TECHNICAL RESEARCH REPORT

Minimum Chi-Square vs Least Squares in Grouped Data

by B. Kedem and Y. Wu

T.R. 97-37

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Minimum Chi-Square vs Least Squares in Grouped Data

by

Benjamin Kedem and Yougui Wu

Mathematics Department and Institute for Systems Research
University of Maryland, College Park
Maryland 20742, USA

April 1997 [1]

[1]Work supported by NASA grant NAG52783

Abstract

Estimation of parameters from grouped data is considered using a least squares estimator popular in scientific applications. The method minimizes the square distance between the empirical and hypothesized cumulative distribution functions, and is reminiscent of a discrete version of the Cramér-von Mises statistic. The resulting least squares estimator, is related to the minimum chi-square estimator, and likewise is asymptotically normal. The two methods are compared briefly for categorized mixed lognormal data with a jump at zero.

Key Words: Maximum likelihood, asymptotic normality, relative efficiency, mixed lognormal.

1 Introduction

In scientific applications data are often grouped or categorized due to instrument limitations, such as the inability to measure very large or very small quantities, and the inability to produce precise measurements. In such cases, suitable estimation methods are used in the estimation of the parameters of the parent distribution of the ungrouped from grouped data. Meneghini and Jones (1993), encountering attenuation problems with a spaceborne precipitation radar, suggested the use of a least squares method, related to minimum chi-square, that minimizes the squared distance between the empirical and hypothesized cumulative distribution functions at a few "thresholds". This amounts to minimizing a parametric discrete version of the W^2 statistic of Anderson and Darling (1952). The apparent similarity between the least squares and minimum chi-square methods has prompted a limited empirical comparison between them using real data in Kedem et al. (1997), however no theoretical study of the least squares method has been attempted there. The present work is motivated by the need to assess the variability of the least squares estimator, an adopted algorithm for the Tropical Rainfall Measuring Mission (TRMM) of the National Aeronautics and Space Administration to measure rainfall from space via a radar and several other instruments. See Simpson et al. (1996) for a comprehensive overview of TRMM.

The least squares method is described in section 2, and in section 3 we study the asymptotic distribution of the least squares estimator. A brief comparison with minimum chi-square is carried out in section 4 assuming the parent distribution is mixed lognormal with a jump at 0.

2 Estimation in Grouped Data

Let $\{A_j\}_{j=1}^r$ be a partition of the support of a random variable X with cumulative distribution function (cdf) $F(x, \boldsymbol{\theta})$, and let $X_1, ..., X_n$ be a random sample from X. Define n_j as the number of X_j's that fall in cell j, and put $\pi_j \equiv \pi_j(\boldsymbol{\theta}) = P(X \in A_j)$, $j = 1, ..., r$. The count data follow the "parametric" multinomial model

$$P(\mathbf{n}|\boldsymbol{\theta}) = \frac{n!}{n_1! \cdots n_r!} \pi_1^{n_1}(\boldsymbol{\theta}) \cdots \pi_r^{n_r}(\boldsymbol{\theta}) \tag{1}$$

with r cells, cell counts $\mathbf{n} = (n_1, ..., n_r)'$, and cell probabilities $\pi_1^{n_1}(\boldsymbol{\theta}), ..., \pi_r^{n_r}(\boldsymbol{\theta})$, depending on a vector parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)'$, such that $\sum_{j=1}^r n_j = n$ and

$\sum_{j=1}^{r} \pi_j(\boldsymbol{\theta}) = 1$.

There are several methods to estimate $\boldsymbol{\theta}$ from the "grouped" or count data modeled as (1) including:

(a) Maximum likelihood (ML) whereby the log-likelihood

$$\log L(\boldsymbol{\theta}) = \text{constant} + \sum_{j=1}^{r} n_j \log \pi_j(\boldsymbol{\theta})$$

is maximized with respect to $\boldsymbol{\theta}$.

(b) Minimum chi-square (MCS) whereby the discrepancy between the observed counts and their expectation is minimized with respect to $\boldsymbol{\theta}$ using the quadratic form

$$\chi^2(\boldsymbol{\theta}) = \sum_{j=1}^{r} \frac{[n_j - n\pi_j(\boldsymbol{\theta})]^2}{n\pi_j(\boldsymbol{\theta})}$$

(c) Modified minimum chi-square (MMCS) where $\chi^2(\boldsymbol{\theta})$ is replaced by

$$\chi_1^2(\boldsymbol{\theta}) = \sum_{j=1}^{r} \frac{[n_j - n\pi_j(\boldsymbol{\theta})]^2}{n_j}$$

where n_j is replaced by 1 if it vanishes.

Several additional methods that minimize the discrepancy between observed and expected counts such as Kullback-Leibler distance are described in Rao (1973), p. 352, and a general approach is reviewed in Hsiao (1985).

It is well known, under fairly general conditions, the asymptotic distribution of the ML estimator $\hat{\boldsymbol{\theta}}$ is normal,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}))$$

where $\mathbf{I}(\boldsymbol{\theta})$ is Fisher information "per observation" corresponding to (1),

$$\mathbf{I}(\boldsymbol{\theta}) = \sum_{j=1}^{r} \frac{1}{\pi_j(\boldsymbol{\theta})} \nabla \pi_j(\boldsymbol{\theta}) \nabla' \pi_j(\boldsymbol{\theta})$$

and ∇ denotes the column gradient operator,

$$\nabla f(\boldsymbol{\theta}) \equiv \left(\frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_2}, \cdots, \frac{\partial f}{\partial \theta_m} \right)'$$

and that the three methods ML, MCS, and MMCS are asymptotically equivalent, all producing best asymptotically normal estimates (Sen and Singer 1993, Sec. 6.3). Thus, for sufficiently large n, the MCS estimator is normal with mean $\boldsymbol{\theta}$ and $m \times m$ covariance matrix $\mathbf{I}^{-1}(\boldsymbol{\theta})/n$.

There is yet another closely related "grouped data" estimator that minimizes the squared distance between the empirical and hypothesized cdf's. Denote the empirical cdf by $F_n(x)$, and barring the two ends of the support, let $R_T(1), ..., R_T(k)$, $k = r - 1$, be the cell boundaries. Consider a discrete version of the Cramér-von Mises statistic,

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{k} [F_n(R_T(i)) - F(R_T(i), \boldsymbol{\theta})]^2 \tag{2}$$

Meneghini and Jones (1993) estimate $\boldsymbol{\theta}$ by minimizing $S(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, and in Kedem et al. (1997) it is found empirically the resulting estimate–from now on "least squares" (LS) estimate–is close to the MCS estimate in a special case.

The LS estimator that minimizes (2) belongs to the family of minimum-distance estimators, that minimize a certain distance $\delta(F_n(x), F(x, \boldsymbol{\theta}))$, and its asymptotic distribution can be found by appealing to the general theory of such estimators. General approaches are studied in Bolthausen (1977) and Parr and Schucany (1981). However, the asymptotic distribution of the LS estimator can be obtained by much simpler means.

The purpose of this paper is to derive the asymptotic distribution of the LS estimator in general, and then compute its efficiency relative to the MCS estimator, with the same class boundaries, in a special case when estimating the mean of a mixed lognormal distribution with a jump at 0, the case considered in Meneghini and Jones (1993) and Kedem et al. (1997).

3 Distribution of the LS Estimator

With $k = r-1$, $\mathbf{I}_k$ the $k \times k$ identity matrix, $\boldsymbol{\pi}_k^{1/2} \equiv \boldsymbol{\pi}_k^{1/2}(\boldsymbol{\theta}) = (\pi_1^{1/2}(\boldsymbol{\theta}), ..., \pi_k^{1/2}(\boldsymbol{\theta}))'$, and $Z_j = (n_j - n\pi_j(\boldsymbol{\theta}))/\sqrt{n\pi_j(\boldsymbol{\theta})}$, $j = 1, ..., k$, a fact to be used below is that as $n \to \infty$ (e.g. Sen and Singer 1993, p. 251),

$$\mathbf{Z} \xrightarrow{\mathcal{D}} \mathrm{N}\left(\mathbf{0}, \mathbf{I}_k - \boldsymbol{\pi}_k^{1/2}(\boldsymbol{\pi}_k^{1/2})'\right) \tag{3}$$

Observe $\sum_{i=1}^k \pi_i^{1/2}(\boldsymbol{\theta}) < 1$.

The LS estimator can be studied more conveniently by noticing that,

$$\begin{aligned} S(\boldsymbol{\theta}) = \sum_{i=1}^k \left[\frac{1}{n}\sum_{j=1}^i n_j - \sum_{j=1}^i \pi_j(\boldsymbol{\theta})\right]^2 &= \frac{1}{n^2}\sum_{i=1}^k \left[\sum_{j=1}^i n_j - n\sum_{j=1}^i \pi_j(\boldsymbol{\theta})\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^k \left[\frac{M_i - nq_i(\boldsymbol{\theta})}{\sqrt{n}}\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^k Y_i \end{aligned} \tag{4}$$

where $M_i = \sum_{j=1}^i n_j$, $q_i(\boldsymbol{\theta}) = \sum_{j=1}^i \pi_j(\boldsymbol{\theta})$, and $Y_i = [M_i - nq_i(\boldsymbol{\theta})]/\sqrt{n}$, $i = 1, ..., k$. Let $\mathbf{Y} = (Y_1, ..., Y_k)'$, and define a $k \times k$ matrix $\mathbf{G} \equiv \mathbf{G}(\boldsymbol{\theta})$,

$$\mathbf{G} = \begin{pmatrix} \pi_1^{1/2}(\boldsymbol{\theta}) & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ \pi_1^{1/2}(\boldsymbol{\theta}) & \pi_2^{1/2}(\boldsymbol{\theta}) & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \pi_1^{1/2}(\boldsymbol{\theta}) & \pi_2^{1/2}(\boldsymbol{\theta}) & \cdot & \cdot & \cdot & \cdot & \pi_k^{1/2}(\boldsymbol{\theta}) \end{pmatrix}$$

It follows that $\mathbf{Y} = \mathbf{GZ}$, or from (3)

Lemma 3.1 *As* $n \to \infty$,

$$\mathbf{Y} \xrightarrow{\mathcal{D}} \mathrm{N}\left(\mathbf{0}, \mathbf{G}(\mathbf{I}_k - \boldsymbol{\pi}_k^{1/2}(\boldsymbol{\pi}_k^{1/2})')\mathbf{G}'\right)$$

Define next a $k \times m$ matrix $\mathbf{B}$ by

$$\mathbf{B} = (\nabla q_1(\boldsymbol{\theta}), ..., \nabla q_k(\boldsymbol{\theta}))'$$

and denote the LS estimator by $\boldsymbol{\theta}^*$.

Theorem 3.1 *Assume* $k = r - 1$ *and*
(i) $\mathbf{B}'\mathbf{B}$ *has full rank.*
(ii) $\dfrac{\partial q_i(\boldsymbol{\theta})}{\partial \theta_j}$ *and* $\dfrac{\partial^2 q_i(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l}$, $i = 1, ..., k$, $j, l = 1, ..., m$, *are continuous.*
Then

$$\sqrt{n}(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{G}\left\{\mathbf{I}_k - \boldsymbol{\pi}_k^{1/2}(\boldsymbol{\theta})(\boldsymbol{\pi}_k^{1/2}(\boldsymbol{\theta}))'\right\}\mathbf{G}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}$$

Proof: The method of proof follows that of ML. Define

$$H(\boldsymbol{\theta}) = -\frac{1}{2n}\sum_{i=1}^{k}[M_i - nq_i(\boldsymbol{\theta})]^2$$

Then $\boldsymbol{\theta}^*$ maximizes $H(\boldsymbol{\theta})$. Let $||\mathbf{u}|| < K$, $0 < K < \infty$, $\mathbf{u} \in \mathcal{R}^k$, and consider the Taylor expansion,

$$H(\boldsymbol{\theta} + n^{-1/2}\mathbf{u}) = H(\boldsymbol{\theta}) + \frac{1}{\sqrt{n}}\mathbf{u}'\nabla H(\boldsymbol{\theta}) + \frac{1}{2n}\mathbf{u}'\nabla\nabla' H(\tilde{\boldsymbol{\theta}})\mathbf{u}$$

where $\tilde{\boldsymbol{\theta}}$ lies on the line segment connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \mathbf{u}/\sqrt{n}$. Introduce the function of $\mathbf{u}$,

$$\lambda(\mathbf{u}) \equiv H(\boldsymbol{\theta} + n^{-1/2}\mathbf{u}) - H(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\mathbf{u}'\nabla H(\boldsymbol{\theta}) + \frac{1}{2n}\mathbf{u}'\nabla\nabla' H(\tilde{\boldsymbol{\theta}})\mathbf{u}$$

and put,

$$\mathbf{U} = \nabla H(\boldsymbol{\theta}),\ \mathbf{V} = \nabla\nabla' H(\boldsymbol{\theta}),\ \mathbf{W} = \nabla\nabla' H(\tilde{\boldsymbol{\theta}}) - \nabla\nabla' H(\boldsymbol{\theta})$$

so that

$$\lambda(\mathbf{u}) = \frac{1}{\sqrt{n}}\mathbf{u}'\mathbf{U} + \frac{1}{2n}\mathbf{u}'\mathbf{V}\mathbf{u} + \frac{1}{2n}\mathbf{u}'\mathbf{W}\mathbf{u}$$

Now observe that,
(i)

$$\frac{1}{\sqrt{n}}\mathbf{U} = \frac{1}{\sqrt{n}}\sum_{i=1}^{k}\left[M_i - nq_i(\boldsymbol{\theta})\right]\nabla q_i(\boldsymbol{\theta}) = \mathbf{B}'\mathbf{Y}$$

Therefore by Lemma (3.1), as $n \to \infty$,

$$\frac{1}{\sqrt{n}}\mathbf{U} \xrightarrow{\mathcal{D}} \mathrm{N}\left(\mathbf{0}, \mathbf{B}'\mathbf{G}(\mathbf{I}_k - \boldsymbol{\pi}_k^{1/2}(\boldsymbol{\pi}_k^{1/2})')\mathbf{G}'\mathbf{B}\right) \tag{5}$$

(ii) By the law of large numbers, as $n \to \infty$,

$$\begin{aligned}\frac{1}{n}\mathbf{V} &= \frac{1}{n}\sum_{i=1}^{k}\left[-n\nabla q_i(\boldsymbol{\theta})\nabla' q_i(\boldsymbol{\theta})\right] + \frac{1}{n}\sum_{i=1}^{k}\left[M_i - nq_i(\boldsymbol{\theta})\right]\nabla\nabla' q_i(\boldsymbol{\theta}) \\ &= -\mathbf{B}'\mathbf{B} + \frac{1}{n}\sum_{i=1}^{k}\left[M_i - nq_i(\boldsymbol{\theta})\right]\nabla\nabla' q_i(\boldsymbol{\theta}) \xrightarrow{p} -\mathbf{B}'\mathbf{B}\end{aligned}$$

(iii) By the assumed continuity of the derivatives, as $n \to \infty$,

$$\frac{1}{n}\mathbf{W} \xrightarrow{p} \mathbf{0}$$

Consequently for sufficiently large n, $\lambda(\mathbf{u})$ is quadratic,

$$\lambda(\mathbf{u}) = \frac{1}{\sqrt{n}}\mathbf{u}'\mathbf{U} - \frac{1}{2}\mathbf{u}'\mathbf{B}'\mathbf{B}\mathbf{u} + o_p(1)$$

with maximum at

$$\hat{\mathbf{u}} = \frac{1}{\sqrt{n}}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{U} + o_p(\mathbf{1})$$

This implies the maximum of $H(\boldsymbol{\theta})$ occurs at

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} + \frac{1}{\sqrt{n}}\hat{\mathbf{u}} = \boldsymbol{\theta} + \frac{1}{n}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{U} + o_p(\mathbf{1})$$

and this together with (5) gives,

$$\sqrt{n}(\boldsymbol{\theta}^* - \boldsymbol{\theta}) = \frac{1}{\sqrt{n}}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{U} + o_p(\mathbf{1}) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \qquad \square$$

From the proof of Theorem (3.1), k may be less than the case of interest $r-1$, however, the efficiency of the LS estimate deteriorates rapidly for an excessively small k.

4 Relative Efficiency

Since the MCS and ML estimators are asymptotically equivalent for grouped data, the MCS is denoted here by $\hat{\boldsymbol{\theta}}$ as well. The preceding discussion concerns then the MCS and LS estimators $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, respectively. One way to compare $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, is to compare the asymptotic variances of $g(\hat{\boldsymbol{\theta}})$ and $g(\boldsymbol{\theta}^*)$, where $g(\boldsymbol{\theta})$ is a sufficiently smooth real valued function of $\boldsymbol{\theta}$. Let $\boldsymbol{\gamma} = \nabla g(\boldsymbol{\theta})$. Then the delta method (e.g. Sen and Singer 1993, pp. 131-133) gives for the MCS estimator,

$$\sqrt{n}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0, \boldsymbol{\gamma}'\mathbf{I}^{-1}(\boldsymbol{\theta})\boldsymbol{\gamma}\right)$$

and for the LS estimator

$$\sqrt{n}(g(\boldsymbol{\theta}^*) - g(\boldsymbol{\theta})) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0, \boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma}\right)$$

A relative efficiency is defined by the ratio

$$\mathrm{Eff}(\boldsymbol{\theta}) \equiv \frac{\boldsymbol{\gamma}'\mathbf{I}^{-1}(\boldsymbol{\theta})\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma}} \tag{6}$$

4.1 The Mixed Lognormal Case

Consider now the special case, the one that triggered the present investigation, of mixed lognormal with $m = 3$ parameters p, μ, σ,

$$F(x, \boldsymbol{\theta}) = (1 - p) + p\Lambda(x; \mu, \sigma), \quad x \geq 0$$

and $F(x, \boldsymbol{\theta}) = 0$ otherwise, where $0 < p < 1$, $\Lambda(x; \mu, \sigma)$ is the lognormal cdf with parameters μ, σ, and $\boldsymbol{\theta} = (p, \mu, \sigma)'$. The function of $\boldsymbol{\theta}$ of interest is the mean of the mixed lognormal distribution,

$$g(\boldsymbol{\theta}) = p\exp(\mu + \frac{1}{2}\sigma^2)$$

with

$$\boldsymbol{\gamma} = \nabla g(\boldsymbol{\theta}) = \exp(\mu + \frac{1}{2}\sigma^2)(1, p, p\sigma)'$$

Calculation of the relative efficiency (6) with $\mathbf{I}(\boldsymbol{\theta})$, given explicitly in Kedem et al. (1997) and $\boldsymbol{\Sigma}$ as in Theorem 3.1, is shown in Figure 1 for various choices of p, μ, σ corresponding to some real data situations described in Kedem et al. (1997). Evidently the two methods, MCS and LS produce very similar results.

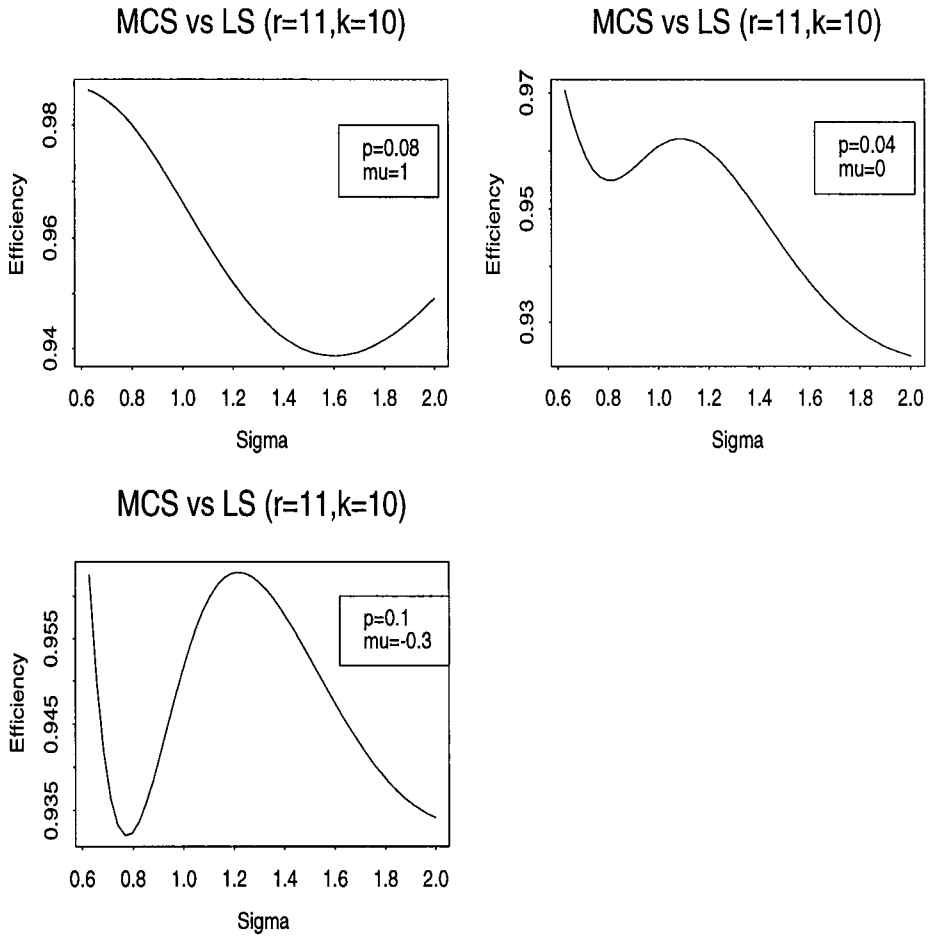

Figure 1: Relative efficiency of (6) for mixed lognormal with fixed p, μ and variable σ. $r = 11, k = 10$

References

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

Bolthausen, E., 1977: Convergence in distribution of minimum-distance estimators, *Metrika*, **24**, 215-227.

Hsiao, C., 1985: Minimum chi-square, in *Encyclopedia of Statistical Sciences*, S. Kotz and N.L. Johnson, eds., Vol. 5, Wiley & Sons, 518-522.

Kedem, B., R. Pfeiffer, and D. A. Short, 1997: Variability of Space-Time Mean Rain Rate, *J. Appl. Meteor.*, **36**, 443-451.

Meneghini, R., and J. A. Jones, 1993: An approach to estimate the areal rain-rate distribution from spaceborne radar by the use of multiple thresholds, *J. Appl. Meteor.*, **32**, 386-398.

Parr, W. C., and W. R. Schucany, 1980: Minimum distance and robust estimation, *J. Amer. Statist. Assoc.*, **75**, 616-624.

Rao, C.R., 1973: *Linear Statistical Inference and Its Applications*, Wiley and Sons, 625 pp.

Sen, P., and J.M. Singer, 1993: *Large Sample Methods in Statistics an Introduction With Applications*, Chapman & Hall, 382 pp.

Simpson, J., C. Kummerow, W. -K. Tao, and R. F. Adler, 1996: On the Tropical Rainfall Measuring Mission (TRMM) satellite, *Meteorol. Atmos. Phys., 20*, 36-46.