KINE 601

Data Types – Relationships Among Data

Reading: Huck pp 17 - 74

Types of Data

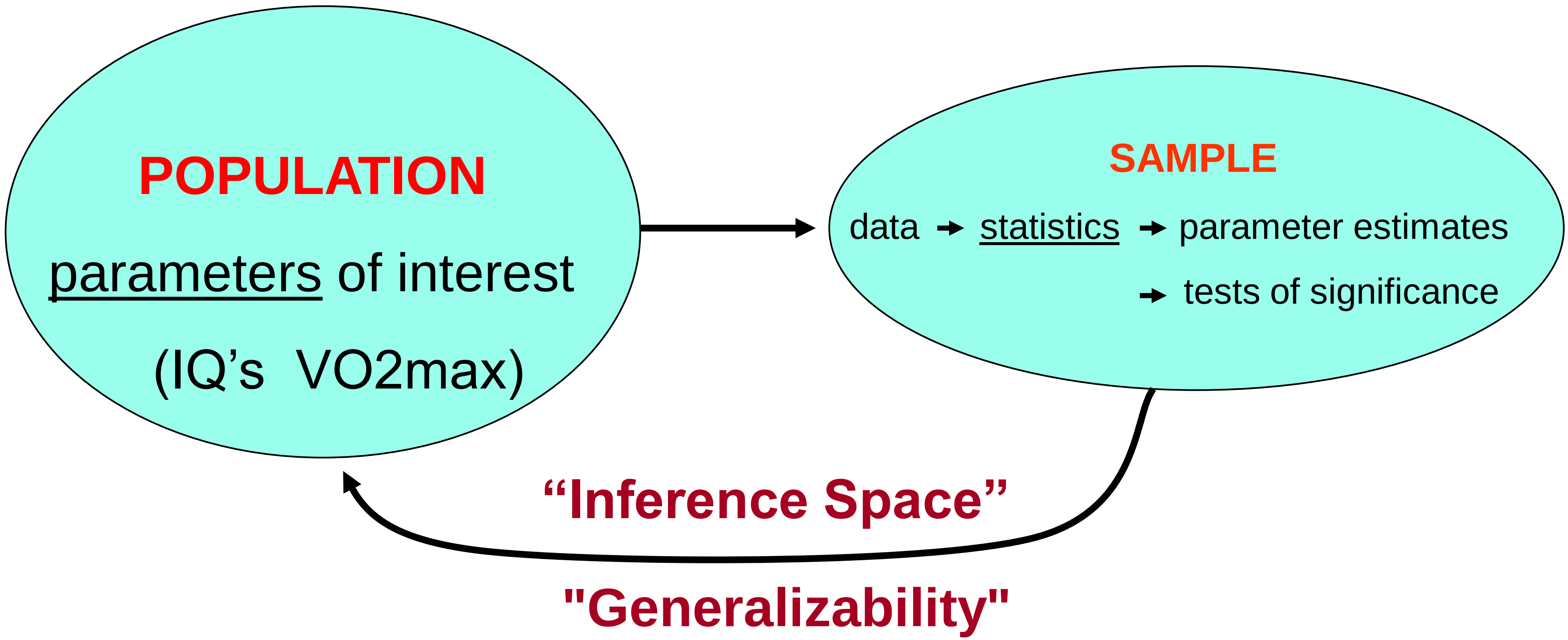
- **Continuous** - infinite subdivisions possible
  - **Interval** - no absolute zero (absence of characteristic) possible
    - Examples: IQ scores, hematocrit, VO2max, time to complete a task
    - Distance between 1 and 2 is the same as between 3 and 4
  - **Ratio** - interval data that may have an absolute zero
    - allows for more precision - ratios are technically possible
    - Examples: age, academic test scores, PSA
    - Ratio statements like 18 lbs. Is 3 times heavier than 6 lbs.
- **Discrete** - no subdivisions possible - finite number of values
  - Examples: number of siblings, number of bullets in a gun
- **Dichotomous** - two mutually exclusive polar extremes
  - Examples: yes-no or true-false responses on a questionnaire
- **Categorical (Nominal)** - arbitrary or systematic classifications
  - Examples: religious preference, age brackets, Likert scores (scaling) often treated as continuous
- **Ordinal** – rankings
  - Examples: rankings of football players according to their 40 yd dash time: 1st, 2nd, 3rd, etc.

What are Statistics ?

- **Statistics:** a **tool** of research
  - **Webster's Definition:**
    - 1. a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data (this definition associated with **inferential statistics**)
    - 2. a collection of such numerical data (associated with **descriptive statistics**)
  - **Descriptive Statistic:**
    - an index number used to **describe** or summarize sample data or a particular place in that data
    - mean, median, mode, percentile rank
  - **Inferential Statistic:**
    - a value resulting from a method of analysis of sample data that takes "chance" into account when samples are used to derive conclusions **(inferences)** about populations
    - allows for making decisions **(inferences)** from incomplete data (samples) …. hence the term **"inference space"**

Important Statistical Terms & Concepts

- **Population:** all members of a specified group
- **Sample:** a defined subset of the population
- **Parameter:** a numerical characteristic of a population
  - Parametric statistics: used when parameter from sample data comes from a population in which the 1. parameter is normally distributed and 2. sample group variance is homogeneous (obviously, the data must be **CONTINUOUS**).
- **Statistic:** a numerical characteristic of a sample

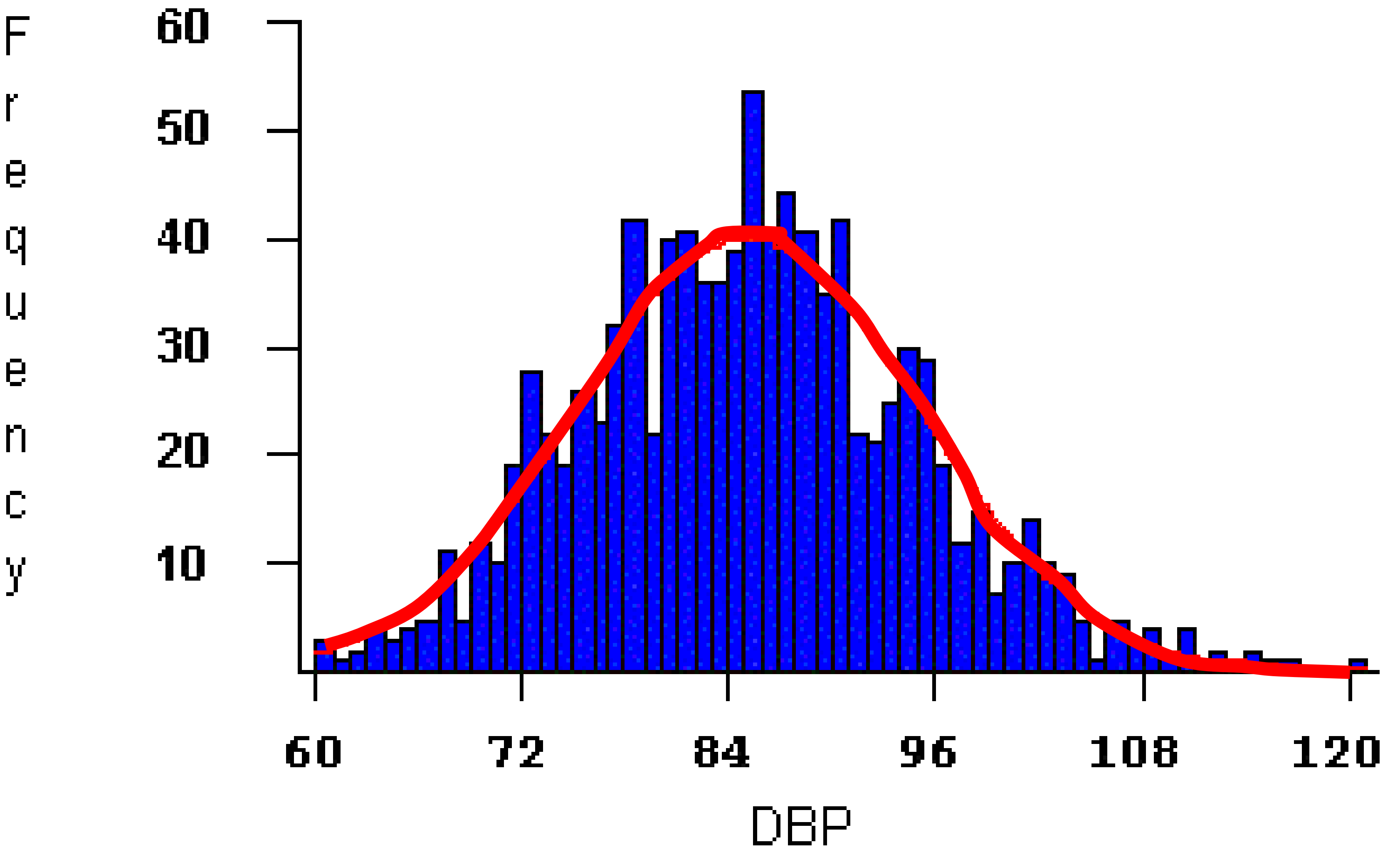


Statistical Terms & Concepts

- **univarite:** pertaining to only one dependent variable
- **bivariate:** pertaining to two dependent variables
- **multivariate:** pertaining to two or more dependent variables

- **distribution:** a group of dependent variable scores
- **frequency distribution:** distribution of dependent variable scores grouped into various types of frequency categories
  - cumulative frequency distribution - each value represents an accumulated or summed frequency
- **normal distribution:** a distribution of scores (or frequency distribution) in which most of the scores are clustered around the mean with a gradual symmetric decrease in frequency of scores in both directions away from the mean

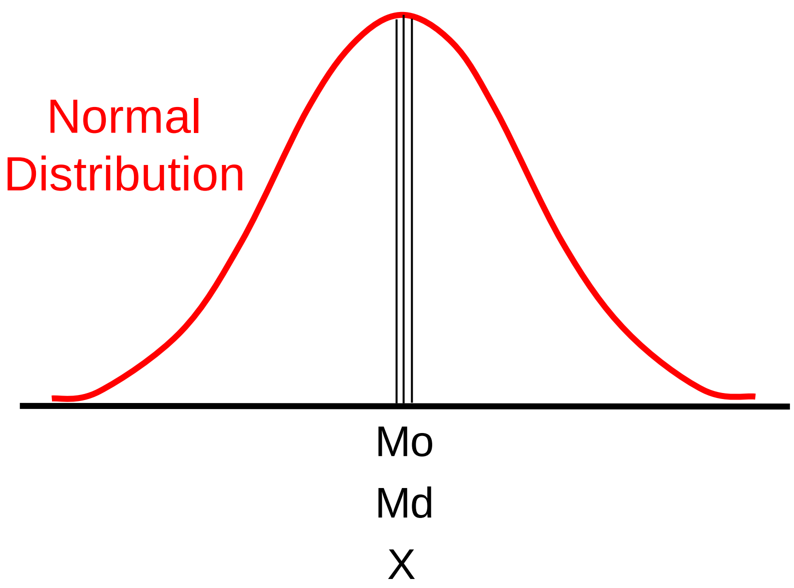
Example of a Near Normal Distribution

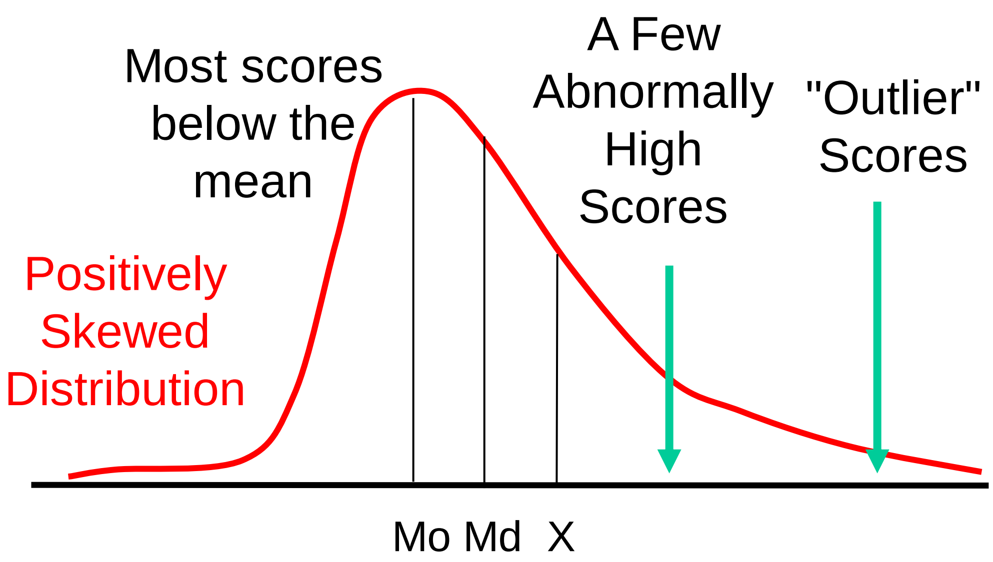


Frequency distribution of diastolic blood pressure reading for 1000 people

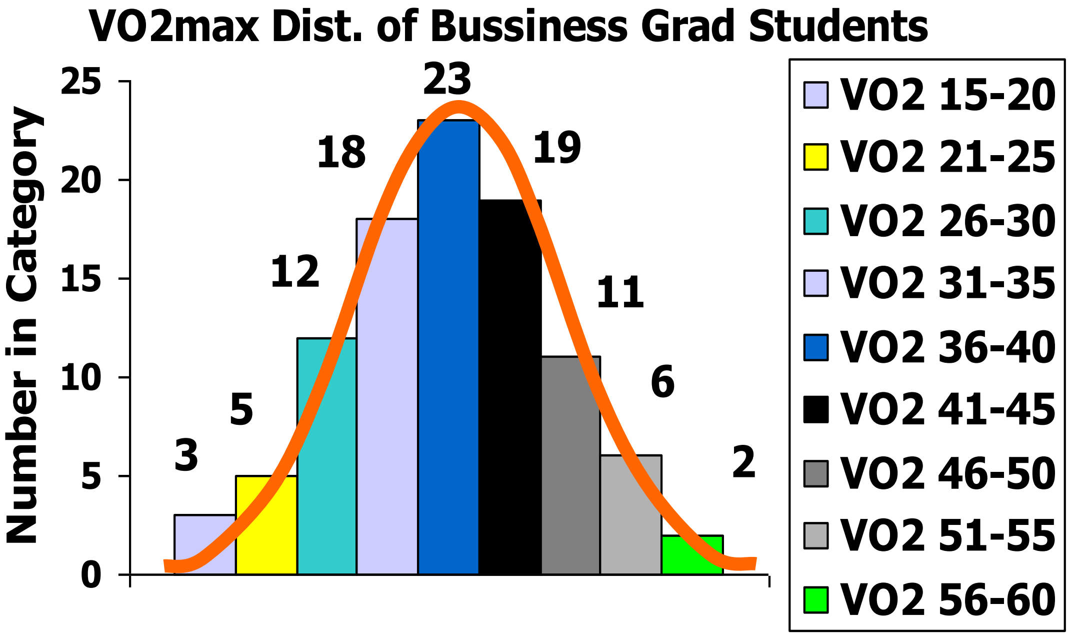
Descriptive Statistics

- **Measures of Central Tendency**
  - **Mean** - average (denoted by **x** or **y** for a sample, **m** for population)
  - **Median** - middle score - score that divides distribution into equal halves
  - **Mode** - score that occurs most often
  - In a 100% normal distribution: mean, median, & mode are the same
  - Having a few scores "strung out" on one side of the distribution or many of the scores on particular side of the mean may significantly "skew" a distribution, making it non-normal.



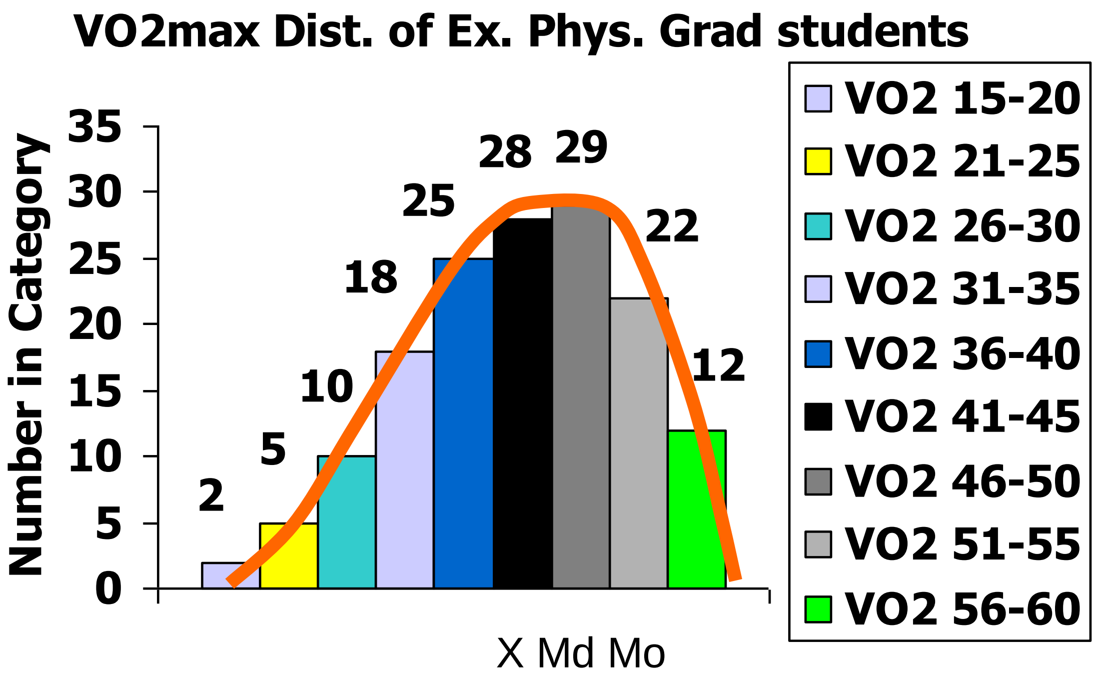


Distribution Examples



Normal Distribution Example

Negatively Skewed Distribution Example



Descriptive Statistics

- **Measures of Variability** (Spread of Scores - Score Dispersion)
  - Consider a group of "n" scores (n = number of scores): 0 1 2 3 4
    - **Range:** difference between lowest & highest score (4 - 0 = 4)
      - interquartile range: 75th %tile - 25th %tile (3 - 1 = 2)
    - **Variation:** the sum of the squared deviations of scores from the mean

      $\Sigma (X - \mathbf{X})^2$ called "Sum of Squares" "SS"

$\mathbf{X} = 2$	$0 - 2 = -2$	$(-2)^2 = 4$
	$1 - 2 = -1$	$(-1)^2 = 1$
	$2 - 2 = 0$	$(0)^2 = 0$
	$3 - 2 = 1$	$(1)^2 = 1$
	$4 - 2 = 2$	$(2)^2 = 4$
		10

Descriptive Statistics

- **Measures of Variability for distribution: 0 1 2 3 4**

  - **Variance**: the "average variation"
    - denoted by s^2 for a sample, σ^2 for a population

$$\frac{SS}{n-1} = \frac{10}{5-1} = 2.5$$

note that division would be by n for a "population". In sample statistics, you loose one **"degree of freedom"**.

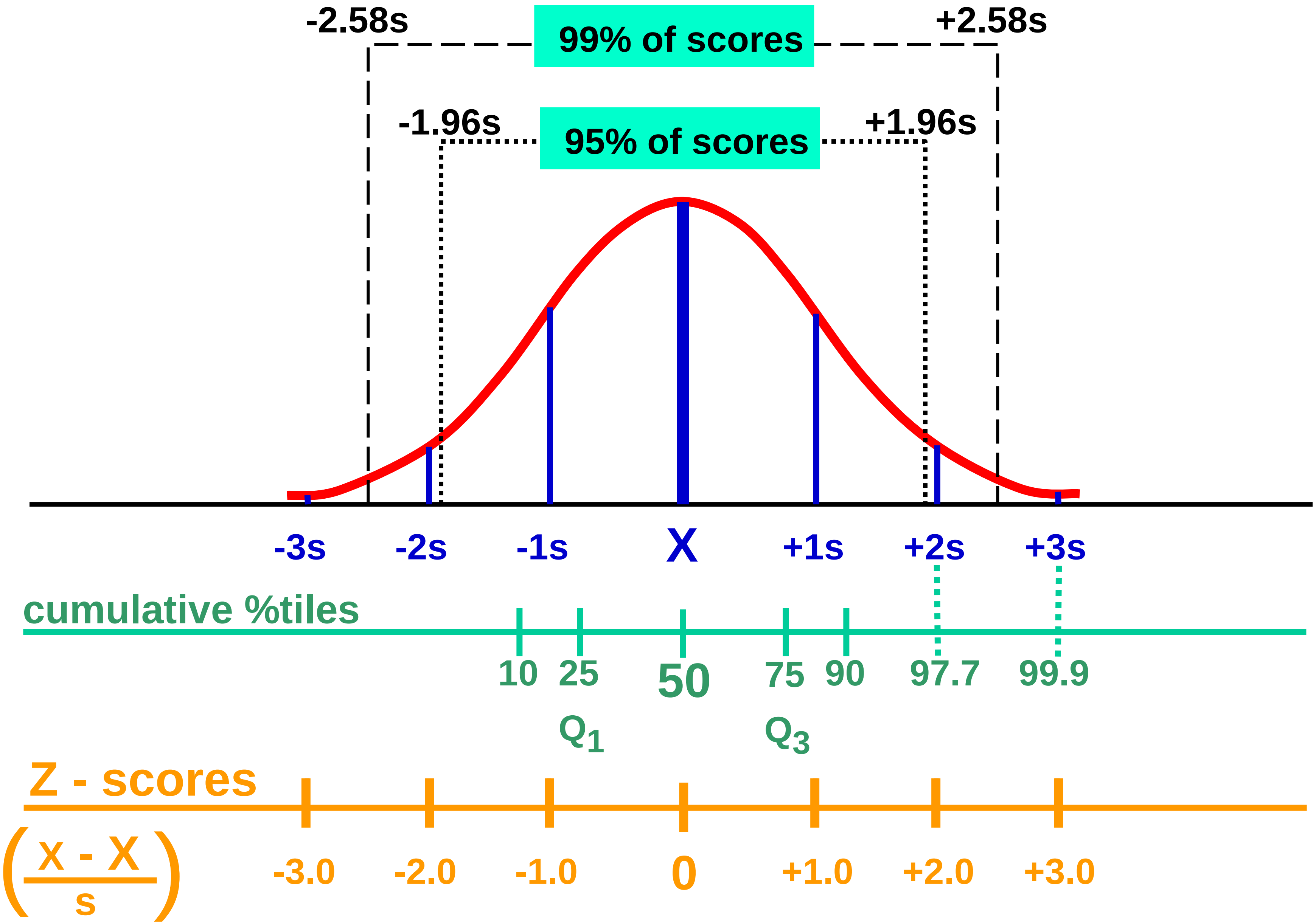
  - **Standard Deviation**: the positive square root of the variance
    - denoted by s for a sample, σ for a population

$$\sqrt{\frac{SS}{n-1}} = \sqrt{\frac{10}{5-1}} = 1.6$$

  - **Coefficient of variation**: standard deviation divided by the mean
    - used to compare the variability of two distributions with different units

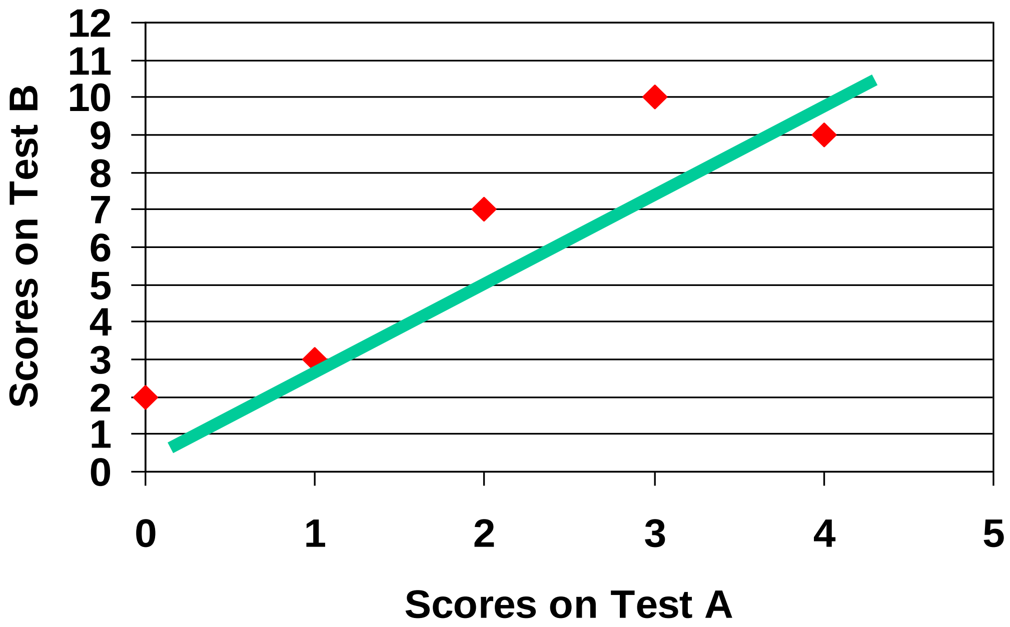
$$\frac{s}{\mathbf{X}} = \frac{1.6}{2} = .8$$

the Normal Distribution



Relationships Among Data

- **Pearson correlation coefficient: " r "**
  - requires data to be continuous and come from a normal distribution
  - a “-1 to 1" representation of the degree of **relationship**.
  - consider 2 sets of scores: x = 0 1 2 3 4 and y = 2 3 7 10 9
    - suppose the first set of scores represent a score on Test A and the 2nd set of numbers represent scores on Test B. Is there a relationship between the scores on Test A and the scores on Test B? One way this can be determined is by examining a scatter plot



Relationships Among Data

- the numerical representation of the relationship is found by calculating the Pearson correlation coefficient (**r**):

$$r = \frac{\text{co-variation of } x \text{ and } y}{\sqrt{(\text{variation of } x)(\text{variation of } y)}}$$

(how variation "corresponds") / (total variation in both variables)

$$r = \frac{\Sigma (X - \mathbf{X})(y - y)}{\sqrt{\Sigma (X - \mathbf{X})^2 \; \Sigma (y - \mathbf{y})^2}} = .93$$

- this correlation coefficient can be positive or negative
  - negative indicates inverse relationship
- strength of the relationship depends on the value of **r**
  - 0-.2 slight .2-.4 weak .4-.6 moderate .6-.8 substantial >.8 high

Relationships Among Data

- the correlation matrix
  - used to display bivariate relationships among numerous variables

	WEIGHT	SBP	TCHOL
WEIGHT	1.00000 **.000**	r → 0.85934 significance → **.014**	-0.94059 **.002**
SBP	0.85934 **.014**	1.00000 **.000**	-0.85822 **.027**
TCHOL	-0.94059 **.002**	-0.85822 **.027**	1.00000 **.000**

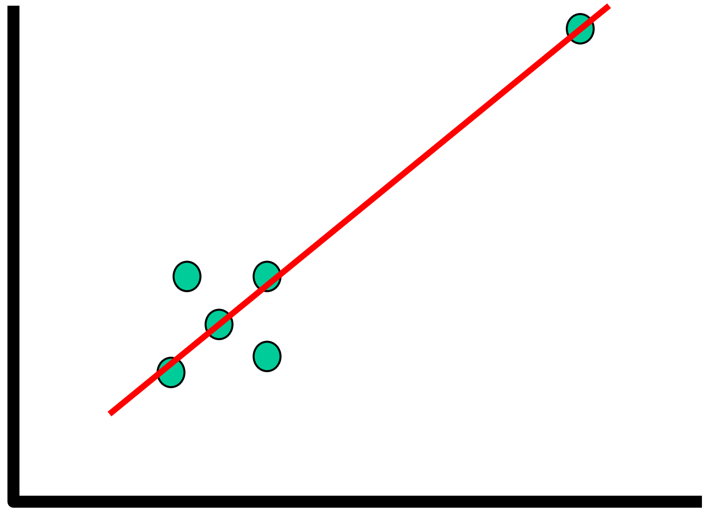
- correlation does not imply cause - effect
  - consider x and y to be related:

x → y   **y → x**   **z → x & y**

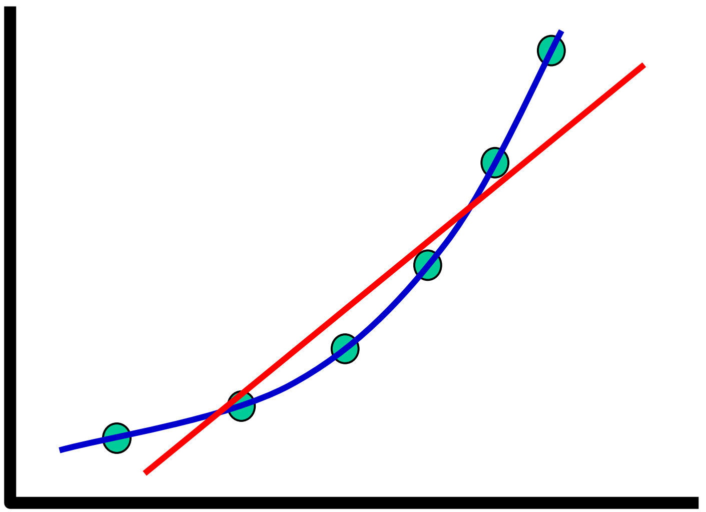
- coefficient of determination: r^2 (R^2)
  - the amount of variability in one variable "explained" by the other
  - represents strength of association (other similar measure: ω^2)

Relationships Among Data

- **the danger of "outliers"**
  - outliers: distribution values located far away from the bulk of values
  - may cause **r** values to be exaggerated or underestimated
  - can be checked by scatter plot
- **Pearson r will underestimate curvilinear relationships**



outlier exaggerating relationship



perfect curvilinear relationship underestimated by Pearson **r**

Non-Parametric Correlation Statistics

(does not require continuous or normally distributed data)

- **Spearman's rho or rank order** **(r_s or r)** - used for ordinal data
  - used when both sets of data are ranked (listed as ranks: 1st, 2nd....etc.)
- **Kendall's tau** **(t)**
  - same as Spearman's, but does a better job with tied ranks
- **Point biserial** **(r_{pb})**
  - used when one variable is truly dichotomous - other continuous
- **Biserial** **(r_{bis})**
  - used when one variable is artificially dichotomous - other continuous
- **Phi correlation** **(ϕ)**
  - used when both variables represent true dichotomies
- **Tetrachoric correlation**
  - used when variables represent artificial dichotomies
- **Cramer's V**
  - used when both variables are nominal (categorical) data