International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 12, 2012, pp 47-52

Stem and Leaf Analysis and Its Validation for Moments

Md. Aminul Islam, Md. Ayub Ali and Md. Abul Basher Mian
Department of Statistics
University of Rajshahi.
Rajshahi-6205, Bangladesh.

[Received October 20, 2008; Revised February 20, 2012; Accepted March 11, 2012]

Abstract

The stem and leaf plot is a combined tabular and graphical display. A frequency distribution can easily be constructed from stem and leaf display by counting the leaves belonging to each stem noting that each stem defines a class interval. The purpose of the present study is to develop some computing formulae of different statistics for stem and leaf display so that the data set could be displayed and analyzed in grouping format without loosing any information. We have proposed some formulae to compute different raw moments, established relationship between raw and central moments and verified their properties as per frequency data.

Keywords and Phrases: Stem and Leaf, Raw data, Group data and Validity.

AMS Classification: Primary 62E15; Secondary 62G07.

1 Introduction

Stem and leaf display are widely used by researchers (Tukey, 1977: Walpole, 1983; Danial, 1995, Islam, 2001, Islam at el. (2008)). An advantage of the stem and leaf display over the histogram is that this process is an easy and quick way of displaying ungrouped data in a grouped format, which is constructed during the tallying process. To calculate the value of any statistic from grouped data we loss some information as the operation depends only on the mid-value of the class interval (Gupta and Kapoor, 1994; Kapur and Saxena, 1986; Goel, Prakash and Lal, 1991; and Islam, 2001). Thus some researchers used Sheppard's correction to reduced deficiency of information loosing for moments. But it is not possible to remove deficiency completely by Sheppard's correction which is applied under certain restrictions (Weatherburn, 1986; Goel, Prakash and Lal, 1991 and Gupta and Kapoor, 1994). Daniel (1995) pointed out that, stem and leaf enables to represent the whole data set in a grouping manner and helps to compute the statistic (median, percentile, deciles, mode, etc.) with exact precision without loosing any information. Rahman, et al. (2004) has developed computing formulae for the mean, variance, central moments, skewness, kurtosis, etc. for stem and leaf display data with highest precision without loosing information (Islam et al (2008)) . The objective of the present paper is to develop formulae for raw moments followed by the establishment of relation between raw and central moments and to study their properties.

2 Central Moment of Leaves

Suppose the leaves corresponding to k^{th} stem are l_{k1}, l_{k2}, l_{k3}, ..., l_{kn}.

Then r^{th} central moment for leaves of k^{th} Stem is defined as

$\mu_{rk}(l) = \frac{1}{n_k}\sum_{i=1}^{n_k}(l_{ki} - \bar{l}_k)^r; r = 1, 2, 3...$

3 Raw Moments of the Distribution in Terms of Raw Moments of Leaves

Let X be a variable with m stems $S_1, S_2, ..., S_m$ with corresponding numbers of leaves $n_1, n_2, ..., n_m$. For every stem the stem unit is h and leaf unit is 1. Then, base corresponding to kth stem is $B_k = hS_k$. Let T_a be any arbitrary constant such that $T_a \neq \bar{T}$.

Then, r^{th} raw moment of the distribution about any arbitrary constant T_a is defined as

$\mu'_r = \frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k+l_{ki}-T_a)^r}{\sum_{k=1}^{m} n_k}; r = 1, 2, 3...$

Let $T_a = B_a + l_a$ where B_a and l_a are respectively the base part and leaf part of T_a.

Then, $\mu'_r = \frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k}\{(l_{ki}-l_a)+(B_k-B_a)\}^r}{\sum_{k=1}^{m} n_k}$

$$= \frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k}\{(l_{ki} - l_a)^r + {}^r c_1(l_{ki} - l_a)^{r-1}(B_k - B_a) + {}^r c_2(l_{ki} - l_a)^{r-2}(B_k - B_a)^2 + ... + (B_k - B_a)^r\}}{\sum_{k=1}^{m} n_k}$$

$$= \frac{\sum_{k=1}^{m}\{n_k\mu'_{kr}(l) + {}^r c_1\mu'_{k(r-1)}(l)(B_k - B_a) + {}^r c_2\mu'_{k(r-2)}(l)(B_k - B_a)^2 + ... + (B_k - B_a)^r\}}{\sum_{k=1}^{m} n_k}$$

$= \frac{\sum_{k=1}^{m}\sum_{j=0}^{r} {}^{r}c_j n_k \mu'_{k(r-j)}(l)(B_k-B_a)^j}{\sum_{k=1}^{m} n_k}$

$= \frac{\sum_{k=1}^{m}\sum_{j=0}^{r} {}^{r}c_j n_k \mu'_{k(r-j)}(l)d_k^j}{\sum_{k=1}^{m} n_k}; where\ d_k = B_k - B_a.$

In particular, $\mu'_1 = \frac{\sum_{k=1}^{m} n_k\{\mu'_{k1}(l)+d_k\}}{\sum_{k=1}^{m} n_k}$, $\mu'_2 = \frac{\sum_{k=1}^{m} n_k\{\mu'_{k2}(l)+2\mu'_{k1}d_k+d_k^2\}}{\sum_{k=1}^{m} n_k}$,

$\mu'_3 = \frac{\sum_{k=1}^{m} n_k\{\mu'_{k3}(l)+3\mu'_{k2}d_k+3\mu'_{k1}(d_k)^2+d_k^3\}}{\sum_{k=1}^{m} n_k}$ and

$\mu'_4 = \frac{\sum_{k=1}^{m} n_k\{\mu'_{k4}(l)+4\mu'_{k3}d_k+6\mu'_{k2}(d_k)^2+6\mu'_{k1}(d_k)^3+d_k^4\}}{\sum_{k=1}^{m} n_k}$

Corollary 1: Raw moments in terms of central moments of leaves are

$\mu'_r = \frac{\sum_{k=1}^{m}\sum_{j=0}^{r} {}^{r}c_j n_k \mu_{k(r-j)}(l)(d'_k)^j}{\sum_{k=1}^{m} n_k}$; r=1, 2, 3...

Where, $d'_k = B_k + \bar{l}_k - T_a$

In particular, $\mu'_1 = \frac{\sum_{k=1}^{m} n_k d'_k}{\sum_{k=1}^{m} n_k}$, $\mu'_2 = \frac{\sum_{k=1}^{m} n_k\{\mu_{k2}(l)+2\mu_{k1}d'_k+(d'_k)^2\}}{\sum_{k=1}^{m} n_k}$,

$\mu'_3 = \frac{\sum_{k=1}^{m} n_k\{\mu_{k3}(l)+3\mu_{k2}d'_k++(d'_k)^3\}}{\sum_{k=1}^{m} n_k}$

and

$\mu'_4 = \frac{\sum_{k=1}^{m} n_k\{\mu_{k4}(l)+4\mu_{k3}d'_k+6\mu_{k2}(d'_k)^2+(d'_k)^4\}}{\sum_{k=1}^{m} n_k}$

4 Central Moments in Terms of Raw Moments

The r^{th} central moment for stem and leaf display data is

$\mu_r = \frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k}\{(B_k+l_{ki}-\bar{l}_k)\}^r}{\sum_{k=1}^{m} n_k}$; r=1, 2, 3...

and r^{th} raw moment about any arbitrary value T_a is

$\mu'_r = \frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k+l_{ki}-T_a)^r}{\sum_{k=1}^{m} n_k}; r = 1, 2, 3...$

Where, $\bar{T} = \frac{\sum_{k=1}^{m} Su_k+\sum_{k=1}^{m} Lu_k}{\sum_{k=1}^{m} n_k}, Su_k = n_k B_k, Lu_k = \sum_{i=1}^{n_k} l_{ki}$ and n_{ki} is the frequency of i^{th} leaf corresponding to k^{th} stem.

Now, $\mu_r = \frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k}\{(B_k+l_{ki}-\bar{l}_k)\}^r}{\sum_{k=1}^{m} n_k}$

$= \frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k}\{(B_k+l_{ki}-T_a+T_a-\bar{l}_k)\}^r}{\sum_{k=1}^{m} n_k}$

$= \frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k}\{(B_k+l_{ki}-T_a)-(\bar{T}-T_a)\}^r}{\sum_{k=1}^{m} n_k}$

$= \frac{1}{\sum_{k=1}^{m} n_k}\sum_{k=1}^{m}\sum_{i=1}^{n_k}\{(B_k + l_{ki} - T_a)^r - {}^{r}c_1(B_k + l_{ki} - T_a)^{r-1}(\bar{T} - T_a) + {}^{r}c_2(B_k+l_{ki}-T_a)^{r-2}(\bar{T} - T_a)^2 - ... + (-1)^{r-1}(\bar{T} - T_a)^r\}$

$= \frac{1}{\sum_{k=1}^{m} n_k} \sum_{k=1}^{m} \sum_{i=1}^{n_k} (B_k + l_{ki} - T_a)^r - {}^r c_1 (\bar{T} - T_a) \frac{1}{\sum_{k=1}^{m} n_k} \sum_{k=1}^{m} \sum_{i=1}^{n_k} (B_k + l_{ki} - T_a)^{r-1}$
$+ {}^r c_2 (\bar{T} - T_a)^2 \frac{1}{\sum_{k=1}^{m} n_k} \sum_{k=1}^{m} \sum_{i=1}^{n_k} (B_k + l_{ki} - T_a)^{r-2}$
$- ... + (-1)^{r-1} (\bar{T} - T_a)^r \}$
$= \mu'_r - {}^r c_1 (\bar{T} - T_a) \mu'_{r-1} + {}^r c_2 (\bar{T} - T_a^2) \mu'_{r-2} - ... + (-1)^{r-1} (\bar{T} - T_a)^r$

$= \sum_{j=1}^{m} (-1)^{j\, r} c_j \mu'_{r-j} (\mu'_1)^j$

$\Rightarrow \mu_r = \sum_{j=1}^{m} (-1)^{j\, r} c_j \mu'_{r-j} (\mu'_1)^j$; r=1, 2, 3...

In particular, $\mu_1 = 0, \mu_2 = \mu'_2 - (\mu'_1)^2$
$\mu_3 = \mu'_3 - 3\mu'_2 \mu'_1 + 2(\mu'_1)^3$ and
$\mu_4 = \mu'_4 - 4\mu'_3 \mu'_1 + 6\mu'_2 (\mu'_1)^2 - 3(\mu'_1)^4$

Corollary 2: r^{th} raw moments in terms of central moments is
$\mu'_r = \sum_{j=1}^{m} {}^r c_j \mu_{r-j} (\mu'_1)^j$; r=1, 2, 3...

In particular, $\mu'_1 = \bar{T} - T_a, \mu'_2 = \mu_2 + (\mu'_1)^2$
$\mu'_3 = \mu_3 + 3\mu_2(\mu'_1) + (\mu'_1)^3$ and
$\mu'_4 = \mu_4 + 4\mu_3(\mu'_1) + 6\mu_2(\mu'_1)^2 + (\mu'_1)^4$

5 Proposed Theorems

Theorem 1: r^{th} central moment depends on scale but not on origin.

Proof: Considering same notations given section 2.
Let us consider a new variable after changing its scale and origin as $u_k = \frac{B_k - A}{h}$, where A and h are respectively origin and scale.

$\Rightarrow B_k = A + hu_k$
Now, $\bar{T} = \frac{\sum_{k=1}^{m} Su_k + \sum_{k=1}^{m} Lu_k}{\sum_{k=1}^{m} n_k}, Su_k = n_k B_k, Lu_k = \sum_{i=1}^{n_k} l_{ki},$

where, $\sum_{k=1}^{m} Su_k$
$= \sum_{k=1}^{m} n_k B_k$
$= h \sum_{k=1}^{m} n_k U B_k + A \sum_{k=1}^{m} n_k$

Thus, $\bar{T} = \frac{h \sum_{k=1}^{m} n_k u_k + A \sum_{k=1}^{m} n_k + \sum_{k=1}^{m} Lu_k}{\sum_{k=1}^{m} n_k}$
$\Rightarrow \bar{T} = \frac{h(\sum_{k=1}^{m} n_k u_k - h \sum_{k=1}^{m} Lu_k) + A \sum_{k=1}^{m} n_k}{\sum_{k=1}^{m} n_k}$
$\Rightarrow \bar{T} = h\bar{u} - h\bar{l} + A$
$\Rightarrow \bar{T} = h\bar{u} - \bar{l}) + A$

Hence, $\mu_r = \frac{1}{\sum_{k=1}^{m} n_k} \sum_{k=1}^{m} \sum_{i=1}^{n_k} \{hu_k + A + l_{ki} - h(\bar{u} - \bar{l}) - A\}^r$
$\Rightarrow \mu_r = \frac{1}{\sum_{k=1}^{m} n_k} \sum_{k=1}^{m} \sum_{i=1}^{n_k} \{hu_k + l_{ki} - h(\bar{u} - \bar{l})\}^r$

This shows that r^{th} central moments vary with the change of scale but not on origin. Q.E.D.

Corollary 3: β_1 and β_2 are independent of origin but not of scale in stems and leaves display data set.

Theorem 2: For non zero unequal values for stems and leaves display data set $\beta_2 \geq \beta_1 + 1$

Proof: Considering same notations given section 2.

Then r^{th} central moment is $\mu_r = \frac{1}{\sum_{k=1}^{m} n_k} \sum_{k=1}^{m} \sum_{i=1}^{n_k} (B_k + l_{ki} - \bar{T})^r; r = 1, 2, 3...$
$\Rightarrow \mu_2 = \frac{1}{\sum_{k=1}^{m} n_k} \sum_{k=1}^{m} \sum_{i=1}^{n_k} (B_k + l_{ki} - \bar{T})^2$
$\Rightarrow \sum_{k=1}^{m} n_k \mu_2 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} u_{ki}^2$ when $u_{ki} = B_k + l_{ki} - \bar{T}$

similarly, $\sum_{k=1}^{m} n_k \mu_3 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} u_{ki}^3$,
$\sum_{k=1}^{m} n_k \mu_4 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} u_{ki}^4$,
$\sum_{k=1}^{m} \sum_{i=1}^{n_k} u_{ki} = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (B_k + l_{ki} - \bar{T}) = 0$

Now, coefficient of measure of skewnwss is $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ and that of kurtosis is $\beta_2 = \frac{\mu_4}{\mu_2^2}$.
Let us consider three real constants a, b and c such that $au_{ki}^2 + bu_{ki} + c$ is a real term.

So, $\sum_{k=1}^{m} \sum_{i=1}^{n_k} (au_{ki}^2 + bu_{ki} + c)^2 \geq 0$
$\Rightarrow \sum_{k=1}^{m} \sum_{i=1}^{n_k} (a^2 u_{ki}^4 + b^2 u_{ki}^2 + c^2 + 2ab u_{ki}^3 + 2bc u_{ki} + 2ca u_{ki}^2) \geq 0$
$\Rightarrow a^2 \sum_{k=1}^{m} \sum_{i=1}^{n_k} u_{ki}^4 + b^2 \sum_{k=1}^{m} \sum_{i=1}^{n_k} u_{ki}^2 + c^2 \sum_{k=1}^{m} n_k$
$+2ab \sum_{k=1}^{m} \sum_{i=1}^{n_k} u_{ki}^3 + 2bc \sum_{k=1}^{m} \sum_{i=1}^{n_k} u_{ki} + 2ca \sum_{k=1}^{m} \sum_{i=1}^{n_k} u_{ki}^2 \geq 0$
$\Rightarrow a^2 \sum_{k=1}^{m} n_k \mu_4 + b^2 \sum_{k=1}^{m} n_k \mu_2 + c^2 \sum_{k=1}^{m} n_k + 2ab \sum_{k=1}^{m} n_k \mu_3 + 2bc.0$

$+ 2ca \sum_{k=1}^{m} n_k \mu_2 \geq 0$
$\Rightarrow a^2 \mu_4 + b^2 \mu_2 + c^2 + 2ab\mu_3 + 2ca\mu_2 \geq 0$ as $\sum_{k=1}^{m} n_k > 0$

Setting $a = 1$, $b = -\frac{\mu_3}{\mu_2}$ and $c = -\mu_2$ we get
$\mu_4 + (-\frac{\mu_3}{\mu_2})^2 + \mu_2^2 + 2.1.(-\frac{\mu_3}{\mu_2})\mu_3 + 2(-\mu_2).1.\mu_2 \geq 0$
$\Rightarrow \mu_4 + \frac{\mu_3^2}{\mu_2} - 2\frac{\mu_3^2}{\mu_2} - \mu_2^2 \geq 0$
$\Rightarrow \mu_4 - \frac{\mu_3^2}{\mu_2} - \mu_2^2 \geq 0$
$\Rightarrow \frac{\mu_4}{\mu_2^2} - \frac{\mu_3^2}{\mu_2^3} - 1 \geq 0$

$\Rightarrow \beta_2 - \beta_1 - 1 \geq 0$
$\Rightarrow \beta_2 \geq \beta_1 + 1$
Q.E.D.

Corollary 4a: For non zero unequal values for stems and leaves display data set $\beta_2 \geq \beta_1$

Corollary 4b: For non zero unequal values for stems and leaves display data set $\beta_2 \geq 1$

References

[1] Daniel, W. W. (1995). *Biostatics: A Foundation for Analysis in The Health Science*. 6th Edition Wiley Series in Probability and Mathematical Statistics-Applied.

[2] Goel, B. S., Prakash, S. and Lal, R. (1991). *Mathematical Statistics*. Pragati Prakation,inc., Englewood cliffs.

[3] Gupta, S. C. and Kapoor, V. K. (1994). *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons Educational Publishers, New Delhi, India. .

[4] Islam M. A., Mian M. A. B. and Ali M. A. (2008). *Some Properties of Stem and Leaf Display Data Set*, International Journal of Statistical Sciences **8**, 111-118.

[5] Islam, M. N.(2001). *An Introduction to Statistics*. 3rd edition, Book world, Dhaka, Bangladesh.

[6] Kapoor, J. N. and Saxena, H. C. (1986). *Mathematical Statistics*. S. Chand & Company Ltd. New Delhi, India..

[7] Rahman K. B., Mian M. A. B. and Ali M. A. (2004)). *Stem and Leaf Analysis and its validation*, International Journal of Statistical Sciences **3**, 93-102.

[8] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addision-Wesley Publishing Co., Reading Mass.

[9] Walpole, R. E. (1983). *Introduction To Statistics* 3rd edition, Macmillan Publishing Co., New York.

[10] Weatherburn, C. E. (1986). *A First Course in Mathematical Statistics*. S. Chand & Company Ltd., New Delhi, India.