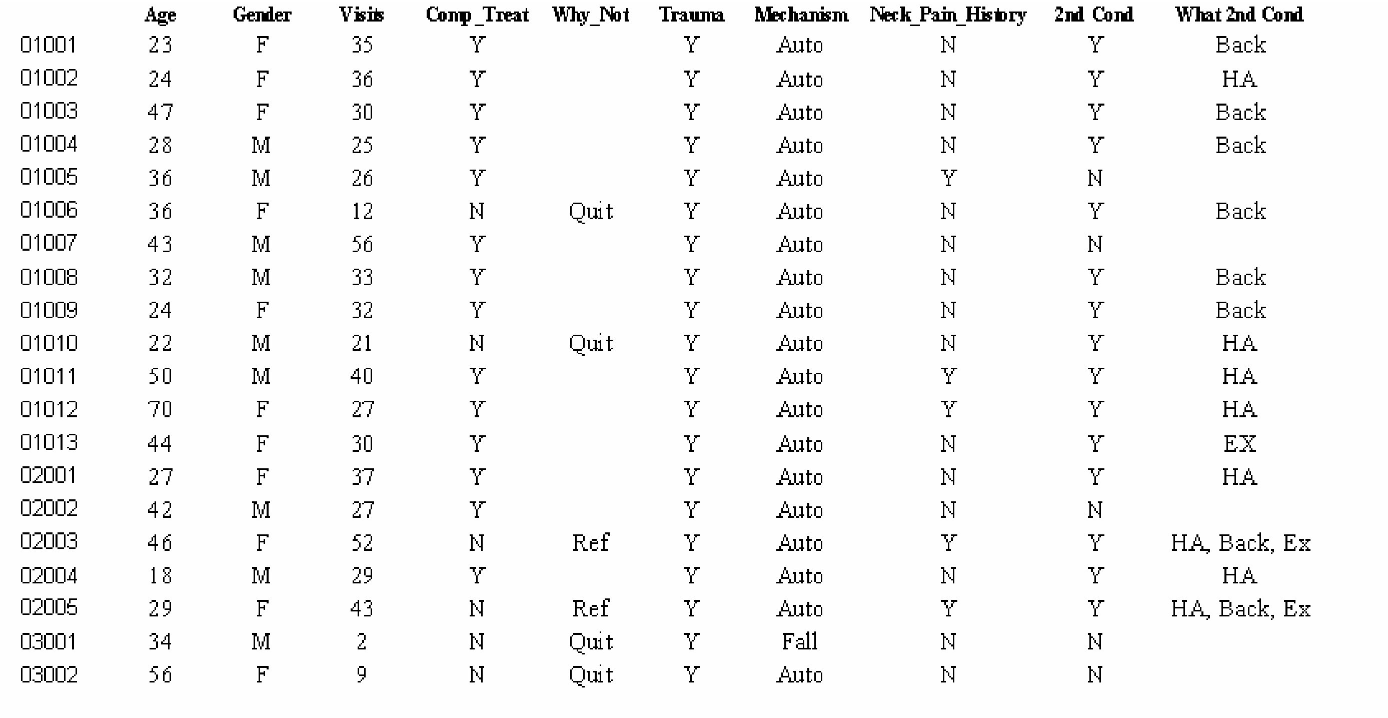
Descriptive Statistics

Michael T. Haneline, DC, MPH

Descriptive statistics (DS) characterize the shape, central tendency, and variability of a set of data. When referring to a population, these characteristics are known as **parameters;** with sample data, they are referred to as **statistics**.

Data (plural) are the measurements or observations of a variable. A **variable** is a characteristic that can be observed or manipulated and can take on different values.

Below is an example study spreadsheet containing the kinds of data typically seen in chiropractic-related research:



	Age	Gender	Visits	Comp_Treat	Why_Not	Trauma	Mechanism	Neck_Pain_History	2nd Cond	What 2nd Cond
01001	23	F	35	Y		Y	Auto	N	Y	Back
01002	24	F	36	Y		Y	Auto	N	Y	HA
01003	47	F	30	Y		Y	Auto	N	Y	Back
01004	28	M	25	Y		Y	Auto	N	Y	Back
01005	36	M	26	Y		Y	Auto	Y	N	
01006	36	F	12	N	Quit	Y	Auto	N	Y	Back
01007	43	M	56	Y		Y	Auto	N	N	
01008	32	M	33	Y		Y	Auto	N	Y	Back
01009	24	F	32	Y		Y	Auto	N	Y	Back
01010	22	M	21	N	Quit	Y	Auto	N	Y	HA
01011	50	M	40	Y		Y	Auto	Y	Y	HA
01012	70	F	27	Y		Y	Auto	Y	Y	HA
01013	44	F	30	Y		Y	Auto	N	Y	EX
02001	27	F	37	Y		Y	Auto	N	Y	HA
02002	42	M	27	Y		Y	Auto	N	N	
02003	46	F	52	N	Ref	Y	Auto	Y	Y	HA, Back, Ex
02004	18	M	29	Y		Y	Auto	N	Y	HA
02005	29	F	43	N	Ref	Y	Auto	Y	Y	HA, Back, Ex
03001	34	M	2	N	Quit	Y	Fall	N	N	
03002	56	F	9	N	Quit	Y	Auto	N	N	

DS provide summaries about various aspects of a sample or a population and can be distinguished from inferential statistics (where hypotheses may be tested).

DS include measures of:

- **Distribution** - frequencies of the values of the observations
- **Central tendency** - define the middle of a distribution of values
- **Dispersion** - the spread of values around the central region of a distribution

For example, we could use DS to characterize the number of visits provided to patients in the following study:

Case #	**Total Visits**
1	7
2	2
3	2
4	3
5	4
6	3
7	5
8	3
9	4
10	6
11	2
12	3
13	7
14	4

Distribution:

- Defined by the frequencies of each of the values. There are 3 – 2s, 4 – 3s, and so on . . .

  2 – **3**
  3 – **4**
  4 – **3**
  5 – **1**
  6 – **1**
  7 – **2**

- In addition to frequencies, the percentage of each of the values and cumulative percentages further help describe the data. This table is usually provided by statistical software packages.

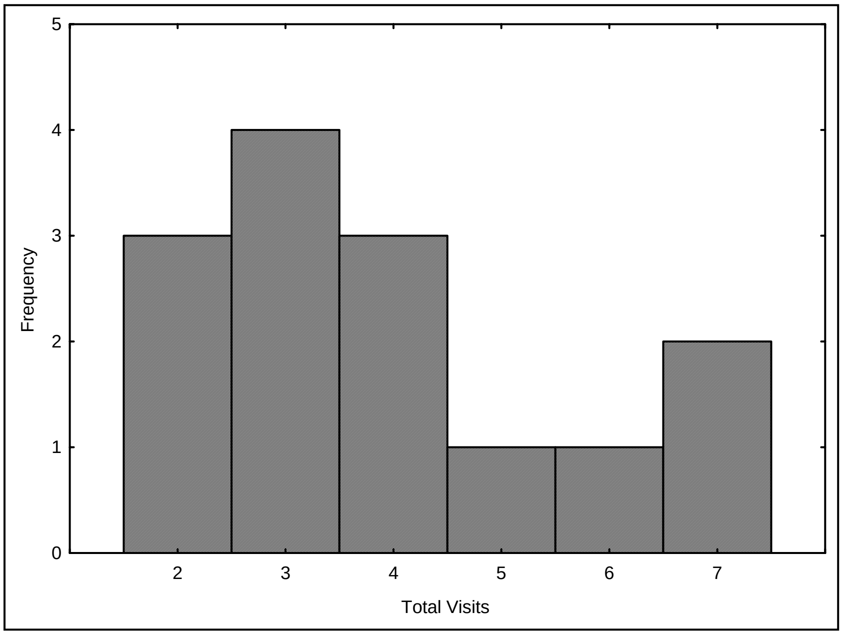
	Frequency	Percent	Cumulative %
2 –	3	21.4	21.4
3 –	4	28.6	50.0
4 –	3	21.4	71.4
5 –	1	7.1	78.5
6 –	1	7.1	85.6
7 –	2	14.3	100.0

- The ranges of values for a study are also frequently given in a research paper.

  Lowest = 2
  Highest = 7

- Histograms (a type of bar graph) are used to visually depict a frequency distribution. This type of graph merely utilizes bars to depict the same numbers generated in the above frequency distribution.

Histogram of the above hypothetical study:



The following table is produced by most statistical programs (from Microsoft Excel in this case):

Descriptive Statistics	
Mean	3.93
Standard Error	0.46
Median	3.50
Mode	3.00
Standard Deviation	1.73
Sample Variance	3.00
Kurtosis	-0.54
Skewness	0.75
Range	5.00
Minimum	2.00
Maximum	7.00
Sum	55.00
Count	14.00

Measures of **Central Tendency**

- **Mean**
  - Most commonly used descriptive statistic (parameter)
  - Also known as the average
  - The mean is easily calculated by adding all values of a series of numbers, and then dividing by the number of elements.
- **Mode**
  - The most frequently occurring value
- **Median**
  - The number that divides a series in half when all elements are listed in order.

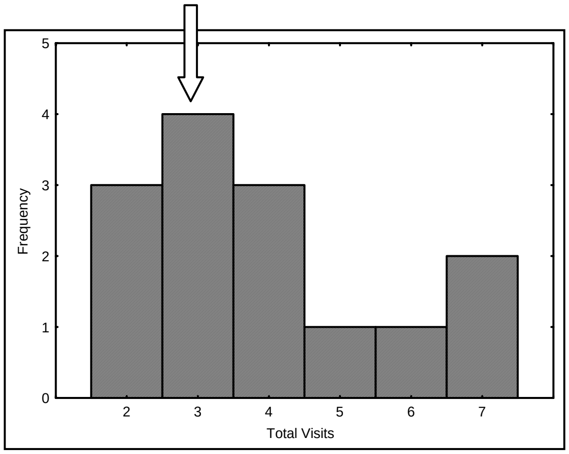
Formulas for calculating the mean ($\overline{X}$ refers to a sample and μ refers to a population)

- Mean of a sample $$\overline{X} = \frac{\Sigma X}{n}$$

- Mean of a population $$\mu = \frac{\Sigma X}{N}$$

The **modal** value is the one that occurs most frequently; 3 in this case, which occurs 4 times:

2 – 3
3 – 4
4 – 3
5 – 1
6 – 1
7 – 2

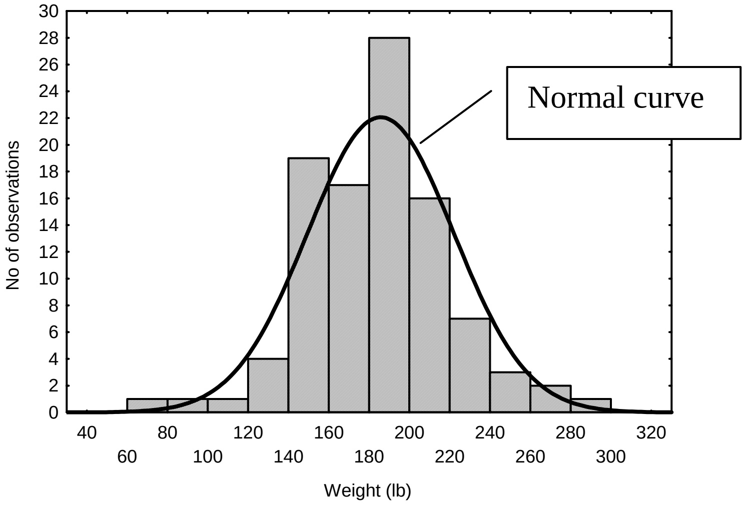


The **median** divides a series in half when all elements are listed in order. When there are an odd number of values, it's merely the middle value. When there are an even number of values, count from each end of the series toward the middle and average the 2 middle values:

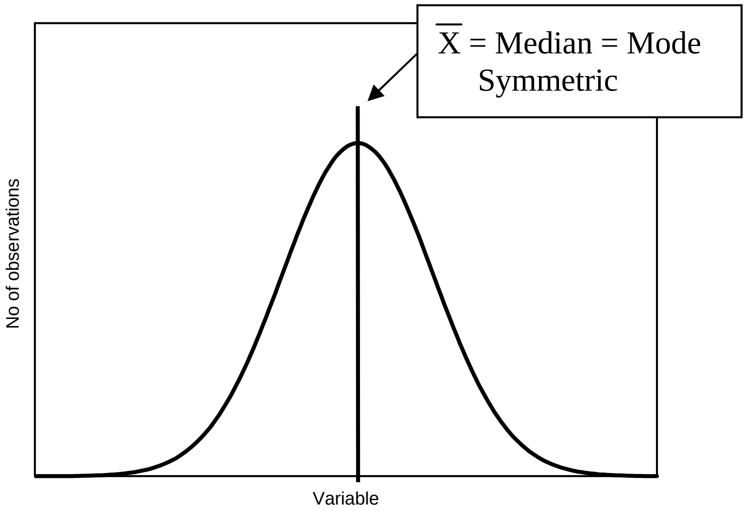
2 2 2 3 3 3 3 4 4 4 5 6 7 7

Median = (3 + 4)/2 = 7/2 = 3.5

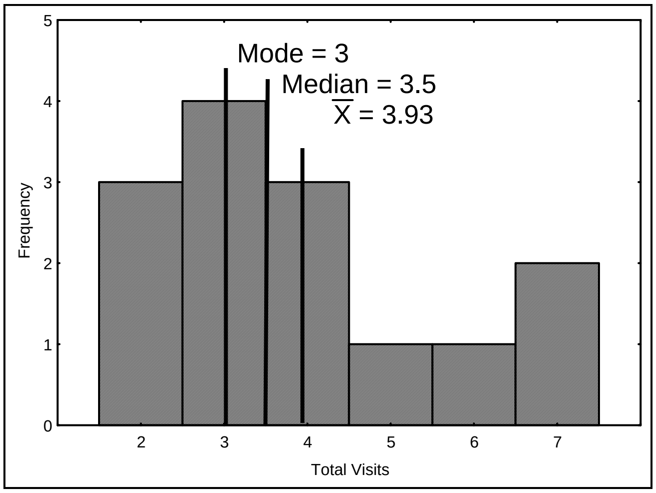
A frequency distribution of most biological data would be symmetrical along the X-axis forming a bell-shaped curve. This is known as a **normal curve** and the associated data has a **normal distribution**. The theory of the normality of data is very important to hypothesis testing.



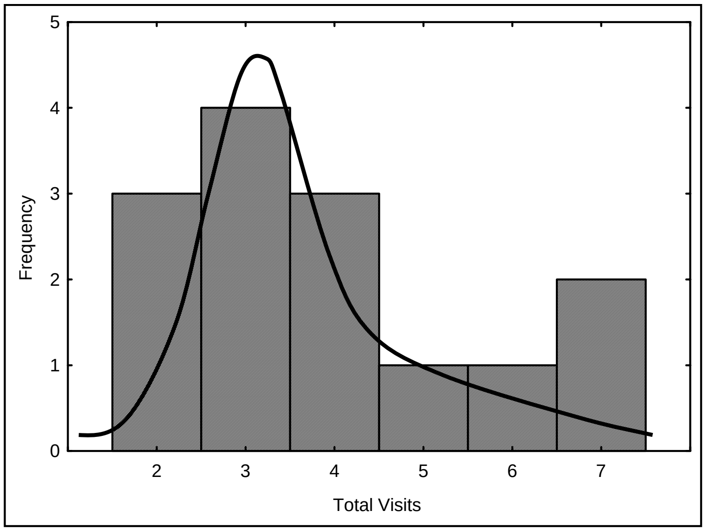
The normal distribution has defined properties with the mean, median and mode all being equal and at the middle of the normal curve. This divides the normal curve into equal halves that have similar properties.



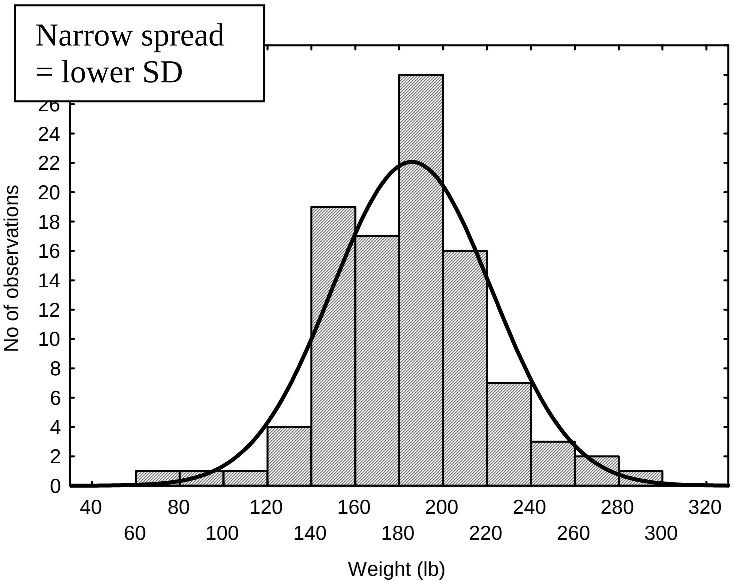
The data from our hypothetical study do not have similar $\overline{X}$, median, and mode, so the data are referred to as being **skewed** (lopsided distribution caused by extreme values).

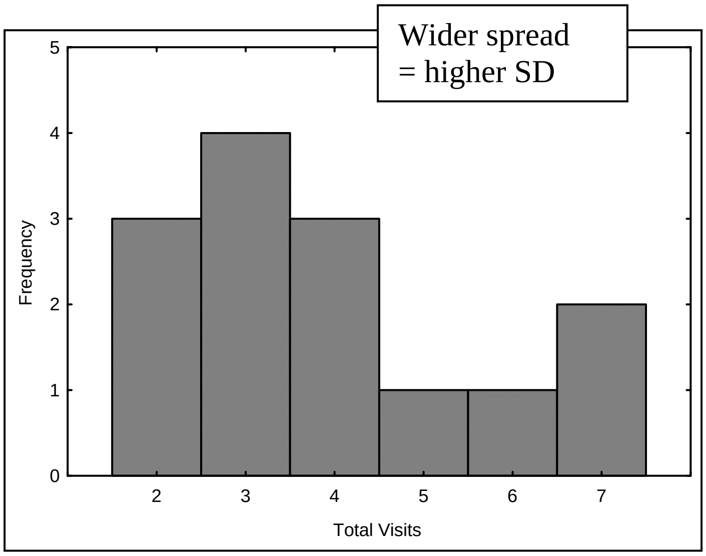


This distribution is skewed to the right (skew is always to the direction of the longer tail)

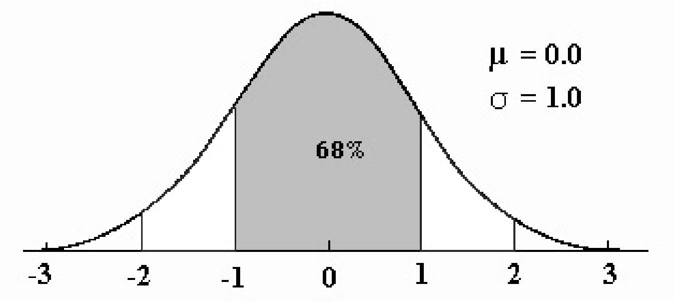


When standard deviations are calculated for the following distributions, data that are more spread out will result in correspondingly higher SDs.

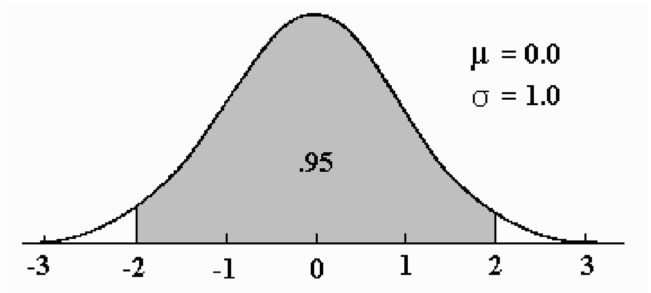

Narrow spread = lower SD


Wider spread = higher SD

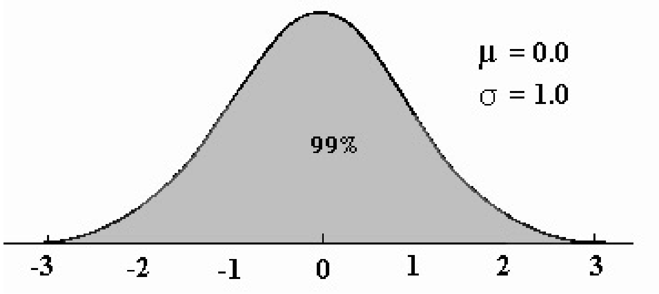
Properties of the normal distribution:



68% of scores will be within ±1 *s* of the mean



95% of scores will be within ±2 *s* of the mean



99% of scores will be within ±3 *s* of the mean

Actually, it's 68.26%, 95.44%, and 99.72%.

Using the properties of a normal distribution, we can make accurate statements about the underlying data. For instance, we can say something like: we are 95% confident that a certain value lies within 2 standard deviations of the mean or that we are 95% confident that 2 given values define the upper and lower limits of the interval (i.e., a 95% confidence interval).

Levels of Measurement

There are **four levels of measurement**, each one having specific rules for the operation and interpretation of associated data. Comprehending and applying these rules is imperative in deciding which mathematical operations are allowed and in determining which statistical test should be used.

- **Nominal measurement –**
    - AKA classificatory scale because individuals or objects are classified into categories
    - A code in the form of a number, name, or letter is assigned to each category
    - Examples
        - Male (may be coded as 0) and female (coded as 1)
        - Agree (0) disagree (1)
        - Hair color – blonde (0), brunette (1), red (2), black (3)
    - Counting of the categories is the only permissible mathematical operation (i.e., 25 males and 30 females).

- **Ordinal measurement –**
    - Includes categories, but they are rank-ordered
    - Examples
        - Lieutenant, captain, major, colonel, general
        - Normal, minimal, moderate, severe
    - Ordinal value only represents position of order, not quantity. Therefore, no mathematical operations are possible.

- **Interval measurement –**
    - Exhibit equal intervals between ordered measurements
    - Does not have a true zero
    - Examples
        - Fahrenheit scale – has zero, but does not signify the absence of heat
        - Height, weight, and age qualify as interval, but also qualify at a higher level of measurement . . . ratio.

- **Ratio measurement –**
    - There are equal intervals and a true zero
    - The most advanced level of measurement
    - Examples
        - Cervical range of motion
        - A patient's capacity for lifting
    - Does a measurement qualify as being ratio?
        - Does a measurement of zero represent an absence of the characteristic being tested?
        - Do differences between consecutive measured numbers represent equal amounts of a characteristic?

Nominal
Ordinal
Interval
Ratio

The mnemonic NOIR can help you remember the order of these terms

When to use mean, median, or mode –

Measurement scale	**Best measure of “middle”**
Nominal	Mode
Ordinal	Median
Interval	Symmetrical – Mean Skewed – Median
Ratio	Symmetrical – Mean Skewed – Median