Far East Journal of Theoretical Statistics, 23(2), 241-259. (2007)

MIXTURE DISTRIBUTIONS AND A TEST STATISTIC

K.O. BOWMAN[1] and L.R. SHENTON[2]

[1]Computational Sciences and Engineering Division
Oak Ridge National Laboratory, P.O.Box 2008, 4500N, MS-6191
Oak Ridge, TN 37831-6191, U.S.A.
e-mail: bowmanko@ornl.gov
[2]Department of Statistics, University of Georgia
Athens, Georgia 30602, U.S.A.

October 31, 2007

Abstract

It is assumed that a moderately large sample is available from a finite mixture distribution, the components being discrete distributions for which moments exist. For s components there will be at least $2s-1$ parameters. Psi function differences create a random variable which itself leads to a test function statistic, this statistic in general being asymptotically normal. Eleven percentage points to the distribution of the standardized test function (location and scale free) are given.

2000 Mathematics Subject Classification: 62E20.
Key words and phrases: goodness of fits test, logarithmic derivatives, maximum likelihood estimator, moment estimator, Psi-function.

1 Introduction

We consider a mixture of discrete distributions defined by s components, namely

$$Pr(X = x) = \pi_1 C_1(x; \underline{p}_1) + \pi_2 C_2(x; \underline{p}_2) + \cdots + \pi_s C_s(x; \underline{p}_s)$$

where $\underline{p}_\lambda$ defines the related parameter space. Each component consists of a denumerable set of point masses with support from the set of non-negative integers $0, 1, \cdots$. If the components relate to the same structures, then we call the mixture homogeneous (all binomial for example). The mean of the mixture is

$$\sum_{x=0}^{\infty} xPr(X = x) = \pi_1 C_{11} + \pi_2 C_{21} + \cdots + \pi_s C_{s1} = \mu_1'.$$

In a previous paper (Bowman and Shenton, 2007) we have studied the Psi-function test statistic $T_1(\underline{n}; \lambda)$ when there is one component.

Upper and lower bounds are set up for the S_1's random variable distribution. Since means are involved there is little surprise in the appearance of near normality, measured by $\sqrt{\beta_1} = \mu_3/\sigma^3$, and β_2 measured by μ_4/σ^4 these being moment ratio. Moments are supposed to exist.

The distribution of the Psi-test statistic $T_1(\underline{n}; \lambda)$ is adumbrated in the bounds; more specific information is available from Pearson percentage points based on the assumption that a Pearson curve (Pearson, 1902) provides a reasonable approximation. The Pearson percentage point approach uses the first four central moments, and the paper by Bowman and Shenton (1979a, 1979b) provide an error assessment. It is readily set up for computing. For the bounds algebraic forms as series are stated.

Our main objective is to test the hypothesis that a given random sample stems from a defined mixture distribution. The basic distribution is discrete, moments assumed to exist, there being support from the class of non-negative integers. The test statistic is a vector of $2s - 1$ parameters and is based on the Psi functions. A sample realization of the test function turns out to be a sample mean, which gives equal weight to all values of the vectors. It is a new goodness of fit without a problem relating to outliers.

2 Bounds for S_1 variables

2.1 Two component Poisson model

The S_1 functions are:

$$\begin{aligned} S_1(x; \theta_1) &= \psi(x + \theta_1) - \psi(\theta_1) \\ &= \frac{1}{\theta_1} + \frac{1}{\theta_1 + 1} + \cdots + \frac{1}{\theta_1 + x - 1}, \qquad (\theta_1 > 0, x = 0, 1, \cdots) \end{aligned}$$

$$S_1(x;\theta_2) = \frac{1}{\theta_2} + \frac{1}{\theta_2+1} + \cdots + \frac{1}{\theta_2+x-1}, \qquad (\theta_2 > 0)$$

and

$$S_1(x;\pi) = \frac{1}{\pi} + \frac{1}{\pi+1} + \cdots + \frac{1}{\pi+x-1}.$$

These are all examples of

$$S_1(x;\alpha) = \frac{1}{\alpha} + \frac{1}{\alpha+1} + \cdots + \frac{1}{\alpha+x-1}, \qquad (x = 0, 1, \cdots)$$

with $\alpha \in \Re^+$.

An upper bound for $S_1(x;\alpha)$ is

$$E\left(\frac{x}{\alpha}\right) = \frac{\pi_1\theta_1 + \pi_2\theta_2}{\alpha} \qquad (\alpha \in \Re^+, \pi_2 = 1 - \pi_1).$$

A lower bound is

$$E\left(\frac{x}{x+\alpha-1}\right) = \sum_{x=1}^{\infty} \frac{x}{x+\alpha-1}\left(\pi_1\frac{e^{-\theta_1}\theta_1^x}{x!} + \pi_2\frac{e^{-\theta_2}\theta_2^x}{x!}\right) \qquad (\theta_1, \theta_2 > 0, \alpha > 0).$$

The right hand-side is a Stieltjies integral and is a generalization of the Stieltjies transform

$$\sum_{x=0}^{\infty} \frac{e^{-\theta}\theta^x/x!}{x+z} \qquad (\Re(z) > 0).$$

See Bowman and Shenton (1989, p.30-32).

A form of the bound follows from writing $\frac{x}{x+\alpha-1}$ or $1 + \frac{1-\alpha}{x+\alpha-1}$ but this modification only hold if $\alpha > 1$.

2.2 Three component Poisson model

The parameters are θ_1, θ_2, θ_3, π_1, π_2 with $\pi_1 + \pi_2 + \pi_3 = 1$, and $0 < \pi_1 < 1$, $0 < \pi_2 < 1$. The probability function for the mixture is

$$Pr(X = x) = \sum_{r=1}^{3} \pi_r Pr(x;\theta_r),$$

where $Pr(x;\theta_r) = \frac{e^{-\theta_r}\theta_r^x}{x!}$, $(\theta_r > 0, \sum \pi_r = 1, 0 < \pi_r < 1, r = 1, 2, 3)$. An upper bound for $S_1(x;\alpha)$ is

$$\frac{\pi_1\theta_1 + \pi_2\theta_2 + \pi_3\theta_3}{\alpha}$$

where α is replace by any one of the five parameters defining the Poisson mixture.

For the lower bounds we compute

$$\sum_{x=1}^{\infty} \frac{x}{x+\alpha-1}\left(\sum_{r=1}^{3} \pi_r \frac{e^{-\theta_r}\theta_r^x}{x!}\right).$$

Lastly

$$\begin{aligned}
E[S_1(x;\alpha)] &= \sum_{x=1}^{\infty}\left(\frac{1}{\alpha}+\frac{1}{\alpha+1}+\cdots+\frac{1}{\alpha+x-1}\right)\left\{\sum_{r=1}^{s}\pi_r P_r(x;\theta_r)\right\}\\
&= \sum_{x=1}^{\infty}\int_0^1 (t^{\alpha-1}+t^{\alpha}+\cdots+t^{x-1})\left\{\sum_{r=1}^{s}\pi_r P_r(x;\theta_r)\right\}dt\\
&= \sum_{x=1}^{\infty}\int_0^1 \frac{t^{\alpha-1}(1-t^x)}{1-t}\left\{\sum_{r=1}^{s}\pi_r P_r(x;\theta_r)\right\}dt\\
&= \sum_{x=0}^{\infty}\int_0^1 \sum_{r=1}^{s} t^{\alpha-1}\left\{\frac{\pi_r(1-e^{\theta_r(t-1)})}{1-t}\right\}dt.
\end{aligned}$$

The coefficient of π_r is

$$\int_0^1 t^{\alpha-1}\left\{\sum_{m=1}^{\infty}\frac{(-1)^{m-1}\theta_r^m(1-t)^{m-1}}{m!}\right\}dt = \frac{\theta_r}{\alpha}-\frac{\theta_r^2}{2\alpha(\alpha+1)}+\frac{\theta_r^3}{3\alpha(\alpha+1)(\alpha+2)}-\cdots,$$

so

$$E[S_1(x;\alpha)] = \sum_{r=1}^{s}\pi_r\left\{\frac{\theta_r}{\alpha}-\frac{\theta_r^2}{2\alpha(\alpha+1)}+\frac{\theta_r^3}{3\alpha(\alpha+1)(\alpha+2)}-\cdots\right\},$$

a series of alternating terms; note that the general form (sign not important here) is

$$\frac{1}{r}\left(\frac{\theta_r}{\alpha}\right)\left(\frac{\theta_r}{\alpha+1}\right)\left(\frac{\theta_r}{\alpha+2}\right)\cdots\left(\frac{\theta_r}{\alpha+r-1}\right)$$

and the ratio of successive terms is

$$\left(\frac{r-1}{r}\right)\frac{\theta_r}{\alpha+r-1}\to 0,\quad r\to\infty$$

and so the series converges. See T. J. I. A Bromwich (1926, p.55).

2.3 A four component binomial model

The probability function is

$$Pr(X=x) = \sum_{r=1}^{4}\pi_r\binom{n}{x}p_r^x q_r^{n-x},\quad (x=0,1,\cdots,n)$$

with $0<p_r<1$, $q_r=1-p_r$, $r=1,2,3,4$, and the proportions less than unity but summing to unity. For the bounds we consider

$$\begin{aligned}
S_1(x;p_\lambda) &= \frac{1}{p_\lambda}+\frac{1}{p_\lambda+1}+\cdots+\frac{1}{x+p_\lambda-1},\quad (\lambda=1,2,3,4)\\
S_1(x;\pi_\lambda) &= \frac{1}{\pi_\lambda}+\frac{1}{\pi_\lambda+1}+\cdots+\frac{1}{x+\pi_\lambda-1},\quad (\lambda=1,2,3)\\
S_1(x;n^+) &= \frac{1}{n}+\frac{1}{n+1}+\cdots+\frac{1}{x+n-1},\quad (n=1,2,\cdots)\\
S_1(x;n^-) &= \frac{1}{n}+\frac{1}{n-1}+\cdots+\frac{1}{n-x+1},\quad (n=1,2,\cdots)
\end{aligned}$$

each S_1 being zero with $x = 0$. Upper bounds are

$$E[S_1(x;\alpha)] = \frac{n(\pi_1 p_1 + \pi_2 p_2 + \pi_3 p_3 + \pi_4 p_4)}{\alpha}$$

for the various values of α, excepting

$$E[S_1(x;n^-)] = \sum_{x=1}^{n} \frac{x}{n-x+1} \sum_{r=1}^{4} \pi_r \binom{n}{x} p_r^x q_r^{n-x}.$$

The coefficient of π_λ is

$$\begin{aligned}\sum_{x=1}^{n} \left(-1 + \frac{n+1}{n-x+1}\right) \binom{n}{x} p_\lambda^x q_\lambda^{n-x} &= -(1-q_\lambda^n) + (n+1) \sum_{x=1}^{n} \int_0^1 t^{n-x} \binom{n}{x} p_\lambda^x q_\lambda^{n-x} dt \\ &= -(1-q_\lambda^n) + (n+1) \left\{ -\frac{q^n}{n+1} + \int_0^1 (p_\lambda + q_\lambda t)^n dt \right\} \\ &= -1 + \frac{1-p_\lambda^{n+1}}{1-p_\lambda} = p_\lambda + p_\lambda^2 + \cdots + p_\lambda^n.\end{aligned}$$

Hence, for the general case of s binomial distributions

$$\sum_{r=1}^{s} \pi_r p_r < E[S_1(x;n^+)] < \sum_{r=1}^{s} \pi_r \sum_{t=1}^{n} p_r^t \qquad (s = 1, 2, \cdots; n = 1, 2, \cdots; 0 < p_r < 1).$$

By a modification of the above, we find

$$E[S_1(x;n^-)] = \sum_{r=1}^{s} \pi_r \left(p_r + \frac{p_r^2}{2} + \frac{p_r^3}{3} + \cdots + \frac{p_r^n}{n} \right)$$

providing some interesting inequality in algebra. The architectural symmetry displayed is surely surprising.

2.4 An additional binomial example

It is interesting to note the binomial case

$$E[S_1(x;np)] = E\left\{ \frac{1}{np} + \frac{1}{np+1} + \cdots + \frac{1}{x+np-1} \right\}$$

for a single mixture component; clearly

$$E[S_1(x;np)] < \frac{E(x)}{np} = 1.$$

For a mixture of s binomial component, this becomes

$$E[S_1(x;p_\lambda)] < \sum_{r=1}^{s} \pi_r p_r / p_\lambda.$$

For the lower bound of a mixture of s binomial components we have the expression

$$\sum_{x=1}^{n} \frac{x}{x+np_\lambda-1} \sum_{r=1}^{s} \pi_r \binom{n}{x} p_r^x q_r^{n-x}, \qquad (\lambda = 1, 2, \cdots, s).$$

2.5 Negative binomial mixture model

A single component has probability generating function $(p+1-pt)^{-k}$, $p>0$, $k>0$. For a mixture of s components

$$\sum_{x=1}^{\infty}\frac{x}{x+k_\lambda-1}P_s(x) < E[S_1(x;k_\lambda)] < \frac{\pi_1k_1p_1+\pi_2k_2p_2+\cdots+\pi_sk_sp_s}{k_\lambda}$$
$$\sum_{x=1}^{\infty}\frac{x}{x+p_\lambda-1}P_s(x) < E[S_1(x;p_\lambda)] < \frac{\pi_1k_1p_1+\pi_2k_2p_2+\cdots+\pi_sk_sp_s}{p_\lambda}$$

where $P_s(x)=\sum_{r=1}^{s}\pi_r(p_r+1)^{-k_r}\frac{\Gamma(k_r+x)}{x!\Gamma(k_r)}$. Moreover

$$E[S_1(x;k_\lambda)]=\sum_{x=1}^{\infty}\int_0^1\frac{t^{k_\lambda-1}(1-t^x)}{1-t}P_s(x)dt$$

and the coefficient of π_r is

$$\sum_{x=1}^{\infty}\int_0^1\frac{t^{k_\lambda-1}(1-t^x)}{1-t}P_s(x)dt=\int_0^1\frac{t^{k_\lambda-1}[1-(p_r+1-p_rt)^{-k_\lambda}]}{1-t}dt.$$

If there is only one component then this reduces to $\ln(p+1)$, as mentioned in Bowman and Shenton (2007). Thus

$$E[S_1(x;k_\lambda)]=\sum_{r=1}^{s}\pi_r\int_0^1\frac{t^{k_\lambda-1}[1-(p_r+1-p_rt)^{-k_\lambda}]}{1-t}dt.$$

2.6 Poisson-Poisson distribution

Here the probability function is

$$Pr(X=x)=\frac{\theta(\theta+\lambda x)^{x-1}e^{-\theta-\lambda x}}{x!}\qquad(\theta>0,0<\lambda<1,x=0,1,\cdots),$$

and is due to P. C. Consul (1989).

The mean is $\theta/(1-\lambda)$. Percentage points and bounds for Poisson-Poisson mixtures are readily set up using the descriptions given in previous paragraphs.

3 Computing bounds for $S_1(x;\alpha)$

3.1 Bounds

$$S_1(x;\alpha)=\frac{1}{\alpha}+\frac{1}{\alpha+1}+\cdots+\frac{1}{\alpha+x-1}\qquad(\alpha>0)$$

s component $\psi_1(x)$, $\psi_2(x),\cdots,$ $\psi_s(x)$ with mean values $\bar{\psi}_1$, $\bar{\psi}_2,\cdots,$ $\bar{\psi}_s$.

$$E[S_1(x;\alpha)] < \frac{\pi_1\bar{\psi}_1 + \pi_2\bar{\psi}_2 + \cdots + \pi_2\bar{\psi}_s}{\alpha},$$
$$E[S_1(x;\alpha)] > \sum_{x=1}^{\infty} \frac{x}{x+\alpha-1}(\pi_1\psi_1(x) + \cdots + \pi_s\psi_s(x)),$$
$$E[S_2(x;\alpha)] = \sum_{x=1}^{\infty} \left(\frac{1}{\alpha} + \frac{1}{\alpha+1} + \cdots + \frac{1}{\alpha+x-1}\right)(\pi_1\psi_1(x) + \cdots + \pi_s\psi_s(x)).$$

For the 4 component binomial mixture model, parameters are
$\alpha \rightarrow p_1,\ p_2,\ p_3,\ p_4,\ \pi_1,\ \pi_2,\ \pi_3$.

3.2 Binomial case

For the binomial there are two more cases, the first is

$$S_1(x;n^+) = \frac{1}{n} + \frac{1}{n+1} + \cdots + \frac{1}{n+x-1}. \quad (n = 1, 2, \cdots)$$

The second is the case $\alpha = n^-$.

$$S_1(x;n^-) = \frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + \frac{1}{n-x+1}, \quad (n = 1, 2, \cdots)$$

$$E[S_1(x,n^-)] = \sum_{x=1}^{n} \left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-x+1}\right)(\pi_1\psi_1(x) + \cdots + \pi_s\psi_s(x)).$$

This does not subscribe to an α form.

A Pearson Type I (Beta) appears to be the appropriate distribution to approximate the distribution of T_1 the probability function being $C(x-a)^{p-1}(b-x)^{q-1}$. The lower and upper bound determine a and b. In standard measure

$$p = m_1'(m_1' - m_2)/m_2, \quad q = (1 - m_1')(m_1' - m_2)/m_2.$$

4 Illustrative examples and percentage points for the approximate distribution of the random variable S_1

4.1 Pearson percentage points for S_1

As mentioned in the introduction we have a computer program to assess the percentage points (eleven basic percentages) of a Pearson curve (Pearson, 1902) based on a first order differential equation; some readers regard the Pearson curve as a manifold. As approximating distributions, given the existence of μ_1', σ, and moment ratios $\sqrt{\beta_1} = \mu_3/\sigma^3$, and $\beta_2 = \mu_4/\sigma^4$, the system has had wide applications, in part due to the papers (Bowman and

Shenton, 1979a, 1979b). The reader may refer to Elderton and Johnson (1969). For the two component Poisson model, percentage points are given in Table 3; the upper bound for S_1 is computed for comparison, quite satisfactorily.

4.2 A two component Poisson model

The data relates to the number of women over 80 years of age reported in the Times (London), and is taken from Hasselblad (1969). For the three parameters, moment estimators are $\theta_1 = 1.10209$, $\theta_2 = 2.58164$, and $\pi = 0.28705$,

Table 1 Data of two component Poisson mixture

Observed death count	Observed frequency	Single Poisson Expected frequency	Mixture Expected frequency
0	162	126.79	163.62
1	267	273.47	267.78
2	271	294.92	260.46
3	185	212.04	192.84
4	111	114.34	115.83
5	61	49.32	57.91
6	27	17.73	24.57
7	8	5.46	9.00
8	3	1.47	2.89
9	1	0.35	0.83
		$\chi^2_6 = 26.97$	$\chi^2_4 = 1.52$

The second column refers to the data frequency, the third to a Poisson model, and the forth to a two component Poisson model, the latter clearly indicating a good fit. The new test functions relating to θ_1, θ_2, π will indicate whether the assumption regarding the basic model is justified. Note that sample size $N = 1096$. Our computer output follows.

Table 2 Values of test functions

	θ_1	θ_2	π
Parameter value	1.10209	2.58164	0.28705
$E(S_1)$	1.25050	0.63305	3.68252
σ	0.05035	0.02578	0.14981
$\sqrt{\beta_1}$	-0.03500	-0.03053	-0.03858
β_2	2.99850	2.99842	2.99854
Test value $\bar{T}_1$	0.02922	0.02362	0.04025
Upper bound	1.957	0.835	7.514

The skewness and kurtosis, moment ratio estimators suggest normality or Pearson Type I distribution. The values of $\bar{T}_1(\theta_1)$, $\bar{T}_1(\theta_2)$, $\bar{T}_1(\pi)$ are approximately $3/10^2$, $2/10^2$, and $4/10^2$ respectively. These being scale and location free indicate acceptance of the basic model structure. Note that sample size is scarcely an issue.

Table 3 Percentage points of S_1

%	1	2.5	5	10	25	50	75	90	95	97.5	99
$S_1(x;\theta_1)$	1.135	1.153	1.168	1.186	1.216	1.250	1.284	1.315	1.334	1.350	1.369
$S_1(x;\theta_2)$	0.574	0.583	0.591	0.600	0.616	0.633	0.650	0.666	0.676	0.684	0.694
$S_1(x;\pi)$	3.338	3.392	3.437	3.491	3.581	3.682	3.783	3.875	3.931	3.979	4.036

4.3 A three component Poisson mixture

The parameters of the mixture are θ_1, θ_2, θ_3, and π_1, π_2. Everitt and Hand (1981) produced a random sample of $N = 500$ from the mixture $\theta_1 = 0.5$, $\theta_2 = 3.0$, $\theta_3 = 6.0$, $\pi_1 = 0.3$, $\pi_2 = 0.3$.

$$Pr(X = x) = \sum_{r=1}^{3} \pi_r \frac{e^{-\theta_r}\theta_r^x}{x!} \qquad (\theta_r > 0, \sum \pi_r = 1, 0 < \pi_r < 1, r = 1, 2, 3).$$

The example is given to show whether the five test functions will indicate rejection of base hypotheses. Here is the computer output.

Table 4 Three component Poisson Model ($N = 500$)

	θ_1	θ_2	θ_3	π_1	π_2
Parameter value	0.5000	3.0000	6.0000	0.3000	0.3000
$E(S_1)$	6.9990	0.7660	0.3812	6.4372	2.3131
σ	0.2277	0.0651	0.0385	0.3342	0.3342
$\sqrt{\beta_1}$	-0.0371	-0.0314	-0.0286	-0.0377	-0.0377
β_2	2.9953	2.9952	2.9952	2.9952	2.9952
Test values $\bar{T}_1$	0.4338	0.2553	0.2049	0.4787	0.4787
ml estimators	0.1430	2.9210	7.1650	0.1580	0.6040
$E(S_1)$	6.9990	0.7660	0.3812	6.4372	2.3131
σ	0.6110	0.0661	0.0335	0.5609	0.1973
$\sqrt{\beta_1}$	-0.0402	-0.0324	-0.0275	-0.0403	-0.0391
β_2	2.9954	2.9953	2.9952	2.9954	2.9954
Test values $\bar{T}_1$	0.1574	0.0320	0.0322	0.1526	0.0771

Comment: From the moment ratios $\sqrt{\beta_1}$, β_2, we may assume the approximating distribution to be near normal. The standard deviates associated with the maximum likelihood

estimators $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, $\hat{\pi}_1$, $\hat{\pi}_2$ are no longer small and acceptable. See the list line in the computer output. The rejection may be traced to

(i) sample size not large enough,

(ii) estimation procedures,

(iii) error in assumption regarding the basic model.

(iv) There is also a discrepancy it should noted. By random generation (Everitt and Hand, 1981) the parameters are $\theta_1 = 0.5$, $\theta_2 = 3.0$, $\theta_3 = 6.0$, $\pi_1 = 0.3$, and $\pi_2 = 0.3$. Maximum likelihood estimators are $\hat{\theta}_1 = 0.143$, $\hat{\theta}_2 = 2.921$, $\hat{\theta}_3 = 7.165$, $\hat{\pi}_1 = 0.158$, and $\hat{\pi}_2 == .604$. In particular, for θ_1, π_1 and π_2 can not be ignored, so the sample is not a good representative of the population assumed.

A graph of the population distribution against that estimated by maximum likelihood is shown in Fig. 1. There is excellent agreement for $x > 4$, but disagreement for $x < 4$, particularly for $2 < x < 4$. Can this be explained? Note the proportion in the population is 0.3; the corresponding maximum likelihood proportion is 0.15. Again 0.3 for population, 0.6 by maximum likelihood. Again in the population $\theta_1 = 0.5$, whereas the maximum likelihood value is 0.14. Taking into account these discrepancies supports the hypothesis that the maximum likelihood fit is only moderately acceptable. From Table 4 and values of the skewness ($\sqrt{\beta_1}$) and kurtosis (β_2) for the standardized test statistic $\bar{T}_1$, approximate normality is suggested, the highest value being 0.16 for $\hat{\theta}_1$. Note that for the 2 component Poisson mixture, corresponding values of $\bar{T}$ are one tenth of this value.

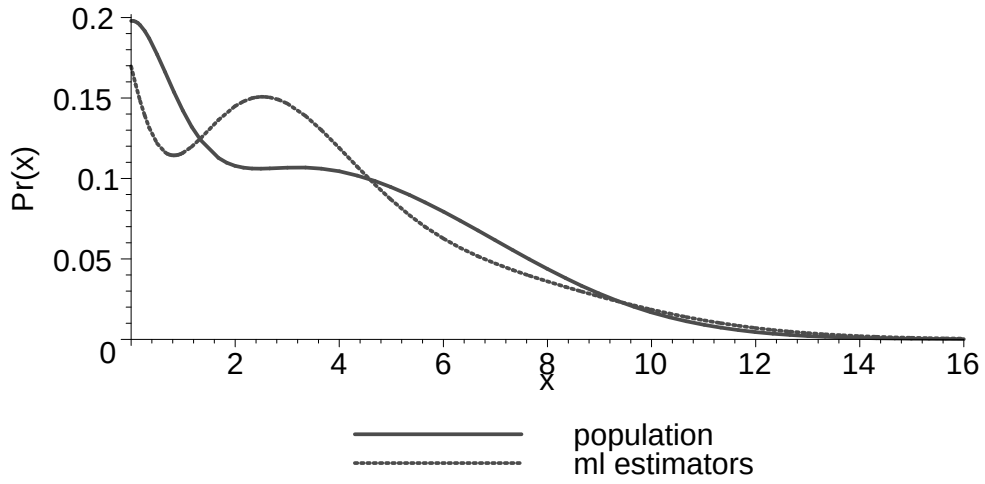

Figure 1: 3 component mixture of Poisson model

4.4 A four component binomial model

Everitt and Hand (1981, pp. 92-93) constructed a 4 component binomial mixture for a sample of $N = 200$ from the mixture $\theta_1 = 0.1$, $\theta_2 = 0.2$, $\theta_3 = 0.6$, $\theta_4 = 0.9$, $\pi_1 = 0.2$, $\pi_2 = 0.2$, $\pi_3 = 0.2$, and $n = 30$.

The 200 sample values were

Table 5 Four component binomial mixture $N = 200$

7	25	27	27	24	25	12	6	29	5
7	29	6	2	24	15	15	27	4	5
26	28	5	16	29	7	28	28	27	29
10	28	29	30	1	30	5	28	7	2
3	28	28	22	21	30	27	4	6	25
2	28	16	7	2	2	17	3	3	3
29	27	4	3	1	23	23	28	28	27
25	18	21	28	26	22	28	16	26	29
19	20	24	17	3	17	26	18	6	28
28	28	29	6	27	4	27	20	28	5
30	29	29	27	26	3	10	28	22	7
29	28	24	28	12	22	2	27	5	28
28	14	13	27	22	3	7	7	28	3
8	2	27	5	4	29	29	2	1	3
6	1	3	6	21	17	6	10	29	4
1	25	0	7	28	29	28	28	29	1
28	28	26	28	24	8	19	28	25	3
17	19	10	0	9	15	15	26	20	26
6	27	22	2	27	14	14	4	9	28
27	28	28	3	28	28	27	24	27	27

Table 6 Frequency distribution for Table 4.

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Freq.	2	6	9	13	7	7	9	9	2	2	4	0	2	7	3	4
x		16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Freq.		3	5	2	3	3	3	6	2	6	6	8	19	34	16	4

From Table 6 there is evidence of 3 peaks in the data. Computer output now follows in three parts displaying the values of $\bar{T}_1$ (test statistics being location and scale free) for the seven parameters $(\theta_1, \theta_2, \theta_3, \theta_4; \pi_1, \pi_2, \pi_3)$. Using

(i) the base set of these from the distribution sampled,

(ii) maximum likelihood estimators,

(iii) moment estimators.

Table 7 four component Model ($n = 30$, $N = 200$)

	θ_1	θ_2	θ_3	θ_4	π_1	π_2	π_3
Parameter value	0.1000	0.2000	0.6000	0.9000	0.2000	0.2000	0.2000
$E(S_1)$	12.7008	7.6278	3.9726	3.2339	7.6278	7.6278	7.6278
σ	3.4557	2.0526	1.0521	0.8525	2.0526	2.0526	2.0526
$\sqrt{\beta_1}$	-0.0368	-0.0372	-0.0372	-0.0370	-0.0372	-0.0372	-0.0372
β_2	2.9860	2.9860	2.9861	2.9861	2.9860	2.9860	2.9860
Test values $\bar{T}_1$	0.0284	0.0503	0.0980	0.1182	0.0503	0.0503	0.0503
ml estimators	0.0909	0.2197	0.6052	0.9123	0.1832	0.1780	0.1756
$E(S_1)$	13.7827	7.2495	4.0443	3.3004	8.1879	8.3500	8.4303
σ	3.9113	2.0219	1.1053	0.8952	2.2923	2.3391	2.3623
$\sqrt{\beta_1}$	-0.0353	-0.0359	-0.0363	-0.0363	-0.0358	-0.0358	-0.0358
β_2	2.9859	2.9860	2.9860	2.9860	2.9860	2.9860	2.9860
Test values $\bar{T}_1$	0.0054	0.0083	0.0123	0.0139	0.0076	0.0075	0.0074
Moment est.	0.0670	0.1777	0.5694	0.9090	0.0959	0.2488	0.1687
$E(S_1)$	17.6784	8.3631	4.1832	3.3168	13.2003	6.6982	8.6698
σ	5.0674	2.3522	1.1472	0.9012	3.7603	1.8699	2.4413
$\sqrt{\beta_1}$	-0.0349	-0.0357	-0.0362	-0.0363	-0.0352	-0.0359	-0.0356
β_2	2.9859	2.9860	2.9860	2.9860	2.9859	2.9860	2.9860
Test values $\bar{T}_1$	0.0077	0.0060	0.0032	0.0021	0.0072	0.0052	0.0061

Remarks: (i) Notice that the skewness values are small, and that $\beta_2(\mu_4/\mu_2^2)$ is near to 3 suggesting normality; (ii) the standard deviates $\bar{T}_1$ are less than 1/10 for maximum likelihood and moment approach, but increase when the theoretical values are used; and (iii) the test function results give some support for accepting the validity of the basic assumptions that sampling from a binomial mixtures.

Note that closeness of the means of the component is an obvious warning of estimation procedure problems. In this connection, Matusita introduced the concept of distance between discrete distributions (Matusita, 1955). For example, if $p_q(x)$ and $p_s(x)$ are two discrete distributions then the distance between them is defined as

$$||D||_2 = [\sum(\sqrt{p_q(x)} - \sqrt{p_s(x)})^2]^{\frac{1}{2}}.$$

The affinity is $\left(\sum \sqrt{p_q(x)}\sqrt{p_s(x)}\right)^{\frac{1}{2}}$.

For Poisson distributions, the affinity is

$$e^{\frac{1}{2}(\sqrt{\theta_1}-\sqrt{\theta_2})^2}$$

involving square roots. For $\theta_1 = 1$, $\theta_2 = 9$, closeness depends on 1 and 3, not 1 and 9. For examples, see Bowman and Shenton (2004).

5 Some mathematical forms

In some cases mathematical forms may be found for the moments of S_1 and similar terms. For a binomial variable, $0 < p < 1$, $n = 1, 2, \cdots$,

$$E[S_1(x, n^+)] = E\left\{\frac{1}{n} + \frac{1}{n+1} + \cdots + \frac{1}{n+x-1}\right\} = \int_0^1 \frac{(1-(pt+q)^n)t^{n-1}}{1-t}dt$$
$$E[S_1(x, p)] = \int_0^1 \frac{(1-(pt+q)^n)t^{p-1}}{1-t}dt$$

and

$$E[S_1(x, n^-)] = E\left\{\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-x+1}\right\}$$
$$= \int_0^1 \frac{(qt+p)^n - t^n}{1-t}dt = p + \frac{p^2}{2} + \cdots + \frac{p^n}{n}.$$

Similar expressions can be found for S_2, $S_3, \cdots$. In particular,

$$E[S_j(x, n^+)] = E\left\{\frac{1}{n^j} + \frac{1}{(n+1)^j} + \cdots + \frac{1}{(n+x-1)^j}\right\}$$
$$= \frac{1}{(j-1)!}\int_0^1 \frac{\ln(\frac{1}{t})(1-(pt+q)^n)t^{n-1}}{1-t}dt$$
$$E[S_j(x, p)] = E\left\{\frac{1}{p^j} + \frac{1}{(p+1)^j} + \cdots + \frac{1}{(p+x-1)^j}\right\}$$
$$= \frac{1}{(j-1)!}\int_0^1 t^{p-1}(\ln\frac{1}{t})^{j-1}\frac{(1-(pt+q)^n)}{1-t}dt.$$

The generating function of $S_1(x; k)$ for the negative binomial distribution is

$$\frac{1}{\ln(p+1)}\sum_{x=1}^{\infty}(p+1)^{-k}\left(\frac{p}{p+1}\right)^x \frac{\Gamma(k+x)}{x!\Gamma(k)}\left(\frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{k+x-1}\right)t^x$$
$$= (p+1-pt)^{-k}[1 - \ln(p+1-pt)^{-k}],$$

the logarithmic term suggesting the deviation from normality when $k \to \infty$.

6 Conclusion

The random variable $\psi(k+x) - \psi(k)$, with $k > 0$, $x = 0, 1, \cdots$, has a distribution if the distribution of x is known; cases when x is binomial, Poisson, negative binomial have been

considered. The test function $T_1(\underline{n};k)$ has arisen from the sample available and is the mean of $S_1(x;k)$; higher moments therefore follow the pattern of those for the mean. The single component case has been studied by Bowman and Shenton (2007). Properties for a single component apply to mixtures since these are linear in the components.

John (1970) treats the two component mixture and asymptotics for moment estimators. The book by Johnson, Kotz and Kemp (1981) has an excellent account of mixtures (Chapter 8) and a treatment of the calculus of probabilities. We believe the book is mainly the work of Dr. A. Kemp (see Read (2004)). An early account of discrete distributions is given in Johnson and Kotz (1969). Karlis and Xekalaki (2005) give an extensive review of literature on Poisson mixtures. (Note here, Poisson is used as a general descriptive).

References

Bowman, K.O. and Shenton, L.R. (1965). *Asymptotic Covariances for the Maximum Likelihood Estimators of the Parameters of a Negative Binomial Distribution*, K-1643, Union Carbide Corporation, Nuclear Division, Oak Ridge, TN.

Bowman, K.O. and Shenton, L.R. (1979a). Approximate percentage points for Pearson Distributions, *Biometrika*, 66, 1, 147-51.

Bowman, K.O. and Shenton, L.R. (1979b). Further approximate Pearson percentage Points and Cornish-Fisher, *Commu.Statist.-Simula.Computa.*, B8(3), 231-244.

Bowman, K.O. and Shenton, L.R. (1989). *Continued Fractions in Statistical Applications*, Marcel Dekker, Inc., New York, New York.

Bowman, K.O. and Shenton, L.R. (2004). Mixtures, hybrid mixtures, canonical forms, and Matusita's distance, *Far East Journal of Theoretical Statistics*, 12(1), 69-88.

Bowman, K.O. and Shenton, L.R. (2007). Binomial test statistics using psi-functions, *Far East Journal of Theoretical Statistics*, accepted for publication.

Bromwich, T.J.I.A. (1926). *An Introduction to the Theory of Infinite Series, 2nd ed.*, Macmillan and Co., London.

Consul, P.C. (1989). *Generalized Poisson Distributions, Properties and Applications*, Marcel Dekker Inc., New York.

Elderton, W.P. and Johnson, N.L. (1969). *System of Frequency Curves*, Cambridge University Press, London.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*, Chapman and Hall, London.

Fisher, R.A. (1941). The negative binomial distribution, *Annals of Eugenics*, XI(II), 182-187.

Hasselblad, V. (1969). Estimation of parameters for a mixture of Normal distributions, *Technometrics*, 8, 43-44.

John, S. (1970). On analyzing mixed samples, *JASA*, 65, 751-762.

Johnson, N.L. and Kotz, S. (1969). *Discrete Distributions*, Houghton Mifflin Company Boston.

Johnson, N.L., Kotz, S. and Kemp, A.W. (1981). *Univariate Discrete Distributions*, Second Ed., John Wiley & Sons.

Karlis, D. and Xekalaki, E. (2005). Mixed Poisson distributions, *International Statistical Review*, 73(1), 35-58.

Matusita, K. (1955). Decision rules, based on the distance, for problems of fit, two samples, and estimation, *Amer.Math.Soc.*, 26, 631-640.

Pearson, K. (1902). On the systematic fitting of curves to observations and measurement, *Biometrika*, 1, 265-303; 2, 1-23.

Read, C.B. (2004), A conversation with Normal L. Johnson, *Statistical Science*, 19(3), 544-560.