Managing Describing analyzing data- Notes and summary

A.M- All material obtained from DTSA5900

2023-09-15

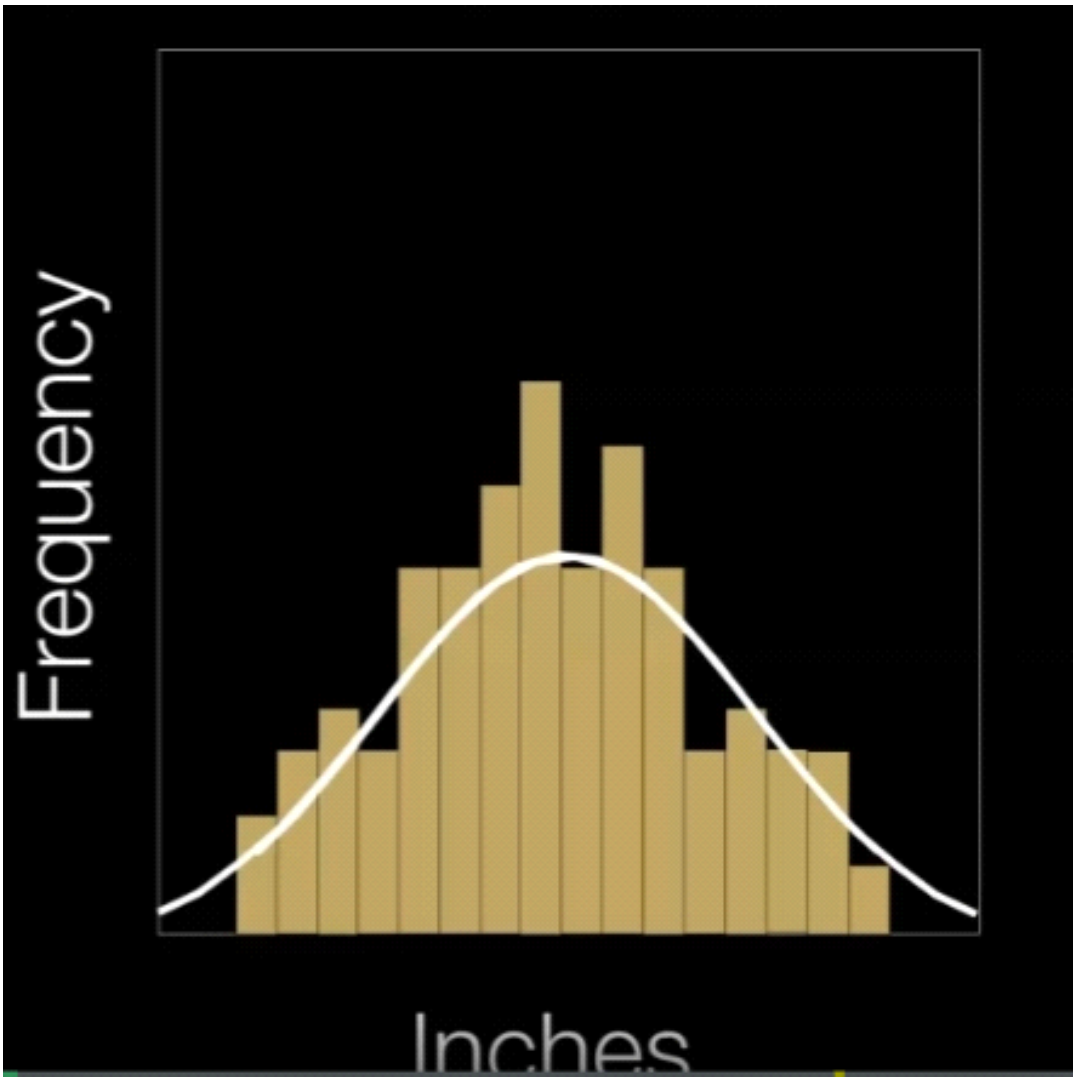
Types of data and Measurement Scales

Data can fall into two types of categories, Qualitative and Quantitative

Quantitative data - Is obtained by measuring along a numerical scale and is *continuous* meaning that it is possible to take infinite measurements, for example height can be measured in a way that finer and finer measurements can be captured.

some examples

- Dimensions
- Temperature
- Speed
- Volume of Sales

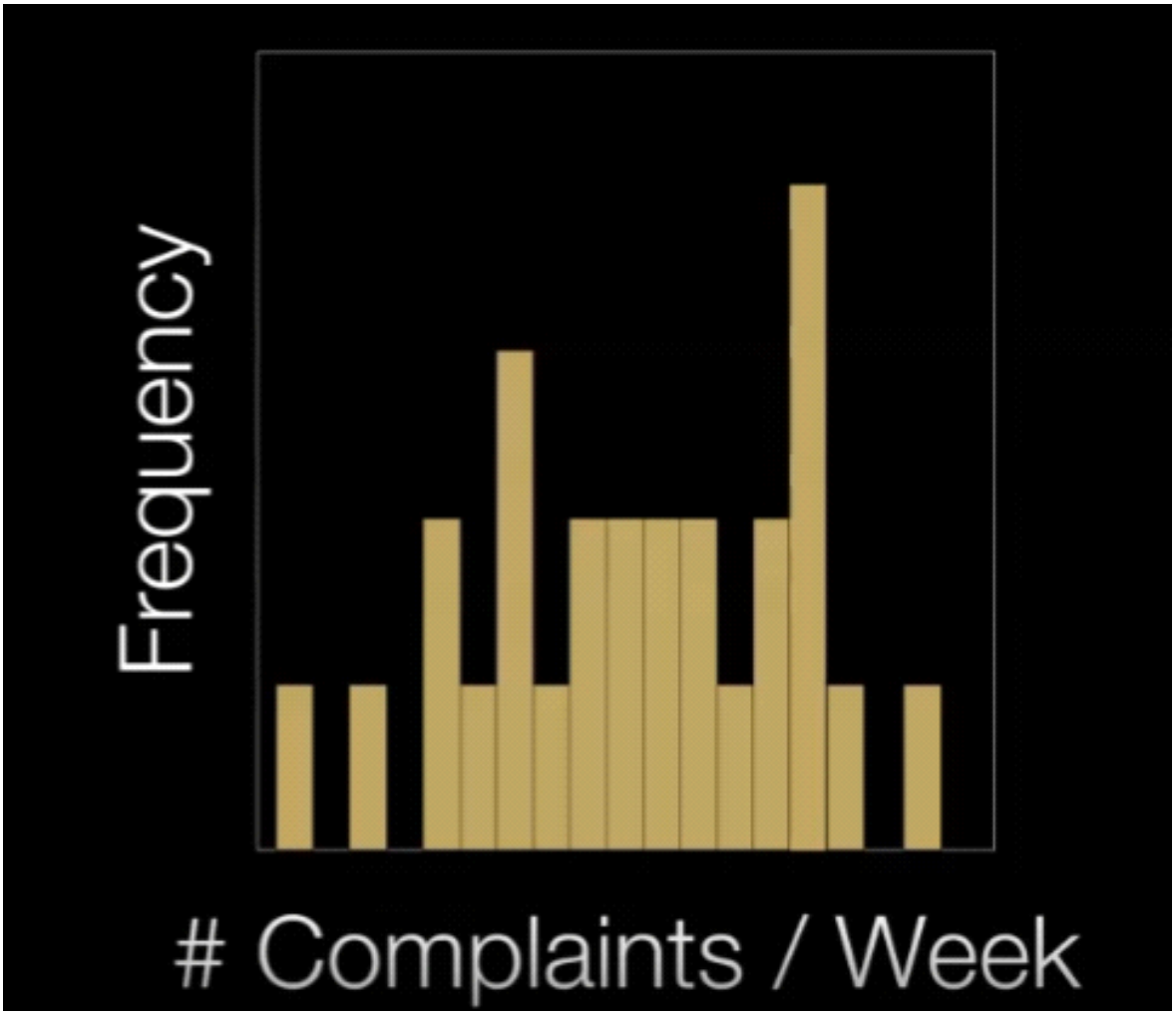


An example of Continous data(Quantitative).

Qualitative data- Data that falls into categories, or can be categorized. How many people prefer red shirts over blue shirts is an example of discrete data. This is known as *discrete* data.

some examples

- Complaints per sales period
- Number of defects per unit
- Percent defective units
- Number of orders shipped on time



An example of Discrete Data (Qualitative).

Underlying properties we wish to study are called dependent variables

Criterion Measure is the way we choose to measure or understand the underlying property. (The dependent variable)

Data Is the result of process measurement

Measurement Scales Continued: Nominal and Ordinal Data

Nominal Scale Numbers are assigned to categorize, identify or name attributes. Nominal scale values can only be used to indicate equal or not equal. Usually considered qualitative. Main statistic used- mode with the highest frequency of occurrence.

some examples

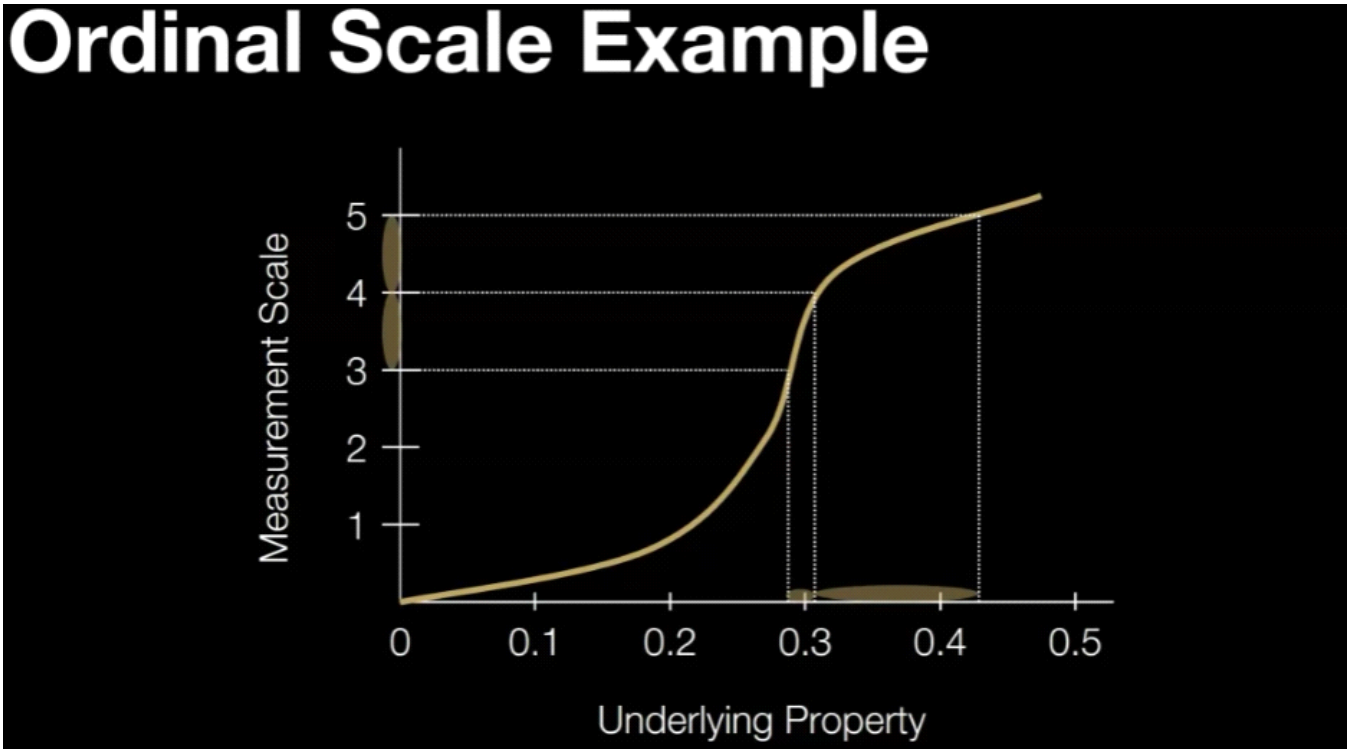
- Zip codes
- Area Codes
- Numbers Assigned to types of Nonconformity in products
- Numbers assigned to presence or absence of an attribute(0,1)
- Numbers assigned to sales territories
- Political party affiliation
- Position of a light switch
- Hair color
- Geographical Location
- Responses on a survey

Ordinal Scale

Numbers are assigned to observations such that the order of the numbers corresponds to the order of the underlying property studied. Ratings and Rankings are ordinal scale measurements. Ratings assign a score. Rankings are the result of sorting items.Rankings have a higher resolution than ratings.But rankings usually require more effort. Median or mode can be used to indicate the center of this data. For the dispersion or spread, look at the range or interquartile range. Ordinal scale values can be used to determine = or != and > or < but not the magnitude of measurements.

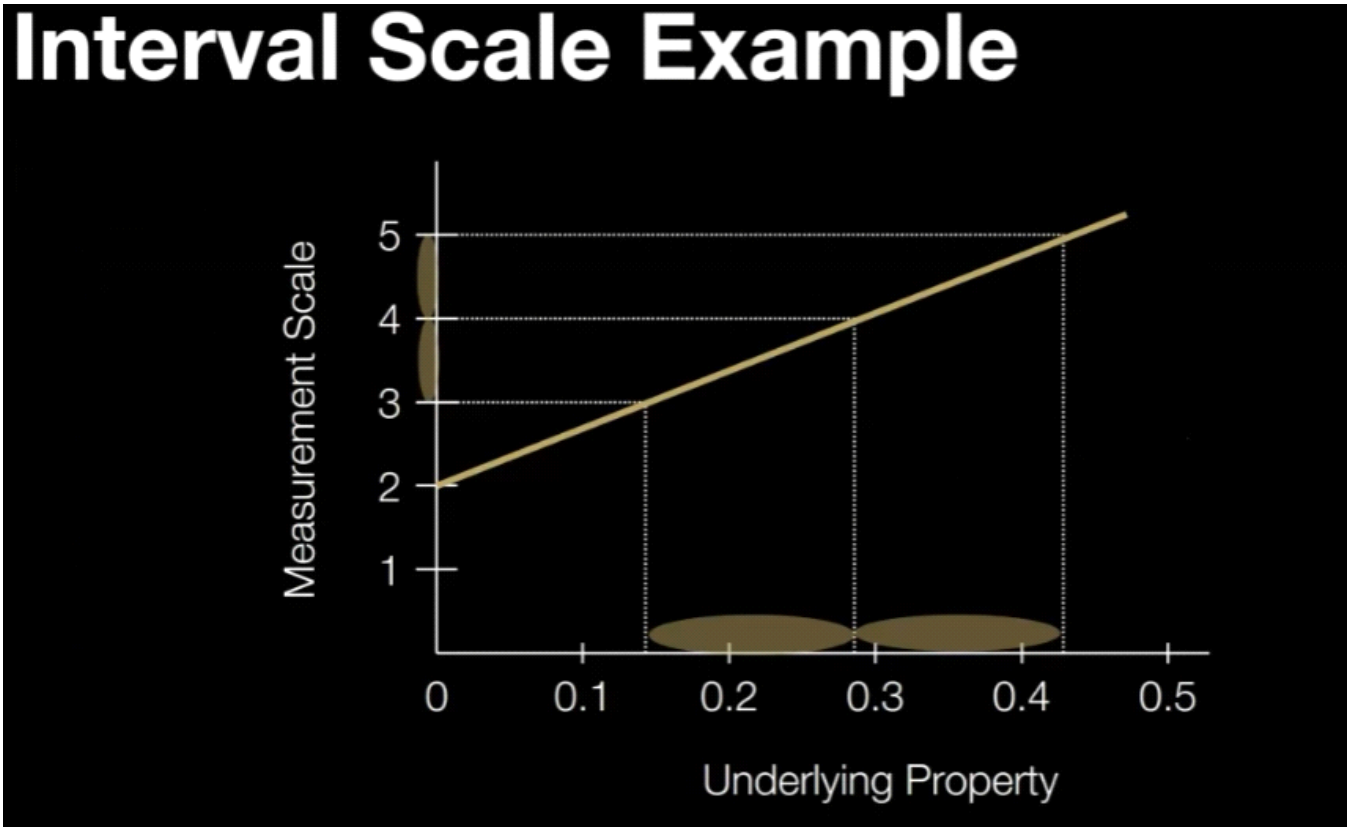
some examples - Satisfaction scale for customer surveys - Letter grades - Sound intensity measured in decibels - The number of scratches on the surface of a roll of aluminum.

The intervals between measurement scales and underlying property may not be equal. Order preserving transformations may be used. And before working with ordinal data, it usually must be ranked.



Interval Scale Numbers are assigned to observations such that differences between any two numbers correspond to proportional differences in the underlying property being studied.There are equal intervals along the scale. 0 is a value on the scale, so negative numbers are possible. Differences on the measurement scale are the same as the differences in the underlying property. To work with interval data best, often a linear transformation $y=mx+b$ is performed. Statistics such as mean and standard deviation apply. And Interval scale data can be used to determine = or !=, > or < and sums or differences can be used.

some examples - Temperature (F or C) - Distance from a reference point - Calendar date - Height above sea level



Ratio Scale An interval scale has been reached, and there is a zero point which corresponds to zero, null or absence. The plot of Ratio scale data looks similar to the interval, except the line now crosses through 0, meaning 0 is a set value. Coefficient of variation and geometric mean. As well as any parametric statistical test can be used.

some examples - Dimensions (length, height) - Volume or weight - Cycle time, and time to repair - Price of an item per day - Chemical concentration in a tank - Monthly salary - Volume of a beverage container

Ratio Scale Data

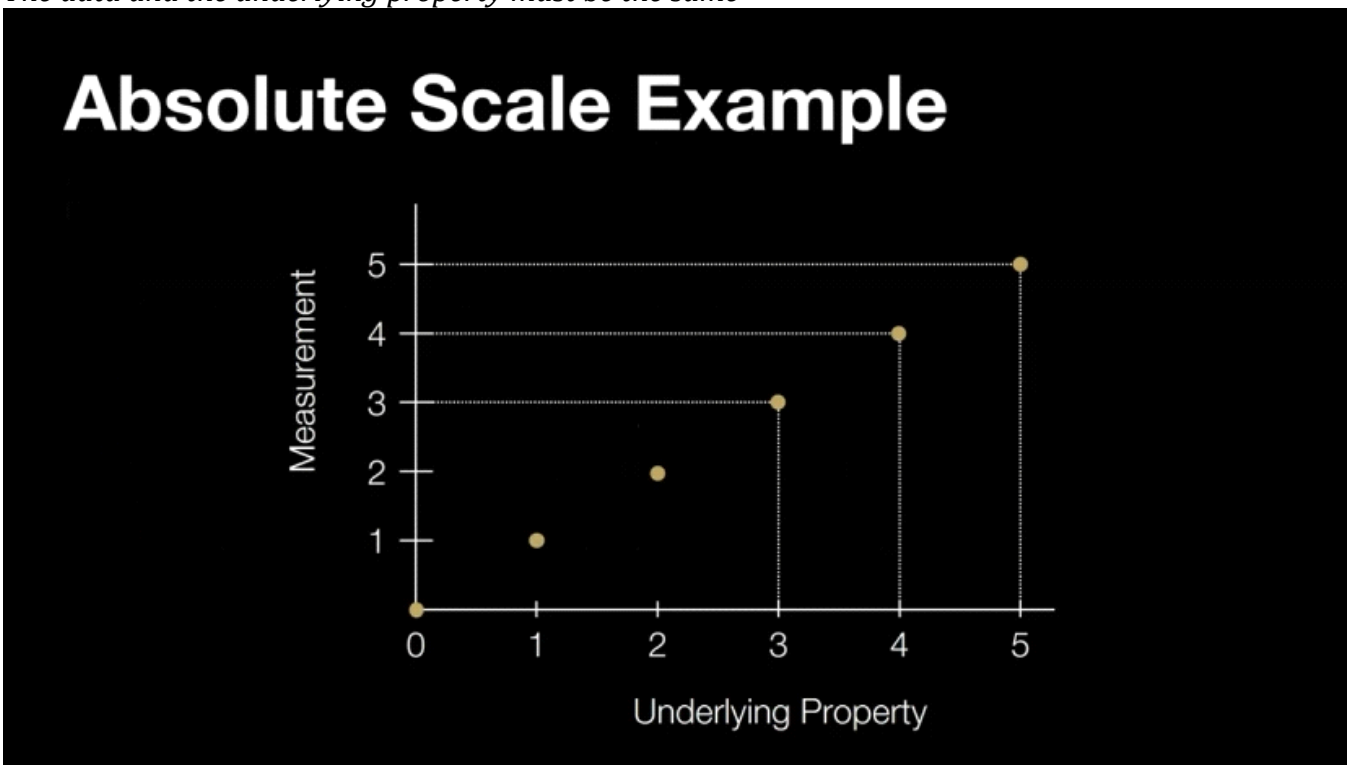
So is it ratio or interval?

If there is a 0 point that is meaningful, more than likely the data is ratio. If there is no 0 point, say the data starts at 10, then the data is now interval.

Absolute scale (Count data) Numbers assigned to observations such that the numbers directly correspond to the property being studied. Mean, median and mode must be used. Data on the absolute scale have some of the properties of ratio data. Standard parametric methods must be used and in some cases non-paramteric methods. (When data or normally distributed- use parametric, and non-parametric if not) Standard parametric method example : The Central limit Theorem. Non parametric method: Spearman's correlation test.

some examples - Number of defects - Number of flights per day - Number of parts made - Number of safety accidents - Number of customer complaints

The data and the underlying property must be the same



Big 5 aspects of Data

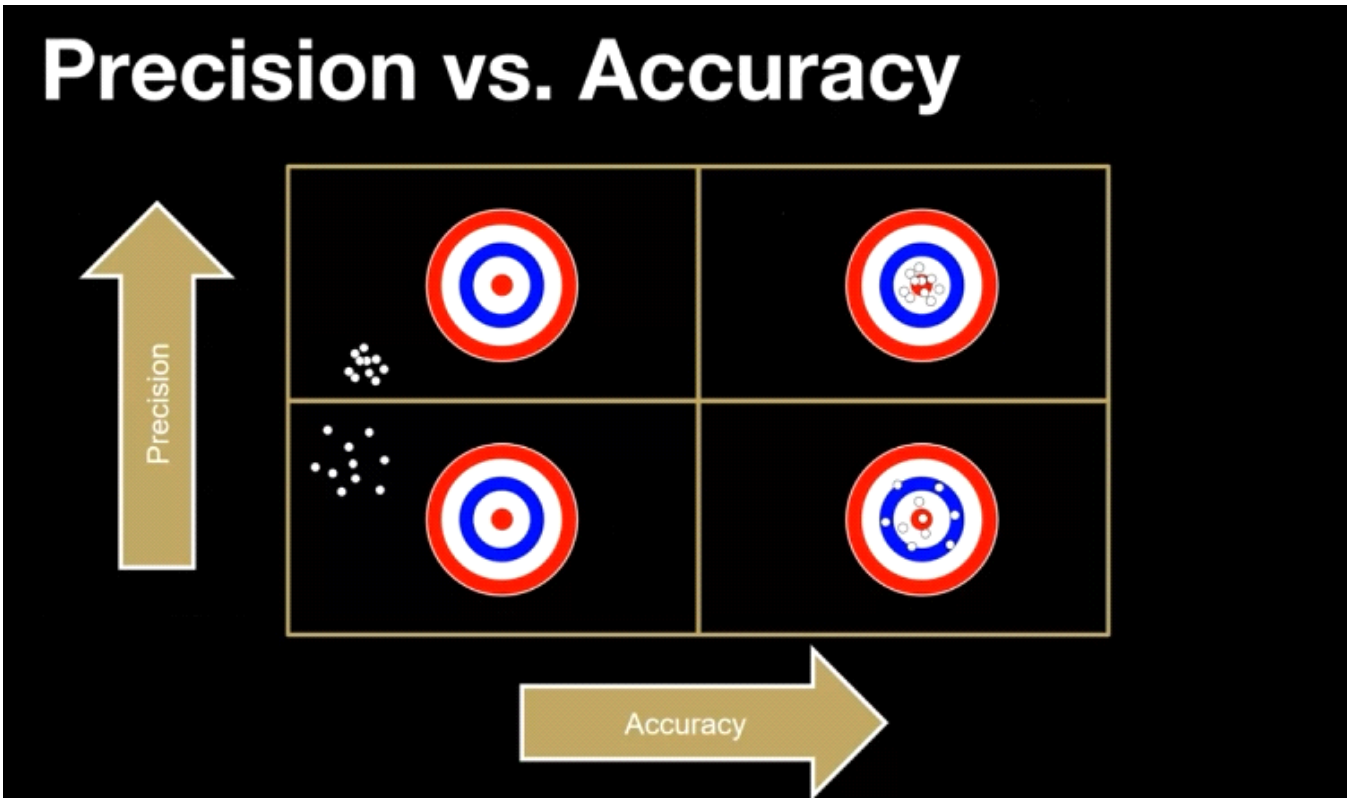
Data must be efficient and effective Process of measurement must be capable and acceptable

Measurement can be subject to many sources of variation: Standard used for calibration Operator Equipment itself

Measurement systems must demonstrate: Stability through time The ability to generate repeatable and reproducible measures. The ability to generate valid measurements.

Reliability is a measure of the *precision* of the device. **Validity** is a measure of the *accuracy* of the device or method.

The time honored image below depicts accuracy vs precision. *Precision* is how well a measure is obtained again and again. *Accuracy* is how well a measure meets some spec or target.



Accuracy VS precision

Tools for understanding data

Prob and stats- To perform calculations, summarize and well, quantify the data.

Control Charts- To determine if a process is stable.

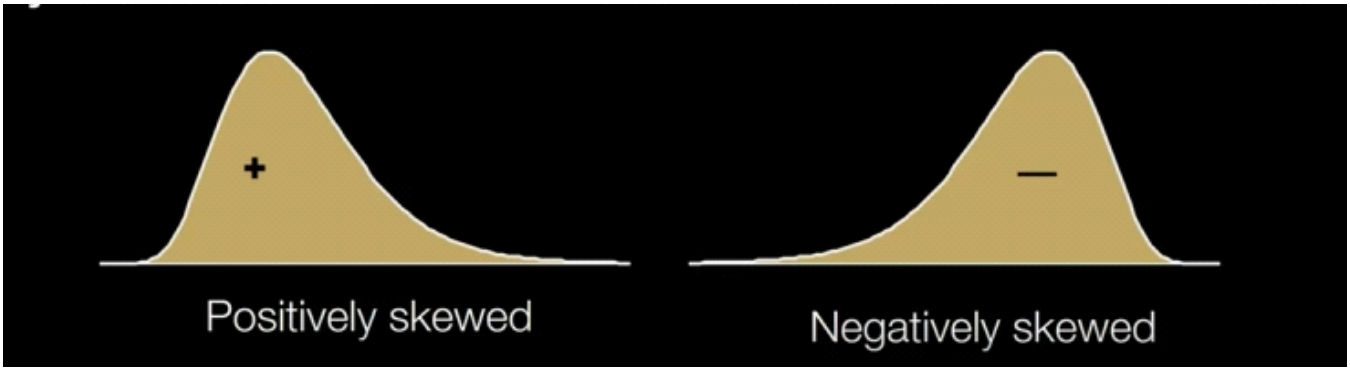
Experimental design- To allow us to see the root cause of a problem so we can manipulate special and common causes of variability to improve or optimize the process output.

5 most important features of data

- Location - Mean, median, mode (Central tendency)
- Spread - Range, SD, Var
- Shape - Skewness and kurtosis
- Time Sequence
- Relationship

If the distribution is symmetrical, usually mode, mean, and median will be similar, but if the data is skewed, often the mode is well.. the mode. The point being that the median and mean will be the center of the data and the mean might not be in the center.

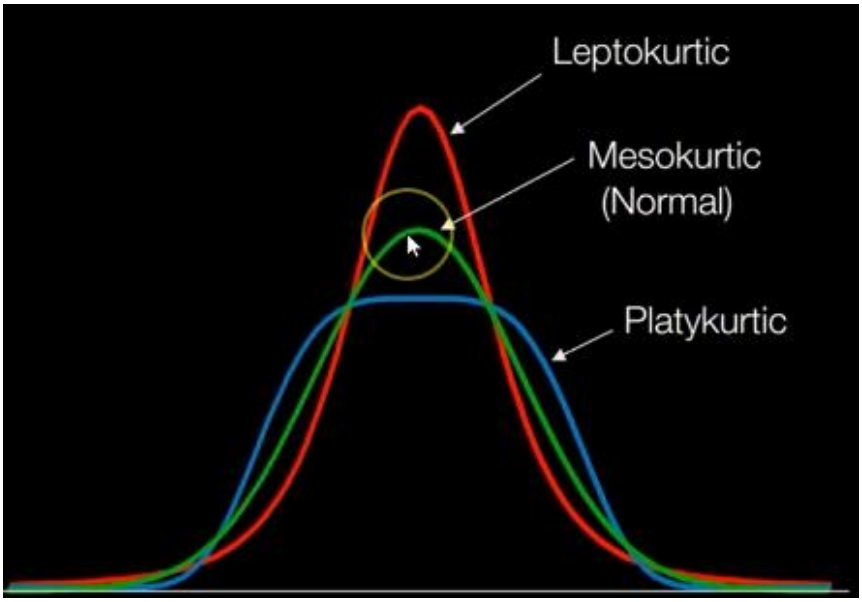
skewness is concerned with the symmetrical nature of the distribution and the is the degree of departure from symmetry.



skewness

Kurtosis How peaky the data is.

Mesokurtic Data is normal **Leptokurtic** Data has positive kurtosis and a high peak, with heavy tails. **Platykurtic** Data has negative kurtosis with a lower peak, and lighter tails.



Examples of Kurtosis

Time sequence

Helps indicate the stability of a process through time.

Measures of relationship

Correlation If variables are continuous **Association** If variables are discrete

Populations and Samples

Population is the entire group of subjects that we pull from for the study. Always really big. A sample is a subset of a population which will be used to draw conclusions.

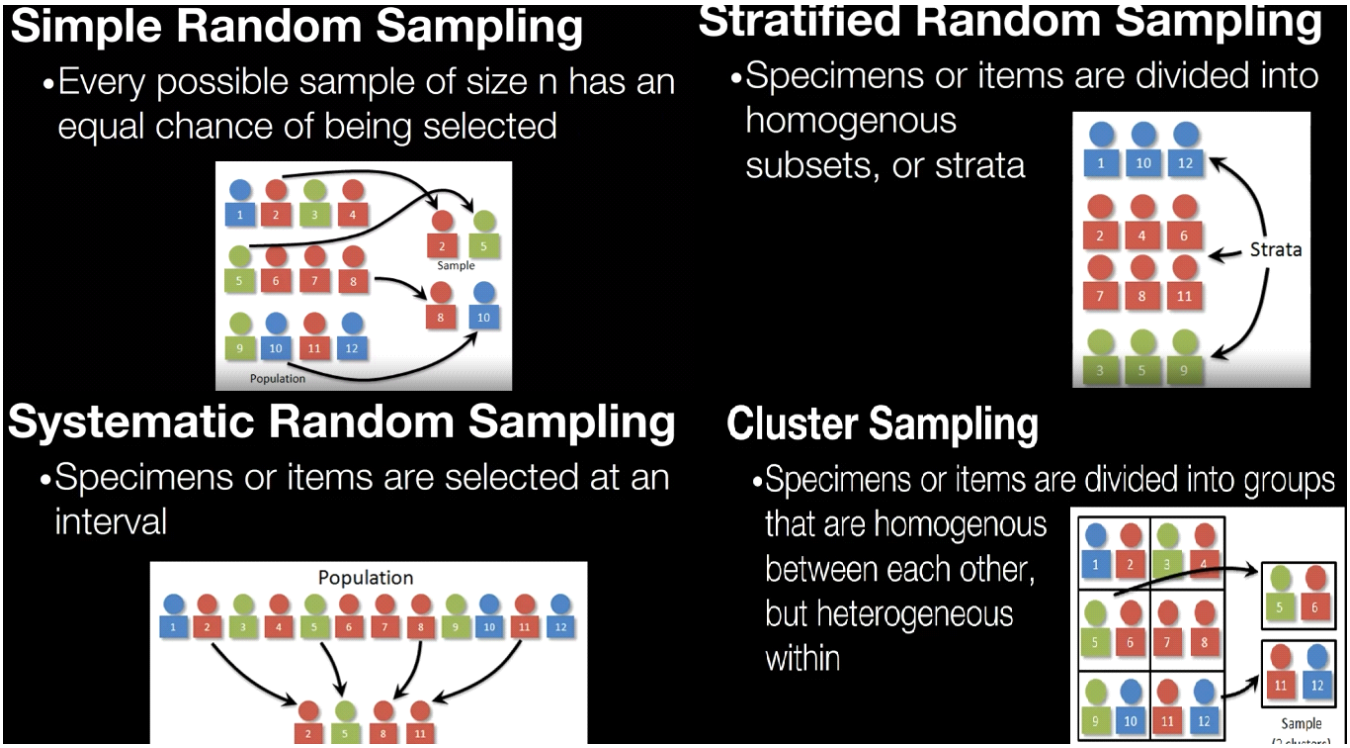
Random sampling must be used, to help keep data accurate.

examples of sampling - Non- Random (judgement sampling) - Can lead to bias - Random sampling - All specimens have a chance of being included and within random sampling:

Simple random sampling - every possible sample has an equal chance of being selected. Can be with or without replacement.

Systematic Random sampling - Selected at an interval (ex. every 3rd person) There is some small potential for bias. **Stratified Random sampling** - Specimens are divided into homogeneous groups. Then specimens are selected from that.

Cluster Sampling Specimens are divided into groups that are homogeneous between each other but heterogeneous within.



Sampling_Types

Statistics and parameters

Remember! Some stuff is for populations and some stuff is for samples

Statistics and Parameters

Sample Statistics	Population Parameters	Description
$\bar{X}$	μ	Mean
$\tilde{X}$	M	Median
s	σ	Standard Deviation
s^2	σ^2	Variance
R	NT'	Range / Natural Tolerance
p	π	Count Per Unit
g_3	γ_3	Skewness
g_4	γ_4	Kurtosis