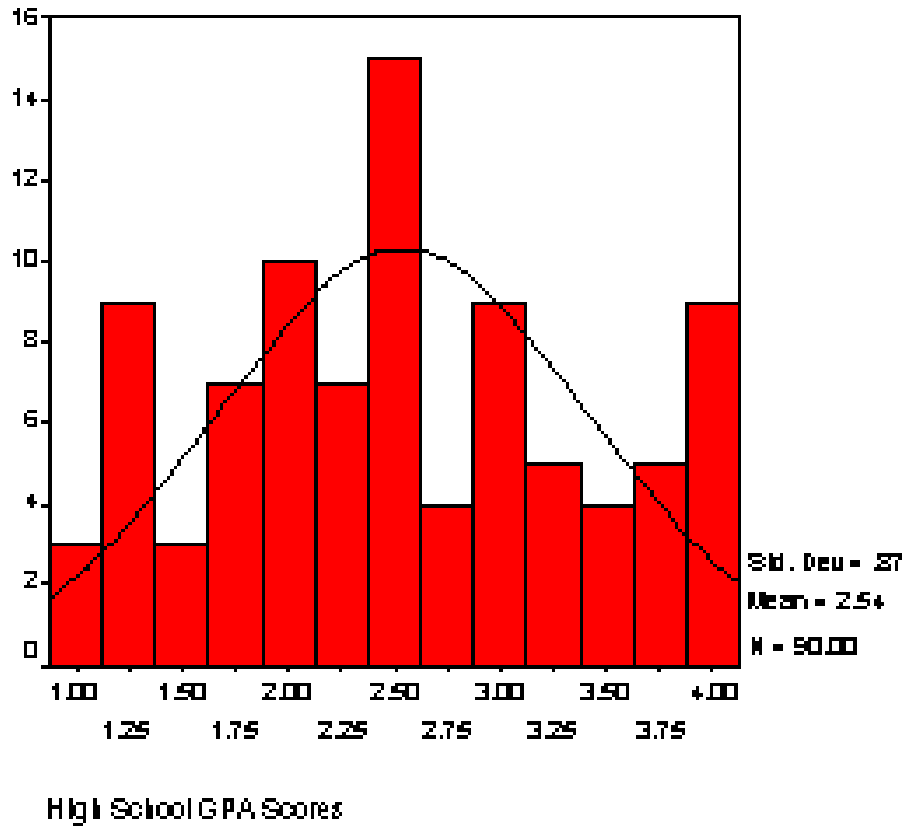
Examples of Distributions and Descriptive Graphs

Frequency Histogram

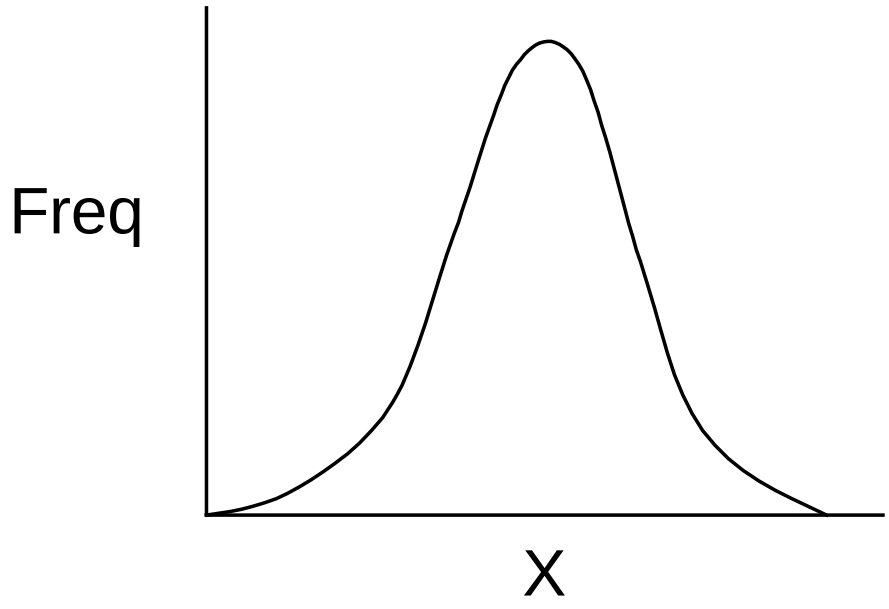
Discrete values of categories are used on the horizontal (x) axis, and the number of scores that fall into that category (i.e., frequency) appears on the vertical (y) axis. A normal curve can be overlaid so that one can easily see how it departs from normal (see next topic). This distribution looks close to normal here, although there is a pretty high peak in the middle and a some high frequencies in the right and left tails. Note that the appearance of the normal curve will sometimes look wider or more narrow than it really should depending on how SPSS chooses the categories for the histogram (this normal curve one looks a little too flat to me).



The shape of a distribution, sometimes illustrated with a connected dots or smooth curve line (“frequency polygon”), can be described in terms of how it relates to the normal curve.

Normal Curve

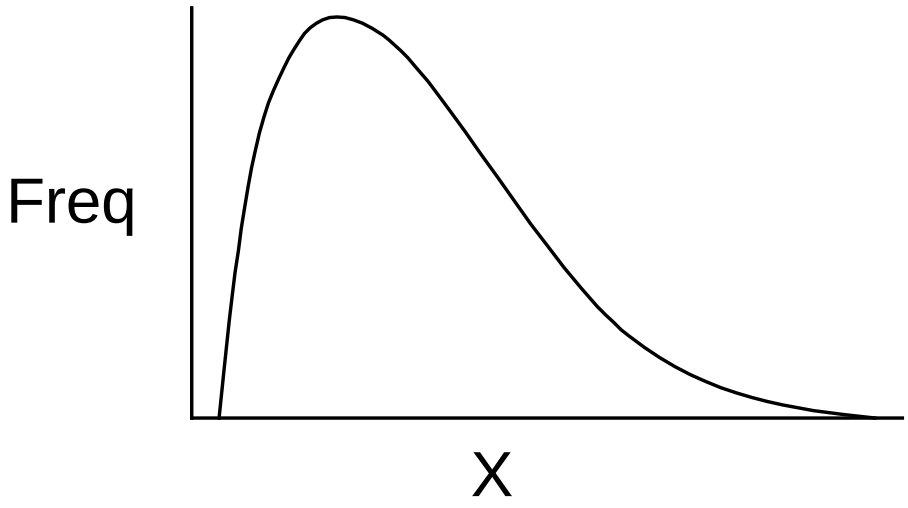
The standard normal curve is a particular shape of frequency distribution commonly found in nature. The “normal” is a kind of statistical ideal that can be described precisely mathematically and looks something like the shape below.



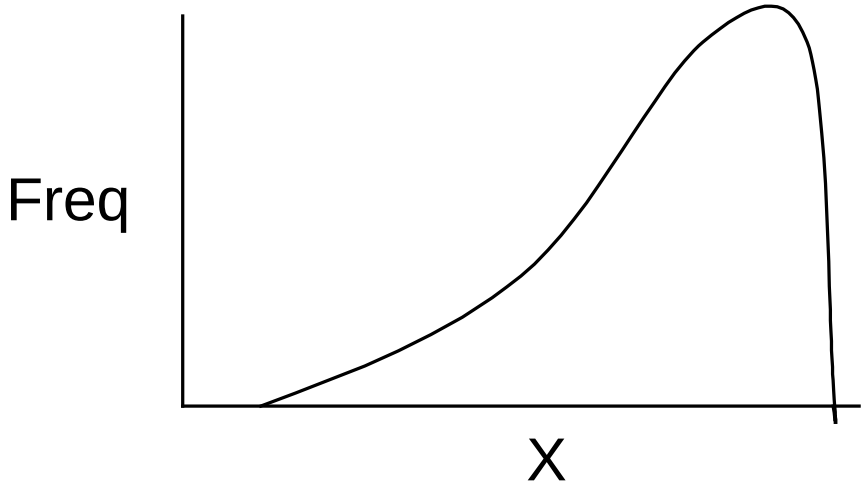
Skewness

Skewness refers to how asymmetric the shape of the distribution is. That is, are there more extreme values out to the right or left.

Right (Positive) Skew



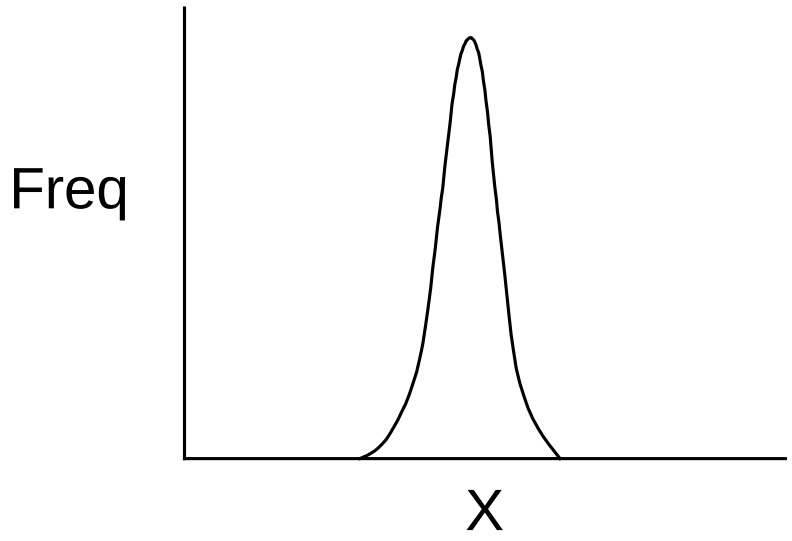
Left (Negative) Skew



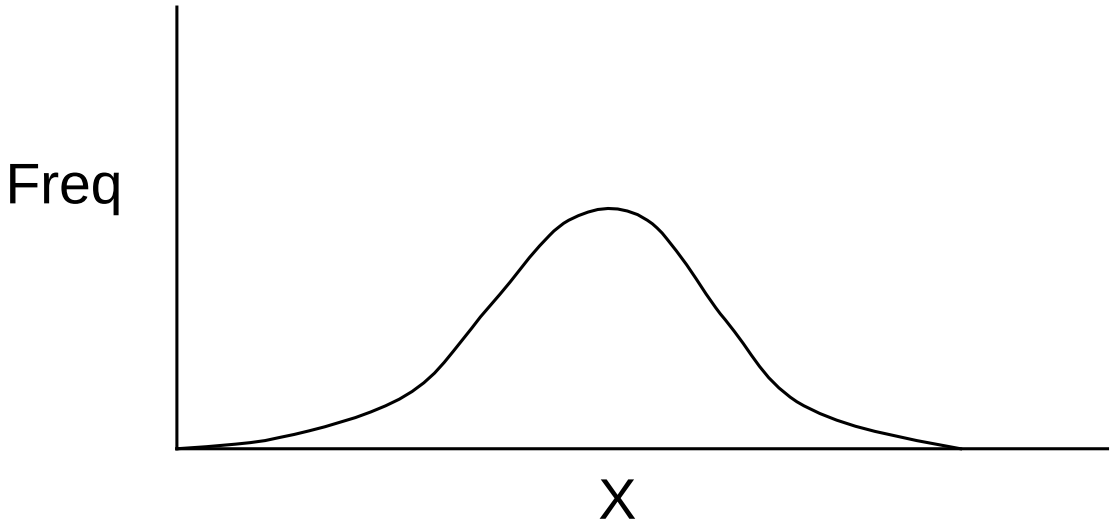
Kurtosis

A frequency distribution can also be described in terms of how flat to narrow it is.

Leptokurtic

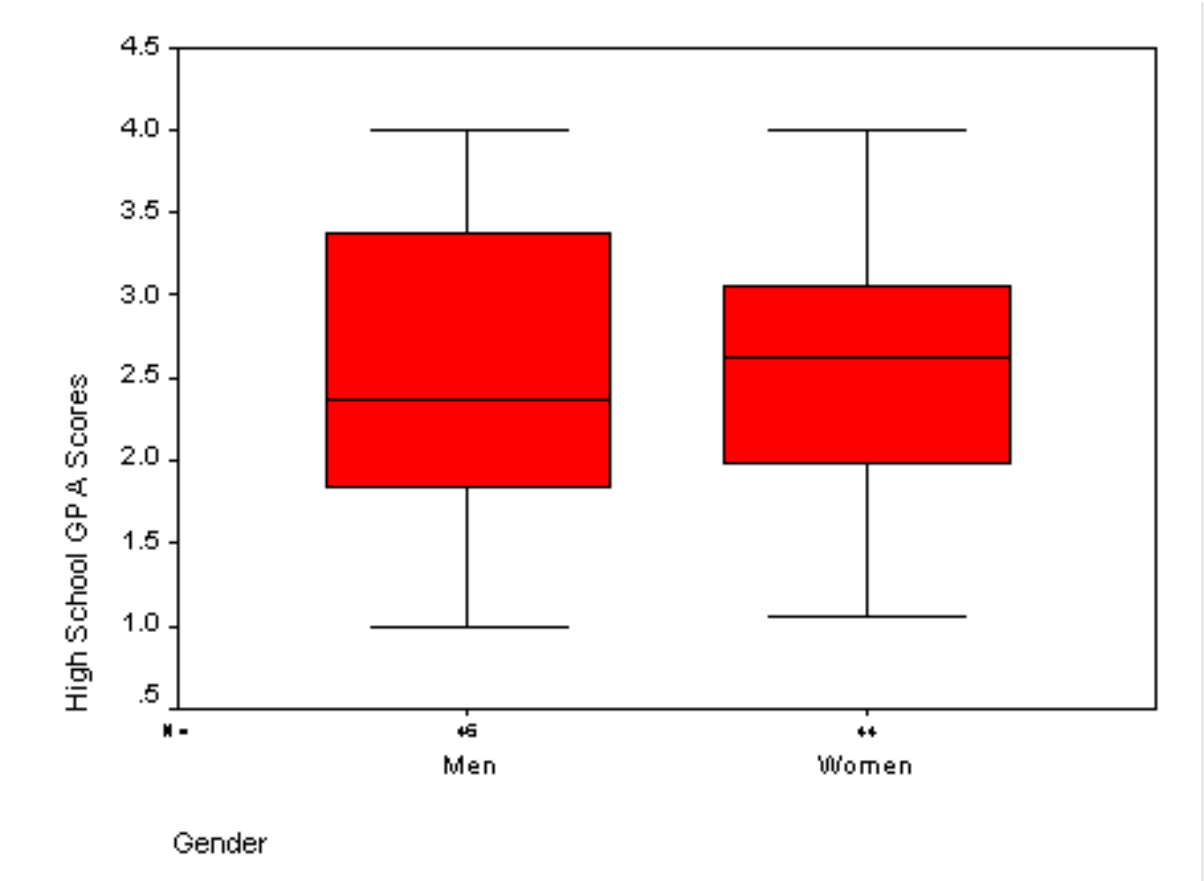


Platykurtic



Box Plots

Below are box plots or sometimes called “box and whiskers.” The center line is the median, the top and bottom of the box is the upper and lower “hinge” (approximately the upper and lower quartiles), and the endpoints of the whiskers represent the approximate range. SPSS will sometimes print “outliers” and “extreme points” (denoted by circles and asterisks) that lay outside the end of the whiskers, in which case the whiskers are not really the lowest and highest scores. Defining an outliers and extreme points is rather subjective. SPSS uses z-scores greater than a certain value to define these. However, identifying outlying points is probably best left to the researcher to define—you may want to consider some points outliers even though SPSS does not or SPSS may identify something as extreme that you do not. Box plots are probably the most useful when you are comparing two groups.



Stem and Leaf

The Stem and Leaf combines a histogram with actual data points. If you imagine turning the figure on its side, it resembles a frequency histogram. The "stems" are categories for the scores, usually identified by the first digit of the score, and the "leaves" are the last digit of the score. Often, some type of rounding is used, depending on whether there are decimals or the number of digits is large. The intention is to allow you to identify exact scores in the figure and also get a sense of the shape of the distribution.

High School GPA Scores Stem-and-Leaf Plot for GENDER= Men

```
 Frequency    Stem &  Leaf
     6.00        1 .  011123
     8.00        1 .  56778889
     9.00        2 .  000112222
     9.00        2 .  555556668
     3.00        3 .  023
     5.00        3 .  55788
     6.00        4 .  000000
 Stem width:      1.00
 Each leaf:       1 case(s)
```

High School GPA Scores Stem-and-Leaf Plot for GENDER= Women

```
Frequency    Stem &  Leaf
     8.00        1 .  01233334
     3.00        1 .  699
     6.00        2 .  000134
    13.00        2 .  5556666779999
     8.00        3 .  00011112
     4.00        3 .  5789
     2.00        4 .  00
 Stem width:      1.00
 Each leaf:       1 case(s)
```

SPSS Descriptives and Frequencies Procedure Examples

Two procedures in SPSS that allow you to obtain some specific descriptive statistics are the Descriptives and Frequencies commands. Descriptives is generally used for continuous data in which mean and standard deviations are useful for describing the data. The Frequencies command gives the number and percentage of cases for each value of the variable and is probably most often used for categorical data. I routinely used both procedures to examine all of my variables prior to use in other analyses to check for errors in coding or data entry and to get summary information for each variable.

Chilean Plebocite Data

Descriptive Statistics

```
descriptives vars= age income statquo.
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
age Age in Years	2699	18.00	70.00	38.5487	14.75642
income Monthly Income in Pesos	2602	2500.00	200000.00	33875.86	39502.86712
statquo Scale of Support of Status Quo	2683	-1.80	2.05	.0000	1.00019
Valid N (listwise)	2590				

FrequencyTable

```
frequencies vars=sex educ.
```

sex Sex of respondent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00 female	1379	51.1	51.1	51.1
	1.00 male	1321	48.9	48.9	100.0
	Total	2700	100.0	100.0	

educ Education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 primary	1107	41.0	41.2	41.2
	2.00 secondary	1120	41.5	41.7	82.8
	3.00 post-secondary	462	17.1	17.2	100.0
	Total	2689	99.6	100.0	
Missing	System	11	.4		
Total		2700	100.0		

An example of frequencies command syntax with no table printed (might be used if there are many values), descriptive statistics, and histogram.

```
frequencies vars=age
   /format=notable
   /statistics=all
   /histogram=normal.
```

Statistics

age Age in Years

N	Valid	2699
	Missing	1
Mean		38.5487
Std. Error of Mean		.28404
Median		36.0000
Mode		21.00
Std. Deviation		14.75642
Variance		217.752
Skewness		.473
Std. Error of Skewness		.047
Kurtosis		-.860
Std. Error of Kurtosis		.094
Range		52.00
Minimum		18.00
Maximum		70.00
Sum		104043.00