

2: Frequency Distributions

Stem-and-Leaf Plots (Stemplots)

The **stem-and-leaf plot (stemplot)** is an excellent way to begin an analysis. Consider this small data set:

218 426 53 116 309 504 281 270 246 523

To construct a stemplot, start by drawing the **stem**. Stem-values represent *either* the first *or* first-two significant digits of each value. As a rule-of-thumb, we want between 3 and 15 stem-values on the stem. You are forming a number line, so stem-values must be evenly spaced and no stem-value can be skipped. Values for our data range from 53 to 523, so a reasonable first approximation for the stem is:

```
0 |  
1 |  
2 |  
3 |  
4 |  
5 |  
×100
```

A **stem-multiplier** is included to allow the reader to decipher the magnitude of values. The stem-multiplier of $\times 100$ on this stem shows that a stem value of 2 represents about two hundred (and not two or twenty, etc.). Values between 200 and 299 will be stored next to the “stem bin” of 2.

We plot the subsequent significant digit of each value, so we plot the “tens” place, truncating remaining significant digits, if any.* For example, 218 is plotted as:

```
0 |  
1 |  
2 | 1  
3 |  
4 |  
5 |  
×100
```

The remaining leaves are plotted in rank order:

```
0 | 5  
1 | 1  
2 | 1478  
3 | 0  
4 | 2  
5 | 02  
×100
```

Note the value of 53 in the above plot is shown as 0 | 5 (0 in the hundreds-place and 5 in the tens-place).

* These rules differ from the simplified rules taught in California public school and are instead based on the original intention and intent of John W. Tukey in the groundbreaking book *Exploratory Data Analysis* (1977).

Shape, location, and spread

Shape

I'm going to rotate the plot 90 degrees to display the distribution in a more familiar way.

```
      8
      7
      4      2
5 1 1 0 2 0
-----
0 1 2 3 4 5 (x100)
-----
```

This is now just a **histogram**. Note that the batch of numbers forms a **distribution** with a shape, location, and spread.

The **shape** of a distribution is seen as a “skyline silhouette”:

```
      X
      X
      X      X
X X X X X X
-----
0 1 2 3 4 5 (x100)
-----
```

Describe the shape narratively. While it is difficult to make reliable statements about the shape of a distribution when the data set is small, you can still get a general impression of whether (A) a mound is present, (B) data are symmetrical, and (C) if there are any data that separate from the rest of the distribution (i.e., outliers). The current stemplot is mound-shaped and is [almost] symmetrical. There are no clear outliers.

Note:

- (1) When the data set is this small, keep a soft focus when describing “shape.” If a shift in just a few data points can change your impression of the shape, *avoid overstatements*.
- (2) Identification of outliers depends on context and can be subjective. There is no uniform definition of an “outlier.” See <http://www.tufts.edu/~gdallal/out.htm> for further remarks (optional for now).

Location

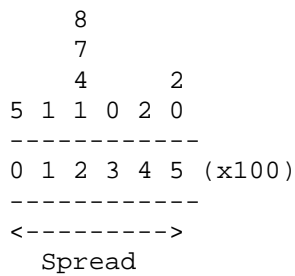
We will simply **summarize the central location** of a distribution by its median. The **median** is the value that is greater than or equal to half of the values in the data set. It's easier to see the median if we stretch-out the data in rank-order:

```
053    116    218    246    270    281    309    426    504    523
                        ^
                        Median
```

The median will have a **depth** of $(n + 1) / 2$. For the current data, $n = 10$ and median has a depth of $(10 + 1) / 2 = 5.5$. Count in from either the top or bottom of the ordered array to the depth of the median. When n is even, the median falls between two values, as it is here. Under these circumstances, **interpolate** the median as the average of the two adjacent values. In this case, the median = average(270, 281) = 275.5.

Spread

The spread of the distribution is its dispersion around its center:



The data above spread from around 50 to 520. Do not be too concerned about the loss of the 3rd significant digit. Most of the information in is in the first two significant digits and, at this point in time, we are just exploring the data.

The technical term for “spread” is **variability**.

Second stemplot example

The next illustration shows how to draw a stemplot for data that might not immediately lend itself to plotting. Consider:

1.47 2.06 2.36 3.43 3.74 3.78 3.94 4.42

Since data spread from about 1.4 to 4.4, we use a stem-multiplier of $\times 1$ and draw the stem:

```
1 |
2 |
3 |
4 |
×1
```

We *truncate* extra digits before plotting, so that 1.47 is truncated to 1.4, 2.06 is truncated to 2.0, and so on. The decimal point should not be displayed because the next digit is automatically in the tenths place.

```
1 | 4
2 | 03
3 | 4779
4 | 4
×1
```

Narrative interpretation: This stemplot is [almost] symmetrical, mound-shaped, and has no apparent outliers. The median has a location of $(8+1)/2 = 4.5$ and is between 3.4 and 3.7 (that’s good enough for now). Data spread from around 1.4 to 4.4.

Third stemplot example (split stem-values)

The following pollution levels were found in water samples {1.4, 1.7, 1.8, 1.9, 2.2, 2.2, 2.3, 2.4, 2.6, 2.6, 2.7, 2.8, 2.9, 3.0, 3.0, 3.0, 3.1, 3.2, 3.3, 3.4, 3.4, 3.5, 3.6, 3.7, 3.8}. Here is our first cut at the stemplot:

```
1 | 4789
2 | 223466789
3 | 000123445678
×1
```

The above stemplot is too squashed to display the distribution's shape. Let's try "splitting" the stem-values so that values between 1.0 and 1.4 are listed on the first stem-value of 1 and values between 1.5 and 1.9 are listed on the second stem-value of 1:

```
1 | 4
1 | 789
2 | 2234
2 | 66789
3 | 00012344
3 | 5678
×1
```

Narrative interpretation: This distribution has a tail toward its lower values. This "left tail" is called a "left" or "negative skew." The median has a depth of $(25+1) / 2 = 13$. Count the leaves from either end of the stemplot and you will see that the median is approximately equal to 2.9. Data spread from 1.4 to 3.8.

Fourth stemplot example (quintuple split)

Sometimes splitting stem-values in two still leaves an unclear picture of shape. Consider these 9 values:

17 35 40 49 55 70 74 81 84

You may be tempted to start the process as follows:

```
1 | 7
2 |
3 | 5
4 | 09
5 | 5
6 |
7 | 04
8 | 14
×10
```

This is too spread out to draw out the shape, so you try a stem-multiplier $\times 100$ with split the stem-values. We'll put values between 0 and 49 on the first 0 and values between 50 and 99 on the second 0:

```
0 | 1344
0 | 57788
×100
```

Now try a quintuple-split. This means you will divide the range 0 to 99 into five class intervals, each 20 units in width. The first interval will contain values between 0 and 19, the second interval will contain values between 20 and 39, and so on. The stem will store the hundreds place and the leaves represent the tens place. The ones-place will be truncated. Thus:

```
0 | 1
0 | 3
0 | 445
0 | 77
0 | 88
×100
```

This distribution is relatively symmetrical with no outliers.

Frequency Tables

Frequency table example #1

A more traditional way to explore a distribution is in tabular form. It's easier to show you how to construct a frequency table than to provide formulas. In addition, it is best *not to be mechanical* in our approach toward statistics.

If you've already created a stemplot, the hard work is behind you. Here's the stemplot for the first illustrative data set in this chapter:

```
0 | 5
1 | 1
2 | 1478
3 | 0
4 | 2
5 | 02
×100
```

The stemplot has already sorted data into **class intervals**. The class intervals are 100 units in width. The first class interval contains values from 0 to 99, the second class interval contains values of 100 to 199, and so on.

Count the number of observations that fall into each class interval. This is the **frequency**. Also determine the **relative frequency** of each count. The relative frequency is just the proportion. It doesn't matter if you report the relative frequency as a proportion or percentage, as long as its labeled clearly.

Class interval	Frequency	Relative frequency
0 – 99	1	10%
100 – 199	1	10%
200 – 299	4	40%
300 – 399	1	10%
400 – 499	1	10%
500 - 599	2	20%
Total	10	100%

An additional concept worth noting is called cumulative frequency. The **cumulative frequency** is the frequency up to and including the current interval. This can be reported as a count or as a proportion of the total (**cumulative relative frequency**), as shown in the table on the next page.

	Cumulative Frequency	Cumulative relative frequency
<= 99	1	10%
<= 199	2	20%
<= 299	6	60%
<= 399	7	70%
<= 499	8	80%
<= 599	10	100%

Frequency table example #2 (large dataset)

A frequency table of AGES (years) from a childhood health survey is:

AGE	Freq	Rel.Freq	Cum.Freq.
3	2	0.3%	0.3%
4	9	1.4%	1.7%
5	28	4.3%	6.0%
6	37	5.7%	11.6%
7	54	8.3%	19.9%
8	85	13.0%	32.9%
9	94	14.4%	47.2%
10	81	12.4%	59.6%
11	90	13.8%	73.4%
12	57	8.7%	82.1%
13	43	6.6%	88.7%
14	25	3.8%	92.5%
15	19	2.9%	95.4%
16	13	2.0%	97.4%
17	8	1.2%	98.6%
18	6	0.9%	99.5%
19	3	0.5%	100.0%
Total	654	100.0%	

Narrative description and interpretation. You might be asked to make narrative interpretations based on data in a frequency table. The proper interpretation will depend on context, so it must be reasonable. There are no mechanical rules. For the above data, we see that ages range from 3 to 19. The most common in this data set is 9, representing 14.4% of the kids. The most common value (9, in this instance), is called the mode.

Frequency table example #3 (Non-uniform class intervals)

It is occasionally useful to create a frequency table with **non-uniform class intervals**. For example, here are the ages grouped into intervals according to school level.

AGE RANGE, YRS (SCHOOL AGE)	Freq	RelFreq	CumFreq
3 - 4 (PRE)	11	1.7%	1.7%
5 - 11 (ELEM)	469	71.7%	73.4%
12 - 13 (MIDDLE)	100	15.3%	88.7%
14 - 19 (HIGH)	74	11.3%	100.0%
Total	654	100.0%	