ReDI School of Digital Integration

Statistics

Lecture 3

We use tech to connect human potential and opportunity with dignity & humility

Statistics

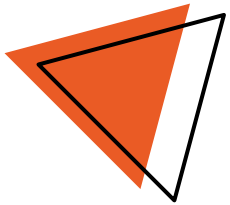
Statistics

“Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data.”

- Models the real world into mathematical objects (lines, curves, etc.)
- reason new knowledge from existing evidence (inference) and state how confident we are about that reasoning (probability)

Descriptive Statistics: Summarizing/describes the characteristics of a population or a sample of a population.

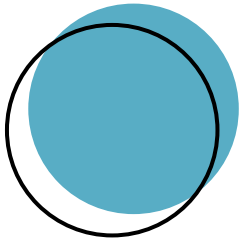
Inferential Statistics: Predict behavior of population based on a sample. Uses probability to find out the confidence of the predictions.



Descriptive Statistics

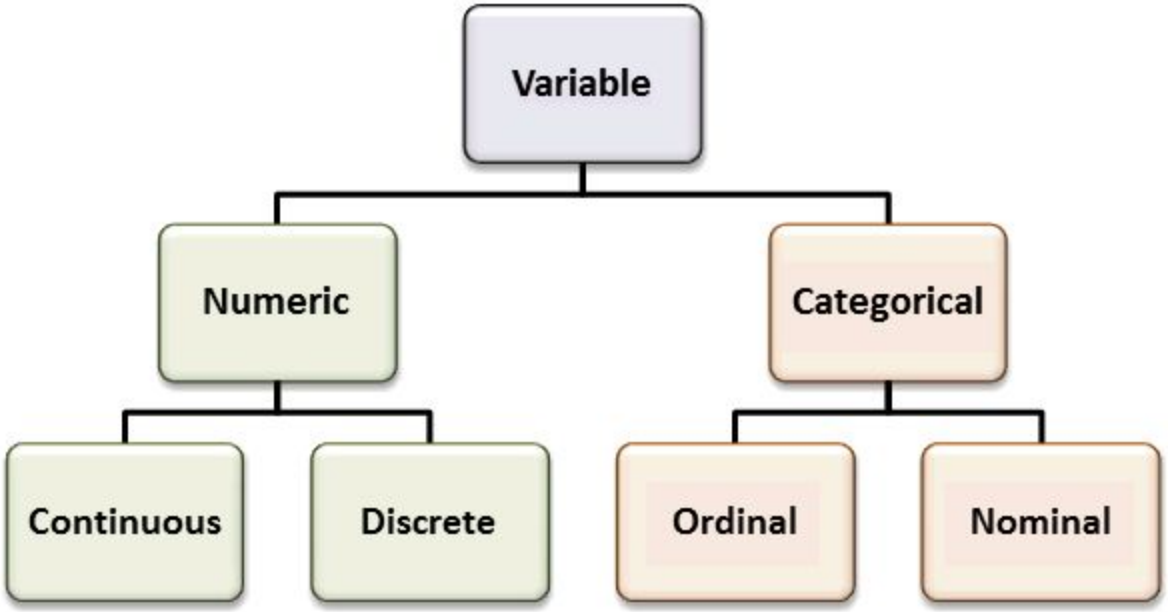
Variable Types

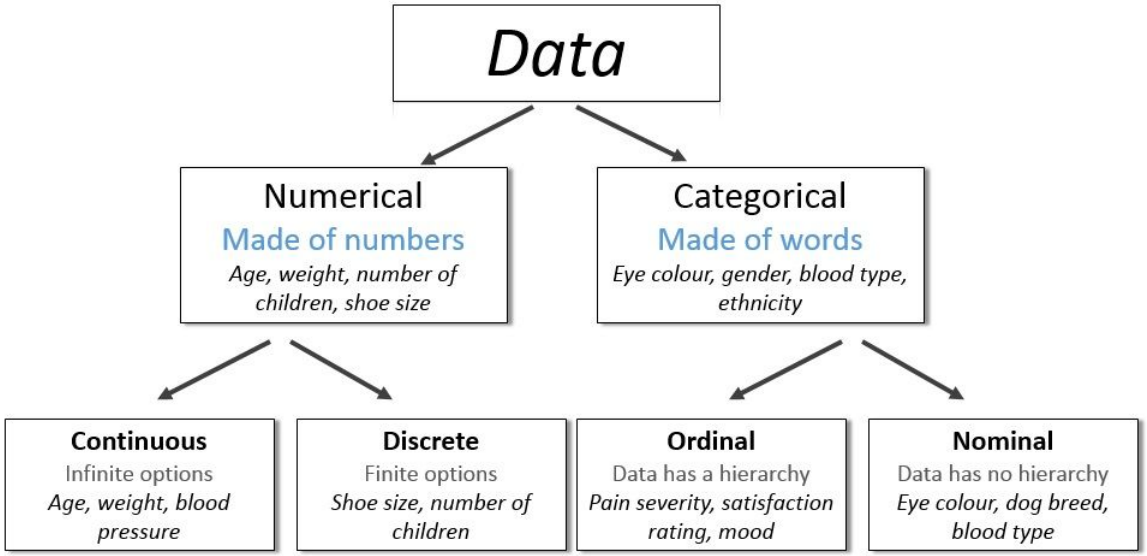
In a quantitative study, it is very important to know what types of data are being studied, for different types of data different analyzes are performed.
The two main data types: quantitative and qualitative.



Qualitative (or categorical) variables are characteristics that cannot be measured numerically, for example sex, gender, color and education.
Qualitative data can be nominal or ordinal.

Quantitative variables represents numerical values and can be measured on a quantitative scale, for example height, mass, income and age.

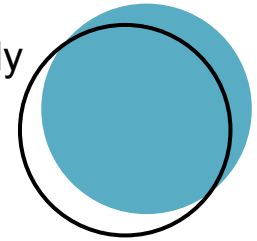

Variable
Numeric
Categorical
Continuous
Discrete
Ordinal
Nominal


Data
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type, ethnicity
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

Numerical

Numerical values or observations can be measured and placed in ascending or descending order. Scatter plots and line graphs are used to graph numerical data.

DISCRETE values or observations are counted as distinct and separate and can only take particular values. They cannot be measured, for instance number of threads in a sheet, number of stars given for an energy rating.

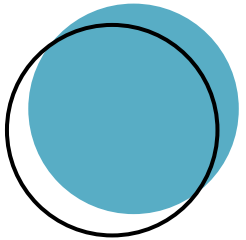


CONTINUOUS data represents measurements and therefore their values cannot be counted, but they can be measured. Values or observations may take on any value within a finite or infinite interval. Examples: height, time and temperature.

Categorical data

Categorical data can be sorted into groups or categories. Bar charts, histograms and pie graphs are used to graph categorical data.

NOMINAL values represent discrete units and are used to label variables that have no quantitative value. Just think of them as “labels”.
You can count but not order or measure nominal data.
Examples: sex, eye color.



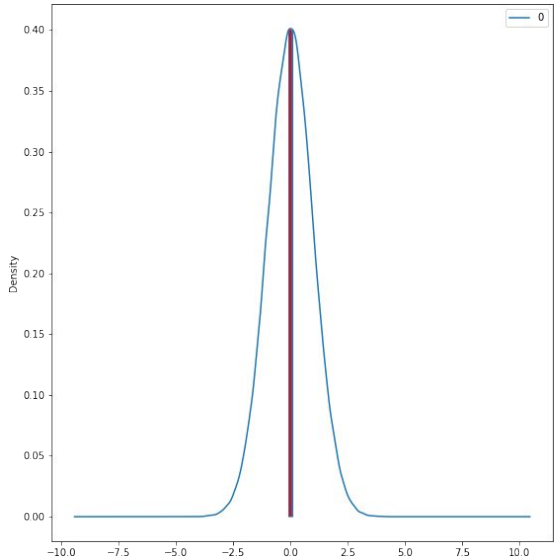
ORDINAL values or observations that can be ranked (put in order) or have a rating scale attached.
You can count and order, but not measure ordinal data.
Example: house numbers and swimming level.

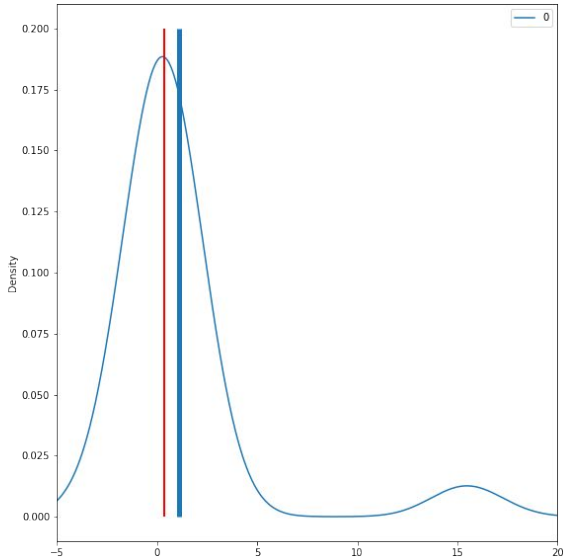
Measure of Central Tendency

Mean: average, sum of observations divided by number of observations

Median: observation that falls in the middle of ordered sample, splits sample into 2 when ordered.

Mode: most frequent value

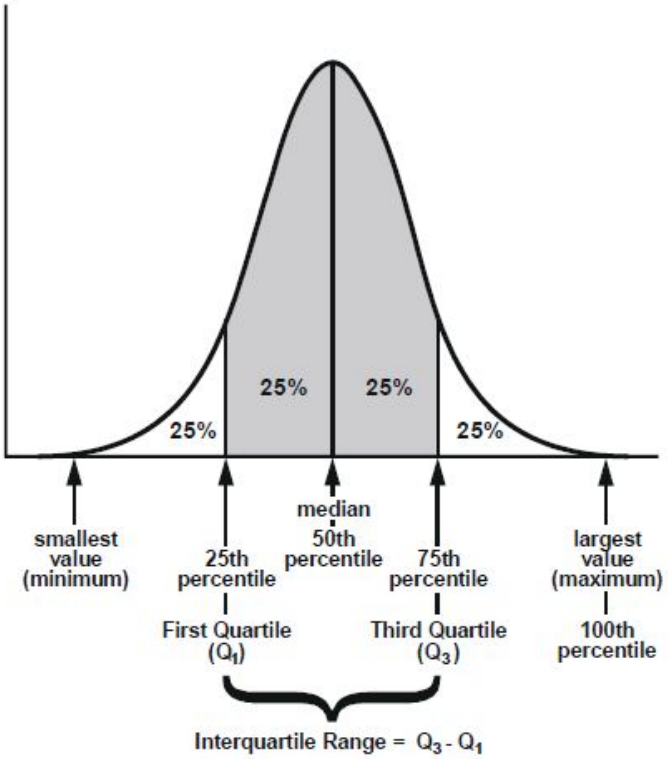


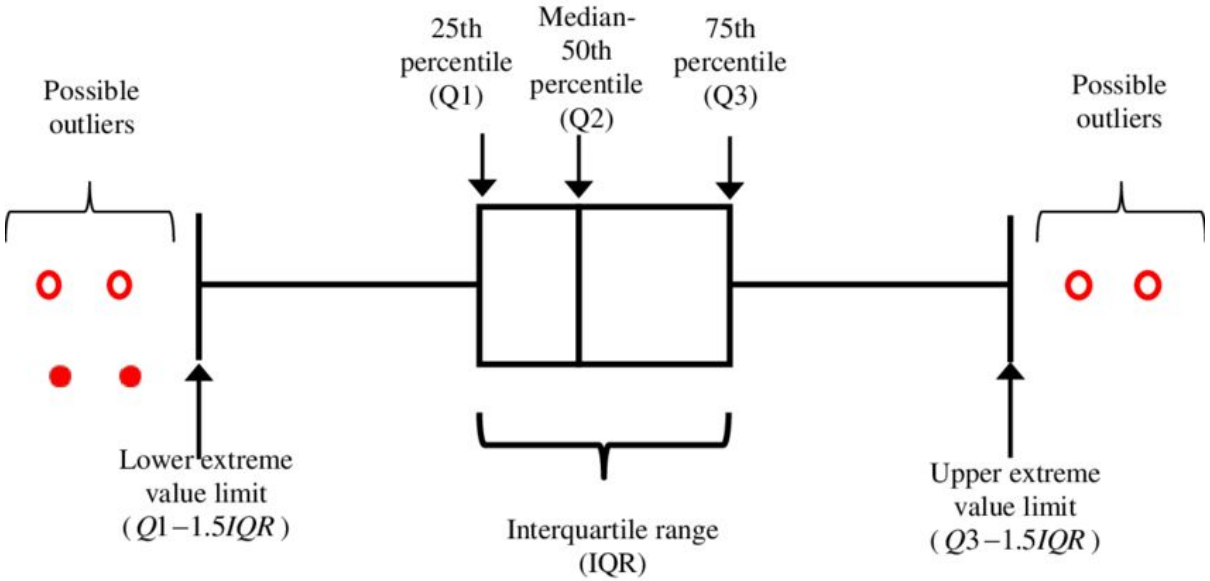


Measure of Position

point at which a given percentage of data fall below or above that point

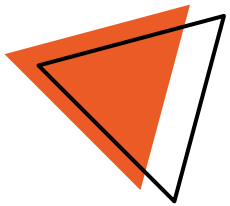
- **Percentile**: the p^{th} percentile is the point such that p% of the observations fall below or at that point and (100-p%) fall above it. The median is the 50^{th} percentile
- **Lower (upper) quartile**: 25^{th} (75^{th}) percentile
- **Outlier**: data point that differs significantly from other observations. Not all outliers are errors!


25%
25%
25%
25%
smallest value (minimum)
25th percentile
median 50th percentile
75th percentile
largest value (maximum)
First Quartile (Q_1)
Third Quartile (Q_3)
100th percentile
Interquartile Range = $Q_3 - Q_1$


Possible outliers
25th percentile (Q1)
Median-50th percentile (Q2)
75th percentile (Q3)
Possible outliers
Lower extreme value limit (Q1 − 1.5IQR)
Interquartile range (IQR)
Upper extreme value limit (Q3 − 1.5IQR)

Measure of Dispersion

- **Range**: difference between largest and smallest observation
- **Standard Deviation**: deviation of data from the mean value
- **Variance**: measure of how far observations are spread out from their mean value. Calculated as square of standard deviation

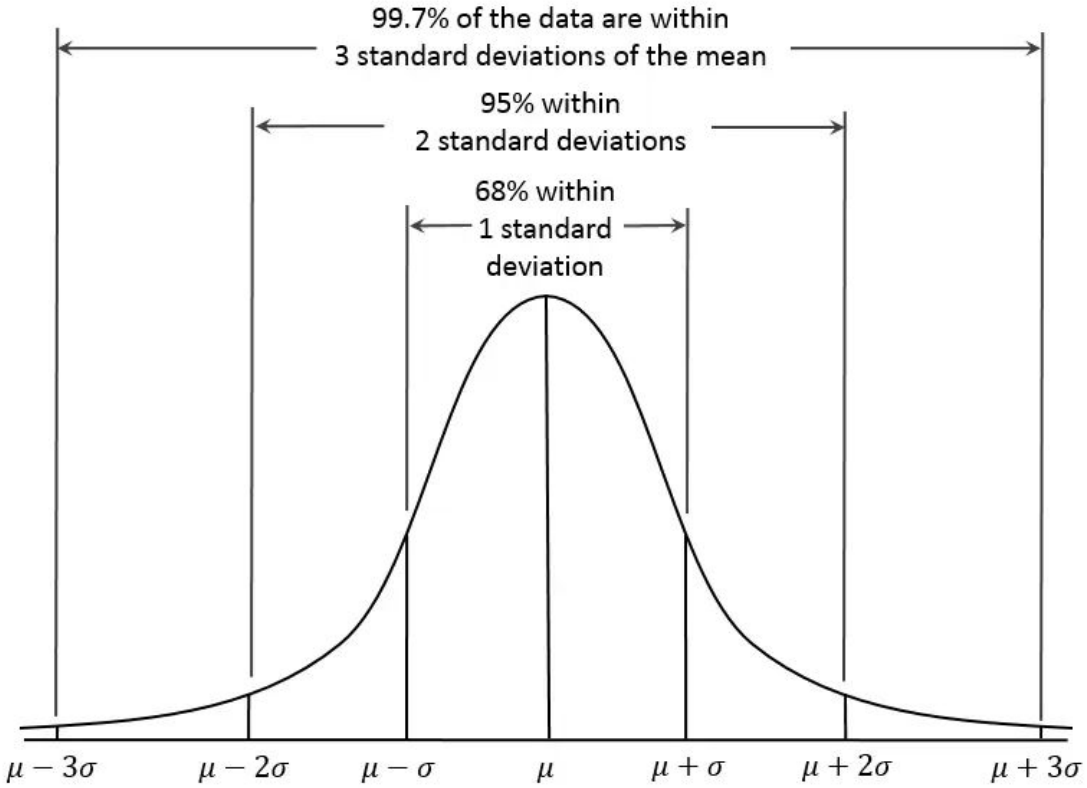


Standard Deviation

- The **deviation** of an observation y_i from the sample mean $\bar{y}$ is $(y_i - \bar{y})$, the difference between them.
- The **standard deviation** s of n observations is:

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}}$$

- y_i = value of an observation
- $\bar{y}$ = sample mean


99.7% of the data are within
3 standard deviations of the mean
95% within
2 standard deviations
68% within
1 standard
deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

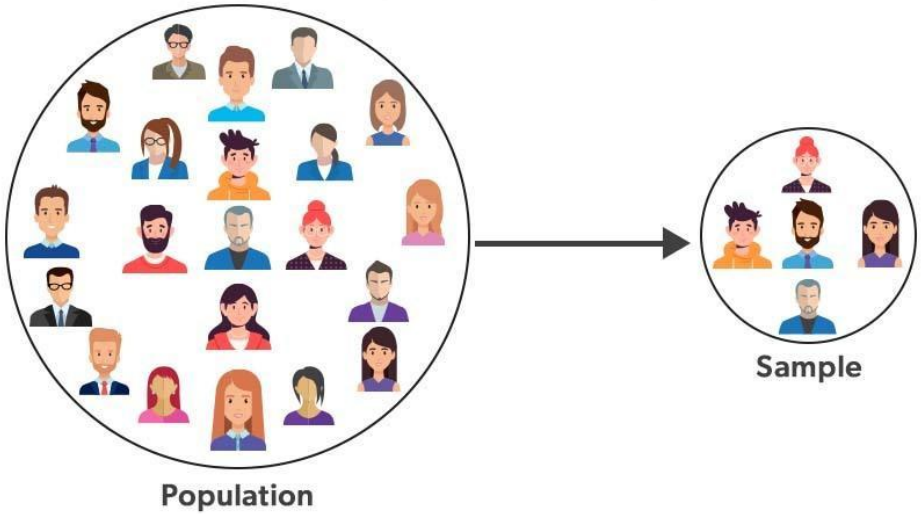
Population & Sample

Population

The entire group that we want to infer. (not necessarily a group of people)

Sample:

Specific group of the population that will be collected data from.



Representative samples

Is the sample representative?

What is the probability that the sample is equal to the population

If x% of a sample of people have Y, it does **NOT** automatically mean that x% of people have Y.

A sample is representative if the variation is the same in the sample as in the population.

How do you measure variation?

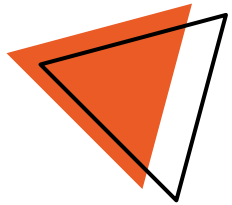
- Mean
- Deviation from mean

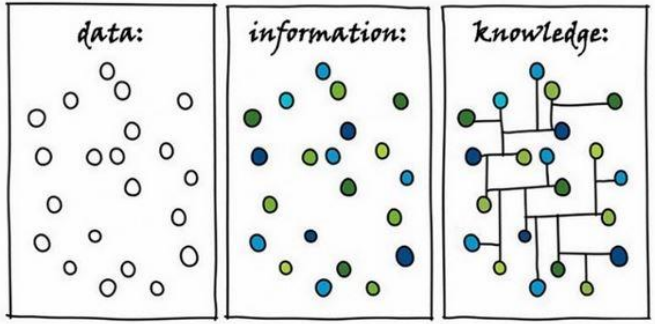
From data to wisdom

Information consists of data, but data is not necessarily information.

Also, wisdom is knowledge, which in turn is information, which in turn is data, however knowledge is not necessarily wisdom.

Wisdom is a subset of knowledge, which is a subset of information, which is a subset of data.

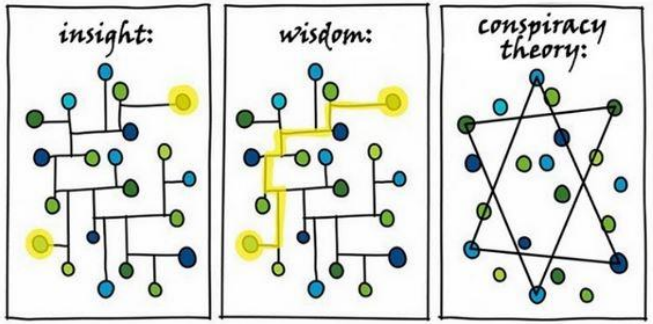




Data: information such as numbers, text, images and sounds, in a form that is suitable for storage or processing by a computer, used as a basis for making calculations or drawing conclusions.

Information: data that has been organized and presented in a systematic fashion to clarify the underlying meaning.

Knowledge: awareness or possession of information, facts, ideas, truths, or principles, understanding gained through experience or study.



Insight: understanding of relationships that sheds light on or helps to solve a problem.

Wisdom: knowledge and experience needed to make sensible decisions and judgments.

Conspiracy Theory: a theory that rejects the standard explanation for an event and instead credits a covert group or organization with carrying out a secret plot.

WE DID IT!