DESCRIPTIVE/SUMMARY STATISTICS

Discipline of quantitatively describing the main features of a collection of data ⟷ Numerical and graphical summaries used to characterize a dataset

The three main measures are:

- CENTER — measure of central tendency --- the typical or average value --- (mean, median, mode)
- SPREAD — measure of dispersion or variability of the data --- (standard deviation, variance, min, max, range)
- SHAPE — symmetric or skewed data --- (bell-shaped, normal curve, left/negative skewed, right/positive skewed)

The tools used for describing a collection of data are dependent on the nature of the data ----- Two main data types:

CATEGORICAL DATA (aka… qualitative) or QUANTITATIVE DATA (aka...numeric or measurement)

CATEGORICAL DATA (aka… qualitative)

Categorical Data Fit into Defined Groups ----- Two types of categorical data:

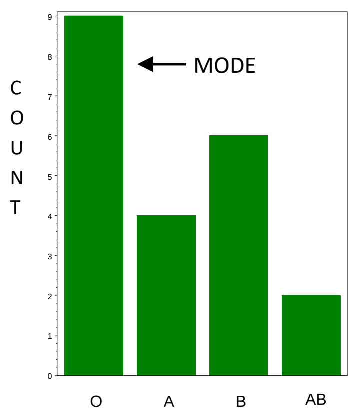
NOMINAL DATA or ORDINAL DATA (aka…ranked data)

NOMINAL DATA

GROUPS HAVE NO NATURAL ORDERING

Examples: gender, race, blood type, eye color, political affiliation, country of residence

Measures of Center: **MODE** = category w/ largest count
Measures of Spread – not germane with nominal data
Shape – not germane with nominal data



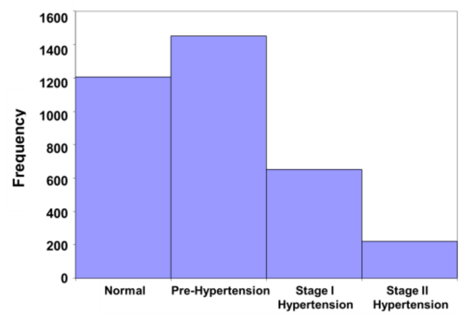
Bar Chart / Bar Graph for Blood Type

ORDINAL DATA (aka…ranked data)

GROUPS HAVE A NATURAL ORDERING

Examples: satisfaction level (Likert scale), educational level, shirt size, medical condition (good, fair, serious, critical)

Measures of Center: **MEDIAN** = category containing middle value **MODE** = category with largest count
Measures of Spread: **MIN** = minimum category **MAX** = maximum category **RANGE** = min cat. to max cat. **IQR** = middle 50 percent of the data
Shape – seldom used (can be problematic due to possible unequal or unquantifiable changes/differences in magnitude among/between categories)



Histogram for Blood Pressure Classification

QUANTITATIVE DATA (aka...numeric or measurement)

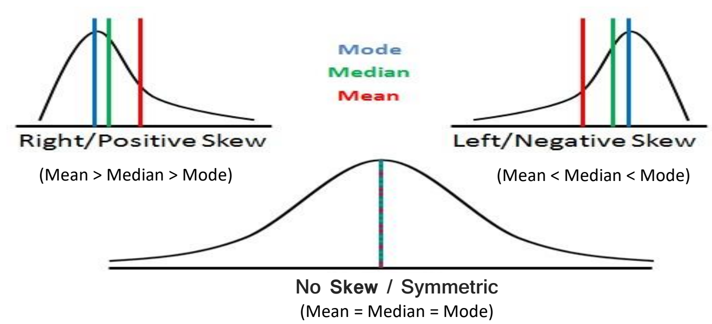
Continuous - data that have an infinite number of real values and there are <u>no</u> spaces/gaps between values (rounded to a specified precision)

EXAMPLES: BP, temperature, BMI, height, weight, blood serum level

Discrete - data that have a finite number of values within a given interval and there are spaces/gaps between values (typically counts)

EXAMPLES: test score, pages in a book, population of a country, # of trees in a forest

Measures of Center: **MEAN** = arithmetic average **MEDIAN** = middle value **MODE** = most numerous value
Measures of Spread: **STANDARD DEV** = average distance from center **VARIANCE** = $(\text{Standard deviation})^2$ **MIN** = minimum value **MAX** = maximum value **RANGE** = maximum - minimum
Shape: **SYMMETRIC** – bell-shaped? if yes, is it normal? **SKEWED** – left/negative right/positive



INFERENTIAL STATISTICS

Inference Examines/Investigates a Possible Relationship between Variables

Representative Sample(s) of Data are used to make Conclusions about a Broader Population

Two most common procedures making up inferential statistics:

- Hypothesis Testing
  - Calculate a test statistic which is then used to determine a p-value
  - Significance ⟷ if calculated p-value is ≤ level of significance (**α**) usually = .05
- Confidence Intervals (CI)
  - CI ⟷ point estimate ± margin of error (confidence level usually 95%)
  - Significance ⟷ if one CI does **not** capture a null value or if two CIs do **not** overlap

In most cases, the variables of interest can be assigned generic names that help define the relationship being examined – these two variable types are:

Explanatory Variable (aka… Independent or Predictor Variable) **AND** **Response Variable** (aka… Dependent or Outcome Variable)

The simplest type of inferential statistics is univariate analysis which involves ONE EXPLANATORY variable and ONE RESPONSE variable

EXAMPLES: height predicts weight? --- blood type explains cholesterol level? --- aspirin use explains occurrence of heart attack?

One Quantitative Response Variable / One Quantitative Explanatory Variable

Simple Linear regression (SLR)

Used for prediction and to measure how much one variable increases/decreases per unit of change in the other variable

H_0: $\beta_1 = 0$ (slope = 0 , so y and x **not** linearly related)

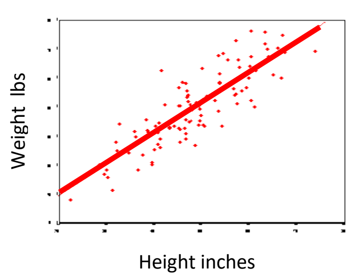
H_a: $\beta_1 \neq 0$ (slope ≠ 0 , so y and x linearly related)

Regression equation

$E(Y) = \beta_0 + \beta_1 x$
- β_0 *is the y intercept*
- β_1 is slope of the regression line

Example: weight = -97.2 + 3.72 (height)

(Scatterplot with regression line)



$r = +.85$

Correlation coefficient -- direction and strength of a linear relationship -- usually represented by r or ρ (*Rho*) $[-1 \le r \le +1]$

Positive correlation: $r > 0 \leftrightarrow y \uparrow$ as $x \uparrow$

Negative correlation: $r < 0 \leftrightarrow y \downarrow$ as $x \uparrow$

- $r \le |.3| \leftrightarrow$ *weak* (*none if* $r = 0$)
- $|.3| < r < |.7| \leftrightarrow$ *moderate*
- $r \ge |.7| \leftrightarrow$ *strong* (*perfect if* $r = \pm 1$)

One Quantitative Response Variable / One Categorical Explanatory Variable

ANOVA — 3 or more groups/categories
T-test — 1 or 2 groups/categories

Generic hypothesis for 2 or more samples:

H_0: The means (μ) for the categories are equal

H_a: At least one mean (μ) for the categories differs

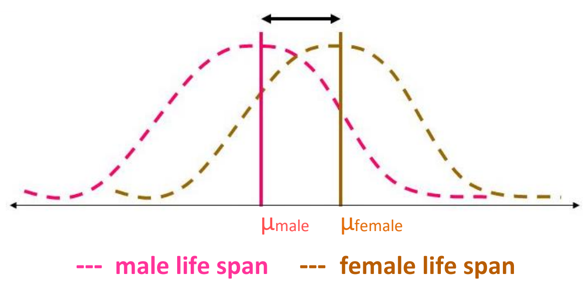
EXAMPLE: One-way ANOVA (**AN**ALYSIS **O**F **VA**RIANCE)

Test if there is a difference in mean cholesterol levels between 4 different blood types (O, A, B, AB)

EXAMPLE: Two-sample T-test

Test if there is a difference in mean life spans between sexes (i.e. male *vs.* female)

Population Distributions for Male and Female Life Span



H_0: $\mu_{male} = \mu_{female}$

H_a: $\mu_{male} \neq \mu_{female}$

One Categorical Response Variable / One Categorical Explanatory Variable

Chi-square test of a relationship/association between two variables

H_0: The two variables are **not** related/associated

H_a: The two variables are related/associated

EXAMPLE: Chi-square test for a relationship between aspirin use and heart attack (MI)

Statistic	DF	Value	Prob
Chi-Square	1	25.0139	.0001

Since p-value = .0001, there is strong evidence of a statistically significant relationship (at the .05 level) between aspirin use and MI

(two-way or contingency table)

Treatment	Heart Attack (MI)?		
	Yes	No	Total
Aspirin	104	10933	11037
Placebo	189	10845	11034
Total	293	21778	22071

Risk of MI w/ Aspirin

104 ⁄ 11037 = .0094

Odds of MI w/ Placebo

189 ⁄ 10845 = .0174

Odds ratio for MI

Placebo *vs.* Aspirin

$$\frac{189/10845}{104/10933} = 1.8321$$

Hence, the odds of MI w/ Placebo trt are ≈ 1.8 times greater than w/ Aspirin trt