

2: Frequency Distributions

[Stem-and-Leaf Plots](#) | [What to Look for In Distributions](#) | [Frequencies Tables](#) | [Frequency Charts](#)

Stem-and-Leaf Plots

The stem-and-leaf plot is a simple way to organize and present data in a histogram-like display. It is an excellent way to begin an analysis. To illustrate this technique, let us consider a data set with the following AGE values:

21 42 05 11 30 50 28 27 24 52

To construct the plot, we divide each value into a “stem values” and a “leaf value.” In this example, the digit in the tens place becomes the stem value and the digit in the units place becomes the leaf value. For example, the value “21” has a stem value of 2 and leaf value of 1.

Stem-values are listed in numerical order as a quasi-axis. A vertical line is drawn to separate these stem-values from future leaf-values. Here’s the stem:

```
| 5 |  
| 4 |  
| 3 |  
| 2 |  
| 1 |  
| 0 |  
(x10)
```

An *axis multiplier* ($\times 10$) is included to allow the viewer to decipher the magnitude of values (e.g., the stem value of 5 here represents 5×10 , or 50).

The right-most digit of each value is now plotted as a “leaf” on its proper axis location. For example, 21 is plotted as:

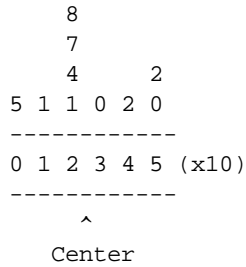
```
| 5 |  
| 4 |  
| 3 |  
| 2 | 1  
| 1 |  
| 0 |  
(x 10)
```

The remaining data points are plotted:

```
| 5 | 02  
| 4 | 2  
| 3 | 0  
| 2 | 1874  
| 1 | 1  
| 0 | 5  
(x 10)
```

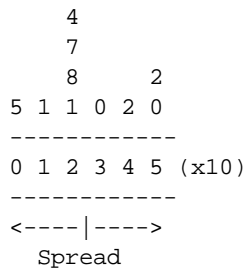
Data resemble a histogram on its side, with the distribution’s *shape*, *location*, and *spread* now visible.

I'm now going to rotate the stem-and-leaf plot 90 degrees to display these features in a more familiar way. The *location* of the data set is summarized by its center. For example, the central location of the current stem-and-leaf plot is somewhere between 20 and 30:



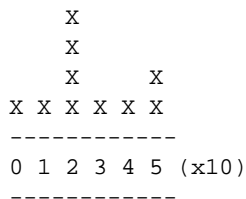
The center can more precisely be described by the distribution's *median*. The median has a *depth* of $(n + 1)/2$ from either end of the data set. For the current illustrative data, $n = 10$. Thus, the median has a depth of $(10 + 1)/2 = 5.5$. We count in from either end of the data set to the 5th and 6th value and average the values associated with these depths determine the median. In this instance, the value with a depth of 5 is 27. The value with a depth of 6 is 28. The median is the average of these two values = $(27 + 28) / 2 = 27.5$.

The *spread* of the data set is seen as the dispersion of values around the distribution's center:



We will learn measures of spread in the next chapter.

The *shape* of the distribution is seen as a "skyline silhouette":



Notice the "skyscraper" in the middle of this distribution. This peak represents the distribution's *mode*. In describing a distribution's shape, you should note the extent to which it is mound-shaped and symmetrical. Describing the shape of a small data sets is often difficult and may be unwarranted when the data set is small.

Second Illustration of a Stem-and-Leaf Plot: The next illustrative example shows how to draw a stem-and-leaf plot 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

Values have 3 significant digits and a decimal point. In such instances, we *truncate*[†] the data to include the first two significant digits. Thus, a value of 1.47 becomes 1.4, a value of 2.06 becomes 2.1, and so on. The stem-and-leaf plot of this truncated data looks like this:

```
|1|4
|2|03
|3|4779
|4|4
(x 1)
```

Third Illustrative Example: Suppose the following pollution levels are observed in a river: 2.2 3.4 3.0 2.6 3.8 1.8 2.8 3.2 3.7 1.4 2.7 3.6 1.9 2.2 3.0 3.3 2.3 1.7 2.6 3.5 3.0 2.9 3.4 3.1 2.4. Using stem-values of 1, 2, and 3 we get the following plot:

```
|1|8497
|2|268723694
|3|408276035041
(x 1)
```

This above plot is squashed, thus hiding the shape of the distribution. We may better show the distribution by using *double stem-values* with the first stem value reserved for leaf-values 0 to 4 and the second stem-value reserved for leaf-values 5 to 9. Here is the same data shown with double stem-values, thus providing a better idea of distributional shape:

```
|1|4
|1|789
|2|2234
|2|68739
|3|40203041
|3|8765
(x1)
```

In summary, always use judgment when constructing stem-and-leaf plots. You want to be able to see the distribution's shape, location, and spread.

In summary, to create a stem-and-leaf plot,

- (A) Draw a *stem-like axis* that covers the range of values. A good rule-of-thumb is to start with between 3 and 12 stem-values to act as "bin's" for the leaves, and then to see what develops. (You may have to redraw the plot if it turns out to be too squished or too spread out.)
- (B) Truncate the data to two or three significant digits.
- (C) Separate each data-point into its stem-component and leaf-component.
- (D) Place each leaf adjacent to its associated stem-component, one leaf on top of the other.

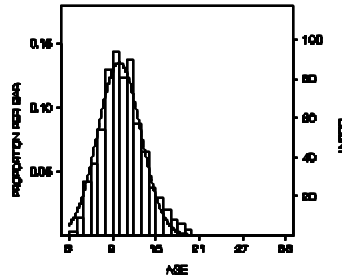
[†] You may also round the digit before plotting; for uniformity, we will truncate the data.

SPSS: To create a stem-and-leaf plot in SPSS, click on Analyze > Descriptive Statistics> Explore. Then put the name of the variable you want to plot into the "Dependent List" box. The stem-and-leaf plot will be shown near the bottom of the output.

What to Look for In Distributions

Distributional Characteristics

As noted, we are interested in describing a distributions shape, location, and spread. **Shape** refers to the configuration of values when plotted on a graph. **Location** refers tot he position of data points on the graph. **Spread** refers to the dispersion of values around a central point of reference. Stem-and-leaf plots and histograms provide insights into these distributional dimensions. One such histogram is shown below.



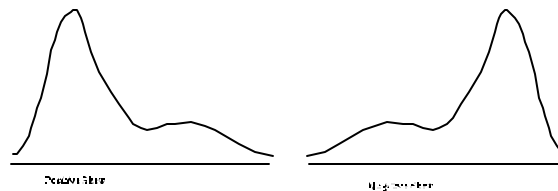
The above histogram is more-or-less bell-shaped, is centered around 10, and most values fall between 5 and 15 (10 ± 5), thus there is a spread of about 5 on either side of 10).

Superimposed on the above histogram is a **curve**. The degree to which the curve fits the data is not critical for the current discussion, but it is clear that it is a pretty good fit. More importantly, the curve provides a convenient convention for discussing distributional shape, location, and spread.

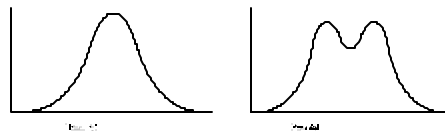
Shape Descriptors

Distributional shape may be described in terms of symmetry, modality, and kurtosis. **Symmetry** refers to the degree to which a distribution reflects a mirror-image of itself around its center. **Modality** refers to the number of peaks found on the distribution. **Kurtosis** refers to how peaked or flat the distribution appears.

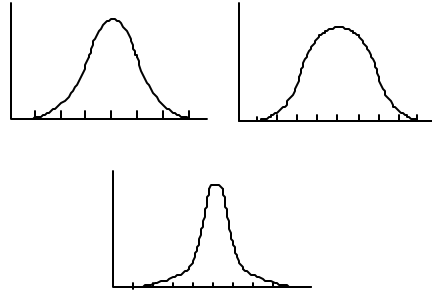
Asymmetrical distributions are described by the position of their longer tail. A distribution with a long right tail is said to have a **positive skew**. A distribution with a long left tail is said to have a **negative skew**.



Modality is defined by the number of peaks on the distribution. Distributions may be **unimodal**, **bimodal**, or **multimodal**.



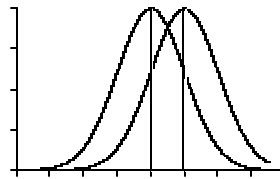
Finally, the shape of a distribution may be mesokurtotic (moderate steepness, top left), platykurtotic (flat like a platypus, top right), or leptokurtotic (steeply peaked, bottom).



Location

The location of a distribution is usually describe in terms of its **center**. The most common measure of central location is the **mean**, which will be covered in the next chapter. Other measures of central location include the **median** and **mode**. The mean, median, and mode are the same when the distribution is symmetrical and unimodal.

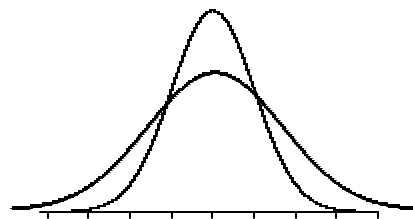
A common statistical practice is to compare the centers of two distributions to see if they differ. For example, we might compare the cholesterol distribution of a population with a healthy diet (curve on the left) and one with a high-fat diet (curve on the right).



Although the two distributions overlap, the population on the left (health diet) has a lower value on the average (a specified by the central line of the distribution).

Spread

The spread of a distribution is the variability of value (how much values disperse in the population). Common measures of spread include the variance, standard deviation, and inter-quartile range. (These statistics are covered in the next chapter.) The curves immediately below represent two distributions with the same central location but different spreads.



Frequency Tables

Frequency Tables For Raw Data

A common way to begin an analysis to count frequencies of values. There are three different 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 rank-ordered value.

An example of a frequency table of ages from a large 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%	

Notice how the frequency column sums to n and the relative frequency column sums to 100%.

To construct a frequency table for raw data:

- List all value in ascending order. (If a value appears more than once, list it once only. You'll tally frequencies as a separate step.)
- Tally frequencies (f_i) with tick marks or some other accounting mechanism. List this information in the **Freq** column of the table.
- Sum the frequency counts to determine the total sample size: $n = \sum f_i$
- Calculate the relative frequency (p_i) of each value as the proportion of the total: $p_i = f_i / n$.
- Determine cumulative frequencies (c_i) by adding the cumulative frequency from the prior level to the relative frequency of the current level ($c_i = p_i + c_{i-1}$).

A frequency table for the small data set {21, 42, 5, 11, 30, 50, 28, 27, 24, 52} is:

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% --

Frequency Tables Based on Uniform Class Intervals

Because the data set in the above table is small ($n = 10$) and has a large range (5 to 52), frequencies of raw values are not particularly useful. In such instances, you should condense the data into class intervals (“groupings”) before tallying results.

There are no hard-and-fast rule for determining the number of class intervals you should use, but here are some general ideas on how to proceed:

- (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 consideration 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) **Tabulate the data:** Once boundaries are established, data are tabulated in the usual manner. A frequency table for the data {21, 42, 5, 11, 30, 50, 28, 27, 24, 52} using 15-year class-intervals is:

Range	Tally	Freq.	RelFreq	CumFreq
0-14	//	2	20%	20%
15-29	////	4	40%	60%
30-44	//	2	20%	80%
45-59	//	2	20%	100%

TOTAL		10	100%	--

Nonuniform Class Intervals

You might, at times, want to use *nonuniform* class-intervals when describing data. In such instances you should use boundaries that have meaning. For example, you may want to look at the age distribution of children with ages grouped into pre-school age (2-4), elementary school age (5-11), middle-school age (12-13), and high-school age (14-19). The data from the initial table in this chapter can now be displayed as follows:

AGEGRP	Freq	RelFreq	CumFreq
-----+-----			
PRESCHOOL	11	1.7%	1.7%
ELEMENTARY	469	71.7%	73.4%
MIDDLE	100	15.3%	88.7%
HIGH	74	11.3%	100.0%
-----+-----			
Total	654	100.0%	

Notice that 72% of the children in this survey are in elementary school.

In the end, the best frequency table is the one that sheds the most light on the information you want to know.

SPSS: To create a frequency table in SPSS, click Analyze > Descriptive Statistics > Frequencies.

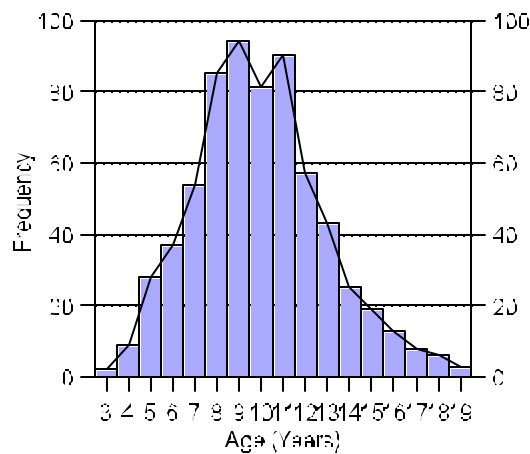
Other Frequency Charts

The stem-and-leaf plot introduced on the first page of this chapter is an modern way to display frequencies. The more traditional way to display frequencies uses histograms or frequency polygons.

Histograms are bar charts with contiguous (touching) bars, with the height of each bar proportional to the frequency of occurrence. The X axis contains values or class-intervals for the variable. The Y axis is the frequency or relative frequencies of occurrence. Because the bars of histograms are touching, histograms are normally reserved for scale measurements. When working with ordinal and nominal data, frequencies should be plotted in the form of a bar chart with non-contiguous (non-touching) bars.

Frequency polygons are like histograms except instead of plotting bars they show frequencies with a line. Like histograms, frequency polygons should be reserved for continuous measurements.

A histogram with an overlying frequency polygon showing the age distribution participants in a school survey is:



SPSS: For histograms, click Graphs > Histogram. For frequency polygons click Graphs > Line > Simple and use the variable as the category axis.

Notation and Vocabulary

n = sample size

f_i = frequency interval i

p_i = relative frequency interval i

c_i = cumulative relative frequency interval i

Cumulative frequency: the accumulation of relative frequencies up to and including the current rank-ordered value or class.

Frequency: the number of times a particular item occurs in a data set; a count.

Histogram: a bar graph of frequencies or relative frequencies in which bars touch.

Relative frequency: frequencies expressed as a percentage of the total.

Stem-and-leaf plot: a data display in which data points are divided into a stem component and leaf component, and are then plotted in a fashion to resemble a histogram on its side.