

2: Frequency Distributions

Stem-and-Leaf Plots

The **stem-and-leaf plot (stemplot)** is an excellent way to begin an analysis. Consider this small data set containing 10 AGE values:

21 42 05 11 30 50 28 27 24 52

To construct a stemplot, start by dividing each value into a **stem component** and **leaf component**. For these data, digits in the tens-place become stem components and digits in the units-place become leaf components. For example, "21" has a stem component of 2 and leaf component of 1.

Stem-values are listed in numerical order to form an axis. Vertical lines may be drawn to outline the stem:

```
0 |  
1 |  
2 |  
3 |  
4 |  
5 |  
×10
```

An **axis-multiplier** is included to allow the reader to decipher the magnitude of values. Here, the multiplier ($\times 10$) is used to show that a stem value of 5 represents 50 and not 5, for instance.

The value of each leaf is plotted in its appropriate location. For example, 21 is plotted as:

```
0 |  
1 |  
2 | 1  
3 |  
4 |  
5 |  
×10
```

The remaining leaves are plotted, preferably in rank order:

```
0 | 5  
1 | 1  
2 | 1478  
3 | 0  
4 | 2  
5 | 02  
×10
```

This plot resembles a histogram on its side. 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 (x10)
-----

```

Central location: The central location of the data may be described in one of two ways: by its gravitational center or middle-point. Here, the gravitational center is shown between 20 and 30.

```

      8
      7
      4      2
5 1 1 0 2 0
-----
0 1 2 3 4 5 (x10)
-----
      ^
Gravitational
Center

```

This is only the approximate **gravitational center**. The exact gravitation is the **arithmetic average** of the data set, which in this case is 29.0.

The center of the distribution can also be described in terms of its **middle point** or **median**. It's easier to see this middle point if we stretch-out the dataset with data in rank-order. Here's the data stretched-out in rank order:

```

      5      11      21      24      27      28      30      42      50      52
                        ^
Middle Point

```

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

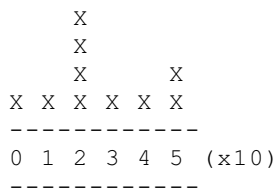
```

      8
      7
      4      2
5 1 1 0 2 0
-----
0 1 2 3 4 5 (x10)
-----
<----|---->
Spread

```

As a rough description, we may say data spread from 5 to 52. More precise ways to quantify spread are discussed in the next chapter.

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

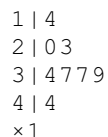


While it is difficult to make many stable statements about shape when n is small, you can still often determine whether data are more-or-less symmetrical and if there are any outliers. For the current stemplot, the data is more-or-less symmetrical and there are no apparent outliers.

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

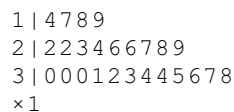
1.47 2.06 2.36 3.43 3.74 3.78 3.94 4.42

Because values have three digits and we want to plot only two (a stem and a leaf), we *truncate* data points before plotting. Truncation is a cutting-off of extra digits. For example, the value 1.47 is truncated to 1.4, 2.06 is truncated to 2.0, and so on. Let us use an axis multiplier of $\times 1$. The plot is:

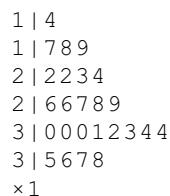


The median of this distribution is between the entries of 3.4 and 3.7 (location), data spread from 1.4 to 4.4, and the distribution is mound-shaped with no apparent outliers (shape).

Third illustration of a stemplot: The following pollution levels were found in a river 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 the regular stemplot:



The above plot is too squashed. To spread out the display, we split the stem-values in two. Here is the data with split stem-values:



The first 1s on the stem, reserved for values between 1.0 to 1.4, the second 1 is reserved for values

between 1.5 to 1.9, the first 2 is reserved for 2.1 to 2.4. (and so on).

Fourth illustration of a stemplot: In an experiment involving 9 healthy men, subjects drank half a bottle of red wine each day for two weeks. The level of polyphenols in the blood of subjects was measured at the beginning and end of the experiment. Here are the percent changes in polyphenols levels:

3.5 8.1 7.4 4.0 0.7 4.9 8.4 7.0 5.5

Here, we can also split the stem-values into five intervals. Here's the coding system used to tag the stem:

* for leaves of zero and one
t for leaves of two and three
f for leaves of four and five
s for leaves of six and seven
. for leaves of eight and nine

Plotting these data using this system makes a nice picture:

```
0* | 0  
t | 3  
f | 445  
s | 77  
. | 88  
×1
```

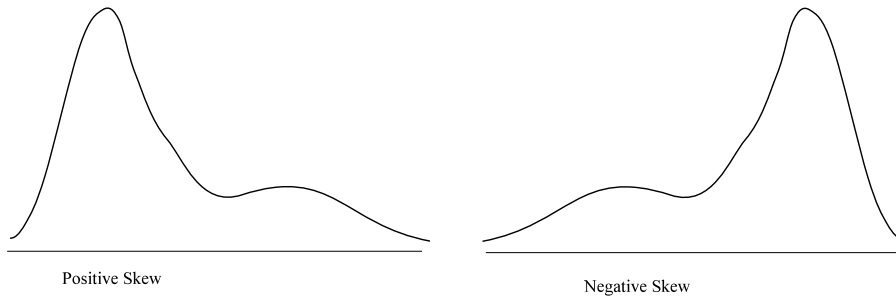
How many stem-values?

A good rule-of-thumb is to start with between 4 and 12 stem-values. Then, if the plot appears too squished, split the stem. If the plot is too spread out, use a bigger stem multiplier. When you are done with your plot, you want to be able to see the its shape, central location, and spread.

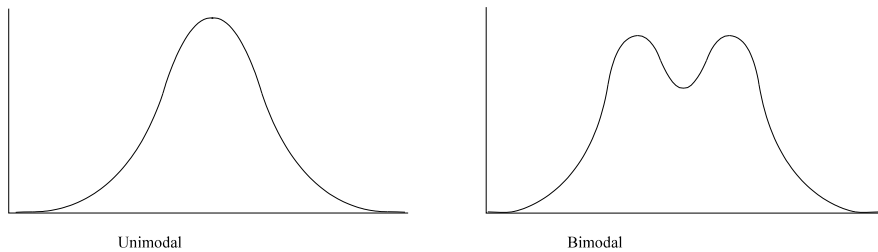
What to Look for in a Distribution

Shape

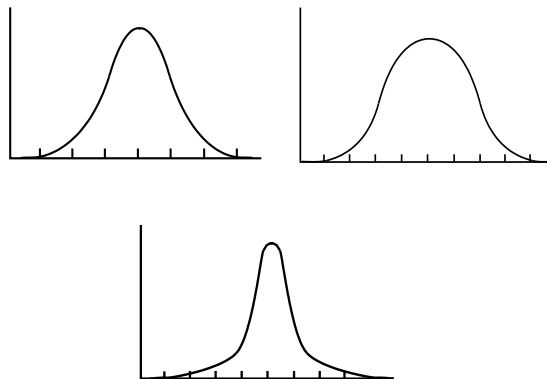
Let's consider the general characteristics of distributions. To discuss shape, let's look at the idealized shapes of distributions. A distribution's shape is described in terms of its symmetry, modality, and kurtosis. **Symmetry** refers to the degree to which a distribution reflects a mirror-image of itself around its center. Asymmetrical distributions are described by the position of their long tail. A distribution with a long right tail is said to have a **positive skew**. A distribution with a long left tail is said to have a **negative skew**.



Modality refers to the number of peaks on the curve. **Unimodal** and **bimodal** curves are shown below:

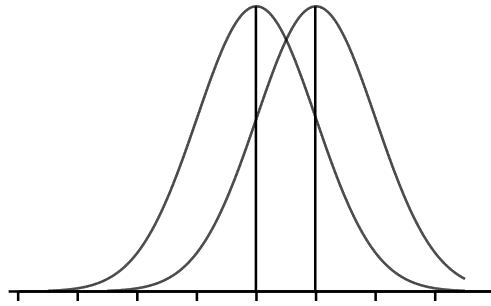


Kurtosis quantifies the flatness of the distribution. Distributions may be **mesokurtotic** (moderate, upper left), **platykurtotic** – (like a platypus, top right), or **leptokurtotic** – (steep, bottom).



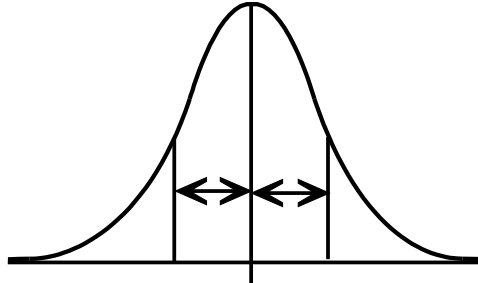
Location

A common statistical practice is to compare the central location of two distributions. The curves below compare locations of two populations. The curves overlap, but their central locations differ.

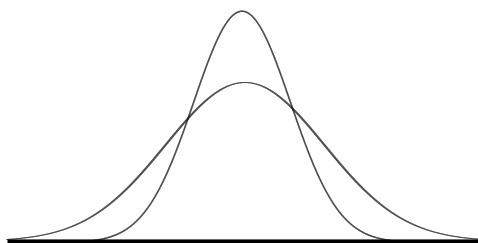


Spread

Spread address the extent to which data vary around the center of the distribution.



These curves have the same central location but different spreads:



Frequency Tables

We are interested in three types of frequencies:

Frequency counts (f_i): The number of times a value occurs in a data set

Relative frequencies (p_i): Frequency counts expressed as percentages of the total.

Cumulative [relative] frequencies (c_i): Relative frequencies up to and including the current value

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%	

To construct a **frequency table** by hand:

- (1) List value in ascending order. If a value appears more than once, list it only once.
- (2) Tally frequencies counts (f_i).
- (3) Sum counts to determine the total sample size: $n = \sum f_i$
- (4) Calculate the relative frequency (p_i) as the proportion of the total: $p_i = f_i / n$.
- (5) Determine cumulative relative frequencies (c_i) by summing cumulative frequencies from prior levels to the current level ($c_i = p_i + c_{i-1}$).

Here's the frequency table for this data set {21, 42, 5, 11, 30, 50, 28, 27, 24, 52}

Value	Tally	Freq.	RelFreq	CumFreq
5	/	1	10%	10%
11	/	1	10%	20%
21	/	1	10%	30%
24	/	1	10%	40%
27	/	1	10%	50%
28	/	1	10%	60%
30	/	1	10%	70%
42	/	1	10%	80%
50	/	1	10%	90%
52	/	1	10%	100%
TOTAL		10	100%	--

Grouped Data—Class Intervals

Sometimes you want to group data into class-intervals before putting them into a frequency table. Here are some guidelines when using uniform class-intervals:

(A) Decide on an appropriate number of class-interval groupings: The optimum number of class groupings will depend on the range of values and the size of the data set. In general, large data sets can support a large number of class groupings and small data sets can support fewer class groupings. Deciding on a suitable number of class-intervals, therefore, may require some trial and error. To start, try creating class-intervals that are of equal and convenient length (e.g., 10-year age intervals). Normally, 3 to 12 such class-intervals are sufficient.

(B) Determine the class interval width. This can be determined with the formula:

$$\text{Interval width} = \frac{\text{maximum} - \text{minimum}}{\text{no. of class groupings}}$$

For example, to create 4 class groupings for a data set with a maximum of 52 and minimum of 5, the class interval width = $(52 - 5) / 4 = 11.75$, which for the current purpose can be “rounded” down to 10 or rounded up to 15.

(C) Set endpoint conventions. If an observation falls on the boundary between two class intervals, we need to know in which class interval it will be counted. The two choices are to: (a) include the left boundary and exclude the right boundary or (b) include the right boundary and exclude the left boundary. When faced with this choice, we will use the option (a). For example, when considering the 15 unit class-interval of 15 to 30, we will exclude the right boundary of 30, so that the interval is really 15 to 29.99.... For convenience, this may be written 15–29.

(D) Count and tabulate: Once boundaries are established, data are tabulated in the usual manner.

Here’s a frequency table for the data {21, 42, 5, 11, 30, 50, 28, 27, 24, 52} using 15-year class-intervals:

Range	Tally	Freq.	Rel.Freq	Cum.Freq
0-14	//	2	20%	20%
15-29	////	4	40%	60%
30-44	//	2	20%	80%
45-59	//	2	20%	100%
TOTAL		10	100%	--

We can also use **non-uniform class intervals** for the frequency table if the purpose suits us. For example, here’s a data set with ages grouped into interval according to school-age”

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%	