ADDITIONAL LITERATURE

Besag, J. (1981). "On resistant techniques and statistical analysis," *Biometrika*, **68**, 463–469.

Hampel, F. R. (1971). "A general qualitative definition of robustness," *Annals of Mathematical Statistics*, **42**, 1887–1896.

Tukey, J. W. (1972). "Data analysis, computation, and mathematics," *Quarterly of Applied Mathematics*, **30**, 51–65.

——— (1979). "Robust techniques for the user." In R. L. Launer and G. N. Wilkinson (Eds.), *Robustness in Statistics*. New York: Academic, pp. 103–106.

——— (1980). "We need both exploratory and confirmatory," *The American Statistician*, **34**, 23–25.

CHAPTER 1

Stem-and-Leaf Displays

John D. Emerson
Middlebury College

David C. Hoaglin
Harvard University and Abt Associates Inc.

The most common data structure is a batch of numbers. Even this simple structure of data may have characteristics not easily discerned by scanning or studying the numbers. The stem-and-leaf display enables us to organize the numbers graphically in a way that directs our attention to various features of the data. This basic but versatile exploratory technique is the most widely used; we call upon it in later chapters, for example, to compare batches and to examine residuals.

The stem-and-leaf display enables us to see the batch as a whole and to notice such features as:

How nearly symmetric the batch is.
How spread out the numbers are.
Whether a few values are far removed from the rest.
Whether there are concentrations of data.
Whether there are gaps in the data.

In exposing the analyst to these features of data, the stem-and-leaf display has much in common with its close relative, the histogram. By using the digits of the data values themselves instead of merely enclosing area, this display offers advantages in some situations. When we work by hand, it is easier to construct, and it takes a major step in sorting the data. Thus we can readily find the median and other summaries based on the ordered batch. It can help us to see the distribution of data values within each interval, as well as patterns in the data values. For example, we might discover that all values are multiples of 3 or that the person recording motor vehicle speeds reported them as multiples of 5 miles per hour. By preserving more early digits of the data values, the stem-and-leaf display also shortens the link back to the individual observation and any identifying information that accompanies it.

The stem-and-leaf display does not involve any elaborate theory. Rather, because the interaction between analyst and display is a personal one, we face considerations of taste and esthetics, particularly in choosing the number of intervals or the interval width. As often happens in exploratory data analysis, we aim to provide flexibility through a number of variations. We begin with the basic stem-and-leaf display and the steps in its construction, and then we add variations. For historical interest, we reproduce a close predecessor of the stem-and-leaf display. We also briefly pursue the connections with sorting to put the data in order. Choosing the number of lines (or the interval width) provides a point of contact with work (some of it more theoretical) on these choices for histograms.

1A. THE BASIC DISPLAY

The stem-and-leaf display (Tukey, 1970, 1972) provides a flexible and effective technique for starting to look at a batch or sample of data. The most significant digits of the data values themselves do most of the work of sorting the batch into numerical order and displaying it.

To explain the display and how one constructs it, we begin with an example. Royston and Abrams (1980) give the mean menstrual cycle length and the mean preovulatory basal body temperature for 21 healthy women who were using natural family planning. We reproduce these data in Table 1-1 and show one stem-and-leaf display for the cycle lengths in Figure 1-1.

If we follow the first data value, 22.9, we see that it appears in the display as 22 | 9. To construct this simplest form of stem-and-leaf display, we proceed as follows. First we choose a suitable pair of adjacent digits in the data—in the example, the ones digit and the tenths digit. Next we split each data value between these two digits:

<table>
<tr><td>data value</td><td></td><td>split</td><td></td><td>stem</td><td>and</td><td>leaf</td></tr>
<tr><td></td><td></td><td>↓</td><td></td><td></td><td></td><td></td></tr>
<tr><td>22.9</td><td>→</td><td>22 | 9</td><td>→</td><td>22</td><td>and</td><td>9</td></tr>
</table>

TABLE 1-1. Mean menstrual cycle length and mean preovulatory basal body temperature (BBT) for 21 women.

i	Cycle Length (days)	BBT (°C)
1	22.9	36.44
2	26.3	36.21
3	26.6	36.71
4	26.8	36.13
5	26.9	36.25
6	26.9	36.53
7	27.5	36.41
8	27.6	36.45
9	27.6	36.53
10	28.0	36.31
11	28.4	36.63
12	28.4	36.54
13	28.5	36.52
14	28.8	36.62
15	28.8	36.40
16	29.4	36.48
17	29.9	36.39
18	30.0	36.37
19	30.3	36.77
20	31.2	36.76
21	31.8	36.50

Source: J. P. Royston and R. M. Abrams (1980). "An objective method for detecting the shift in basal body temperature in women," *Biometrics*, **36**, 217–224 (data from Table 1, p. 221, used by permission).

Then we allocate a separate line in the display for each possible string of leading digits (the *stem*)—for Figure 1-1, the necessary lines (10 in all) run from 22 to 31. Finally we write down the first trailing digit (the *leaf*) of each data value on the line corresponding to its leading digits.

The finished display, Figure 1-2, includes a reminder that all data values are in units of .1 day, as well as a column of depths (which we will define shortly) to the left of the stems. When the raw data values have not been sorted, the initial display will not usually have its leaves in increasing order. As an option, the final display can then sort the leaves. This can happen

(unit = .1 day)	
22	9
23	
24	
25	
26	3 6 8 9 9
27	5 6 6
28	0 4 4 5 8 8
29	4 9
30	0 3
31	2 8

Figure 1-1. One stem-and-leaf display for mean menstrual cycle length.

automatically when a computer produces the display (Velleman and Hoaglin, 1981). In overall appearance the display resembles a histogram with an interval width of 1 day; the leaves add numerical detail, and in this instance they preserve all the information in the data.

Figures 1-1 and 1-2 show that two-thirds of the women have mean cycle length between 26.3 and 28.8 days. All but one of the other third have longer mean cycle lengths. One woman, who appears unusual, has a mean cycle length of only 22.9 days. We regard such an unusual data point as an outlier, and we might well treat it separately from the other data points. If we do, then a stem-and-leaf display for the remaining cycle lengths can give more detail by choosing the stems differently. We return to this example in Section 1B.

Depths of Data Values

A data value can be assigned a *rank* by counting in from each end of the ordered batch. For example, in Figure 1-2, 26.3 has rank 2 when counting up from 22.9 and rank 20 when counting down from 31.8. The *depth* of the data value is the smaller of these two ranks, 2 in the example. Because a number of summary values (such as the median and the quartiles or fourths, see Chapter 2) can easily be defined in terms of their depths, it is helpful to present a set of depths with the display. Except for one middle line, the number in the depth column is the maximum depth associated with data values on that line. Thus the depth of 29.4 is 6.

The "middle line" includes the median, and the depth column shows in parentheses the number of leaves on this line. In the example, this number is 6. (When the batch size is even *and* the median falls between lines, we do not need this special feature.) If the display has been prepared by hand,

Depths	(unit = .1 day)	
1	22	9
	23	
	24	
	25	
6	26	3 6 8 9 9
9	27	5 6 6
(6)	28	0 4 4 5 8 8
6	29	4 9
4	30	0 3
2	31	2 8

Figure 1-2. A stem-and-leaf display with sorted leaves for mean menstrual cycle length ($n = 21$).

adding the count on the middle line and the depths on the two adjacent lines provides a simple check that no data values have been omitted. In the example, $9 + 6 + 6 = 21$, the total number of women in the study. (Incidentally, care is needed when determining the depths of numbers above the median: in the example one might tend, erroneously, to give 4, not 3, as the depth of 30.3.) Chapters 2 and 3 use the depths further in forming other summaries of data.

Organizing the Display

An effective choice of the number of lines in a stem-and-leaf display involves the number of data values in the batch and the range to be covered, as well as some judgment. To get started, we set a maximum number of lines:

$$L = [10 \times \log_{10} n], \tag{1}$$

where n is the number of data values and $[x]$ is the largest integer not exceeding x. This rule seems to give values of L that produce effective displays over the range $20 \leqslant n \leqslant 300$, where most applications fall. For the example, which has $n = 21$, it gives

$$\begin{aligned} L &= [10 \times \log_{10} 21] \\ &= [10 \times 1.32] \\ &= 13. \end{aligned}$$

Thus unless we decide to treat the value 22.9 as special (as we often would), the 10-line stem-and-leaf display of Figures 1-1 and 1-2 seems satisfactory.

Values of n smaller than 20 may need special treatment. These are also more likely to arise when comparing several batches in parallel stem-and-leaf displays, a situation we would want to handle differently anyway (see Section 1B and Chapter 3). Batches of 300 or so are usually cumbersome in a stem-and-leaf display, but the rule should still cope with them reasonably well.

Using L as a rough limit on the number of lines in the display, we must now determine the interval of values corresponding to each line. The simple way to do this uses a power of 10 as the interval width. (We discuss other interval widths in Section 1B.) Thus we divide R, the range of the batch, by L and round the quotient up, if necessary, to the nearest power of 10. In the example, the range $R = 31.8 - 22.9 = 8.9$ and $L = 13$, so that $R/L = .68$. Rounding up to the nearest power of 10 gives the value 1 as the interval width. This is the value used in Figures 1-1 and 1-2, which actually require only 10 lines.

1B. SOME VARIATIONS

A segment of stems for the basic stem-and-leaf display might look like this:

```
0 |
1 |
2 |
3 |
```

and each line may receive leaves 0 through 9. But sometimes this format is too crowded, having too many leaves per line. One effective response is to split lines and repeat each stem:

```
0 * |
0 · |
1 * |
1 · |
2 * |
2 · |
```

putting leaves 0 through 4 on the * line and 5 through 9 on the · line. In such a display, the interval width is 5 times a power of 10.

EXAMPLE: HARDNESS OF ALUMINUM DIE CASTINGS

Shewhart (1931, p. 42) gives the hardness of 60 aluminum die castings. For the first 30 of these, the data values and two stem-and-leaf displays appear in Figure 1-3. We note that $L = [10 \times \log_{10} 30] = [14.77] = 14$, $R = 95.4 - 50.7 = 44.7$, and $R/L = 44.7/14 = 3.19$. If we rounded 3.19 up to the nearest power of 10, we would obtain 10 as the indicated interval width. This width is used for the basic stem-and-leaf display on the left in Figure 1-3. Because this display has relatively few lines, we split the lines and repeat each stem. This, of course, corresponds to rounding 3.19 up to 5. The result, which we feel is an improvement, appears at the right in Figure 1-3.

The stem-and-leaf displays of Figure 1-3 also illustrate how we handle digits that come after those that serve as the leaves in the display. Such low-order digits of data values are preferably truncated rather than rounded. This practice makes it easier to recover the original data value corresponding to a leaf in the display. Thus the values 55.3, 55.7, and 55.7 are all truncated at the decimal point and appear as 5s on the 5 · line in the display; the values at 55.7 are not rounded up to 56. To recover the three data values, we simply locate the three numbers in the raw data set whose first two digits are 55.

Sometimes the display is still too crowded with two lines per stem and too straggly with one line per stem at the next lower power of 10. To cure

Data: Hardness of aluminum die castings

53.0	70.2	84.3	55.3	78.5	63.5	71.4	53.4
82.5	67.3	69.5	73.0	55.7	85.8	95.4	51.1
74.4	54.1	77.8	52.4	69.1	53.5	64.3	82.7
55.7	70.5	87.5	50.7	72.3	59.5		

Displays:

```
(n = 30)                                    (n = 30)
 Depths     (unit = 1)                       Depths     (unit = 1)
  11        5 | 0 1 2 3 3 3 4 5 5 5 9          7        5 * | 0 1 2 3 3 3 4
  (5)       6 | 3 4 7 9 9                     11        5 · | 5 5 5 9
  14        7 | 0 0 1 2 3 4 7 8               13        6 * | 3 4
   6        8 | 2 2 4 5 7                     (3)       6 · | 7 9 9
   1        9 | 5                             14        7 * | 0 0 1 2 3 4
              |                                8        7 · | 7 8
              |                                6        8 * | 2 2 4
              |                                3        8 · | 5 7
              |                                         9 * |
              |                                1        9 · | 5
```

Figure 1-3. Splitting to get two lines per stem. *Source*: W. A. Shewhart (1931). *Economic Control of Quality of Manufactured Product*. Princeton, NJ: D. Van Nostrand, Inc. [data from Table 3 (specimens 1 through 30), p. 42].

these troubles, we have a third form, five lines per stem:

```
0 * |
  t |
  f |
  s |
0 · |
```

with leaves 0 and 1 on the * line, 2 (two) and 3 (three) on the **t** line, 4 (four) and 5 (five) on the **f** line, 6 (six) and 7 (seven) on the **s** line, and 8 and 9 on the · line. As a reminder in starting to place leaves, the three lettered lines contain leaves whose words begin with that letter. Here the interval width is 2 times a power of 10.

EXAMPLE: TUMOR PROGRESSION IN PATIENTS WITH GLIOBLASTOMA

Dinse (1982) gives the times until tumor progression for 172 patients with glioblastoma, a brain tumor. The times for 83 of the patients are *censored* times in that the tumors for these patients had not progressed at the time the study was concluded. These times are indicated by "+" in the data portion of Figure 1-4, which also shows them in a stem-and-leaf display with 5 lines per stem. The display has a total of 19 lines, and each line has width 2 months. Note that for this example, $L = [10 \log_{10} 83] = 19$ lines, $R = 37 - 1 = 36$ months, and $R/L = 36/19 = 1.89$. When this quotient is rounded up to 1, 2, or 5 times the nearest power of 10, we obtain 2×10^0, or 2, as the interval width. This is consistent with the width adopted for the display.

In this display we immediately see the asymmetry of the main part of the data, the clump of values from 30 months to 37 months, and the single value at 25 months. The large number of small censoring times most likely indicates that these patients have not been participating in the study long enough to suffer a progression of their tumors. The practice of recording these times in whole months, so that times within the first month are recorded as 1, explains why there are no values at 0, and it may mean that progression times and censoring times are both rounded up. The times at 25 months and beyond do not fit into a regular pattern with the rest of the censoring times; they may represent an initial bulge in the rate of entry of patients into the study.

Data: Time to tumor progression (in months) for patients with glioblastoma

```
9+, 3+, 6+, 6, 5, 34+, 10, 22, 9, 2, 14+, 3, 9, 6+, 8, 8+, 3, 3, 11,
4+, 9, 5, 17+, 9+, 17, 13+, 3, 5, 3+, 14, 11+, 3+, 9+, 13, 15+, 3,
3+, 4+, 11, 3, 1+, 9, 16, 14+, 6, 2+, 24, 22, 10, 34+, 10, 4, 1, 3, 15+,
6+, 28, 3+, 4, 31+, 6, 2, 9+, 4+, 13+, 21, 8+, 11+, 37+, 6, 1+, 4+, 15+,
7, 4, 3, 19, 2+, 18+, 9+, 6, 9+, 9, 10+, 35+, 23, 33+, 16+, 5+, 34+,
13+, 2, 12+, 3+, 10, 8, 3+, 4+, 1+, 7, 3, 8, 9+, 10+, 10, 6+, 10, 3, 1+, 5, 4, 2,
1, 5, 4+, 5+, 1+, 2, 6, 3, 7, 1+, 7+, 10+, 6, 2+, 11+, 5, 10+, 9+, 18, 3+, 6, 4+,
2, 7, 25+, 2, 30+, 2+, 4, 13+, 5, 19, 9, 5+, 4, 32+, 23, 19, 10+, 5+, 6, 9, 13+,
5, 13, 1, 15, 4+, 8, 9+, 20+, 16+, 19, 8+, 4, 7+, 5, 5, 7+, 6
```

Stem-and-leaf: Times at which patients' observations were censored ($n = 83$)

```
Depths      Censoring time (unit = 1 month)

   6        0 * | 1 1 1 1 1 1
  18          t | 2 2 2 2 3 3 3 3 3 3 3 3
  30          f | 4 4 4 4 4 4 4 4 5 5 5 5
  37          s | 6 6 6 6 7 7 7
 (12)       0 · | 8 8 8 9 9 9 9 9 9 9 9 9
  34        1 * | 0 0 0 0 0 1 1 1
  26          t | 2 3 3 3 3 3
  20          f | 4 4 5 5 5
  15          s | 6 6 7
  12        1 · | 8
  11        2 * | 0
              t |
  10          f | 5
              s |
            2 · |
   9        3 * | 0 1
   7          t | 2 3
   5          f | 4 4 4 5
   1          s | 7
```

Figure 1-4. A stem-and-leaf display with five lines per stem. *Source*: G. E. Dinse (1982). "Nonparametric estimation for partially-complete time and type of failure data," *Biometrics*, **38**, 417–431 (data from Table 1, p. 426, used by permission).

Another variation in the stem-and-leaf display accommodates data that include both positive and negative values. Generally, residuals (to which we devote Chapter 7) are centered at 0, and some other types of data take both signs. When we start to display such data, we see that they require a −0 stem. The only tricky detail is that values exactly equal to 0 belong to either or both the −0 stem and the +0 stem, so that we usually share them roughly equally between the two.

To illustrate this variation, we use the residuals from fitting a straight line (by the "resistant line" method discussed in Chapter 5) to the basal temperature data of Table 1-1. Here

$$y = \text{basal body temperature},$$
$$x = \text{cycle length},$$

and the fitted line is

$$\hat{y} = .02813x + 35.68.$$

Figure 1-5 shows the stem-and-leaf display, with the residuals in units of .01°C. We notice some tendency for the values to pile up around zero, and the value at −.30 attracts some attention. Together with the one at +.28, it probably deserves a closer look.

In Figure 1-5 and in other stem-and-leaf displays that involve both positive and negative values, we have chosen, perhaps arbitrarily, to have the data values increase from the top of the display toward the bottom. We could equally well handle plus and minus by having the entries increase from bottom to top, and others may prefer this direction. In any one book or paper, however, we usually adopt one of these directions.

Resistance

Resistant methods are little affected by a small fraction of unusual data values and are an important part of exploratory data analysis. Thus it is unwise for the scale of a stem-and-leaf display to depend on the largest and smallest data values. (The stem-and-leaf display of the mean cycle lengths in Figure 1-2 is clearly influenced by the unusually small value at 22.9 days.) Instead, we often begin by setting aside any unusual data values, and we then base the choice of scale for the display on the rest of the data. (One rule of thumb for setting aside low and high values appears in Section 2C.) We list those outlying values on the lines labeled "low" and "high," beyond the set of stems. To emphasize further the separate treatment of these values, we may place parentheses around the lists. A comma after each value in the list serves as a reminder that these entries are data values (represented as multiples of the unit in the display) and not strings of leaves.

Depths (n = 21)	unit = .01°C	
1	−3	0
2	−2	0
6	−1	3 5 5 8
(5)	−0	0 2 4 7 9
10	0	3 6 7 9
6	1	1 2 5
3	2	0 3 8

Figure 1-5. Stem-and-leaf display of residuals from a line for mean basal body temperature against mean cycle length.

EXAMPLE: MENSTRUAL CYCLE LENGTHS

The clear separation between the rest of the data and the value at 22.9 days, which we saw in Figure 1-2, suggests that we place that one value on a line labeled "low." To avoid confusion, we recommend putting the entries on such lines in parentheses, as well as separating individual entries by commas. The first display in Figure 1-6 does this. By avoiding the three empty lines in Figure 1-2, we focus more attention on the bulk of the data at the expense of no longer showing just how far the value at 22.9 stands off from

One Line per Stem with "Low" Line

Depths (n = 21)	(unit = .1 day)	
1	low	(229,)
6	26	3 6 8 9 9
9	27	5 6 6
(6)	28	0 4 4 5 8 8
6	29	4 9
4	30	0 3
2	31	2 8

Two Lines per Stem with "Low" Line

Depths (n = 21)	(unit = .1 day)	
1	low	(229,)
2	26 *	3
6	26 ·	6 8 9 9
	27 *	
9	27 ·	5 6 6
(3)	28 *	0 4 4
9	28 ·	5 8 8
6	29 *	4
5	29 ·	9
4	30 *	0 3
	30 ·	
2	31 *	2
1	31 ·	8

Figure 1-6. Two stem-and-leaf displays for mean cycle length.

the rest. We can now calculate the scale that formula (1) suggests for the remaining 20 data values:

$$L = [10 \times \log_{10} 20] = 13$$

$$R = 31.8 - 26.3 = 5.5$$

$$\frac{R}{L} = .42$$

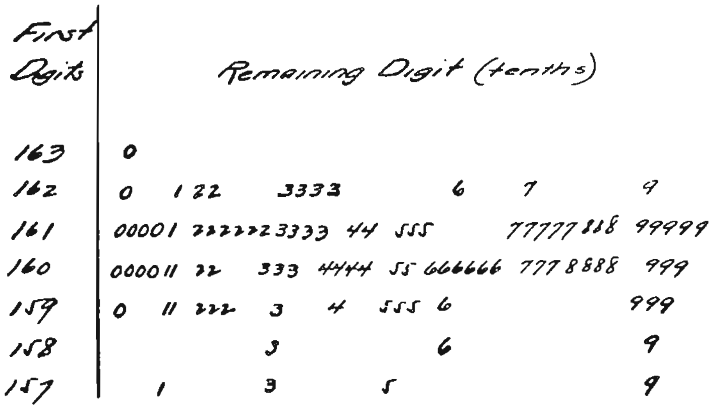
and we round R/L up to .5—a display with two lines per stem. The second part of Figure 1-6 shows the result, which probably has too many lines for some viewers. Both displays list the 22.9 separately, and the choice between them is a matter of taste.

With its variations, the stem-and-leaf display has proved to be a versatile technique for the analyst's first look at a batch of numbers. The three ways of factoring 10 (1 × 10, 2 × 5, and 5 × 2) provide adequate control over scaling (1, 2, or 5 lines per stem) especially when combined with "low" and "high" lines for unusual data values. Setting aside potential outliers in this way often leads to a more detailed and more effective display; it also focuses attention on the unusual data values so that we do what we can to probe their surrounding circumstances for clues.

1C. AN HISTORICAL NOTE

The histogram, the better known relative of the stem-and-leaf display, has been in use for many years. Beniger and Robyn (1978) trace the origin of the term to Karl Pearson in 1895, and they mention earlier examples going back to the bar chart published by William Playfair in 1786 in his *Commercial and Political Atlas.*

Of course, the primary difference between the histogram and the stem-and-leaf display is the use of a digit from each data value to form the display. Without any systematic search for possible predecessors of the stem-and-leaf display, we report finding a digit-based display (Figure 1-7) in the text by Dudley (1946, p. 22). In the way it groups matching "leaves" together and spreads the groups out over the line, this technique serves more as a sorting and tallying device than as a semigraphic display. Still, its similarity to the stem-and-leaf display emphasizes the convenience of working with digits from the data values.



First Digits	Remaining Digit (tenths)
163	0
162	0 1 22 3333 6 7 9
161	00001 2222223333 44 555 77777888 99999
160	000011 22 333 4444 55 666666 7778888 999
159	0 11 222 3 4 555 6 999
158	3 6 9
157	1 3 5 9

Note: When necessary, allow two or more lines for each item of left-hand column.

Figure 1-7. Dudley's transcription of data in order of magnitude. From *Examination of Industrial Measurements* by John W. Dudley, Jr. Copyright © 1946 by McGraw-Hill Book Company, Inc. Used with permission of McGraw-Hill Book Company.

The role and development of graphical methods in statistics have received considerable attention in recent years. Exploratory data analysis has contributed several novel displays. Fienberg (1979) discusses the history and use of graphical methods and includes examples of several recent innovations. [See also Wainer and Thissen (1981).]

1D. SORTING

Because many exploratory techniques work with the ordered observations—the simplest way of gaining resistance—we devote this section to a discussion of sorting, the mechanical process of putting a set of numbers in order (usually, from smallest to largest). Readers who care little about these details may skip this section without loss.

In hand work, as we mentioned in Section 1A, constructing a stem-and-leaf display accomplishes much of the task of sorting the data. All that remains is to rearrange the leaves on each line. Technically, it is a form of "bucket sort" or "radix sort," with the lines playing the role of the "buckets." The skeleton structure of stems is easy to set up after a glance at the data establishes the range, and then one quickly places each leaf on its