International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 3, 2004, pp 93–101

Stem and Leaf Analysis and It's Validation

Kh. Bokhtier Rahman
Professional Service Representative
Diabetes & Cardiology, Renata Limited
16, M.C. Mazumder Road, Guptapara
Rangpur, Bangladesh

M. A. B. Mian
Department of Statistics
University of Rajshahi
Rajshahi, Bangladesh

Md. Ayub Ali
Laboratory of Growth and Egronomics
Otsuma Women's University
Chiyodaku, Tokyo 102-8357, Japan

[Received October 1, 2003; Revised December 19, 2003; Accepted March 23, 2004]

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 the 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 mean, variance, moments, skewness, and kurtosis for this Stem and Leaf display.

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

1 Introduction

Stem and Leaf display is used to represent a strong resemblance of a Histogram and to serve the same purpose (Tukey, 1977; Walpole, 1983; 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 is constructed during the tallying process. To calculate the value of any statistic from group data we loss 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 protect the deficiency of information losing for moments. But Sheppard's correction is unable to remove the deficiency and is applied under certain restrictions (Weatherburn, 1986; Goel, Prakash and Lal, 1991, and Gupta and Kapoor; 1994).

According to Daniel (1995), Stem and Leaf display enables to present the whole data set in a grouping manner and helps to compute the statistic (median, percentile, deciles, mode, etc.) with highest precision without losing information but the computation of mean, variance, moments, skewness, kurtosis, etc., are still undone to the researchers. Thus objective of the present study is to develop formulae for various statistics, like mean, variance, moments, skewness, kurtosis, etc.

2 Methods and Materials

Stem and Leaf display is a graphical technique of representing quantitative data that can be used to examine the shape of a frequency distribution. In this study some formulae from the Stem and Leaf display are developed for mean and moments of the frequency distribution. Also, the validity of those formulae are checked by arbitrary data set.

• To avoid the repetition of the leaves in the same Stem, the number of leaves are counted and the frequency is presented in the right side within a first bracket ( ); (Table-1).

• To translate the Stem and Leaf display data into statistic, let us define the following mathematical notations:

B_k : Base of original data in the k^{th} Stem = original datum belonging to k^{th} Stem-leaf of the corresponding datum.
n_k: number of leafs corresponding to k^{th} base- Stem.
l_{ki}: observed value of i^{th} leaf in the k^{th} Stem.
C_{n_k}: cumulative number of leaf unit up to the k^{th} Stem. $=\sum_{j=1}^{k} n_j$
$L_k=\sum_{i=1}^{n_k} l_{ki}$: Leaf sum for k^{th} Stem.
$S_k=n_kB_k$: Sum of of the k^{th} base Stem.
$n=\sum_{k=1}^{m} n_k$: Sum of all leaves in the data= Total number of observations; m being the number of Stems.

Statistics such as measures of central tendency, dispersion, skewness and kurtosis can be computed easily without losing any information for the real data set if we use the following formulae.

3 Derived Formulae For Arithmetic Means

Mean: The arithmetic mean, for the leaves of the k^{th} Stem may be defined as $\bar{I}_k = \sum_{i=1}^{n_k} \frac{l_{ki}}{n_k}$
$\bar{S}_k = B_k + \bar{l}_k$ = Mean of the observations in the k^{th} Stem.
$\bar{T} = \sum_{k=1}^{m} \frac{n_k \bar{S}_k}{n}$ = Mean of the distribution. (whole data set)

4 Measurement technique for Median and Partition values

Median is a positional average. It divides a distribution into two equal parts. In Stem and Leaf method, data are in ascending order-tallying process. Hence we can get range, median, quartiles, deciles and percentiles as we get those from raw data.

5 Measurement technique for mode

The mode is that value for data which occurs most frequently. If all the values are different there is no mode; on the other hand, a set of values may have more than one mode. In Stem and Leaf method, we can easily determine mode. The greatest value of the number of similar leaves that presented in the first bracket is considered as the modal Leaf and its corresponding Stem as the modal Stem. The original data corresponding to the modal Leaf in the modal Stem is the mode of the data set.

6 Derived Formulae For moments

Let us consider r^{th} Central moment corresponding to k^{th} Stem.

$$\begin{aligned} \mu_{k,r} &= \frac{\sum_{i=1}^{n_k}[(B_k + l_{ki}) - \bar{S}_k]^r}{n_k} \\ &= \frac{\sum_{i=1}^{n_k}[B_k + l_{ki} - B_k - \bar{I}_k]^r}{n_k} \\ &= \frac{\sum_{i=1}^{n_k}(l_{ki} - \bar{l}_k)^r}{n_k} \quad k = 1, 2, \cdots, m; \quad r = 1, 2, 3, \cdots \end{aligned} \tag{1}$$

$$\mu_{k1} = 0, \quad \text{for all } k$$

Skewness & Kurtosis for the kth stem may be obtained as below:

$$\beta_{1,k} = \frac{\mu_{k,3}^2}{\mu_{k,2}^3} = \frac{n_k \sum_{i=1}^{n_k}(l_{ki} - \bar{l}_k)^3}{\sum_{i=1}^{n_k}(l_{ki} - \bar{l}_k)^2} \tag{2}$$

$$\beta_{2,k} = \frac{n_k \sum_{i=1}^{n_k}(l_{ki} - \bar{l}_k)^4}{\{\sum_{i=1}^{n_k}(l_{ki} - \bar{l}_k)^2\}^2} \tag{3}$$

Thus r^{th} Central moment for all the Stem together, that is for the entire data set

$$\begin{aligned}
\mu_r &= \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n_k}[(B_k + l_{ki}) - \bar{T}]^r \quad r = 1, 2, \cdots \\
&= \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n_k}[(l_{ki} - \bar{l}_k) + B_k + \bar{l}_k - \bar{T}]^r \\
&= \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n_k}[(l_{ki} - \bar{l}_k) - d_k]^r, \quad d_k = B_k + \bar{l}_k - \bar{T} \\
&= \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n_k}[(l_{ki} - \bar{l}_k)^r + \binom{r}{1}(l_{ki} - \bar{l}_k)^{r-1}d_k + \cdots \binom{r}{r}d_k^r] \\
&= \frac{1}{n}\sum_{k=1}^{m}[\sum_{i=1}(l_{ki} - \bar{l}_k)^r + \binom{r}{1}d_k\sum_{i=1}(l_{ki} - \bar{l}_k)^{r-1} + \cdots \sum_{i=1} d_k^r] \\
&= \frac{1}{n}\sum_{k=1}[n_k\{\mu_{k,r} + r\mu_{k,r-1}d_k + \frac{r(r-1)}{2}\mu_{k,r-2}d_k^2 + \cdots + d_k^r\}]
\end{aligned} \tag{4}$$

In particular,

$$\begin{aligned}
\mu_1 &= \frac{1}{n}[\sum_k n_k\mu_{k,1} + \sum_k n_k d_k] \\
&= 0 + \frac{1}{n}\sum_k n_k d_k = \frac{1}{n}\sum_k n_k(B_k + \bar{l}_k - \bar{T}) \\
&= \frac{1}{n}[\sum_k n_k(B_k + \bar{l}_k) - \bar{T}\sum_k n_k] = \frac{1}{n}[n\bar{T} - n\bar{T}] = 0
\end{aligned} \tag{5}$$

$$\mu_2 = \frac{1}{n}\sum_k n_k(\mu_{k,2} + 2\mu_{k,1}d_k + d_k^2) = \frac{1}{n}\sum_k n_k(\mu_{k,2} + d_k^2) \tag{6}$$

$$\mu_3 = \frac{1}{n}\sum_k n_k(\mu_{k,3} + 3\mu_{k,2}d_k + 3\mu_{k,1}d_k^2 + d_k^3) = \frac{1}{n}\sum_k n_k(\mu_{k,3} + 3\mu_{k,2}d_k + d_k^3) \tag{7}$$

$$\mu_4 = \frac{1}{n}\sum_k n_k(\mu_{k,4} + 4\mu_{k,2}d_k + 6\mu_{k,2}d_k^2 + d_k^4) \tag{8}$$

From the moments, we can easily calculate skewness and kurtosis. Also using the following formulae we get skewness and kurtosis as

$$\begin{aligned}\sqrt{\beta_1} &= \frac{\mu_{3s}}{\mu_{2s}^{3/2}} = \sqrt{\frac{(\sum_{k=1}^{m} n_k(\sqrt{\beta_{1k}}\mu_{2k}^{3/2} + 3d_k\mu_{2k} + d_k^3))^2}{(\sum_{k=1}^{m} n_k(\mu_{2k} + d_k^2))^3} \times \sum_{k=1}^{m} n_k} \\ &= \frac{\sqrt{n}\sum_{k=1}^{m} nk(\mu_{k,3} + 3\mu_{k,2}d_k + d_k^3)}{[\sum_{k=1}^{m} n_k(\mu_{k,2} + d_k^2)]^{3/2}} \end{aligned} \tag{9}$$

and

$$\begin{aligned}\beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{\sum_{k=1}^{m} n_k(\beta_{2,k}\mu_{2,k}^2 + 4\sqrt{\beta_{1,k}}\mu_{2,k}^{-1/2}d_k + 6\mu_{2,k}d_k^2 + d_k^4)}{(\sum_{k=1}^{m} n_k(\mu_{2k} + d_k^2))^2} \times \sum_{k=1}^{m} n_k \\ &= \frac{n\sum_{k=1}^{m} n_k(\mu_{k,4} + 4\mu_{k,3}d_k + 6\mu_{k,2}d_k^2 + d_k^4}{[\sum_{k=1}^{m} n_k(\mu_{k,2} + d_k^2)]^2} \end{aligned} \tag{10}$$

respectively. In all the formulae $d_k = (B_k + \bar{l}_k - \bar{T})$ where, $k = 1, 2, 3, \cdots, m$. This method is not only serve overall results but also results for each class interval / width / Stem. We have used an arbitrary data as an example. Marks obtained in a university admission test by 255 student out of 150 marks are shown in Appendix. Using the formula developed in section 3 and 6 mean, variance, moments, skewness and kurtosis of data given in appendix are calculated and presented in Table 1. The Table helps us to understand the calculation procedures of the proposed formula. In Table 1 Stem width is taken as 10.

Results of various statistics like mean, mode, median, quartile, moments, skewness & kurtosis are calculated from raw, stem & leaf and grouped data and displayed in Table 2.

Table 1 Table showing calculation of mean, variance, moments, skewness and kurtosis for the data shown in appendix **(Stem unit=10, Leaf unit=1 with Width=10)**

Stem (S_k)	Leaf (l_{ki})	B_k	n_k	Cn_k	L_k	S_k	$\bar{l}_k$	$\bar{T}$	d_k	μ_{2k}	μ_{3k}	μ_{4k}	$\sqrt{\beta_{1k}}$	β_{2k}
5	0(2)1(2)2(2)3(2)4(3)5(2)6(2)7(2)8(3)9(2)	50	22	22	102	1100	4.636363636		-35.96363636	8.049587	-1.35237	116.5381	0.059215	1.341097
6	0(2)1(3)2(4)2(2)5(2)6(2)7(2)8(4)9(8)	60	24	46	143	1440	5.958333333		-24.64166667	9.373264	-13.6199	147.5074	0.474612	1.295734
7	0(1)2(2)3(2)4(2)5(4)6(4)7(2)8(5)9(2)	70	26	72	136	1820	5.230769231		-15.36923077	6.639053	-6.22394	91.54406	0.363837	1.441149
8	0(2)1(2)2(4)3(2)4(6)5(9)6(17)7(4)8(4)9(2)	80	51	123	264	4080	5.176470588		-5.423529412	4.22376	-5.35925	57.60066	0.617384	1.796861
9	0(1)3(2)9(3)8(4)5(12)6(11)7(2)8(8)9(4)	90	62	185	301	5580	4.85483871	90.6	4.25483871	5.543444	0.336494	63.51021	0.025781	1.437613
10	0(2)1(4)2(4)3(5)4(6)3(7)8(3)9(2)	100	24	209	102	2400	4.25		13.65	8.604167	3.03125	120.9049	0.120104	1.277948
11	0(2)1(2)2(2)3(2)4(5)2(6)2(7)2(8)2(9)2(2)	110	19	228	86	2090	4.526315789		23.92631579	8.67036	-0.67794	127.2557	0.026554	1.301073
12	03(2)5(6)6(4)7(8)9(2)	120	17	245	93	2040	5.470588235		34.87058824	4.602076	-5.07185	78.08898	0.513731	1.920176
13	012(4)345(2)	130	10	255	26	1300	2.6		42	2.44	0.792	12.2992	0.207798	1.437304
Combined Result of the statistic			255		1253	21850				419.84= μ_2	1140.154353= μ_3	449746.3514= μ_4	0.132537242= $\sqrt{\beta_1}$	2.551526= β_2

Here, B_k, l_{ki}, n_k, Cn_k, L_k, S_k, $\bar{l}_k$, $\bar{T}$, d_k, μ_{2k}, μ_{3k}, μ_{4k}, $\sqrt{\beta_{1k}}$, β_{2k}, μ_2, μ_3, μ_4, $\sqrt{\beta_1}$, β_2 represents base of original data in k^{th} Stem, observed i^{th} Leaf in k^{th} Stem, number of leaves corresponding to k^{th} Stem, k^{th} cumulative Leaf unit, Leaf sum for k^{th} Stem , sum of base of original data for k^{th} Stem, leaves averages in k^{th} Stem, deviations in k^{th} Stem, 2^{nd}, 3^{rd} and 4^{th} central moments, skewnes and kurtosis for k^{th} Stem, and over all data set moments, skewness and kurtosis respectively.

Table 2: Results of various statistics using raw, stem & leaf and grouped data for the overall data set

Statistics	Raw data	Stem and Leaf	Group data
		Class int. 10	Class int. 10
Mean	90.6	90.6	90.68627
Median	92	92	90.72580645
Mode	86	86	92.24489796
Q_1	77	77	76.82692308
Q_3	102	102	102.6441667
D_3	82	82	80.88235294
D_7	98	98	98.9516129
P_{38}	85	85	84.88235294
P_{62}	95	95	95.66129032
μ_2	419.84	419.84	426.882(Un adjusted) 418.5486667(adjusted)
μ_3	1140.154353	1140.154353	1526.1943
μ_4	449746.3514	449746.3514	460993.6326(Unadjusted) 439941.1993 (adjusted)
$\sqrt{\beta_1}$	0.132537242	0.132537242	0.173040639
β_2	2.551526	2.551526	2.529759909

7 Discussion

Generally, we use frequency distributions and their corresponding histograms to summarize large set of statistical data; it is unfortunate that the identity of the actual observations within the class interval is lost in the grouping process. But the Stem and Leaf display provides the same information as a frequency distribution and its histogram while presenting the actual recorded numerical values.

Validity and efficiency of the proposed formulae are checked by comparing numerical results from raw data and that from grouped data appended in Table 2 where we used the class interval 50 - 59, 60 - 69, $\cdots$ for grouped data. These results are valid for class intervals of any width and for fractional observations in the data set. Always it is found that the results of our proposed formulae are the same as those obtained from raw data. On the other hand, these results are different from that of grouped data. Thus one can consider Stem and Leaf method superior to that of grouped data analysis. In the stem & leaf method, one may enjoy the group format without loosing any information.

References

[1] Goel B.S.; Prakash, S.; Lal, R. (1991): "Mathematical Statistics" Pragati Prakation, Merut, India.

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

[3] Islam M. N. (2001): "An Introduction to Statistics and Probality" Book world, Dhaka, Bangladesh.

[4] Kapoor, J. N. and Saxena, H. C.(1986): "Mathematical Statistics" S. Chand & Company Ltd., New Delhi.

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

[6] Walpole, R. E.; (1983): Elementary Statistical Concepts, 2nd edition, Macmillan Publishing Co., Inc., New York.

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

[8] Weatherburn, C.E.(1986): "A First Course in Mathematical Statistics" S. Chand & Company Ltd., New Delhi.

Appendix

Marks obtained out of 150 by 255 candidates in a University Admission Test

Serial No.	Marks	Serial No.	Marks	Serial No.	Marks	Serial No.	Marks	Serial No.	Marks
1	123	55	91	109	98	163	95	217	69
2	95	56	96	110	129	164	86	218	86
3	87	57	99	111	119	165	97	219	69
4	52	58	108	112	119	166	58	220	95
5	127	59	93	113	95	167	75	221	93
6	98	60	112	114	88	168	86	222	76
7	102	61	50	115	92	169	57	223	85
8	103	62	88	116	82	170	95	224	94
9	68	63	96	117	99	171	86	225	55
10	113	64	84	118	51	172	102	226	52
11	96	65	76	119	126	173	132	227	114
12	110	66	85	120	83	174	80	228	86
13	105	67	71	121	64	175	76	229	95
14	125	68	93	122	109	176	89	231	94
16	56	70	135	124	84	178	126	232	96
17	107	71	62	125	73	179	85	233	98
18	87	72	54	126	126	180	128	234	116
19	111	73	99	127	96	181	85	235	118
20	130	74	115	128	85	182	89	236	98
21	95	75	95	129	60	183	125	237	78
22	97	76	86	130	68	184	58	238	87
23	120	77	72	131	92	185	88	239	98
24	110	78	108	132	73	186	96	240	78
25	98	79	84	133	54	187	76	241	98
26	93	80	79	134	95	188	123	242	78
27	113	81	83	135	86	189	85	243	85
28	132	82	117	136	61	190	95	244	96
29	125	83	64	137	53	191	58	245	96
30	102	84	92	138	86	192	69	246	85
31	105	85	69	139	51	193	86	247	74
32	92	86	93	140	125	194	75	248	116
33	68	87	115	141	96	195	93	249	96
34	91	88	88	142	134	196	125	250	86
35	131	89	92	143	62	197	101	251	84
36	101	90	56	144	79	198	82	252	75
37	111	91	66	145	70	199	100	253	95
38	101	92	106	146	102	200	86	254	86
39	91	93	71	147	135	201	92	255	101
40	118	94	82	148	65	202	72		
41	81	95	93	149	55	203	86		
42	63	96	80	150	92	204	90		
43	69	97	106	151	78	205	68		
44	92	98	82	152	69	206	93		
45	78	99	117	153	57	207	86		
46	133	100	74	154	77	208	109		
47	92	101	86	155	86	209	112		
48	77	102	50	156	95	210	105		
49	129	103	126	157	99	211	84		
50	54	104	94	158	100	212	108		
51	98	105	53	159	85	213	96		
52	63	106	69	160	132	214	86		
53	94	107	125	161	84	215	59		
54	87	108	62	162	59	216	67		