Statistical Science
2004, Vol. 19, No. 4, 579–587
DOI 10.1214/088342304000000459

The Sign Statistic, One-Way Layouts and Mixture Models

R. T. Elmore, T. P. Hettmansperger and F. Xuan

Abstract. We consider the use of sign statistics in two different types of one-way layouts. The first layout is for data collected to compare several treatments. The second layout is for independent repeated measures on several subjects. In the first case we discuss hypothesis testing and multiple comparisons. In the second case we fit mixture models. We then show how fitting mixture models can be helpful in follow-up multiple comparisons in the first case.

Key words and phrases: Binomial mixtures, cut-point models, EM algorithm, Bayesian information criterion, Mood's test.

1. INTRODUCTION

This article presents some statistical settings in which the simple sign statistic is very useful. We primarily discuss the one-way layout. First, we consider the one-way layout in the context of comparing several treatments and, second, we consider a special case of mixture models in which we have repeated measures. Before turning to the one-way layout we briefly review the use of the sign statistic in a single sample of data.

Suppose we have m independent and identically distributed observations denoted by $x_1, \ldots, x_m$. Define

$$S(t) = \sum_{i=1}^{m} I(x_i \le t), \tag{1}$$

where $I(A)$ is the indicator of the event A, meaning that $I(A) = 1$ if A occurs and 0 otherwise. Hence, $S(t)$ counts the number of observations out of m that are less than or equal to t; $S(t)$ is called the sign statistic.

EXAMPLE 1 (Hypothesis testing with the sign test). In the 1960s psychologists suspected that environment affects the anatomy of the brain. The subjects for this study were a genetically pure strain of rats. From each litter, one rat was selected at random for the treatment group and one for the control group. Both groups got exactly the same food and drink. Each animal in the treatment group lived with 10 others in a large cage furnished with toys which were changed daily. Animals in the control group lived in isolation. After a month the animals were killed and their cortex weights were recorded. We wish to test the hypothesis that the treatment group tended to have higher cortex weights. The data are given in Table 1.

This is a paired-data design in which litter mates determine the pairings. Let θ denote the population median for the difference score distribution. Then we wish to test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$. We reject the null hypothesis if $S(0)$, the number of differences less than or equal to 0, is small. Under the null hypothesis, $S(0)$ has a binomial distribution with parameters $m = 11$ and $p = 0.5$. Since $S(0) = 1$, the p value for the test is $P(S(0) \le 1) = 0.0059$ from the binomial table. Hence, we conclude at reasonable significance levels that environment positively impacts cortex weight.

The sign test and the corresponding point estimate (the sample median) have relative efficiency with respect to the t test and sample mean equal to 0.64 when the underlying distribution is normal. However, if the underlying distribution has heavy tails (such as Laplace or double exponential distribution), then the efficiency can be greater than 1 and the sign test is more efficient. Because they are robust against outliers and gross errors in the data, the sign test and the sample median are excellent for exploratory and rough confirmatory

R. T. Elmore is a former graduate student, T. P. Hettmansperger is Professor and F. Xuan is a graduate student, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, USA (e-mail: tph@stat.psu.edu). R. T. Elmore is currently at The Australian National University.

TABLE 1
Cortex weights

Treatment	689	656	668	660	679	663	664	647	694	633	653
Control	657	623	652	654	658	646	600	640	605	635	642
T − C	32	33	16	6	21	17	64	7	89	−2	11

analyses. See Hettmansperger and McKean (1998) for more discussion.

2. ONE-WAY LAYOUT: COMPARING *n* TREATMENTS

In a basic nonparametric statistics course, the Kruskal–Wallis rank test is introduced to test hypotheses concerning the equality of several distributions. In this section we discuss the corresponding test that can be considered an extension of the sign test for one sample. Mood (1950) discussed this test and Minitab has a command to implement it. The test appears in standard texts as the median test for several samples. Of course, it can also be used to compare two samples.

In the one-way layout we have n samples. In the ith sample we have m_i independent observations, $x_{1i}, \ldots, x_{m_i,i}$ for $i = 1, \ldots, n$. We have $M = \sum m_i$ total observations and we wish to test $H_0 : F_1 = \cdots = F_n$ versus the alternative that they are not all equal. Here is an example.

EXAMPLE 2 (Sulfur content of coal). A study was carried out to ascertain the sulfur content of five major coal seams in Texas. Core samples were taken at random from each of the seams and analyzed. The data consist of the percentage of sulfur per plug and are given in Table 2. The research hypothesis is that the seams differ in sulfur content.

TABLE 2
Sulfur content of coal

Seam				
A	**B**	**C**	**D**	**E**
1.51	1.69	1.56	1.30	0.73
1.92	0.64	1.22	0.75	0.80
1.08	0.90	1.32	1.26	0.90
2.04	1.41	1.39	0.69	1.24
2.14	1.01	1.33	0.62	0.82
1.76	0.84	1.54	0.90	0.72
1.17	1.28	1.04	1.20	0.57
	1.59	2.25	0.32	1.18
		1.49		0.54
				1.30

Let $\hat{\theta}$ denote the combined sample median for the $M = 42$ total observations. Mood's test is built from $S_i(\hat{\theta})$, the number of observations in the ith sample that are less than or equal to $\hat{\theta} = 1.21$ for $i = 1, \ldots, 5$. The test statistic is

$$T = 4\sum_{i=1}^{n} \frac{1}{m_i}\left(S_i(\hat{\theta}) - \frac{m_i}{2}\right)^2, \tag{2}$$

where the constants are introduced in the formula so that, under H_0, T is approximately distributed as chi squared with $n-1$ degrees of freedom. See Appendix 1 for details about the asymptotic distribution. Since under H_0 all permutations of the data are equally likely, it is possible to approximate the permutation distribution of T by repeated sampling of the permutations. For the data in Table 2, $T = 12.33$ and the approximate p value is 0.015 from the chi squared table with 4 degrees of freedom. Based on a sample of 50,000 permutations of the data, the permutation p value is approximately 0.0127, close to the asymptotic approximation. Hence, for significance levels greater than 1.5% we can reject the null hypothesis $H_0 : F_1 = \cdots = F_5$ and claim that there is a difference in sulfur content across the five seams. This immediately raises the question of multiple comparisons; since, we want to know which distributions are different from the others.

We consider simple pairwise multiple comparisons. Suppose α_F is the assigned family error rate. Then using the Bonferonni inequality, we distribute the error across the family of $n(n-1)/2$ pairwise comparisons and assign an individual comparison rate of $\alpha = 2\alpha_F/n(n-1)$. Since the vector of n sign statistics is approximately multivariate normal, it can be shown using the details in Appendix 1 that the difference in sign statistics is also approximately normally distributed. We declare the ith and jth treatments significantly different at α_F when

$$\frac{2|\overline{S_i} - \overline{S_j}|}{\sqrt{1/m_i + 1/m_j}} \geq z_{\alpha/2}, \tag{3}$$

where $\overline{S}_i = m_i^{-1} S_i(\hat{\theta})$ and $z_{\alpha/2}$ is the upper $\alpha/2$ percentile from a standard normal table. In our exam-

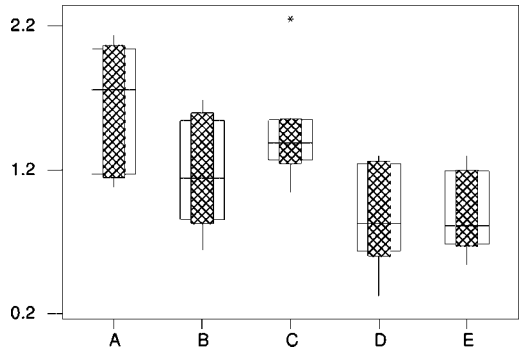

FIG. 1. *Box plots and* 95*% confidence intervals for the coal data.*

ple there are 10 pairwise comparisons. If we assign $\alpha_F = 0.10$, then the individual pairwise comparisons are conducted at $\alpha = 0.01$ and $z_{0.005} = 2.58$. The only two comparisons that are significant are C versus D and C versus E. Hence, we do not have much power to group the seams. Figure 1 shows the 95% confidence intervals (shaded areas) and suggests that A and C are different from D and E, while B cannot be separated from any of the seams.

In the next section we consider analyzing the one-way layout via a mixture model. We return to the coal data later and see how fitting a mixture model to the data may help group the seams.

3. ONE-WAY LAYOUT: MIXTURE MODELS

Recall that the one-way layout has n sets of measurements. In this section we begin by assuming that the number of measurements in each set is m. Later we consider the case when there are different numbers of measurements in the sets. In the previous section, the sets of measurements came from n treatments. In this section, the n sets come from n subjects. Hence, we consider data that come from an experiment in which we have m measurements on each of n subjects. We assume that the measurements within a subject are independent and identically distributed. If we let F_i^* represent the distribution associated with the ith subject, then as in the case of multiple comparisons, we wish to reduce $F_1^*, \ldots, F_n^*$ to a smaller set, say, $F_1, \ldots, F_K$, where $K < n$, and group the subjects into homogeneous groups. That is, we wish to categorize F_i^* as F_k for some k in $1, \ldots, K$. This shortly leads us to a mixture model. We do not wish to assume any distributional form for F_i^*. We first set the context with an example.

EXAMPLE 3 (Rod and frame task). Subjects are seated in a darkened room without visual cues. The subjects are presented with a luminous frame that contains a luminous rod tilted from the vertical. The task is to adjust the rod to a vertical position. Psychological theory suggests that there are two types of subjects: field-independent subjects who can, without much error, adjust the rod to the vertical and field-dependent subjects who tend to make large errors. The measurement is the absolute error from the vertical. The data consist of 83 sets of eight measurements. The 83 subjects were college students. Hence, $n = 83$ and $m = 8$. This is quite different from the one-way layout discussed in the previous section since now we have a very large number for n and a small number for m. The original data set can be downloaded from *http://www.blackwellpublishing.com/rss/Volumes/Bv62p4.htm* and was analyzed by Hettmansperger and Thomas (2000). Thus, we expect to reduce the complete set of distribution functions $F_1^*, \ldots, F_{83}^*$ to perhaps two or so primary components and group the subjects into field-independent and field-dependent groups.

In general, since we do not know the underlying distributions, we transform the data on each subject. Let c be a cut point in the data, and for the ith subject compute $S_i(c) = \sum_j I(x_{ji} \le c)$, the number of measurements on the ith subject that are less than or equal to c. In the previous section we took c to be the combined sample median. In the rod and frame example we take $c = 6°$. If the rod is within 6° of vertical, the subject is considered to have mastered the task.

If we know that the ith subject is associated with, say F_1, then, conditioned on this knowledge, we deduce that $S_i(c)$ is binomially distributed with parameters m and $F_1(c)$, the probability that a measurement on the ith subject will be less than or equal to c. We can then write the binomial in the manner

$$P\big(S_i(c) = s | F_1\big) = \binom{m}{s} F_1(c)^s \big(1 - F_1(c)\big)^{m-s} = b(s; m, F_1(c)), \tag{4}$$

where $b(s; m, p)$ is the binomial mass function with parameters m and p. In fact, the ith subject could come from any of $F_1, \ldots, F_K$. Suppose that λ_k is the probability that a subject is associated with F_k for $k = 1, \ldots, K$. Then, for the ith subject we have

$$P\big(S_i(c) = s\big) = \sum_{k=1}^{K} \lambda_k b(s; m, F_k(c)). \tag{5}$$

This is called the K-component binomial mixture model. It is important that the cut point be close to the center of the data. For general data sets we often take

the combined sample median $\hat{\theta}$. This induces some dependence and the binomial mixture model is then only approximate. Simulations have shown that the binomial mixture still fits quite well; see Hettmansperger and Thomas (2000) for more discussion. For more detailed explanations of all aspects of mixture models, see McLachlan and Peel (2000).

There are $2K-1$ parameters in model (5): $F_1(c)$, $\ldots, F_K(c)$, $\lambda_1, \ldots, \lambda_{K-1}$, since λ_K is determined by the others. The problem is to estimate the parameters for a fixed K. We then vary K and study the value of the (penalized) log-likelihood, choosing that value of K that maximizes the (penalized) log-likelihood. We use a penalty since the dimension of the model changes with K. In this discussion we use the Bayesian information criterion (BIC) of Schwarz (1978). We calculate

$$\text{BIC} = -2\ln(\text{likelihood}) + d\ln(n), \tag{6}$$

where d is the number of parameters that must be estimated. Minimizing BIC is equivalent to maximizing the (penalized) log-likelihood. See Appendix 2 for additional comments on BIC.

Provided that $m \geq 2K-1$, the mixture model is identifiable and there is a unique mixture model representation. If m is smaller than $2K-1$, there may be many different mixture model representations. This puts a limit on the number of components that we can fit to a given data set.

We use an expectation–maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) to fit the binomial mixture model and compute the parameter estimates using maximum likelihood. In addition to the estimates, we also get a set of posterior probabilities for each subject. These are the probabilities that the subject comes from the various component distributions. By using the posterior probabilities as weights in an empirical c.d.f., we can use all of the data to estimate the component c.d.f.'s, $F_1, \ldots, F_K$. The estimator of the component c.d.f. is a weighted version of the sign statistic (1) in which the weights are based on the posterior probabilities. See Appendix 3 for a description of an EM algorithm for binomial mixtures and Appendix 4 for estimates of the component c.d.f.'s.

We provide R/S-Plus functions for binomial mixture model estimation at the website *http://www.stat.psu.edu/~tph/StatScience/*. We note that it is possible to use several cut points; however, this results in a multinomial mixture rather than a binomial mixture. Estimation of $\lambda_1, \ldots, \lambda_K$ and the component c.d.f.'s is more efficient in this case. The R/S-Plus functions mentioned above can also be used to fit a multinomial mixture. See Cruz-Medina, Hettmansperger and Thomas (2004) and Elmore (2003) for additional information regarding multinomial mixtures in this setting. We illustrate these ideas on the rod and frame data.

EXAMPLE 4 (Rod and frame continued from Example 3). The 83 subjects provide $S_1(6), \ldots, S_{83}(6)$, which are integers ranging from 0 to 8. As a preliminary check we compute $T = 295$ from (2) and refer it to a chi squared distribution with 82 degrees of freedom. The resulting p value is 0.000 and we easily reject $F_1^* = \cdots = F_{83}^*$. Hence, we fit binomial mixture models with $m = 8$ and unknown probabilities of success. Since $m = 8$, we can identify up to $K = 4$ components. We next compute the BIC from (6) for various values of K. The values for $K = 2, 3, 4$ are 404, 366 and 375. The minimum occurs at $K = 3$ and so we fit a three-component model to the data. We report in Table 3 the proportions λ_1, λ_2 and λ_3 along with $F_k(6)$, the binomial probability of getting the rod to the vertical position (success) for $k = 1, 2, 3$. In Table 4 we give the observed frequencies of $S_i(6)$ for $i = 1, \ldots, 83$. In addition, we provide the posterior probabilities for each of the possible values $0, 1, \ldots, 8$ along with their respective classification. This classification is related to the "soft" clustering described by Hastie, Tibshirani and Friedman (2001).

First consider Table 3. We have seen from BIC that the 83 subjects can be grouped into three components. The first component, which accounts for 52% of the population, has a success probability of about 0.52. This suggests that the subjects associated with this component are guessing when they try to make the rod vertical, since they have roughly a 50–50 chance of getting it correct. The second component subjects are even worse. They virtually never get it correct. The third component consists of subjects who know exactly how to do the task. Hence, the mixture model consists of two degenerate binomial components and a proper component. We summarize by saying that about 31% of the population know precisely how to do the task

TABLE 3
Parameter estimates for the rod and frame data

	First component	**Second component**	**Third component**
$\hat{\lambda}_k$	0.52	0.17	0.31
$\hat{F}_k(6)$	0.52	0.01	0.94

TABLE 4
Count data and posterior probabilities for the rod and frame data; ξ_k denotes the posterior probability of being in the kth component given the observed count

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Frequency	13	2	5	6	13	13	4	11	16
Rel. freq.	0.16	0.02	0.06	0.07	0.16	0.16	0.05	0.13	0.19
ξ_1	0.01	0.54	0.99	1.00	1.00	0.98	0.76	0.17	0.01
ξ_2	0.99	0.46	0.01	0.00	0.00	0.00	0.00	0.00	0.00
ξ_3	0.00	0.00	0.00	0.00	0.00	0.02	0.24	0.83	0.99
Classification	2	1	1	1	1	1	1	3	3

and we call these subjects field independent. The remaining 69% are field dependent and become confused by the tilt of the frame. The field-dependent population breaks into two further subgroups: one that never gets it correct and one that guesses.

In Table 4 we show the data and the posterior probabilities for assignment to components. For example, if a subject scores four correct out of eight trials, then we estimate that there is roughly a 100% chance that he or she came from the first component in which subjects guess. So far, the analysis is descriptive. We recommend the parametric bootstrap (Efron and Tibshirani, 1993) based on the estimated binomial mixture to estimate standard errors. We do not pursue standard errors further in this article.

We can also analyze the original absolute error data. We wish to estimate the three-component c.d.f.'s. Note that if we assume that a subject comes from, say F_1, then letting X denote the absolute error measurement, we estimate $P_1(X \le x) = F_1(x)$ by $n^{-1}\sum_j I(x_{j1} \le x) = n^{-1}S_1(x)$. However, we do not know from which component a randomly chosen subject is drawn. We estimate the $P_1(X \le x)$ by a weighted average of $S_1(x), \ldots, S_n(x)$, where the weights come from the posterior probabilities computed with the EM algorithm. See Appendix 4 for the formula and a discussion. In Figure 2 we show the estimates of the three component distributions. Note that one distribution has almost all of the distribution below 6 degrees, one is almost completely above 6 degrees and the guessers are more spread out.

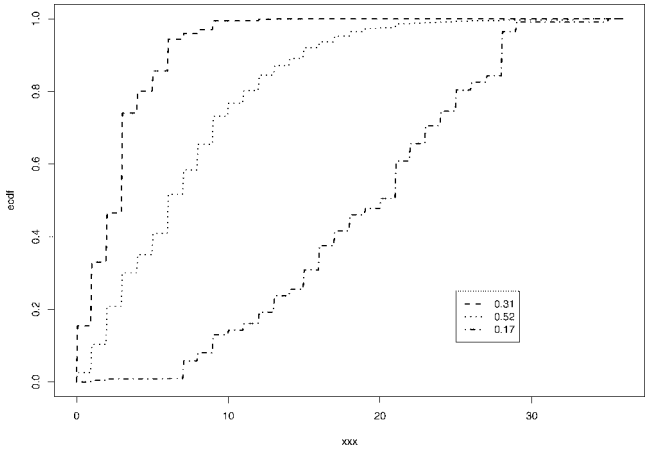

FIG. 2. *The estimated c.d.f.'s and their respective mixing proportions for the rod and frame data.*

We also compute the means and standard deviations from the estimated component c.d.f.'s. They are given in Table 5.

We now wish to return to the coal seam data and consider how mixture models can help us identify the possible different distributions underlying the data. Recall that we rejected the null hypothesis that $F_1 = \cdots = F_5$.

EXAMPLE 5 (Sulfur content of coal continued from Example 2). Now the five coal seams are the subjects. There are different numbers of observations per subject, but that does not present any additional difficulties in the EM algorithm; see Appendix 3 for a discussion of the algorithm. Since we have decided that there are subgroups, the question is how many. Figure 1 along with the multiple comparisons suggest that seams A and C are different from D and E, while B cannot be separated from either of the two groupings. We condition our analysis on the combined sample median 1.21 and use $S_1(1.21), \ldots, S_5(1.21)$, the essential ingredients for Mood's test statistic (2). Our goal is to

TABLE 5
Estimated means and standard deviations for the rod and frame components

	First component (guessers)	**Second component (poor)**	**Third component (excellent)**
$\hat{\mu}_k$	7.29	19.05	2.80
$\hat{\sigma}_k$	5.34	6.83	2.27

TABLE 6
Mixture model analysis for sulfur content; ξ_k denotes the posterior probability of being in the kth component given the observed count

	Seam				
	A	**B**	**C**	**D**	**E**
m_i	7	8	9	8	10
$S_i(1.21)$	2	4	1	6	8
ξ_1	0.11	0.76	0.00	1.00	1.00
ξ_2	0.89	0.24	1.00	0.00	0.00
Classification	2	1	2	1	1

identify the underlying distributions, display the distributions and assign the seams to the distributions. We first compute BIC (6) for the two-, three- and four-component models. We find values of 25.16, 28.37 and 31.59. This suggests that we have two groups and a two-component mixture underlying the combined data. In Table 6 we provide the posterior probabilities for the five seams.

Consistent with the multiple comparisons and Figure 1, A and C are grouped and D and E are grouped. In addition, B is assigned to the group with D and E on the basis of the posterior probability. Using the posterior probabilities from the EM algorithm as weights, we can estimate the component distributions as discussed earlier. The component c.d.f. estimate is described in Appendix 4. Finally, using the estimator for the component distributions, we can plot the two components; see Figure 3. We also compute the means and standard deviations of the two components in Figure 3. These are given in Table 7. Note that the component distributions for sulfur content have roughly the same standard

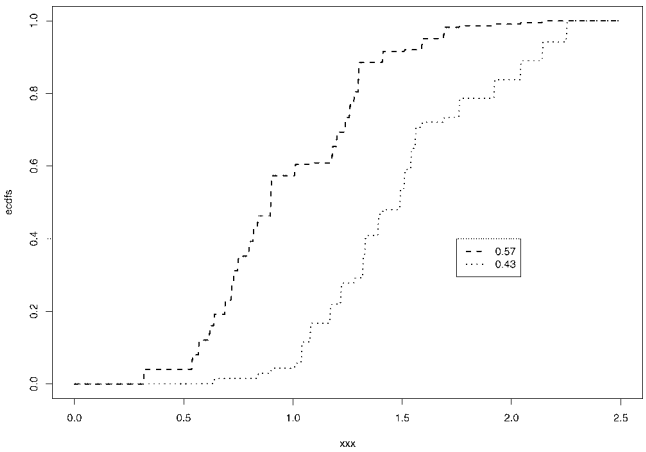

FIG. 3. *The estimated c.d.f.'s. and their respective mixing proportions for the coal data.*

TABLE 7
Estimated mixing proportions, means and standard deviations for the components in the coal data

	First component	**Second component**
$\hat{\lambda}_k$	0.57	0.43
$\hat{\mu}_k$	0.98	1.50
$\hat{\sigma}_k$	0.36	0.38

deviation and means separated by roughly 0.5% of sulfur content.

Thus we have a more complete followup analysis. We do not advocate mixture models as a replacement for multiple comparisons, only as a supplement which provides additional insight.

APPENDIX 1: ASYMPTOTIC DISTRIBUTION FOR SIGN STATISTICS

In this appendix we sketch the derivations for the asymptotic distributions that underlie formulas (2) and (3) in Section 2. See Section 2 for notation. We assume that $F_1 = F_2 = \cdots = F_n = F$, say, that F is continuous and the density function $f(\theta) > 0$, where θ is the true median of the common c.d.f. Furthermore, we assume that $M = \sum_{i=1}^{n} m_i$ and $M \to \infty$ in such a way that $m_i/M \to \pi_i$, where $0 < \pi_i < 1$ for $i = 1, 2, \ldots, n$. Let

$$\begin{aligned}\hat{T}_i &= \frac{2}{\sqrt{m_i}}\left(S_i(\hat{\theta}) - \frac{m_i}{2}\right)\\ &= \frac{2}{\sqrt{m_i}}\left(S_i(\theta) - \frac{m_i}{2}\right)\\ &\quad + 2f(\theta)\sqrt{m_i}(\hat{\theta} - \theta) + o_p(1),\end{aligned}$$

where $o_p(1)$ are terms that converge to zero in probability. This expansion was given by Hettmansperger and McKean (1998, Section 1.5.2). Let $T_i = 2/\sqrt{m_i} \cdot (S_i(\theta) - m_i/2)$. Then

$$\hat{T}_i = T_i + 2f(\theta)\sqrt{m_i}(\hat{\theta} - \theta) + o_p(1). \tag{7}$$

Furthermore, from the definition of the median $\hat{\theta}$ and applying the expansion again, we have

$$\begin{aligned}o_p(1) &= \frac{2}{\sqrt{M}}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(x_{ji} \le \hat{\theta}) - \frac{M}{2}\right]\\ &= \frac{2}{\sqrt{M}}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} I(x_{ji} \le \theta) - \frac{M}{2}\right]\\ &\quad + \sqrt{M}2f(\theta)(\hat{\theta} - \theta) + o_p(1).\end{aligned}$$

Hence

$$
\begin{aligned}
&2f(\theta)\sqrt{M}(\hat{\theta}-\theta)\\
&\quad=-\sum_{k=1}^{n}\sqrt{\frac{m_k}{M}}\frac{2}{\sqrt{m_k}}\left(S_k(\theta)-\frac{m_k}{2}\right)+o_p(1)\\
&\quad=-\sum_{k=1}^{n}\sqrt{\pi_k}T_k+o_p(1).
\end{aligned}
\tag{8}
$$

Substitute (8) into (7) and we have

$$
\hat{T}_i=T_i-\sqrt{\pi_i}\sum_{k=1}^{n}\sqrt{\pi_k}T_k+o_p(1). \tag{9}
$$

Let $\mathbf{T}=(T_1,T_2,\ldots,T_n)^T$. Then by the central limit theorem,

$$
\mathbf{T}\xrightarrow{D}\mathbf{Z}\sim\mathrm{MVN}(\mathbf{0},I),
$$

where $\mathrm{MVN}(\mathbf{0},I)$ means a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix the $n\times n$ identity matrix I. Let

$$
A=\begin{Bmatrix}
1-\pi_1 & -\sqrt{\pi_1\pi_2}\cdots & -\sqrt{\pi_1\pi_{n-1}} & -\sqrt{\pi_1\pi_n}\\
-\sqrt{\pi_2\pi_1} & 1-\pi_2\cdots & -\sqrt{\pi_2\pi_{n-1}} & -\sqrt{\pi_2\pi_n}\\
\vdots & \ddots\ \ldots & \ddots & \vdots\\
-\sqrt{\pi_n\pi_1} & -\sqrt{\pi_n\pi_2}\cdots & -\sqrt{\pi_n\pi_{n-1}} & 1-\pi_n
\end{Bmatrix}.
$$

Then

$$
\hat{\mathbf{T}}=A\mathbf{T}+o_p(1)\xrightarrow{D}A\mathbf{Z}\sim\mathrm{MVN}(\mathbf{0},AA^T).
$$

However, $AA^T=A^2=A$, idempotent, with rank $n-1$. Hence from (2),

$$
T=\hat{\mathbf{T}}^T\hat{\mathbf{T}}\xrightarrow{D}\mathbf{Z}^TA\mathbf{Z}\sim\chi^2(n-1).
$$

The asymptotic distribution for the multiple comparisons (3) also follows from the limiting multivariate normal distribution of $\hat{\mathbf{T}}$.

APPENDIX 2: BAYESIAN INFORMATION CRITERION

Employing finite mixture model methodology as a multiple comparisons diagnostic requires choosing the number of components that the mixture model contains. As we mentioned in Section 3, we use a penalized likelihood approach to this problem, namely, the Bayesian information criterion. See Schwarz (1978) for the seminal article of the BIC and McLachlan and Peel (2000) for a discussion of the BIC, as well as other penalized likelihood approaches applied to finite mixture models.

Let $\mathbf{\Psi}_K$ denote the parameter vector associated with the K-component mixture model given in (10) and let $l(\mathbf{\Psi}_K)$ be the log-likelihood of a sample from this model. The BIC for this situation is given by

$$
\mathrm{BIC}_K=-2l(\hat{\mathbf{\Psi}}_K)+d\ln n,
$$

where d is the dimension of the parameter space and $\hat{\mathbf{\Psi}}_K$ represents the maximum likelihood estimator (MLE) of $\mathbf{\Psi}_K$. Choose the value of K which minimizes BIC.

Our motivation for using a penalized form of the likelihood is due to the following reason. Notice that the parameter space $\mathbf{\Omega}_K$ for the K-component mixture model is a subset of $\mathbf{\Omega}_{K+1}$, the parameter space for the $(K+1)$-component mixture model. Therefore, the value of the likelihood at the MLE will not decrease as we increase the number of components in the mixture. The penalty term is designed to penalize the likelihood based on the complexity of the model. In the case of BIC, the penalty is primarily due to the dimension of the parameter space.

APPENDIX 3: EXPECTATION–MAXIMIZATION FOR BINOMIAL MIXTURES

Let $S_1,S_2,\ldots,S_n$ be a sample of observations from the K-component binomial mixture distribution of the form

$$
P(S_i=s_i)=\sum_{k=1}^{K}\lambda_k b(s_i;m_i,p_k), \tag{10}
$$

where $\sum_{k=1}^{K}\lambda_k=1$ and $b(s_i;m_i,p_k)$ is the binomial mass function with m_i trials and probability of success p_k. We describe an expectation–maximization algorithm for finding maximum likelihood estimators of the parameters in (10): $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_{K-1})$ and $\mathbf{p}=(p_1,\ldots,p_K)$. The standard reference on EM algorithms is Dempster, Laird and Rubin (1977); however, a more comprehensive account was given by McLachlan and Krishnan (1997).

An EM algorithm formulates the problem as a missing-data problem and then iterates between an expectation (E) step and a maximization (M) step until convergence is attained. These steps are outlined below for this problem.

A.3.1 Complete-Data Formulation

The missing data are defined as the multinomial indicator vectors of component membership, $\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_n$, where $\mathbf{Z}_i=(Z_{i1},\ldots,Z_{iK})^T$. If the observation S_i is actually from the kth component, then the

vector $\mathbf{Z}_i$ has a 1 in the kth position and 0's elsewhere. For example, if the first observation is from the third component, then $\mathbf{Z}_1 = (0, 0, 1, 0, \ldots, 0)^T$. The complete-data distribution (i.e., joint distribution of S_i and $\mathbf{Z}_i$) can be written as

$$f_c(s_i, \mathbf{z}_i) = \prod_{k=1}^{K} [\lambda_k b(s_i; m_i, p_k)]^{z_{ik}} \tag{11}$$

with complete-data log-likelihood

$$\begin{aligned} l_c(\boldsymbol{\Psi}) &= \ln \prod_{i=1}^{n} f_c(s_i, \mathbf{z}_i) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} [\ln \lambda_k + \ln b(s_i; m_i, p_k)], \end{aligned} \tag{12}$$

where $\boldsymbol{\Psi} = (\boldsymbol{\lambda}, \mathbf{p})^T$. Since the $\mathbf{Z}_i$ are unknown, we cannot maximize (11) directly. Instead, we replace $l_c(\boldsymbol{\Psi})$ by its conditional expectation in the E step below. We then maximize the conditional expectation in the M step. McLachlan and Peel (2000) showed that this iterative process leads to a sequence of estimates that does not decrease the original likelihood.

A.3.2 E Step

The $(r+1)$st E step of the algorithm requires taking the conditional expectation of $l_c(\boldsymbol{\Psi})$ given the observed data and the current value of the parameter, say $\boldsymbol{\Psi}^{(r)}$. In this case, the conditional expectation of (12) reduces to taking the expectation of Z_{ik} given s_i. Note that Z_{ik} given s_i is a Bernoulli random variable with conditional probability of success given by

$$\begin{aligned} \hat{Z}_{ik}^{(r)} &= E_{\boldsymbol{\Psi}^{(r)}}(Z_{ik} | \mathbf{s}_i) \\ &= P_{\boldsymbol{\Psi}^{(r)}}[Z_{ik} = 1 | \mathbf{s}_i] \\ &= \frac{\lambda_k^{(r)} b(s_i; m_i, p_k^{(r)})}{\sum_{k=1}^{K} \lambda_k^{(r)} b(s_i; m_i, p_k^{(r)})} \end{aligned} \tag{13}$$

from Bayes' theorem. Notice that $\hat{Z}_{ik}^{(r)}$ is the posterior probability that the ith sample member belongs to the kth component, given $\mathbf{s}_i$ and $\boldsymbol{\Psi}^{(r)}$, at the $(r+1)$st iteration of the algorithm.

A.3.3 M Step

The M step is so named because we are performing a maximization at this stage of the problem. We begin by defining the objective function

$$\begin{aligned} &Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(r)}) \\ &\quad = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{Z}_{ik}^{(r)} [\ln \lambda_k + \ln b(s_i; m_i, p_k)]. \end{aligned} \tag{14}$$

The $(r+1)$st iteration of the M step requires the maximization of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(r)})$ to obtain updated estimates of the parameter vector $\boldsymbol{\Psi}^{(r+1)}$. Upon differentiating and simplifying the resulting expressions, we have $\lambda_k^{(r+1)} = \sum_i \hat{Z}_{ik}^{(r)} / n$. In other words, each observation contributes its respective posterior probability of being in the kth component to the estimate of the probability of membership in this component. In addition, it can be shown that the updated parameter estimates of the binomial probabilities for all k are given by

$$p_k^{(r+1)} = \frac{\sum_{i=1}^{n} \hat{Z}_{ik}^{(r)} S_i}{\sum_{i=1}^{n} \hat{Z}_{ik}^{(r)} m_i}.$$

A.3.4 Convergence and Starting Values

For the examples given in this paper, we use a relative difference stopping criterion to assess convergence of an EM algorithm. This is based on the absolute relative difference between parameter estimates at successive iterations of the algorithm. If we let D be the dimension of the parameter vector $\boldsymbol{\Psi}$, then we suggest stopping the algorithm when

$$\frac{|\Psi_d^{(k)} - \Psi_d^{(k+1)}|}{\Psi_d^{(k)}} < \varepsilon \tag{15}$$

for $d = 1, 2, \ldots, D$, given some small, prespecified value of ε (e.g., $\varepsilon = 10\text{e}{-6}$). This stopping rule was discussed by Schafer (1997).

We close this section by noting that this algorithm should converge to at least a local maximum, not necessarily a global maximum. Therefore, we recommend starting the algorithm at several random initial values $\boldsymbol{\Psi}^{(0)}$ to increase the chance that a global maximum is indeed found; see McLachlan and Peel (2000).

APPENDIX 4: EMPIRICAL COMPOUND DISTRIBUTION FUNCTION

As a result of fitting an EM algorithm to a mixture model, the posterior probabilities of being in the kth component ($k = 1, 2, \ldots, K$) are given for each observation. This leads to an empirical cumulative distribution function-like estimator in the setting described in Section 3. That is, suppose we are given a sample of

sign statistics $S_1, S_2, \ldots, S_n$. If we regard these observations as arising from a finite mixture model and fit an EM algorithm, then we can estimate the kth distribution function using

$$\hat{F}_k(x) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\hat{Z}_{ik}^{(\infty)} I(x_{ji} \le x)}{\sum_{i=1}^{n}\hat{Z}_{ik}^{(\infty)} m_i} = \frac{\sum_{i=1}^{n}\hat{Z}_{ik}^{(\infty)} S_i(x)}{\sum_{i=1}^{n}\hat{Z}_{ik}^{(\infty)} m_i} \tag{16}$$

for $x \in \mathbb{R}$, where $\hat{Z}_{ik}^{(\infty)}$ represents the posterior probability (at convergence of the algorithm) of being in the kth component given that we observed S_i.

We can use the function defined in (16) to estimate the mean and the standard deviation of the kth component exactly as one might use the usual empirical c.d.f. That is, the mean and standard deviation of the kth component are estimated using

$$\hat{\mu}_k = \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ik} x_{ji},$$

$$\hat{\sigma}_k^2 = \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ik} x_{ji}^2 - \left(\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ik} x_{ji}\right)^2,$$

respectively, where $w_{ik} = \hat{Z}_{ik}^{(\infty)} / \sum_{i=1}^{n}\hat{Z}_{ik}^{(r)} m_i$. More generally, the qth moment of the kth distribution can be estimated using

$$\hat{E}_q = \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ik} x_{ji}^q.$$

ACKNOWLEDGMENT

This research was supported in part by National Science Foundation Grant SES-01-15619.

REFERENCES

CRUZ-MEDINA, I. R., HETTMANSPERGER, T. P. and THOMAS, H. (2004). Semiparametric mixture models and repeated measures. The multinomial cut point model. *Appl. Statist.* **53** 463–474.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

ELMORE, R. T. (2003). Semiparametric analysis of finite mixture models with repeated measures. Ph.D. dissertation, Pennsylvania State Univ.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2001). *The Elements of Statistical Learning*: *Data Mining, Inference, and Prediction*. Springer, New York.

HETTMANSPERGER, T. P. and MCKEAN, J. W. (1998). *Robust Nonparametric Statistical Methods*. Arnold, London.

HETTMANSPERGER, T. P. and THOMAS, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 811–825.

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

MCLACHLAN, G. J. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York.

MOOD, A. M. (1950). *Introduction to the Theory of Statistics*. McGraw–Hill, New York.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.