* Variable: attribute of entity (place, person, idea, …). Can vary from one entity to another(variable)
* Variable examples:

Top Speed of different cars

Width of different couches

Average Weight of different types of fish

Population count of different countries

* Variables can be classified as **qualitative** (aka, categorical) or **quantitative** (aka, numeric ,- measurable quantity.)
* **Discrete**(# of head in multiple coin flips) vs. Continuous(weight, height) Variables
* **Univariate**(data contains only one variable) vs. Bivariate(data contains multiple variables) Data
* **Population**(is the entire group that you want to draw conclusions about.) VS **Sample**(the specific group that you will collect data from.)
* A measurable characteristic of a population(**parameter).** measurable characteristic of a sample is called a **statistic**.
* Simple Random Sampling
* Sampling With Replacement and Without Replacement
* **Median** (outlier exist) and the **Mean** (no outlier) and mode
* **measure of variability** : **Range**(max – min), the **interquartile range** (q3-q1), **variance**(*s*2 = Σ ( xi - x )2 / ( n - 1 )), and **standard deviation**(sqrt(*s*2))-> expressed in the same units as the original values (e.g., minutes or meters).
* Measures of Position: **Percentiles**(divide a rank-ordered set of elements into 100 equal parts- P50, and it would be greater than 50 percent of the observations in the set.), **Quartiles** divide a rank-ordered data set into **four** equal parts,  **standard score** ( **z-score**) indicates how many [standard deviations](https://stattrek.com/statistics/dictionary.aspx?definition=standard%20deviation) an element is from the mean.
* Patterns in data are commonly described in terms of: **center(**median-centered over 4**)**, **spread(aka variability -** wide [range](https://stattrek.com/statistics/dictionary.aspx?definition=range) vs narrow range**)**, **shape(see below)**, and **unusual features (see below)**.
* **Shape characteristics:**
* **Symmetry**: half is a mirror image of the other.
* **Number of peaks**. 1 peak- **unimodal**, 2 peaks- **bimodal**. single peak at the center- **bell-shaped**.
* **Skewness**. Fewer observations on the right- **skewed right**. fewer observations on the left- **skewed left**.
* **Uniform**. equally spread across the range of the distribution- **uniform distribution**. A uniform distribution has no clear peaks

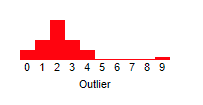
Chart, histogram

Description automatically generated

**Unusual Features:**

* **Gaps**. Areas of no observations.
* Chart, histogram

  Description automatically generated
* **Outliers:** values that differ greatly from the other observations. O.L <1.5\*IQR or >1.5\*IQR



* Plots
* Dotplot: dots that represent frequency. Symmetry & skewness can be used if variable is cont.
* Bar chart: x is categorical, y represented using bars with varying lengths
* Histogram: x is continues(bins), y is frequency represent using connected bars

Chart, bar chart

Description automatically generated

**-boxplot:**

Chart, box and whisker chart

Description automatically generated

**Shape of data, range & IQR can all be inferred.**

Chart, histogram

Description automatically generated

**Scatterplots(data** patterns)**: graphs pairs of numerical data, with one variable on each axis, to look for a relationship between them**

* Linearity refers to whether a data pattern is linear (straight) or nonlinear (curved).
* Slope refers to the direction of change in variable Y when variable X gets bigger. If variable Y also gets bigger, the slope is positive; but if variable Y gets smaller, the slope is negative.
* Strength refers to the degree of "scatter" in the plot. If the dots are widely spread, the relationship between variables is weak. If the dots are concentrated around a line, the relationship is strong.

Chart, scatter chart

Description automatically generated

* **Population**(is the entire group that you want to draw conclusions about.) VS **Sample**(the specific repre subset from the pop fro which that you will collect data from.)
* **Cons 1 -sample is representative**
* **Cons 2 -pop not people representative**
* **Cons 3 - pop not necc big**

**Not feasible andin many cases impossible to to collect data from**

* A measurable characteristic of a population(**parameter).** measurable characteristic of a sample is called a **statistic**.