Intro to Research Methods

Data is gathered from conducting studies.

- **Observation study:** are used to show relationships. measure or survey observation set without affecting them.
  - **good surveys:** have a good sample size (n), are representative of the population and have a sound methodology.
- **Controlled experiment:** are used to show causation. Observation set (randomly selected) are split into groups; some treatment is applied to one group, while the other receives no treatment (known as ***placebo***).

- It is expensive to conduct research studies on entire population, use a sample to learn about the population.
- Each sample may not perfectly (accurately) predict the population parameter but the sample size statistic will give us an interval the population lies as long as the sample is random and unbiased.
  - **population:** total set of observations that can be made. mean is **μ** (mu)
  - **sample:** a subset of the population. mean is **x̄** (x-bar)
  - **random sample size** (**n**): where each subject has an equal chance of selection.
  - Larger sample size leads to a sample statistic being a better approximate to the population parameter.
  - **parameter:** describes a population
  - **statistic:** describes a sample
  - **sampling error:** difference between the population and sample. ***μ - x̄***
- In studies, ***correlation does not imply causation*** due to lurking variables.

In experiments, ***independent variables*** are manipulated to measures changes to dependent variable (***outcome***) while controlling ***lurking*** (***extraneous***) variables.

- **variable**: value that may change or differ between individuals in an experiment. Variables need to be measured. Unmeasure-able variables are called:
  - **construct:** an abstract concept. Ex: happiness level, hunger, age, etc.

If the variable age is not defined in a unit of measurement, it's construct (could be measured in years, maturity level etc). This applies to other variables like distance etc.

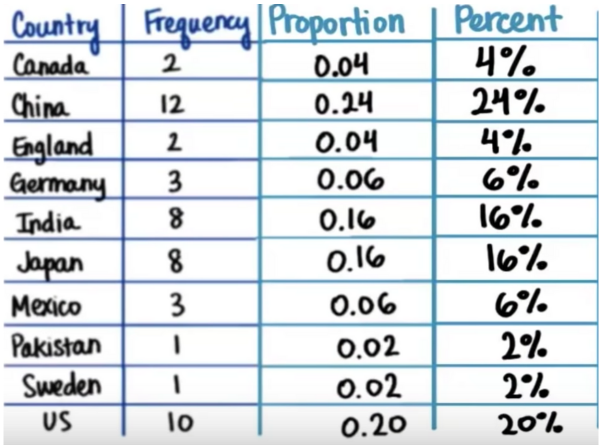
  - **operational definition:** a way of turning a construct into a measure-able form.
- **lurking factors**: unknown factors that are not controlled for; affect variables thereby causing a bias in the relationship between dependent variable and outcome.
- **outcome:** possible result of a study.

It's best to ***visualize*** the data from an experiment to determine the relationship between variables.

- **x-axis:** independent (predictor) variable
- **y-axis:** dependent variable or outcome
- **hypothesis:** statement about relationship between variables. Ex: more hours slept can can lead to better memory.
- **sliding:** prevent biases in experiments:
  - single blinding: participants have no idea of type of treatment received.
  - double blinding: both participants and researchers are unaware of thes treatment received.
- **placebo:** neutral treatment that has no effect on dependent variable. helps identify the presence of lurking variables.

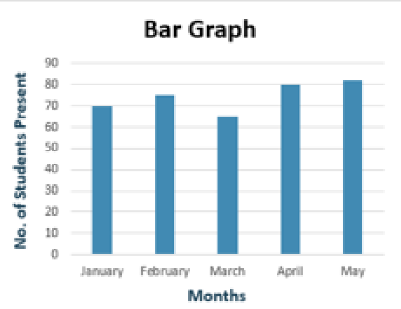
Visualizing Data

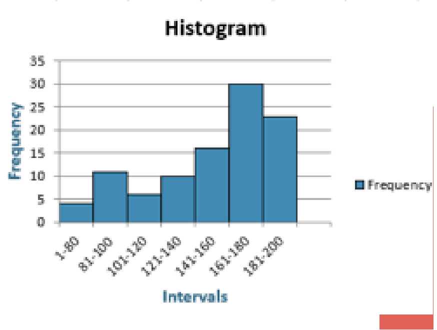
- **Frequency:** the number of times an observation (outcome) occurs in a data.
- **Frequency table:** table that organizes the data (outcome) by its frequency.
    - **relative frequency (proportion):** shows how each outcomes relate to the whole outcome. Divide the frequency of each outcome by the sample size (n).



Country	Frequency	Proportion	Percent
Canada	2	0.04	4%
China	12	0.24	24%
England	2	0.04	4%
Germany	3	0.06	6%
India	8	0.16	16%
Japan	8	0.16	16%
Mexico	3	0.06	6%
Pakistan	1	0.02	2%
Sweden	1	0.02	2%
US	10	0.20	20%

- **Bar graph:** represents **categorical** data with rectangular bars with heights/lengths proportional to the frequency of each outcome.
- **Histogram:** graphical representation of data by equal ***intervals*** (***bin***, ***bucket***). Width of the interval is the ***interval or bin size.***
    - smaller bin size: spread out histogram, fewer observations within each interval and too many details.
    - bigger bin size: compact histogram, more observations within each interval and fewer details.



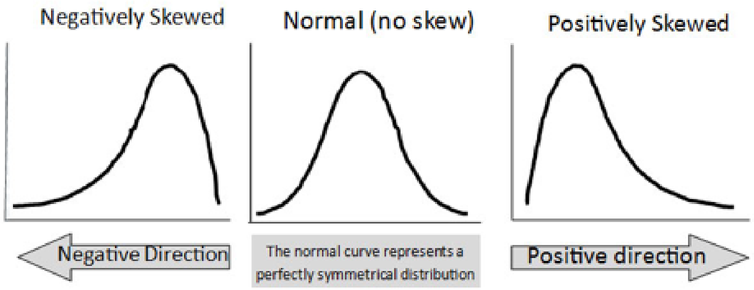


- Histogram displays the shape and spread of continuous data

Looking at the histogram alone won't provide the exact frequency of each outcome due to the use of intervals. There's no way to tell the exact value of each observation.

Histogram	Bar Chart
Can choose any interval (bin size)	Each observation belongs in a distinct category (there's always a space between each category)
x-axis is numerical/quantitative	x-axis is categorical or qualitative
x-axis is ordered from least to greatest	Order in x-axis doesn't matter
Shape of the graph is important in analysis	Shape is arbitrary and not useful for data analysis.

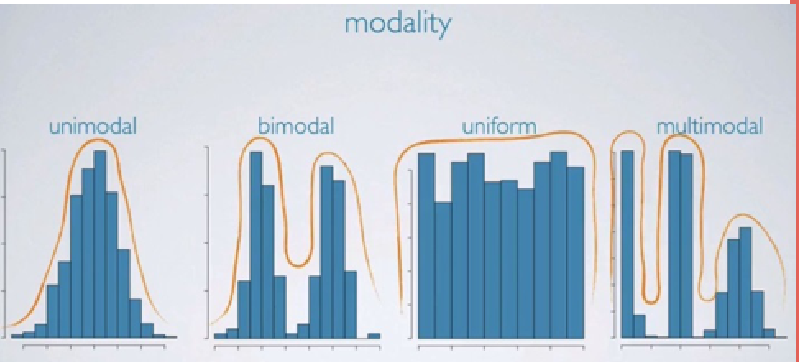
- **Normal (symmetric) distribution:** also known as a bell curve meaning the data observations are symmetric around the mean with one peak (mode).
- **Positively (right) skewed distribution**: most of the observations are on the left tail than the right tail of the distribution making the right tail longer.
- **Negatively (left) skewed distribution:** most of the observations are on the right tail than the left tail of the distribution making the left tail longer.



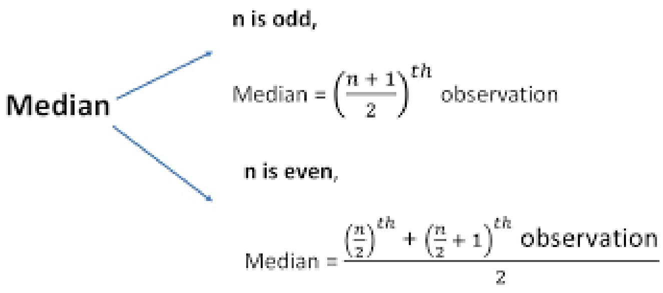
Central Tendency

Three most common measures of central tendency are: mode, median and mean.

- **Mode:** value that has the most frequency (most common value).
  - In a histogram, the mode is the bin/interval with the highest frequency . *Remember: We don't know the exact value of each observation.*
  - Can find the mode or categorical data. It's the outcome with the highest frequency.
  - Taking various samples on same population won't give the same mode.
  - Not a good tool to learn about or describe a population
  - **Uniform distribution:** has no mode.
  - **Multi-modal distribution:** two or more distinct clear trends.



- **Median:** centre (middle) of the distribution.
  - Data needs to be ordered
  - ROBUST: not affected by departures from the norm (i.e. outliers, extreme values) making it the best measure of the centre of tendency for skewed distribution.



Median

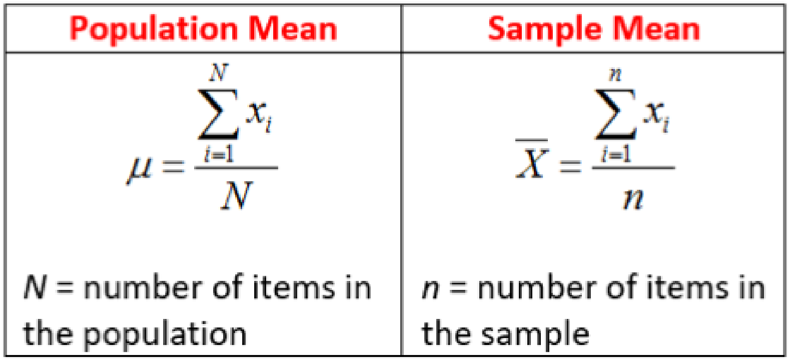
n is odd,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{ observation}$$

n is even,

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{ observation}}{2}$$

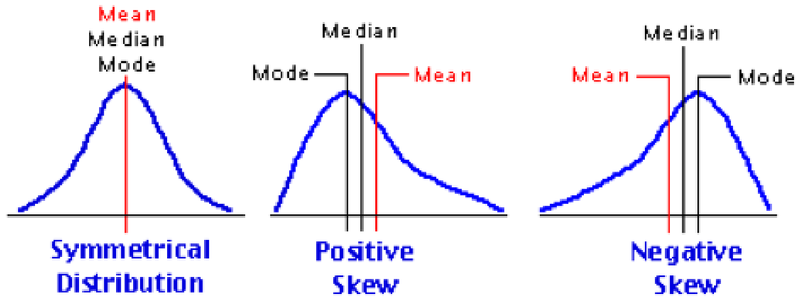
- **Mean:** average of the distribution.
  - All observations in a distribution affect the mean.
  - Many samples from the same population will have similar mean -> can be used to make inferences about the population.
  - Changes with addition of extreme outliers from the dataset skew mean making it irrepresentative of the dataset.



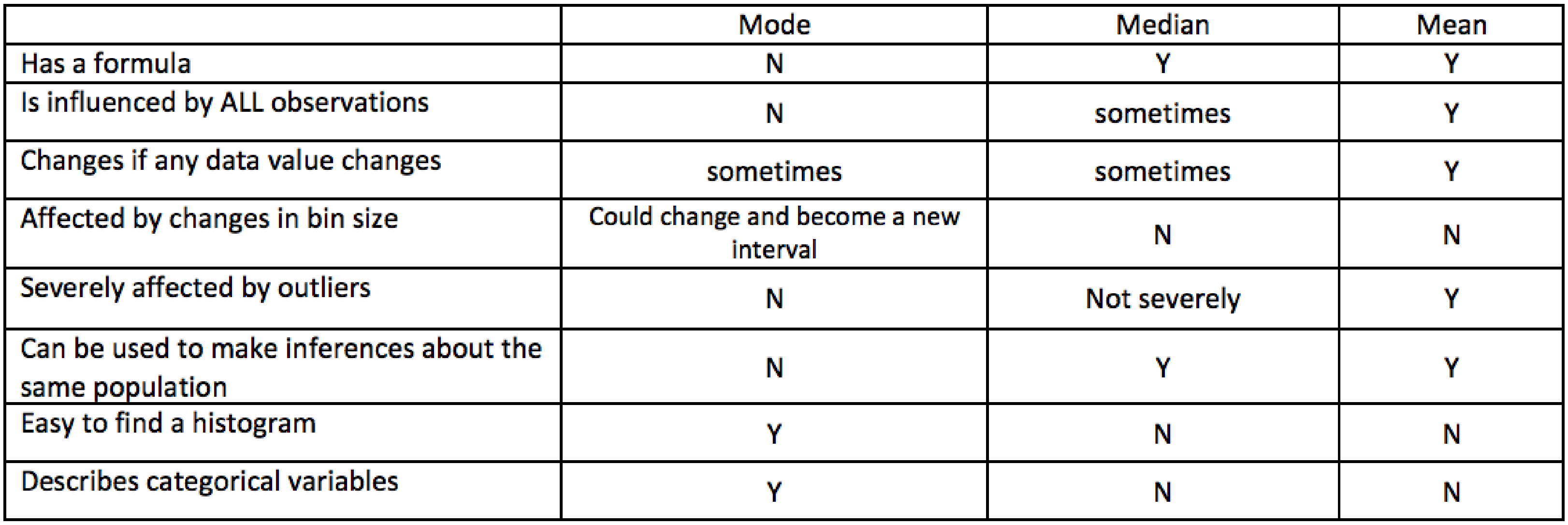
Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	*n* = number of items in the sample

- **Symmetric (Normal)**: mode = median = mean
- **Positive (Right) skew**: mode < median < mean
- **Negative (Left) skew**: mode > median > mean

Skewness mostly affects the mean; it pulls the mean in the direction of the data so it's always the highest or lowest of the three centres of tendency depending on the direction of the skewness.



Central Tendency (cont'd)



	Mode	Median	Mean
Has a formula	N	Y	Y
Is influenced by ALL observations	N	sometimes	Y
Changes if any data value changes	sometimes	sometimes	Y
Affected by changes in bin size	Could change and become a new interval	N	N
Severely affected by outliers	N	Not severely	Y
Can be used to make inferences about the same population	N	Y	Y
Easy to find a histogram	Y	N	N
Describes categorical variables	Y	N	N