International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 8, 2008, pp 111-118

Some Properties of Stem and Leaf Display Data Set

Md. Aminul Islam, Md. Abul Basher Mian and Md. Ayub Ali
Department of Statistics
University of Rajshahi
Rajshahi 6205, Bangladesh
E-mail: ayubali67@yahoo.com

[Received March 24, 2008; Revised August 19, 2008; Accepted November 1, 2008]

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 was to develop a procedure to state and prove some properties in the case of central tendency for the Stem and Leaf display. This study established some valid properties of arithmetic mean in the case of Stem and Leaf method.

Keywords and Phrases: Stem and leaf, Properties, Central tendency.

AMS Classification: 62-07, 62-08, 62-09, 62Q05.

1 Introduction

Stem and Leaf display is used to represent a strong resemblance of a Histogram and to serve the same purpose (Daniel, 1995; Islam, 2001). An advantage of the Stem and Leaf display over the histogram is the fact that Stem and Leaf display is an easy and quick way of displaying ungrouped data in a grouped format, which constructed during the tallying process (Daniel, 1995).

Box-and-Whisker plot and Stem and Leaf displays are examples of what are known as explanatory data analysis techniques (Daniel, 1995; Walpole, 1983). These techniques, made popular as a result of the work of Turkey (1977), allow the investigator to examine data in ways that reveal trends and relationships, identify unique features of data sets and facilitate their description and summarization. In the Stem and Leaf

display, we can see the data as both grouped and ungrouped ways and calculate measures of location such as mean, median and mode (Rahman, et al., 2004). Now it is essential to know important properties of these compared for those existing properties.

Therefore the objective of the present study is to state and established the following Properties.

The important properties of arithmetic mean with their proofs are given below:

Theorem 1: Sum of deviations from arithmetic mean for Stem and Leafs display observations is equal to zero.

$$i.e. \sum_{k=1}^{m} \sum_{i=1}^{n_k} (B_k + l_{ki} - \overline{T}) = 0 \tag{1}$$

where B_k = Base corresponding to the k^{th} stem, $l_{ki} = i^{th}$ leaf corresponding to the k^{th} stem.

Proof: Let us consider that a variable X has m stems as $S_1, S_2, S_3, \cdots, S_m$ and corresponding numbers of leaves are as $n_1, n_2, n_3, \cdots, n_m$.

Suppose the leaves corresponding to k^{th} stem are $l_{k1}, l_{k2}, l_{k3}, \cdots, l_{kn_1}$ its base is B_k and base sum S_{u_k}.

Then

$$B_k = hS_k \tag{2}$$

and

$$Su_k = n_k B_k \tag{3}$$

where h is the stem unit.

Also suppose leaves sum corresponding to k^{th} stem is

$$Lu_k = \sum_{i=1}^{n_k} l_{ki} \tag{4}$$

Hence, sum of all bases and all leaves of the variable is $\sum_{k=1}^{m} Su_k + \sum_{k=1}^{m} Lu_k$ and total number of observations $\sum_{k=1}^{m} n_k$.

According to Rahman et al., 2003

$$\overline{T} = \frac{\sum_{k=1}^{m} Su_k + \sum_{k=1}^{m} Lu_k}{\sum_{k=1}^{m} n_k} \Rightarrow \overline{T} \sum_{k=1}^{m} n_k = \sum_{k=1}^{m} Su_k + \sum_{k=1}^{m} Lu_k \tag{5}$$

Now $\sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - \overline{T})$

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

$= \sum_{k=1}^{m}(Su_k + Lu_k - n_k\overline{T})$ from (2) and (4)

$= \sum_{k=1}^{m} Su_k + \sum_{k=1}^{m} Lu_k - \overline{T}\sum_{i=1}^{m} n_k$

$= \overline{T}\sum_{k=1}^{m} n_k - \overline{T}\sum_{i=1}^{m} n_k = 0$

Hence the proof.

Theorem 2: Sum of squared deviations from arithmetic mean is least for Stem and Leaf display data set.

i.e. $\sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - \overline{T})^2 \leq \sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - a)^2$ where a is any arbitrary constant and other notations are the same as in theorem 1.

Proof: $\sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - a)^2$

$= \sum_{k=1}^{m}\sum_{i=1}^{n_k}\{(B_k + l_{ki} - \overline{T}) + (\overline{T} - a)\}^2$

$= \sum_{k=1}^{m}\sum_{i=1}^{n_k}\{(B_k + l_{ki} - \overline{T})^2 + 2(B_k + l_{ki} - \overline{T})(\overline{T} - a) + (\overline{T} - a)^2\}$

$= \sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - \overline{T})^2 + 2(\overline{T} - a)\sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - \overline{T}) + (\overline{T} - a)^2\sum_{k=1}^{m} n_k$

$= \sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - \overline{T})^2 + 2(\overline{T} - a).0 + (\overline{T} - a)^2\sum_{k=1}^{m} n_k$ from (theorem 1)

$= \sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - \overline{T})^2 + (\overline{T} - a)^2\sum_{k=1}^{m} n_k$

Here $(\overline{T} - a)^2$ is a square of a real term, which is always non-negative, and sum number of leaves also non-negative, and $(\overline{T} - a)^2\sum_{k=1}^{m} n_k$ is also non-negative term.

Hence, $\sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - \overline{T})^2 \leq \sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki} - a)^2$ equality holds only when $a = \overline{T}$.
Hence the proof.

Theorem 3: Arithmetic mean of Stem and Leaf display data set is arithmetic mean of raw data.

Proof: We have $\overline{T} = \frac{\sum_{k=1}^{m} Lu_k + \sum_{k=1}^{m} Su_k}{\sum_{k=1}^{m} n_k}$; (notations are the same as in theorem 1)

$= \frac{\sum_{k=1}^{m} (Lu_k + Su_k)}{\sum_{k=1}^{m} n_k}$

$= \frac{\sum_{k=1}^{m} \sum_{i=1}^{n_k} (l_{ki} + n_k B_k)}{\sum_{k=1}^{m} n_k}$

$= \frac{\sum_{k=1}^{m} \sum_{i=1}^{n_k} (l_{ki} + \sum_{i=1}^{n_k} B_k)}{\sum_{k=1}^{m} n_k}$

$= \frac{\sum_{k=1}^{m} \sum_{i=1}^{n_k} (l_{ki} + B_{k)})}{\sum_{k=1}^{m} n_k}$

$= \frac{\text{Exact Sum of all observations}}{\text{Total number of observations}}$

= Arithmetice mean from raw data
Q.E.D.

Theorem 4: Arithmetic mean of sum of two variables with equal number of observations is equal to the sum of their respective arithmetic means for Stem and Leaf formatted data.

Proof: Let X and Y be two variables with equal number of observations. Consider X have m_x stems $S_1, S_2, S_3, \cdots, S_{m_x}$ and corresponding numbers of leafs are as $n_1, n_2, n_3, \cdots, n_{m_x}$. Again Y has m_y stems $S'_1, S'_2, S'_3, \cdots, S'_{m_y}$ and corresponding numbers of leaves are as $n'_1, n'_2, n'_3, \cdots, n'_{m_y}$.
Then $B_k = hS_k$ and $B'_k = cS_{k'}$; where h and c are the stem unit of X and Y, respectively.
We can write for variable X;
$\overline{T}_x = \frac{\sum_{k=1}^{m_x} Su_k + \sum_{k=1}^{m_x} Lu_k}{\sum_{k=1}^{m_x} n_k}$ and for variable Y; $\overline{T}_y = \frac{\sum_{k'=1}^{m_y} Su_{k'} + \sum_{k'=1}^{m_y} Lu_{k'}}{\sum_{k'=1}^{m_y} n_{k'}}$

Since, both variables have equal number of observations so, $\sum_{k=1}^{m_x} n_k = \sum_{k'=1}^{m_y} n_{k'} = n$ (say).
Let $U = X + Y$

Since both variables have equal number of observations, both variables must have equal numbers of leaves but they may or may not have equal numbers of stems. Thus, for sum of all values of U,

$$\sum_{i=1}^{n} u_i = \sum_{k=1}^{m_x} Su_k + \sum_{k'=1}^{m_y} Su_{k'} + \sum_{k=1}^{m_x} Lu_k + \sum_{k'=1}^{m_y} Lu_{k'}.$$

Now $\overline{T}_u = \frac{\sum_{i=1}^{n} u_i}{n}$

$$= \frac{\sum_{k=1}^{m_x} Su_k + \sum_{k'=1}^{m_y} Su_{k'} + \sum_{k=1}^{m_x} Lu_k + \sum_{k'=1}^{m_y} Lu_{k'}}{n}$$

$$= \frac{\sum_{k=1}^{m_x} Su_k + \sum_{k=1}^{m_x} Lu_k}{\sum_{k=1}^{m_x} n_k} + \frac{\sum_{k'=1}^{m_y} Su_{k'} + \sum_{k'=1}^{m_y} Lu_{k'}}{\sum_{k'=1}^{m_y} n_{k'}}$$

$$= \overline{T}_x + \overline{T}_y.$$

Hence, Arithmetic mean of sum of two variables with equal number of observations is equal to sum of their respective arithmetic means.

Theorem 5: (Mean of the composite series) If $\overline{T}_i$ $(i = 1, 2, 3, \cdots, k)$ are the means of k component series of sizes n_i $(i = 1, 2, 3, \cdots, k)$ respectively, then the mean $\overline{T}$ of the composite series obtained on combining the component series is given by the formula:

$$\overline{T} = \frac{\sum_{i=1}^{r} \sum_{k=1}^{m_k} n_{ik} \overline{T}_i}{\sum_{i=1}^{r} \sum_{k=1}^{m_i} n_{ik}}.$$

Proof: Let us consider if we make Stem and Leaf display for ith subset we get the stems as $S_{i1}, S_{i2}, S_{i3}, \cdots, S_{im_i}$ with respective number of leaves $n_{i1}, n_{i2}, n_{i3}, \cdots, n_{im_i}$ $(i = 1, 2, 3, \cdots, k)$.

So, arithmetic mean of the i^{th} subset is $\overline{T}_i = \frac{\sum_{k=1}^{mi_1} Su_{ik} + \sum_{k=1}^{m_i} Lu_{ik}}{\sum_{k=1}^{m_i} n_{ik}}$; (i=1,2,3,. . . ,k).

$$\Rightarrow \sum_{k=1}^{m_i} n_{ik} \overline{T}_i = \sum_{i=1}^{m_i} Su_{ik} + \sum_{k=1}^{m_i} Lu_{ik}; \text{ (i=1,2,3,. . . , k)}$$

$$\Rightarrow \sum_{k=1}^{m_1} n_{1k} \overline{T}_1 + \sum_{k=1}^{m_1} n_{2k} \overline{T}_2 + \cdots \sum_{k=1}^{m_r} n_{rk} \overline{T}_r$$

$$= \sum_{i=1}^{m_1} Su_{1k} + \sum_{k=1}^{m_1} Lu_{1k} + \sum_{i=1}^{m_2} Su_{2k} + \sum_{k=1}^{m_2} Lu_{2k} + \cdots + \sum_{k=1}^{m_r} Su_{rk} + \sum_{k=1}^{m_r} Lu_{rk}$$

$$= \sum_{k=1}^{m_1} \sum_{i=1}^{n_{1k}} (l_{1ki} + B_{1k}) + \sum_{k=1}^{m_2} \sum_{i=1}^{n_{2k}} (l_{2ki} + B_{2k}) + \cdots + \sum_{k=1}^{m_r} \sum_{i=1}^{n_{rk}} (l_{rki} + B_{rk})$$

= Sum of all observations in r set of data
= number of all observations in r set of data
×Arithmetic mean of the combained r set of data

$$= (\sum_{j=1}^{r}\sum_{k=1}^{m_j} n_{jk})\overline{T}$$

$$\Rightarrow \overline{T} = \frac{\overline{T}_1 \sum_{k=1}^{m_1} n_{1k} + \overline{T}_2 \sum_{k=1}^{m_2} n_{2k} + \overline{T}_3 \sum_{k=1}^{m_3} n_{3k} + \cdots + \overline{T}_r \sum_{k=1}^{m_r} n_{rk}}{\sum_{j=1}^{r}\sum_{k=1}^{m_j} n_{jk}}$$

$$\therefore \overline{T} = \frac{\sum_{j=1}^{r}\sum_{k=1}^{m_j} n_{jk}\overline{T}_j}{\sum_{j=1}^{r}\sum_{k=1}^{m_j} n_{jk}}.$$

Hence, for r set of Stem and Leaf displayed data, arithmetic mean of the combined set is equal to the weighted average.

Corollary 5a: Combined arithmetic mean of two subsets in Stem and leaf formatted data is as $\overline{T}_c = \frac{n_1\overline{T}_1 + n_2\overline{T}_2}{n_1+n_2}$ where n_1 is total number of leaves corresponding to 1^{st} subset with mean $\overline{T}_1$ and n_2 is total number of leaves corresponding to 2^{nd} subset with mean $\overline{T}_2$.

Theorem 6: If there is only one stem of the variable then arithmetic mean of the variable is sum of base of that stem and arithmetic mean of leaves corresponding to that stem.

Proof: Let X be a variable with only one stem S_1 whose leaves are $l_{11}, l_{12}, l_{13}, \cdots, l_{1n}$. Then base of that stem is $B_1 = hS_1$, then sum of base of that stem is $Su_1 = nB_1$ and sum of leaves of that stem is $Lu_1 = \sum_{k=1}^{n} l_{1k}$.

Then mean of the variable is

$$\overline{T} = \frac{Su_1 + Lu_1}{n}$$

Or, $\overline{T} = \frac{Su_1 + Lu_1}{n}$
Or, $\overline{T} = \frac{nB_1 + Lu_1}{n}$
Or, $\overline{T} = B_1 + \frac{\sum_{i=1}^{n} l_{1i}}{n}$
Or, $\overline{T} = B_1 + \bar{l}_1$.
Q.E.D.

Theorem 7: If a and b are two constants and X be a variable then arithmetic mean of $Y = aX + b$ is $\overline{T}_y = a\overline{T} + b$; where $\overline{T}$ is the arithmetic mean of the variable X.

Proof: We have $Y = aX + b$

Suppose the ith value of Y corresponding to kth stem is

$$
\begin{aligned}
&{}^{y}B_k + {}^{y}l_{ki} = a(B_k + l_{ki}) + b \\
&\Rightarrow \textstyle\sum_{k=1}^{m} \sum_{i=1}^{n_k} ({}^{Y}B_k + {}^{y}l_{ki}) = a \sum_{k=1}^{m} \sum_{i=1}^{n_k} (B_k + l_{ki}) + b \sum_{k=1}^{m} n_k \\
&\Rightarrow \sum_{k=1}^{m} {}^{y}Su_k + \sum_{k=1}^{n_k} {}^{y}Lu_k = a(\sum_{k=1}^{m} Su_k + \sum_{k=1}^{n_k} Lu_k) + b \sum_{k=1}^{m} n_k \\
&\Rightarrow \frac{\sum_{k=1}^{m} {}^{y}Su_k + \sum_{k=1}^{n_k} {}^{y}Lu_k}{\sum_{k=1}^{m} n_k} \\
&= a\frac{(\sum_{k=1}^{m} Su_k + \sum_{k=1}^{n_k} Lu_k)}{\sum_{k=1}^{m} n_k} + b; \text{ from theorem 1 and since } \sum_{k=1}^{m} n_k \neq 0 \\
&\Rightarrow \overline{T}_y = a\overline{T} + b
\end{aligned}
$$

Q.E.D.

Corollary 7a: Arithmetic mean of product of a constant and a variable is product of that constant and arithmetic mean of that variable for Stem and Leaf related data.

i.e. $\overline{T}_{aX} = a\overline{T}$.

Corollary 7b: Arithmetic mean of sum of a constant and a variable is sum of that constant and arithmetic mean of that variable for Stem and Leaf related data.

i.e. $\overline{T}_{a+X} = a + \overline{T}$.

Theorem 8: Arithmetic mean depends on both scale and origin in stem and leaves related case.

Proof: Let X be a variable with m stems $S_1, S_2, S_3, \ldots, S_m$ and number of leaves respectively $n_1, n_2, n_3, \ldots, n_m$.

Base corresponding to kth stem is $B_k = hS_k$ when stem unit=h and $Su_k = n_k B_k$.
Let us consider $B'_k = \frac{B_k - a}{h}$ where a= origin and h=scale when both B_k and a are divisible by h.

$\Rightarrow B_k = hB'_k + a$
$\Rightarrow i^{th}$ value of k^{th} stem of X is $B_k + l_{ki} = hB'_k + l_{ki} + a$
$\Rightarrow \sum_{k=1}^{m}\sum_{i=1}^{n_k}(B_k + l_{ki}) = \sum_{k=1}^{m}\sum_{i=1}^{n_k}(hB'_k + l_{ki} + a)$
$\Rightarrow \sum_{k=1}^{m} Su_k + \sum_{k=1}^{m} Lu_k = h\sum_{k=1}^{m} Su'_k + \sum_{k=1}^{m} Lu_k + a\sum_{k=1}^{m} n_k$
$\Rightarrow \sum_{k=1}^{m} Su_k + \sum_{k=1}^{m} Lu_k = h\sum_{k=1}^{m} Su'_k + h\sum_{k=1}^{m} Lu_k - h\sum_{k=1}^{m} Lu_k + \sum_{k=1}^{m} Lu_k + a\sum_{k=1}^{m} n_k$
$\Rightarrow \sum_{k=1}^{m} Su_k + \sum_{k=1}^{m} Lu_k = h(\sum_{k=1}^{m} Su'_k + \sum_{k=1}^{m} Lu_k) - \sum_{k=1}^{m} Lu_k(h-1) + a\sum_{k=1}^{m} n_k$
$\Rightarrow \overline{T} = h\overline{T'} - (h-1)\bar{l} + a$
$\Rightarrow \overline{T} = h(\overline{T'} - \bar{l}) + (\bar{l} + a)$
i.e. Arithmetic mean depends on both scale and origin in stem and leaf analysis.

Acknowledgement

We are very much grateful to the editorial board and to the anonymous reviewers for their nice cooperation, valuable comments and suggestions for the improvement of the manuscript.

References

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

[2] Islam, M. N. (2001). "An Introduction to Statistics and Probability", 3rd edition, Book world, Dhaka, Bangladesh.

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

[4] Tukey, J. W. (1977). Exploratory Data Analysis, Addison-Wesley Publishing Co., Reading, Mass.

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