

Stem-&-Leaf Plots and Frequency Tables



[Stem-and-Leaf Plots](#)
[Frequencies Tables](#)
[Frequency Charts](#)

Stem-and-Leaf Plots

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

21 42 05 11 30 50 28 27 24 52

We note that values range from 5 to 52. To construct the plot, we divide each observation into a “stem value” and a “leaf value.” In this example, the digit in the tens place becomes a stem value and the digit in the units place becomes a leaf value. For example, the value “21” has a stem value of 2 and leaf value of 1.

The 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:

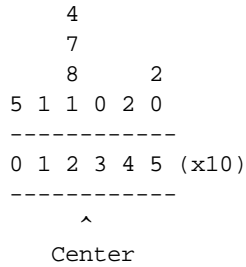
```
| 5 |  
| 4 |  
| 3 |  
| 2 | 1  
| 1 |  
| 0 |  
(x 10)
```

The remaining data points are plotted:

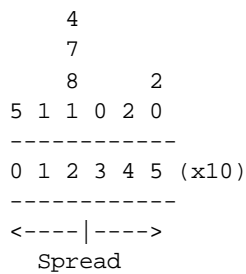
```
| 5 | 02  
| 4 | 2  
| 3 | 0  
| 2 | 1874  
| 1 | 1  
| 0 | 5  
(x 10)
```

Data now resemble a histogram on its side. The distribution’s *shape*, *location*, and *spread* is now visible. I’m now

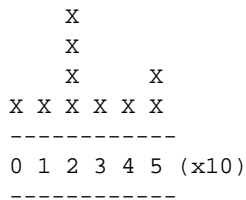
going to rotate the stem-and-leaf plot 90 degrees to display these features in a more familiar way. The *location* of the data set is summarized by its center. For example, the central location of the current stem-and-leaf plot is somewhere between 20 and 30:



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



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

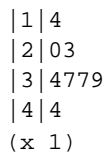


Notice the “skyscraper” in the middle of this distribution. This peak represents the distribution's *mode*. In describing a distribution's shape, you should note the extent to which it is mound-shaped and symmetrical. With small data sets (e.g., $n < 30$), this will be difficult, and may require some imagination with judgments couched in cautious terms.

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 this new data set:

1.47 2.06 2.36 3.43 3.74 3.78 3.94 4.42

These data have 3 significant digits and a decimal point. In such instances, we *truncate* the data to include only two digits. (Truncation means “cut off.”) The stem-and-leaf plot based on the truncated data looks like this:



Third Illustrative Example of a Stem-and-Leaf Plot: 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 stem-and-leaf plot:

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

Realizing that this plot is “squashed” (hiding the distribution’s shape), we spread it out by using *double stem-values*, with the first value reserved for leaf-values between 0 and 4 and the second stem-value reserved for leaf-values between 5 and 9. Here is the same data shown with double stem-values:

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

This gives us a better idea of the distribution’s asymmetrical shape,

Always use judgment in deriving your stem-and-leaf plot. You want to be able to see the distribution’s shape, location, and spread. 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.

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

- (A) Draw a *stem-like axis* that covers the range of values. Start with between 3 and 12 stem-values. (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.

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 appear near the bottom of the output.

Frequency Tables

Frequency Tables For Raw Data

An other useful way to begin an analysis to consider the data in the form of a frequency table. Frequency tables may include up to three different types of frequencies. These are:

Frequency counts: The number of times a value occurs in a data set.

Relative frequencies: Frequencies expressed as percentages of the total.

Cumulative [relative] frequencies: 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 potential 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 of each interval (p_i) as the proportion of the total: $p_i = f_i / n$.
- Sum cumulative frequencies (c_i) by adding the cumulative frequency from the prior level to the relative frequency of the current level ($c_i = p_i + c_{i-1}$).

Building of a frequency table for the small data set {21, 42, 5, 11, 30, 50, 28, 27, 24, 52} is shown on the next page:

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%	--

Because this data set 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 groupings (“class intervals”) before tallying results.

Frequency Tables Based on Uniform Class Intervals

There are no hard-and-fast rule for determining the number of class intervals you should use, but here are some rules-of-thumb:

(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 class-intervals is 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 up to 15.

(C) Set endpoint conventions. If an observation falls on the boundary between two class intervals, we need 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 interval of 15 to 30, we will exclude the right boundary of 30, so that the interval is really between 15 (inclusive) up to 30 (exclusive). This may be written 15–29.

(D) Tabulate the data: Once boundaries are established, the 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.

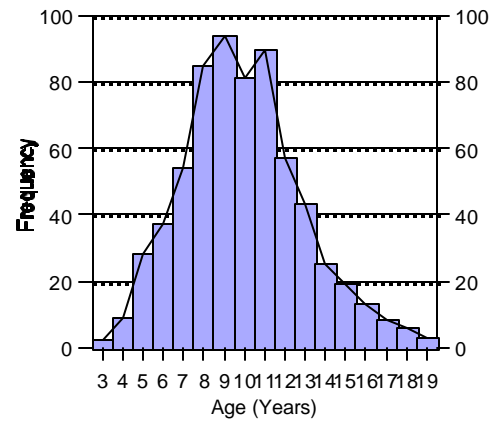
SPSS: To create a frequency table in SPSS, click on Analyze | Descriptive Statistics | Frequencies and select the variable you want to analyze.

Frequency Charts

Histogram or frequency polygon are basic tools of data analysis. *Histograms* are frequency charts with contiguous (touching) bars. Of course the height of each bar is proportional to either a frequency count or relative frequency. Because the bars are touching, histograms are reserved for truly continuous (scale) variables. 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 for the data listed in the first table in this chapter is shown in the figure to the right.

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.