TSWRAFPDCW BHONGIR

INTRODUCTION TO STATISTICS

Department of statistics

Fundamental Applied Statistics

Statistics - deal with the collection, classification, description, presentation, and analysis (interpretation) of data (numerical information).

- It became a method of inquiry during the Biological revolution of the 19th **century**.

- Also considered a study in the ***variation*** in data.

- Based upon ***observations*** or ***measurements*** of data.

- Statistics is ***inductive*** – specific measurements or observations yield a more general conclusion.

- Statistical ***inference*** is also scientific inference. (***Inference*** is a conclusion based upon known facts).

- Statistics relies upon some notion of replication.

- It follows, estimates can be derived, and variation and uncertainty of an estimate understood from repeated observations.

- To promote critical & analytical thinking.

- To describe & summarize data.

- Generalize complex spatial patterns. Clustering, uniform, regular, & random.

- Estimating probabilities of outcomes for an event at a given location.

- Comparative analysis of magnitude & frequency of phenomena from different locations. Population, density of springs (fractures, lithologic similarities), etc.

- Comparative spatial pattern analysis. Does an actual pattern match an expected pattern?

Selected Vocabulary

- **Data** – Numerical information.

- **Data set** – groups of data in tabular format. Consist of, *observations*, *variables*, & *variates*.

    - **Observations** – elements of phenomena (individuals, cases).

    - **Variable -** a property that can be measured, classified, or counted, e.g. male/female, discharge/velocity, etc.

    - **Variate -** a particular value of a variable, e.g. discharge in M^3S^{-1}, velocity in MS^{-1}.

• **Descriptive Statistics -** a concise, numerical, or quantitative summary is reported for a variable or data set.

- Measures of central tendency
- Measures of dispersion & variability.
- Measures of shape or relative position.
- Spatial data.
- Locational Issues.

• **Inferential Statistics -** a reported result (generalization) is derived from a sample of a larger population. Based upon probability theory.

- **Estimation –** based upon confidence intervals.
- **Hypothesis testing – Z** & **t** tests.

• **Sampling Statistics** – a portion of the total set of data.

• **Parameter -** a property descriptive of a population. Expressed by Greek letters: σ is standard deviation, μ is a population mean, and σ^2 is a population variance.

• **Estimate** (statistic) **-** a property of a sample drawn at random from a population. Estimates are expressed by roman letters. Standard deviation is **s**, mean is **x**, and variance is $\mathbf{s^2}$.

•***Function*** – when two variables are related such that the values of one are dependent upon the values of the other.

• If the functional relationship is not known, ***causal*** conclusions can not be inferred.

Characteristics of Data

- **Primary Data –** acquired directly from an original source, i.e. from the field.
- **Secondary Data –** preexisting data from an agency or other source.

Variables of the Data Set

- ***Continuous Variable*** **-** any value within a defined range of values. Values which belong to a continuous series include; height, weight, chronological time, discharge, velocity, etc.

- ***Discontinuous (discrete) Variable*** **-** specific (counted & limited to whole numbers) values only. The size of a family (3) implies the exact size of the group. Other examples include school enrollment & number of books in a library.

Levels of Measurement

- ***Nominal Variable*** - a qualitative property of equality or difference in established categories. Variables must be exhaustive & mutually exclusive. (soil classes, tree species, well locations, etc.)

- ***Ordinal Variable*** - a property of equality or difference & rank order within the data. (soil classes ordered by high, intermediate, or low sand content, well locations ordered by proximity to a town, etc.)

- ***Interval variable*** - a property of equality or difference, order, & no true 0 (starting) point within the data (temp in F^o or C^o).

- ***Ratio variable*** - a property of equality or difference, order, & a true 0 (starting) point within the data (soil classes ordered by percent sand, wells ordered by discharge values).

- Interval data may transformed to ratio data by subtracting the differences of variates which eliminates or cancels out the arbitrary origin.

Principles of Measurement

- **Precision** – Degree of exactness. A measure of repeatability. How close positions are clustered. Based on a relative reference, e.g. a circle 1 inch in diameter.

- **Accuracy** – Closeness of a position to a known absolute reference system.

- **Validity** – based upon operational definitions of acceptance. Subjective parameter, e.g. level of poverty, quality of …, etc.

- **Reliability** – how consistent, repeatable, or stable is the data over changes in spatial pattern over time?

Basic Statistical Properties

Constant (c) – a property common to all members of a group.

• **Property1 -** Multiplying a constant (**c**) by each score is equal to adding all the scores (ΣX_i), and multiplying by a constant (**c**): $c\Sigma X_I$:

$$c\Sigma X_i = cX_1 + cX_2 + cX_3 + ...cX_n$$

• **Property 2 -** If a given constant (**c**) equals 4, and there are 5 variables (**N**), then: $\Sigma C = C + C + C + C + C$, equals **NC**:

4+4+4+4+4 = 20, and **N(5) x c(4) = 20**.

• **Property 3 -** The summation (**Σ**) of the sum of any number of terms is the sum of the summations of these terms taken separately:

$$\Sigma(X_i + Y_i + Z_i) = (X_1 + Y_1 + Z_1) + (X_2 + Y_2 + Z_2) + (X_3 + Y_3 + Z_3)\ldots = \Sigma X_i + \Sigma Y_i + \Sigma Z_i \ \ldots$$

• **Property 4 -** The sum of the products of two sets of paired numbers is:

$$\Sigma X_i Y_i \text{; which equals: } X_1Y_1 + X_2Y_2 + X_nY_n\ldots$$

• **Property 5 -** Given a set of values ($\mathbf{X_n}$), the sum of the squared values ($\mathbf{X_n^2}$) is equal to : $\mathbf{\Sigma X_n^2}$ where,

$\mathbf{X_n}$	$\mathbf{X^2}$
3	9
2	4
5	25
6	36
4	16
$\mathbf{\Sigma X_n = 20}$	$\mathbf{\Sigma X_n^2 = 90}$

• **Property 6 -** Given a set of values ($\mathbf{X_n}$), the square of the sum of the values ($\mathbf{X_n}$'s) is equal to $\mathbf{(\Sigma X)^2}$

where:

$\mathbf{X_n}$
3
2
5
6
4
$\mathbf{\Sigma X_n} = 20$

$\mathbf{(\Sigma X_n)^2 = (20)^2 = 400}$

The Normal Distribution (Curve)

• Frequency curves are conceptualized as extending across the X axis from minus infinity to plus infinity, although they realistically taper off at 0.

• The area under the curve is always finite, and for convenience is taken as unity.

• On occasion it becomes necessary to find the proportion of the total area of the curve between ordinates (X = c, & X = d) erected at particular values of X such that this proportion is the probability that a particular value X drawn at random from the population which the curve describes falls between c and d.

• Therefore, frequency curves are often called **probability curves**.

• The ***Normal Curve*** is written in **standard score** (**Z**) form with a ***mean*** equal to **0**, ***variance*** equal to **1**, and ***standard deviation*** equal to **1**.

• The area under the curve is taken as unity (**1**).

• Thus, the area under the normal curve is divided into standard deviation units such that +1 standard deviation unit accounts for 0.3413 of the total area under the curve.

• Alternatively, -1 standard deviation unit also accounts for 0.3413 of the total area under the curve.

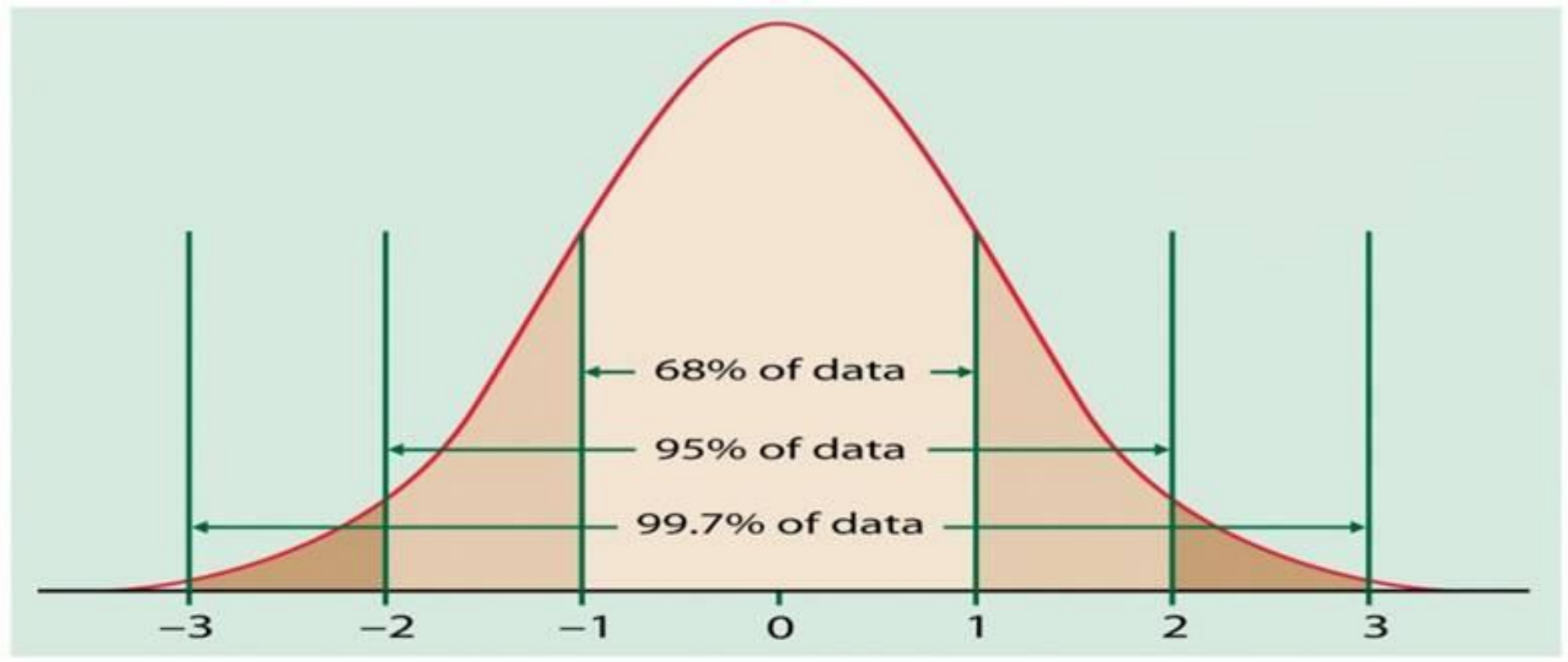
• Therefore plus or minus 1 standard deviation unit accounts for approximately 0.6826 of the total area under the normal curve.

• When considering the area between +1 or -1 standard deviation units to + 2 or - 2 standard deviation units, the area accounted for is 0.1359 on each side.

• When this value is combined with plus or minus 1 standard deviation unit, 0.9544 of the total area under the normal curve is accounted for by plus or minus 2 standard deviation units.

• Three standard deviation units would only account for an additional 0.0214 increase on each side of the mean for a total of 0.9972 of the total area under the curve explained.

The Normal Curve



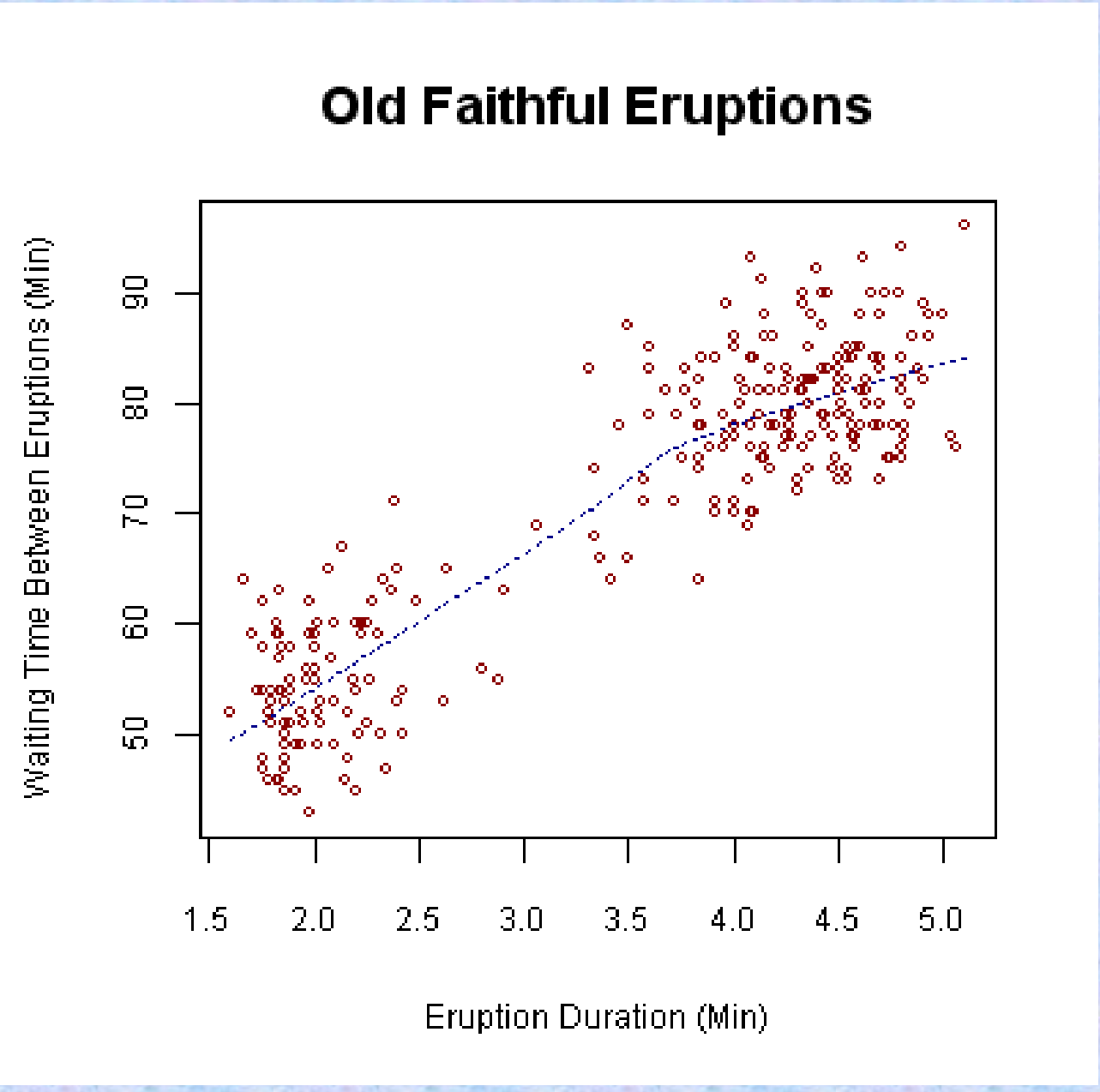
FREQUENCY DISTRIBUTION

• **Scatterplot –** graphic distribution of two variables (points), e.g. temperature/elevation.

• A ***frequency distribution*** shows the number of times each score occurs, and arranges scores from lowest to highest.

• Given a frequency distribution of values (for example brightness values of 56, 57, 67, 99, 120), the ***Ogive*** or ***cumulative frequency*** (**C_f**) is a continuous count of frequencies for each BV at or below a given level.

Scatterplot



• The ***Ogive*** or ***cumulative frequency*** (**C_f**) plot is a continuous count of frequencies for each BV at or below a given level.

• The ***cumulative percentage frequency*** (**C_f%**) is the percentage of a given number of BV's compared to the total (200 in this example, **C_f / 200**).

BV's	**Frequency**	**C_f**	**C_f%**
56	5	5	0.025
57	34	39	0.20
67	100	139	0.70
99	45	184	0.92
120	16	200	1.00

Frequency = 200

- A ***percentile rank*** is a ***percentile*** corresponding to a raw score.

- If one is in the 90 percentile (percentile rank), it would be interpreted as 90 percent of the scores are at or below this value.

To obtain the ***percentile rank*** for a given score, e.g. (67):

• Calculate the lower true limit of the score (67) by subtracting 0.5 unit from the score (66.5).

• Subtract the lower limit (66.5) from the score whose percentile rank is being estimated (67).

• Multiply the result by the frequency of scores with a value of 67 (100).

• Divide the result by the width of the class interval (1).

• Add the result to the cumulative frequency (139).

• Divide the result by the total number of frequencies (200).

(((67-66.5)(100)/1)+139)/200 = .945 x 100 = **95%**.

STANDARD SCORES (Z scores) – a transformation from raw scores to *standard deviation units* used to compare a score with a collection of scores derived from different procedures (English vs. Mathematics test).

• Position is considered rather than the magnitude and measurement of units of scores. $Z\ score = (X_i - X_{bar}) / s$

• Properties of a *Z score*:

If a raw score is $>X_{bar}$, positive Zscore.

If a raw score is $<X_{bar}$, negative Zscore.

If a raw score = X_{bar}, 0 Zscore.

• ***Standard scores*** have a ***mean = 0***, and a ***standard deviation = 1***, thus they are readily amendable to algebraic manipulation.

• After computing for a *standard score*, locate the **Z** value in a table of **Z** scores.

• This will give a value of area between the mean and a **Z** score.

• If the raw score is greater than the mean, add the **Z** value to 50 to obtain the ***percentile rank***.

• If the raw score is lower than the mean, subtract the **Z** value from 50.

• In a given distribution of brightness values (frequency distribution), it is important to recognize a variety of properties about the distribution.

• The four properties include - ***Central Tendency***, ***Variation***, ***Skewness***, *and* ***Kurtosis***.

MEASURES OF CENTRAL TENDENCY

Central Tendency - refers to a method of describing the spread of the distribution of scores around a central measure of the frequency distribution.

• The four properties include – ***mode***, ***median***, & ***arithmetic mean.***

ARITHMETIC MEAN (X_{bar}) - The sum of the value of X_i **(ΣX_i)** multiplied by the frequency of its occurrence (f_i), divided by the number of measurements (**N**): Mean equals the arithmetic average:

$$X_{bar} = \Sigma X_i / N \quad \textbf{also}$$

$$X_{bar} = f_1X_1 + f_2X_2 + ...f_nX_n/N = \Sigma f_iX_i \ / \ N$$

where: $\Sigma X_i = \Sigma f_iX_i$

Deviation from the mean ($X_i - X_{\underline{bar}}$) or x_i) - the difference between a particular score (X_i) and the mean (X_{bar}):

$$x_i = (X_i - X_{\underline{bar}})$$

• **Property 1 of the mean -** The sum of the deviations of all the measurements in a set from their arithmetic mean equals **0**.

$$\Sigma(X_i - X_{bar}) = \Sigma X_i - \Sigma X_{bar} = NX - NX_{bar} = 0$$

since $X_{bar} = \Sigma X_i / N$; **then** $\Sigma X_i = NX_{bar}$

• **Property 2 of the mean -** The sum of squares or the sum of squared deviations from the arithmetic mean, $\Sigma(X_i - X_{bar})^2$ **or** Σx^2 is less than the sum of squares of deviations from any other value.

• **Property 3 of the mean -** The **mean** is that measure of central tendency about which the sum of squares is a *minimum*.

• The **mean** is a measure of central location in the *least square* sense.

MEDIAN - The point on the scale such that half of the observations fall above it and half below it.

MODE - The most frequently occurring value.

• If the frequency of occurrences is equal for each value, there is no mode.

• Where two values have equal frequency, the mode is determined by adding the brightness values of the two that occur equally, and dividing by the total number of repetitive values (2).

• The ***mode*** represents the highest point on a curve (histogram).

MEASURES OF DISPERSION & VARIATION

Note - As the variation between measurements increases, the departure of the observations from their sample mean increases. This is used to define a **measure of variation**.

Variation can be ***absolute*** or ***relative***.

- ***Absolute*** – variation is based upon the magnitude of values in the data set.
- ***Relative*** – variation is a ratio or proportion of the variation to typically the mean of the data set.

RANGE - The difference between the largest and smallest measurements.

- For large samples this is an unstable descriptive measure.
- It is also not normally dependent upon sample size.

Mean Deviation (a measure of ***dispersion***) **-** the sum of the deviations from the arithmetic mean divided by the number of values without regard to sign (absolute values (e.g. |**2**|), or the ***arithmetic mean of the absolute deviations from the arithmetic mean***.

• As an example **- M.D. =** $\Sigma|X_i - X_{bar}| / N$ where: $|X_I - X_{bar}|$ is an absolute value w/out regard to sign.

• This measure is normally avoided, as it does not lend itself well to algebraic manipulations.

SUM OF SQUARES - The sum of the squared deviations about the mean:

$SS = \Sigma(X_i - X_{bar})^2 = \Sigma x^2$ **also** $SS = \Sigma X_i^2 - (\Sigma X_i)^2 / N$

MEASURES OF VARIABILITY

SAMPLE VARIANCE - (Includes *degrees of freedom* (df) defined as the total number of variables (N) minus the number of constraints placed on the data or the number of variables free to vary).

For example given 5 measurements which equal 100, four are free to vary, but the last must equal a value which when combined with the other 4 = 100, thus

df = N-1 or 4.

SAMPLE (unbiased) **VARIANCE** (s^2) - is the *mean (average) of the sum of squared deviations around the mean, & a* preferred method of dealing with negative signs.

-

$$s^2 = \Sigma(X_i - X_{bar})^2 / N-1$$

STANDARD DEVIATION (s) – is a measure of variation in units of original measurements.

• If the deviation represents a statistic in feet, then the variance will be reported in square feet.

• By extracting the square root of the variance, we derive a unit in original measurements (ft.).

$$s = \sqrt{\Sigma(X_i - X_{bar})^2 / N-1}$$

COEFFICIENT OF VARIATION – expresses a relative measure of variability among distributions as a ratio of standard deviation to mean.

$$CV = s/x_{bar} \quad \text{or} \quad (s/x_{bar})100$$

- Removes the absolute magnitude in the data set.

- Allows direct comparison of the variability of different data sets.

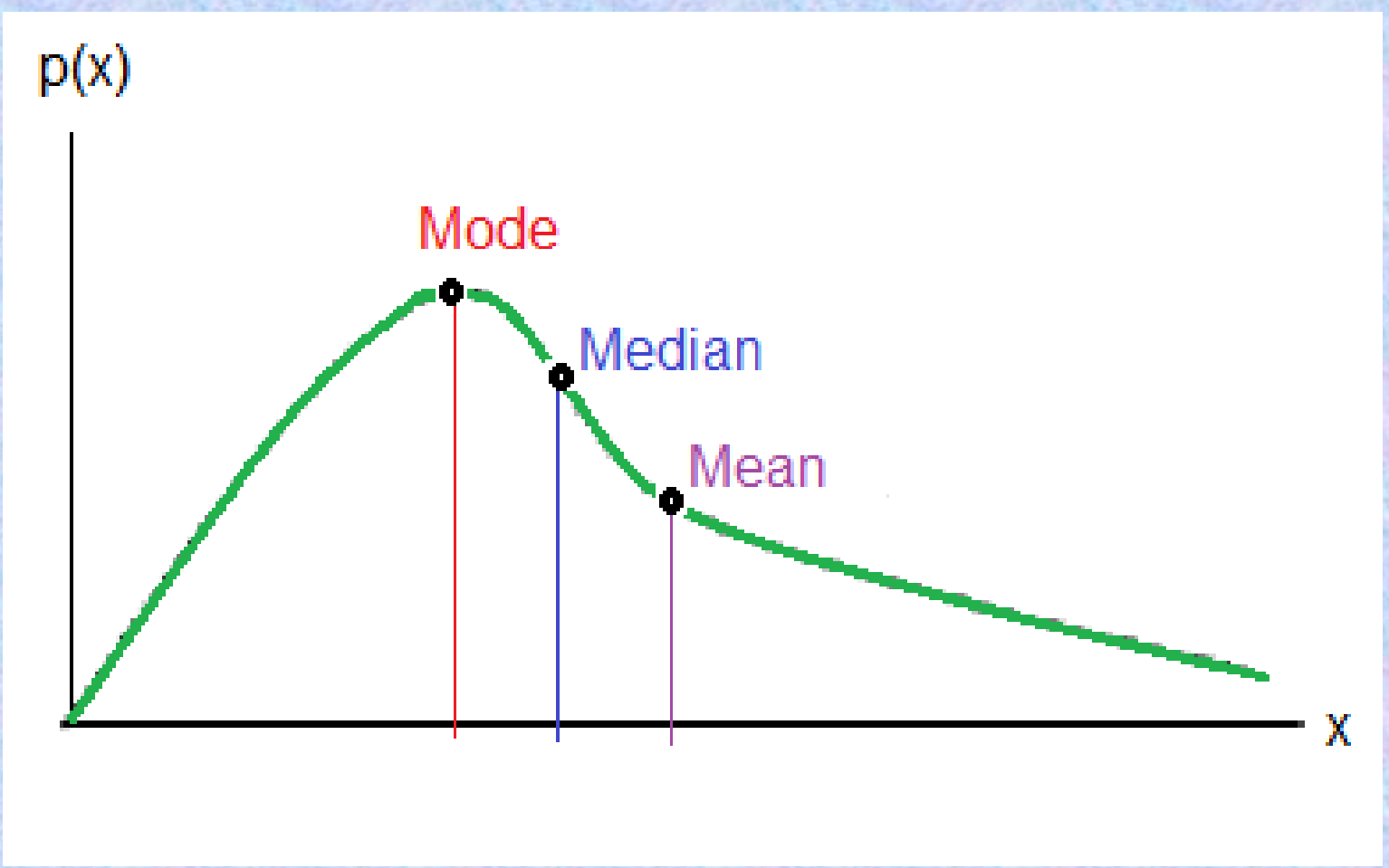
Measure of Shape/Position

Skewness - *Skewness* refers to the symmetry of a distribution.

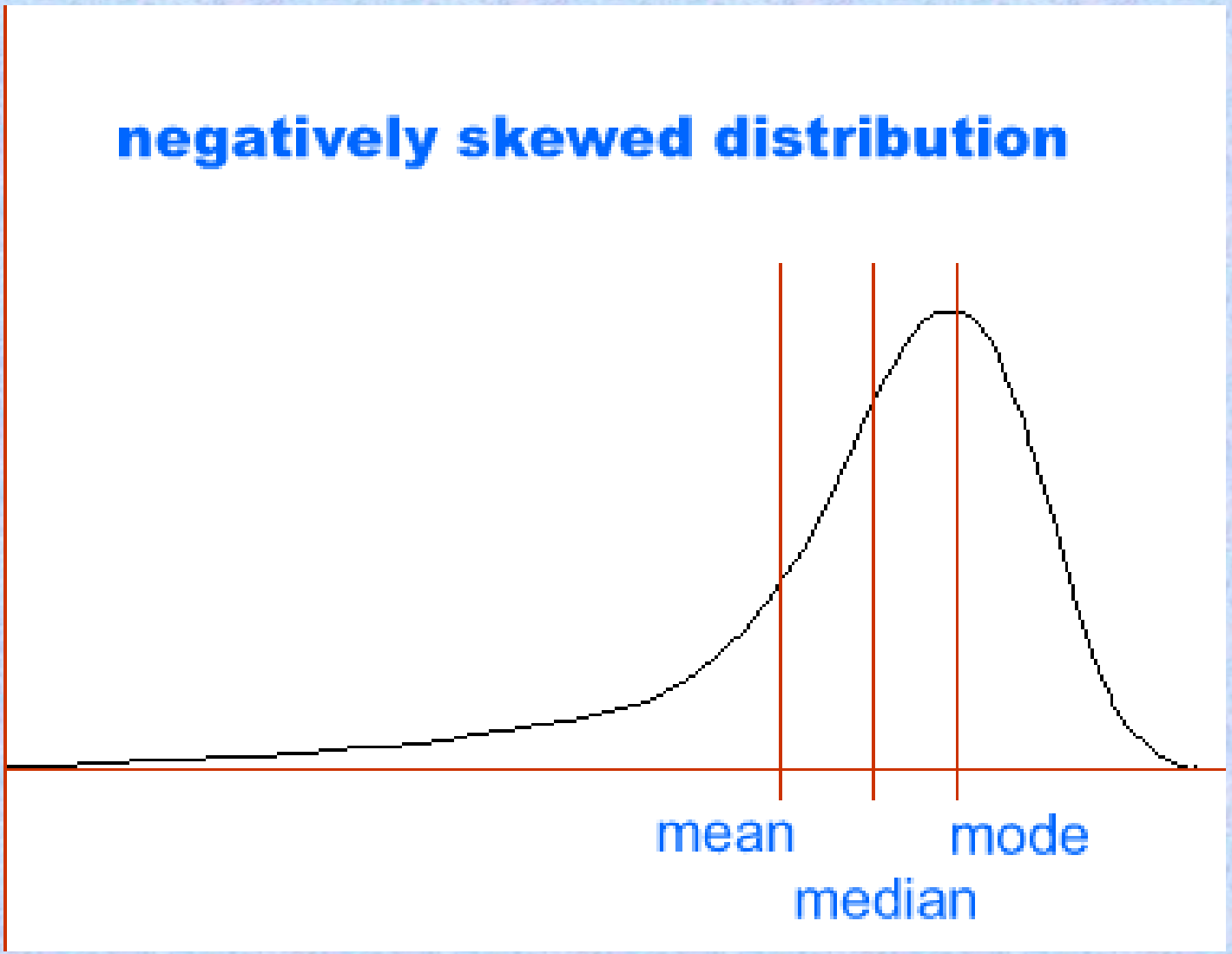
• When a distribution of values is not normal, and the tail extending long to the *right* (***positively skewed*** to the high end of the distribution).

• When a distribution of values is not normal, and the tail trends to the *left* (***negatively skewed*** to the low end of the distribution).

Positive Skewed



Negative Skewed



• The ***skewness*** is related to moments about the mean such that the expressions are as follows:

First moment: $\Sigma(X_i - X_{bar}) / N$ (Avg. of deviations from the mean).

Second moment: $\Sigma(X_i - X_{bar})^2 / N$ (Biased variance).

Third moment: $\Sigma(X_i - X_{bar})^3 / N$ (Avg. cubed deviation from the mean).

Forth moment: $\Sigma(X_i - X_{bar})^4 / N$

• By incorporating the second (M_2) and third (M_3) moments, when $g_1 = 0$, and $M_3 = 0$, the distribution is said to be **normal.**

$$\mathbf{g_1 = M_3/(M_2)^{3/2}}$$

When g_1 is = 0: Symmetrical distribution.

When g_1 is = -: Negatively skewed (to the left).

When g_1 is = +: Positively skewed (to the right).

• In a truly ***symmetrical*** (unimodal) distribution, the ***mean***, ***mode***, and ***median***, are ***equal***.

• In a ***skewed*** distribution the *mode remains unchanged*, however, the *median is displaced in the direction to which the distribution is skewed*.

• The arithmetic mean will also be displaced in the same direction and to the outside (distal end) of the median.

• It follows that in a skewed distribution, the mean is highly sensitive by the degree of ***skewness***, and ceases to describe a "***central value***" or typical value.

KURTOSIS - The degree of steepness or flatness of a distribution is measured by ***kurtosis*** which is derived from the second and forth moments:

$$g_2 = (M_4 / M_2^2) - 3$$

When g_2 is = 0: ***Mesokurtic*** (Normal distribution).

When g_2 is = <0: ***Platakurtic*** (Flat topped distribution).

When g_2 is = >0: ***Leptokurtic*** (Peaked distribution).

Kurtosis