

## Definitions, Scaling, Graphs

### 1. Definitions

**Variable** -- something that can take on more than one value. Syn. - fluctuates, changes, dynamic. Ant. - constant, fixed, unchanging.

Examples of variables	Some possible values
Sex	Female, Male
Eye color	Blue, Brown, Green
Class standing	Freshman, Sophomore, Junior, Senior
Age	6, 19, 25, 55, 70 (years)
Height	60, 65, 72, 84 (inches)
Weight	97, 101, 159, 220 (lbs.)
Visualization (speed of rotation)	1.5, 2.0, 2.5, 3.0 (seconds)
SAT-Verbal	500, 600, 750 (scale scores)
Curiosity (personality scale)	40, 50, 60 (scale scores)

**Independent variable (IV)** - that which explains or causes; usually manipulated (cause).

**Dependent variable (DV)** - that which is explained or is influenced (effect).

### Examples of IVs & DVs

1. Student volunteers given a "study drug." Half get mild dose of caffeine, half get saline (nobody knows which is which). Both take exam the next day.

IV - type of drug; DV - exam score.

2. Want to know whether using a microcomputer simulation can help train pilots to better communicate when they fly. Half the pilots are given flight training with the microcomputer. The other half play video games (Asteroids) with the same microcomputer for the same amount of time. All pilot then fly full motion simulator through same series of problems. Instructor pilots (blind to condition) evaluate all with checklist.

IV=training type; DV=evaluation of flight in full motion sim.

3. Want to know whether SAT scores predict grades in college. Collect SAT scores from entering freshmen. Collect GPAs at end of year. Look for association.

IV=SAT; DV = GPA.

### Continuous vs. Discrete Variables

In statistics and mathematics (Thorne & Slane), continuous variables correspond to real numbers and discrete variables correspond to integers. Real numbers take on an infinite number of values; integers only take whole number values.

Continuous (real)	Discrete (integer)
Time (seconds)	Number of Siblings
Weight (lbs.)	Number of errors on a test
Height (inches)	Number of bar presses

In math/stat, there is an enormous difference between continuous and discrete variables. This distinction is virtually never important in psychology.

### Continuous vs. Nominal Variables

In research design (Smith & Davis) we usually speak of continuous (many valued and ordered) or nominal (categorical, labeling) variables. Nominal means *in name only*.

Continuous (many valued, ordered)	Nominal (categorical, labeling)
Age	Sex
Weight	Eye color
Number of siblings	Major
Test score (# right)	Political party

### Population, Sample, Parameter & Statistic

**Population** - the complete collection; everyone of interest (e.g., adults in the U.S., students at USF); can vary in size, usually a large number of people.

**Sample** - a subset of the population (e.g., students at USF are a sample of students in U.S. colleges (the population); students in this class are a sample of students at USF (the population). In research, we usually take (select, invite) a sample to represent the population. It is usually too difficult or expensive to select the whole population.

**Parameter** - a numerical summary of the population (e.g., the average age of the population of students at USF would be a parameter).

**Statistic** - a numerical summary of a sample (e.g., the average age of students in this class when this class is considered a sample of students at USF).

**Sample statistics are used to estimate population parameters.**

### 2. Scale Types

**Nominal** -- categories. E.g., football numbers, classroom numbers, SSN. or area codes; DSMIII (R) 303 intoxic; 307 stuttering

**Ordinal** -- rank order. E.g., place of finish in race or other contest, bakery numbers for order of arrival.

Moh's scale of rock hardness

(1) Talc	(6) Orthoclase
(2) Gypsum	(7) Quartz
(3) Calcite	(8) Topaz
(4) Fluorite	(9) Corundum
(5) Apatite	(10) Diamond

No info about how much harder, just harder.

**Interval** -- order and interval (difference) have meaning. E.g., Celsius & Fahrenheit temperature scales.  $(75 - 50) = (100 - 75) = 25$  in terms of difference. Cannot say 50 is twice as hot as 25. Many psychological scales thought to be interval (e.g., SAT, personality tests, etc.)

**Ratio** -- order, interval, and ratio have meaning. E.g., Kelvin scale of temperature; reaction time. There is a meaningful zero point in the ratio scale.

### Review: Footrace Example

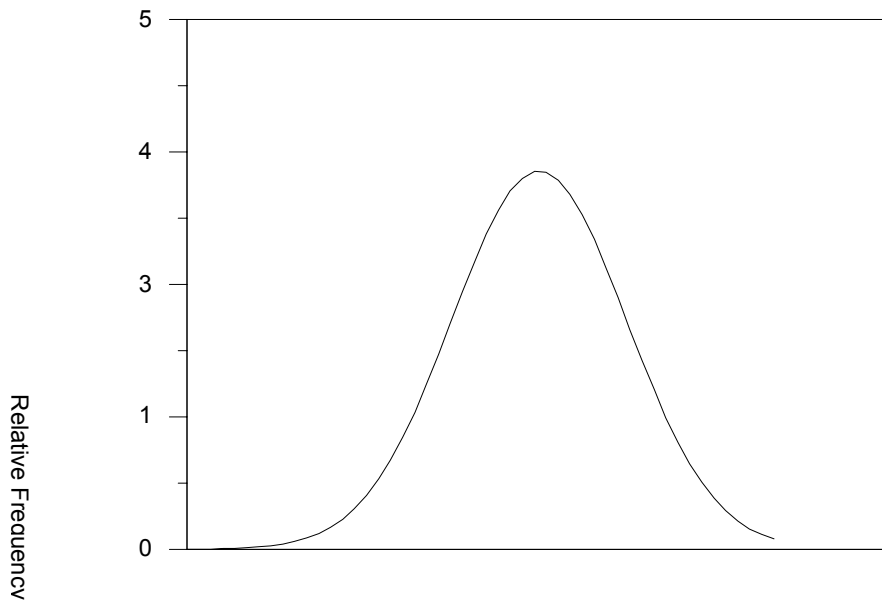
Nominal	Ordinal	Interval	Ratio
Jersey Number (registration ID)	Rank order finish	Time of day	Elapsed time
043	1	10:57 a.m.	4 min
011	2	10:59 a.m.	6 min
136	3	11:01 a.m.	8 min
112	4	11:02 a.m.	9 min
086	5	11:04 a.m.	11 min

## 3. Graphing

Boxplot and Stem-and-Leaf Plot

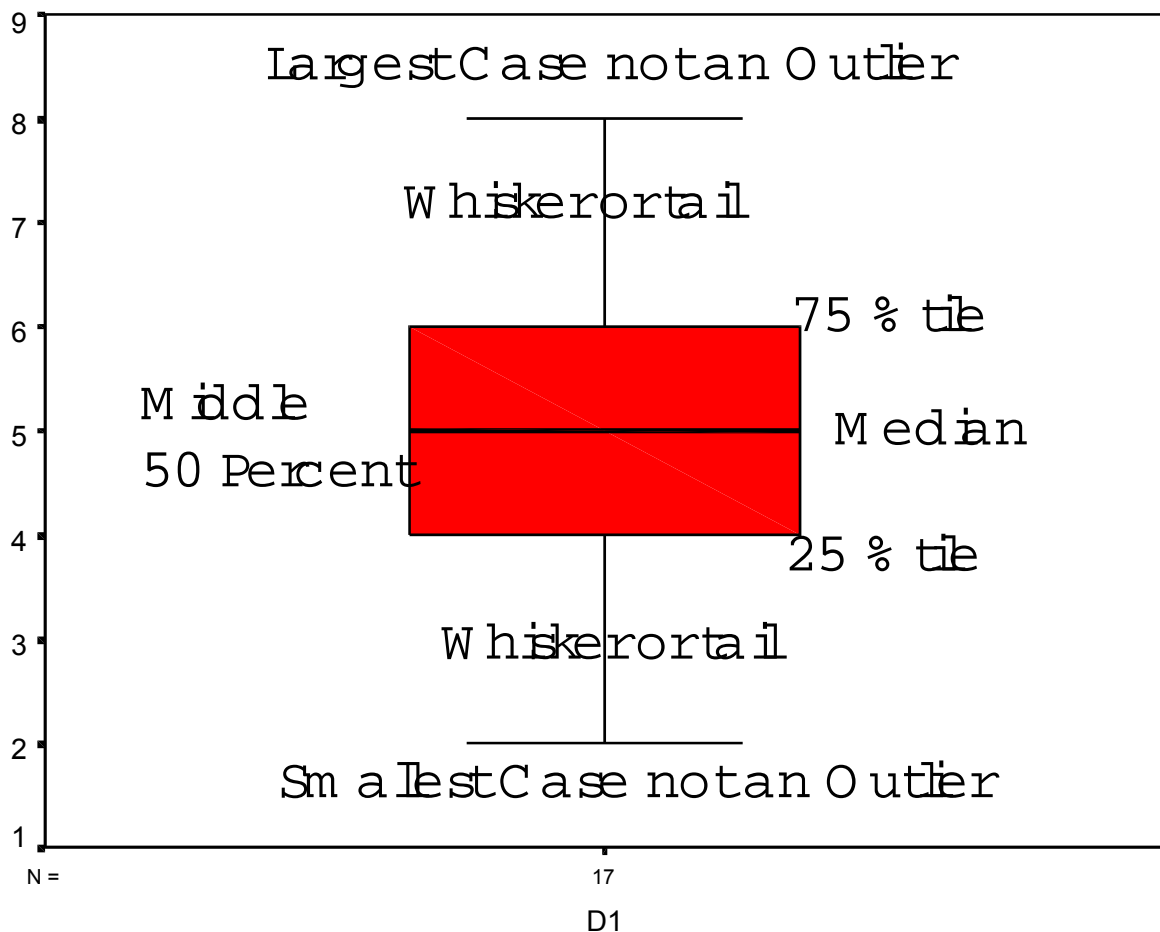
The boxplot and stem-and-leaf (or stem-leaf, for short) are both computer generated graphs that show lots of information about a distribution.

A theoretical graph of a normal distribution (the Bell-Shaped Curve)



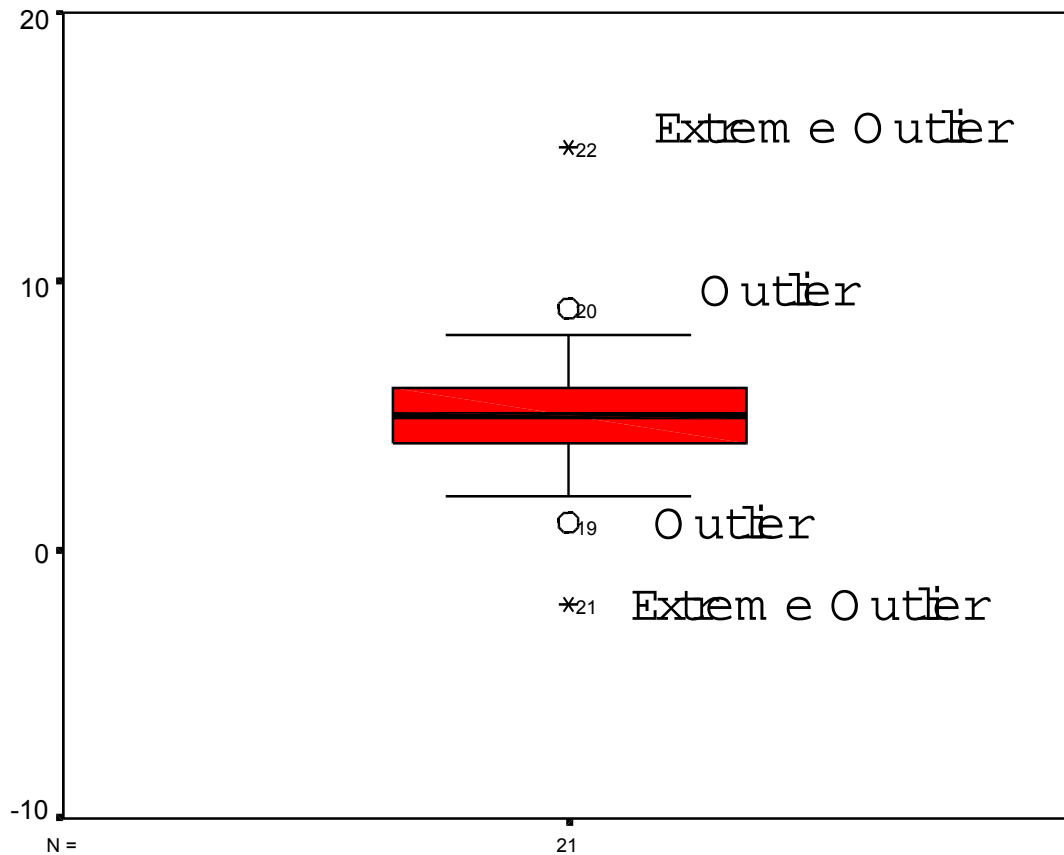
The distribution of height in the adult population looks like this.

Boxplot for a normal distribution.



A **boxplot** is a graphic representation of a distribution. The boxplot shown above represents a normal distribution. When the distribution is normal or approximately normal, the line representing the median of the distribution will appear in the middle of the shaded box at the center of the plot. The shaded box of the boxplot contains the middle 50 percent of the cases. If the distribution is normal, there will be whiskers of equal length, and there will be nothing visible beyond the ends of the tails, that is, no outliers. The above box plot is what we love to see. It has the line for the median right in the middle, whiskers of equal length, and nothing beyond the whiskers.

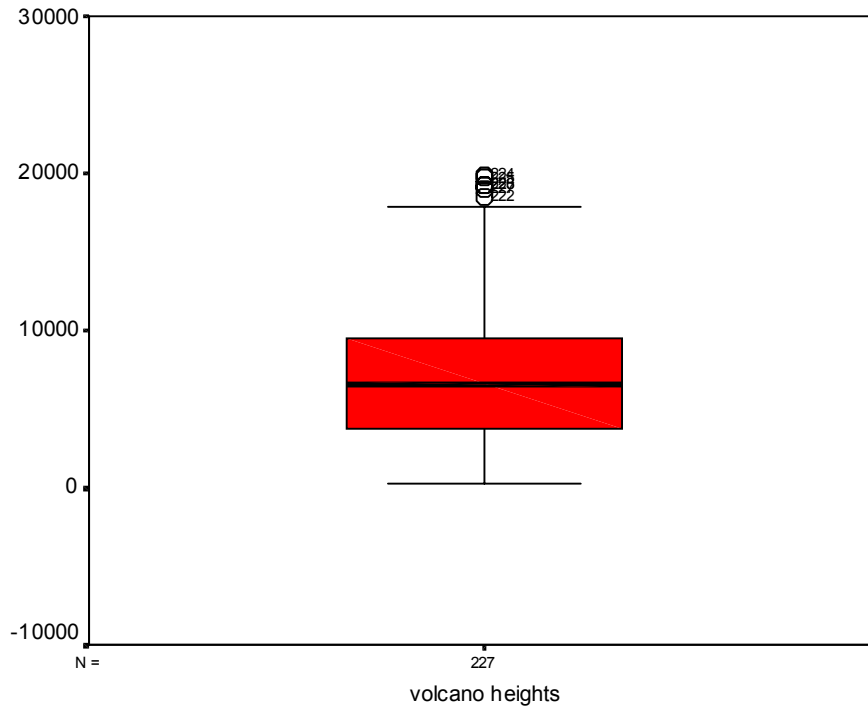
A boxplot with outliers and extreme outliers:



Note that outliers are represented as circles and extreme outliers are represented as asterisks. The numbers next to the outliers tell you where in your data the outlying case is. The extreme outlier at the top of the boxplot has number 22 next to it. If you go to your data (input file), the 22nd case will have this value. This helps you locate problem data in your files.

If the distribution is skewed, one tail will be longer than the other. The median may also be pulled away from the center of the box.

Example of a skewed distribution:



This is a boxplot of a distribution of volcano heights. As you can see, most of the volcanos are not very tall; a few are very tall indeed. The tails are not even; the top one is longer than the bottom, indicating skew. There are outliers on the top but not on the bottom. Also, there is a floor effect because volcanos (on land, which all of these are) cannot be less than zero feet tall, so they are bounded on the bottom, but not on the top. If you look at salary data, that is, what people working full-time make per year, you will see a similar distribution. There is a bottom below which nobody works. There are a few people making truly astounding sums. Most people make a salary near the bottom, just as most volcanos are not very tall. Such a graph is important in at least two ways: (1) it shows that the distribution is loaded toward the bottom, and (2) the tall volcanos have a big impact on the arithmetic mean (average) and on most other analyses we do with these data.

### Stem-and-Leaf Plots

The stem-and-leaf (or stem-leaf for short) plots are basically histograms constructed from the numbers themselves. These give you an extremely accurate view of the shape of a distribution. You get peaks and valleys in the middle of the distribution that you don't get with the boxplot. You also get some idea of the skew, and any floor or ceiling effect. The only thing you don't get is a visual representation of the outliers (in SPSS).

Example of a stem-leaf plot of a normal distribution:

NORM Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	2 .	0
2.00	3 .	00
3.00	4 .	000
4.00	5 .	0000
3.00	6 .	000
2.00	7 .	00
1.00	8 .	0

Stem width: 1.00  
Each leaf: 1 case(s)

This is basically what we see for the very first (normal) boxplot. Note that the big numbers are at the top of the boxplot, but at the bottom of the stem-leaf plot in SPSS. This is because they had different software engineers working on the problem who weren't talking to each other. Be sure to read the numbers. This says that there is one observation (person) with a score of 2. There are two people with a score of 3; there are three with a score of 4 and so on.

A stem-leaf plot with outliers looks like this:

OUT1 Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	Extremes	(= $\leq$ -5)
1.00	0 .	1
3.00	0 .	233
7.00	0 .	4445555
5.00	0 .	66677
2.00	0 .	89
1.00	Extremes	( $\geq$ 15)

Stem width: 10.00  
Each leaf: 1 case(s)

Note the "Extremes" listings.

The volcanos data looks like this:

volcano heights Stem-and-Leaf Plot

Frequency	Stem &	Leaf
8.00	0 .	25666789
8.00	1 .	01367799
23.00	2 .	00011222444556667788999
21.00	3 .	011224445555566677899
21.00	4 .	011123333344678899999
24.00	5 .	001122234455666666677799
18.00	6 .	001144556666777889
26.00	7 .	00000011112233455556678889
12.00	8 .	122223335679
14.00	9 .	00012334455679
13.00	10 .	0112233445689
10.00	11 .	0112334669
9.00	12 .	111234456
5.00	13 .	03478
2.00	14 .	00
3.00	15 .	667
2.00	16 .	25
2.00	17 .	29
6.00	Extremes	(>=18500)

Stem width: 1000.00

Each leaf: 1 case(s)

Note all the detail in this distribution that you don't get from the boxplot.