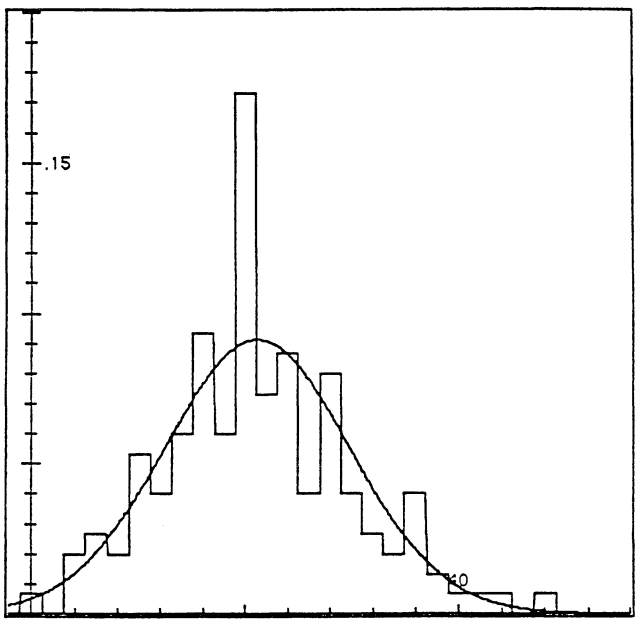

Figure 2. Histogram of Fisher's Iris Data and a Normal Density With the Same Mean and Standard Deviation.

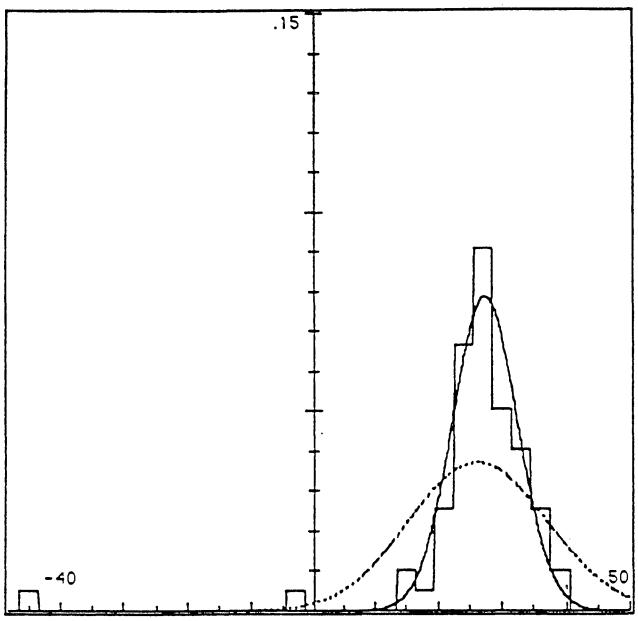

Figure 3. Histogram of Newcomb's Data. Solid curve shows best-fitting normal density. Dotted curve is $N(\bar{x}, s^2)$ density as in Figure 1.

Proof. Note that

$$\varphi_{\mu,\sigma}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-((t-\mu)^2/2\sigma^2)}.$$

It is clear that the function $D(\mu, \sigma)$ is differentiable in both μ and σ and satisfies

$$\lim_{\mu \to \pm\infty} D(\mu, \sigma) = \int g^2(t) \, dt + \int \varphi^2_{0,\sigma}(t) \, dt$$

$$\lim_{\sigma \to \infty} D(\mu, \sigma) = \int g^2(t) \, dt \text{ uniformly in } \mu$$

$$\lim_{\sigma \to 0} D(\mu, \sigma) = \infty \text{ uniformly in } \mu.$$

Also, $\inf\{D(\mu, \sigma): (\mu, \sigma)\} < \int g^2(t) \, dt$. Hence $D(\mu, \sigma)$

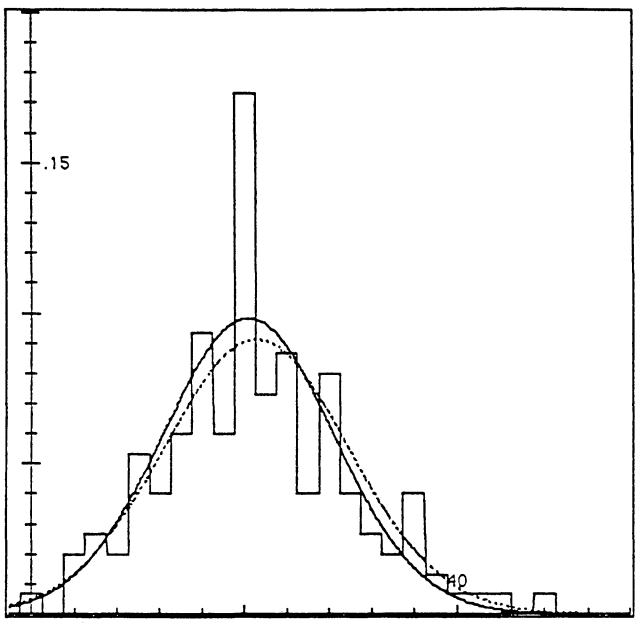

Figure 4. Histogram of Fisher's Iris Data. Solid curve shows best-fitting normal density. Dotted curve is $N(\bar{x}, s^2)$ density as in Figure 2.

achieves its minimum, and all points of minima must satisfy

$$\frac{\partial D}{\partial \mu} = 0 = \frac{\partial D}{\partial \sigma}.$$

The derivatives may be calculated inside the integral sign. Hence

$$0 = \frac{\partial}{\partial \mu} D(\mu, \sigma)$$

is equivalent to

$$0 = \int (t - \mu)\varphi_{\mu,\sigma}(t)(\varphi_{\mu,\sigma}(t) - g(t))\, dt. \quad (3.4)$$

By symmetry,

$$\int (t - \mu)\varphi^2_{\mu,\sigma}(t)\, dt = 0. \quad (3.5)$$

Combining (3.4) and (3.5) yields (3.2). Similarly,

$$0 = \frac{\partial}{\partial \sigma} D(\mu, \sigma)$$

is equivalent to

$$0 = \int \left(\frac{(t-\mu)^2}{\sigma^3} - \frac{1}{\sigma}\right)\varphi_{\mu,\sigma}(t)(\varphi_{\mu,\sigma}(t) - g(t))\, dt. \quad (3.6)$$

Now, $\varphi^2_{\mu,\sigma} = (\varphi_{\mu,\sigma/\sqrt{2}})/2\sigma\sqrt{\pi}$ so that

$$\int \left(\frac{(t-\mu)^2}{\sigma^3} - \frac{1}{\sigma}\right)\varphi^2_{\mu,\sigma}(t)\, dt = -\frac{1}{4\sigma^2\sqrt{\pi}}. \quad (3.7)$$

Substituting (3.7) in (3.6) and rearranging terms yields (3.3).

For a general g Equations (3.2) and (3.3) may be moderately awkward to solve numerically. When g is a histogram, they can be reduced to a more convenient form, as follows.

Corollary 3.1. When g is a scaled histogram (2.1), then Equations (3.2) and (3.3), which describe the optimal μ, σ, become

$$\sum_{j=1}^{M} n_j[\varphi_{\mu,\sigma}(\xi_j) - \varphi_{\mu,\sigma}(\xi_{j-1})] = 0 \quad (3.8)$$

and

$$4\sigma\sqrt{\pi}\sum_{j=1}^{M} Cn_j[\xi_j \cdot \varphi_{\mu,\sigma}(\xi_j) - \xi_{j-1} \cdot \varphi_{\mu,\sigma}(\xi_{j-1})] = 1. \quad (3.9)$$

Proof. The formulas follow from the facts that

$$\int_{\xi_{j-1}}^{\xi_j} (t - \mu)\varphi_{\mu,\sigma}(t)\, dt = [\varphi_{\mu,\sigma}(t_j)]_{\xi_{j-1}}^{\xi_j}$$

and

$$\int_{\xi_{j-1}}^{\xi_j} \frac{(t-\mu)^2}{\sigma^2}\varphi_{\mu,\sigma}(t)\, dt = \int_{\xi_{j-1}}^{\xi_j} \varphi_{\mu,\sigma}(t)\, dt - [(t - \mu)\varphi_{\mu,\sigma}(t)]_{\xi_{j-1}}^{\xi_j}.$$

Equations (3.8) and (3.9) involve only sums and not integrals. They are consequently much more tractable for numerical solutions than are (3.2) and (3.3).

Unfortunately, it appears that the system (3.8)–(3.9) may have multiple roots. If this occurs then not all solutions to the system will correspond to the desired minimum of $D(\mu, \sigma)$. However, multiple roots appear not to be a serious problem in several examples we have investigated. (Furthermore, as a referee points out, one way to ameliorate the multiple root problem is to, during the iteration to a solution of (3.8)–(3.9), take a step only when the distance (3.1) is reduced. Otherwise, cut the step in half repeatedly until there is a reduction. Stop if there is no significant reduction, since this indicates one is already near a local minimum.)

4. AN ASYMPTOTIC SOLUTION

The solution provided by Corollary 3.1 naturally depends on the bar width b used to construct the histogram. However, as b decreases, the values (μ, σ) minimizing $D(\mu, \sigma)$ converge, say to (μ^*, σ^*). Whenever b is not large (as compared to σ), (μ^*, σ^*) can be used in place of the minimizing (μ, σ) in order to provide a satisfactory fit to the histogram. Here is a precise result describing this convergence.

Theorem 4.1. Consider a sample consisting of the values $\{x_i: i = 1, \ldots, n\}$. Let $b_k \to 0$. For the given sample consider the scaled histogram h_k, say, constructed with bar width b_k. Let $D_k(\mu, \sigma)$ denote the corresponding squared error measure (3.1), and let (μ_k, σ_k) denote the parameter values yielding its minimum. Then there is a subsequence k' such that $(\mu_{k'}, \sigma_{k'})$ converges. Let μ^*, σ^* denote the limit of any convergent subsequence. Then μ^*, σ^* satisfy

$$\sum_{i=1}^{n} (x_j - \mu^*)\varphi_{\mu^*,\sigma^*}(x_j) = 0 \quad (4.1)$$

$$4\sigma^*\sqrt{\pi}\sum_{i=1}^{n}\left(1 - \frac{(x_j - \mu^*)^2}{\sigma^{*2}}\right)\varphi_{\mu^*,\sigma^*}(x_j) = 1. \quad (4.2)$$

Proof. Let $G_k(t) = \int_{-\infty}^{t} h_k(t)dt$. Then (except for $t \in \{x_i\}$) $G_k(t) \to \hat{F}_n(t)$, the sample cdf. It follows that

$$0 = \int (t - \mu)\varphi_{\mu_{k'},\sigma_{k'}}(t)h_{k'}(t)\, dt \to \int (t - \mu)\varphi_{\mu^*,\sigma^*}(t)\, d\hat{F}_n(t),$$

which is the left side of (4.1). Similar convergence holds in (3.3), which converges to (4.2). The assertions of the theorem thus follow from Theorem 3.1 and standard convergence arguments.

The solutions to (4.1) and (4.2) are as easily computed numerically as those to (3.8) and (3.9). (4.1) and (4.2) depend only on the data, not on the bar width b of the histogram. Thus one may compute the values μ^*, σ^* directly from the data, and expect the corresponding normal density to provide nearly the best fit

to the histogram produced from the data, unless M is small.

5. EXAMPLES

Figures 1 and 2 illustrate the process of fitting a histogram with a normal density having $\mu = \bar{x}$, $\sigma = s$. Figures 3 and 4 show these same histograms fitted by the least squares solution of Corollary 3.1. For comparison the normal density curves of Figures 1 and 2 are also shown on these plots. [A referee suggests we could also have compared the solution of Corollary 3.1 with that produced from a conventional robust procedure such as a 25% trimmed mean for μ and the appropriate mean absolute deviation (MAD) for σ. Those values turn out to be close to what Corollary 3.1 produces, but not identical; and of course Corollary 3.1 yields a better fit in the sense of (3.1).]

Figure 5 is another illustration; the histogram this time is the population histogram of the Poisson probability function with $\lambda = 2.5$. This histogram of course possesses more regularity than one would expect from a statistical sample. It is also somewhat skewed. Because of the central limit theorem this histogram is customarily compared to that of a normal $\mu = 2.5$, $\sigma^2 = 2.5$ density. Figure 5 also shows the best-fitting normal density as computed from Corollary 3.1.

Table 1 summarizes numerical results related to Figures 3 to 5. For comparison it also gives the values of μ^*, σ^* for the asymptotic (as $b \to 0$) best fit for the Newcomb and Fisher data. Note that in each case μ^*, σ^* are rather close to the optimal μ, σ found from Corollary 3.1.

6. ROBUSTNESS PROPERTY

The estimators $\hat{\mu} = \mu^*$ and $\hat{\sigma} = \sigma^*$ obtained by solving (μ^*, σ^*) in (4.1) and (4.2) seem to be quite robust. To examine $\hat{\mu}$, we carried out a simulation study in which x_i, $1 \le i \le n$, are iid standard Cauchy random variables with $n = 20, 30, 50$. We compared the performance of $\hat{\mu}$ with $\bar{x}$. One thousand n vectors of Cauchy random variables are generated in each case of Table 2. Although the theoretical value of the standard deviation of $\bar{x}$ is infinite, the numerical values are included for comparison. As expected the numerical values of $\bar{x}$ fluctuate a lot, as illustrated by the simulation standard deviations in Table 2.

Since the estimators $(\hat{\mu}, \hat{\sigma})$ are location and scale invariant, the above simulation provides useful information for other parameter configurations as well. If instead of observing the Cauchy random variable x_i, we observe $a + bx_i$, then the bias of $\hat{\mu}$ and its standard deviation will be multiplied by b and $|b|$. This also indicates that $\hat{\mu}$ is an unbiased estimator for this Cauchy example. Indeed it must be unbiased for any symmetric model since the distributions of $a + bx_i$ and $a - bx_i$ are the same.

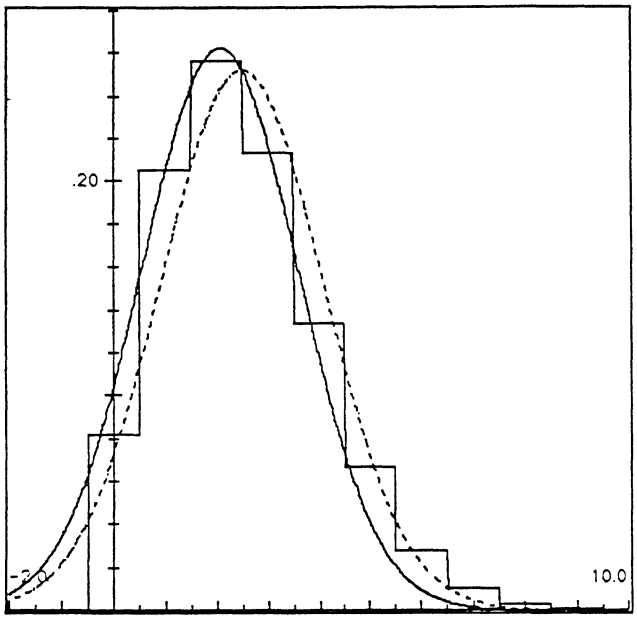

Figure 5. Histogram of the Poisson ($\lambda = 2.5$) Distribution. Solid curve is best-fitting normal density. Dotted curve is N(2.5, 2.5) density.

Technical difficulty arises in solving (4.1) and (4.2). Newton's method, simultaneously applied to both μ and σ (in which a gradient matrix is calculated) does not work well. Instead, by fixing σ^*, we solve μ^* from (4.1) via a one-dimensional Newton method and then plug such a value of μ^* into (4.2) and solve σ^* in (4.2), again via a one-dimensional Newton method. The procedure is iterated until numerical convergence is attained. The initial points for μ^* and σ^* are taken to be the median and interquartile range of the data.

This stepwise Newton method works well because for a fixed σ^*, a precise determination of μ^* is very easy to obtain from (4.1). There was also usually no problem in solving σ^* from (4.2) during the simulation study. A referee also argues that this works well since μ and σ are nearly orthogonal. However, occasionally, it did happen that the solution strayed off to infinity. This is

Table 1. Values of $\bar{x}$, s; of μ, σ From Corollary 3.1; and of μ^*, σ^* From Theorem 4.1 (where appropriate)

Histogram	$\bar{x}$	s	μ	σ	μ^*	σ^*
Newcomb (as in Figs. 1 and 3)	26.2121	10.7453	27.3801	5.0687	27.2946	4.6726
Fisher (as in Figs. 2 and 4)	30.5530	4.3728	30.3892	4.2042	30.1748	4.0583
Poisson ($\lambda = 2.5$) (as in Fig. 5)	2.5	1.58	2.07	1.53		

Table 2. Simulated Bias

m =	20	30	50
μ^*	.00190	−.0069	.014
	(.340)	(.289)	(.223)
$\bar{x}$	46	−1.83	.276
	(40.42)	(59.46)	(37.18)

NOTE: The simulation standard deviations are reported in parentheses.

due to the fact that D has a "local maximum" at infinity, in the sense that it converges to a constant. Therefore in the simulation we restricted the region of σ^* to be $(0, \max x_i - \min x_i)$ and when it overshot we generated a new initial point $U(\max x_i - \min x_i)$ where U is a uniform random number over [0, 1].

The result shows that μ^* has a fairly small standard deviation.

7. RELATION TO ROOTOGRAMS

Velleman and Hoaglin (1981) discussed the "rootogram" procedure derived from Freeman and Tukey (1950). Essentially, this procedure examines the residuals between the square root of the histogram and the square root of a fitted normal density. The resulting differences are called double-root residuals. (Actually, a small correction factor is introduced into the formulae to avoid difficulties with small cell counts.)

Velleman and Hoaglin suggested either visually fitting the best normal density or using some robust estimator of μ and σ to produce this fit. Alternately, it would be possible to extend the ideas of the current article to produce the normal density which minimizes the sum of squares of their double-root residuals. We believe that the method of the present article, based on ordinary residuals, will generally produce a better visual fit. (On the other hand, the double-root residuals may be preferable for other purposes. It may result in more efficient estimation under appropriate assumptions. Also the double-root residual produces a convenient test of goodness of fit. This is because the sampling distribution of this sum of squares is probably very well approximated by a χ^2 distribution.) The rootogram is also related to minimum Hellinger distance estimation. See Simpson (1987) for more on this topic.

[*Received November 1991. Revised September 1992.*]

REFERENCES

Andrews, D. F., and Herzberg, A. M. (1985), *Data*, New York: Springer-Verlag.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7, 179–188.

Freeman, M. F., and Tukey, J. W. (1950), "Transformations Related to the Angular and the Square Root," *Annals of Mathematical Statistics*, 21, 607–611.

Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Density Estimators, *Scandinavian Journal of Statistics*, 9, 65–78.

Simpson, D. G. (1987), "Minimum Hellinger Distance Estimation for the Analysis of Count Data," *Journal of the American Statistical Association*, 82, 802–807.

Stigler, S. M. (1977), "Do Robust Estimators Work With Real Data?" *Annals of Statistics*, 5, 1055–1978.

Velleman, P. F., and Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis*, Belmont, CA: Duxbury Press.