Statistics

It is used to quantitatively describe the attributes of the known data and provides summaries of either the sample or the population. Graphs, charts, and tables can be used to represent descriptive statistics

Descriptive statistics are divided into measures of central tendency and measures of variability (spread)

Qualitative Data or Categorical Data

Quantitative data takes on numeric values that allow us to perform mathematical operations (like the number of dogs , age ,income,height, Temperature, Average Speed).all take on values that we can add, subtract and perform other operations with to gain useful insight

Categorical is used to label a group or set of items (like dog types , Marital Status (Single, Married, Divorced) Ratings on a Survey (Poor, Ok, Great), Letter Grades (A+, A, A-, B+, B, B-, ...) Zip Code etc.).

Categorical Ordinal vs. Categorical Nominal

We can divide categorical data further into two types: Ordinal and Nominal.

Categorical Ordinal data take on a **ranked ordering** (Letter Grades (A+, A, A-, B+, B, B-, ...),ratings on survey (poor ,ok ,good), eduction (primary, secondary,university,masters,phd..))

Categorical Nominal data do not have an order or ranking (like the types of the dog,gender,nationality,types of fruit).

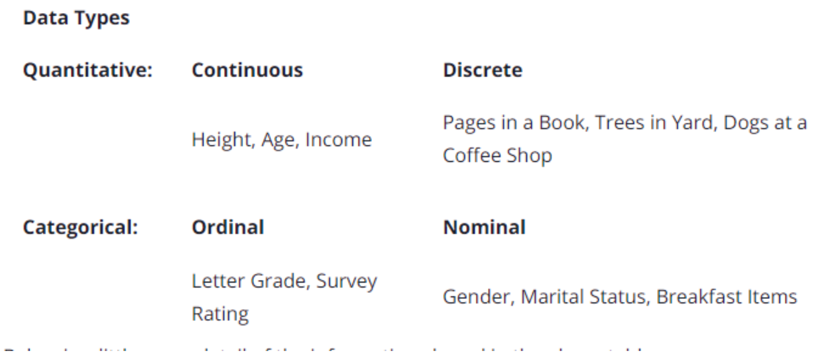
Continuous vs. Discrete

We can think of ***quantitative data*** as being either continuous or discrete.

Continuous data can be split into smaller and smaller units, and still a smaller unit exists.

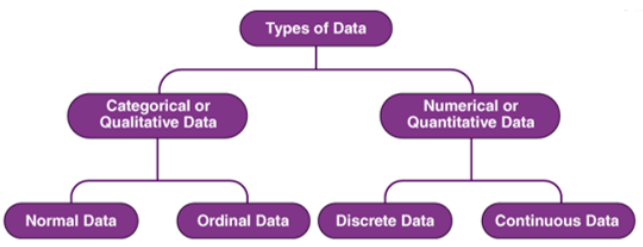
Ex: Amount of Water Consumed in a Day, Time to Run a Mile, Travel Distance from Home to Work

Discrete data only takes on ***countable values.*** Ex:The number of dogs,number of movies watched ,number of pages on a book



Data Types

Quantitative:	**Continuous**	**Discrete**
	Height, Age, Income	Pages in a Book, Trees in Yard, Dogs at a Coffee Shop
Categorical:	**Ordinal**	**Nominal**
	Letter Grade, Survey Rating	Gender, Marital Status, Breakfast Items



There are four main aspects to analyzing Quantitative data

1-Measures of center:

describe the center of the data set (mean, median, mode).

-**The *mean*** :is often called the ***average*** or the ***expected*** value in mathematics. We calculate the mean by adding all of our values together and dividing by the number of values in our dataset.

-**The median:** splits our data so that 50% of our values are lower and 50% are higher. The median is the middle number in a sorted listMedian for Odd Values is simply the number in the direct middle,but for even numbers is the average of the two values in the middle.

5, 8, 15, 7, 10, 22, 3, 1, 15 >>>> 1,3,5,7,8,10,15,15,22 >> mid=8

5, 8, 15, 7, 10, 22, 3, 1, 15, 2 >>>> 1,2,3,5,7,8,10,15,15,22 >> mid(7,8) >> 7.5

-The mode is the most frequently observed value in our dataset. If all observations in our dataset are observed with the same frequency, there is no mode. If we have the dataset:1, 1, 2, 2, 3, 3, 4, 4 There is no mode . same no of repeating.

1, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, 8, 9 >>> two modes >> 3 , 6

Notation is a common language used to communicate mathematical ideas. Think of notation as a universal language used by academic and industry professionals to convey mathematical ideas('+' for adding , '/' for division)

Before collecting data, we usually start with a question, or multiple questions, that we would like to answer. The purpose of data is to help us in answering these questions.

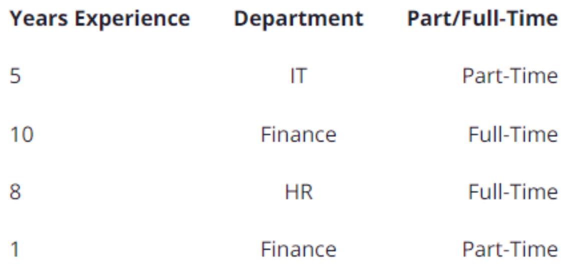
Random Variables

A random variable is a placeholder for the possible values of some process.

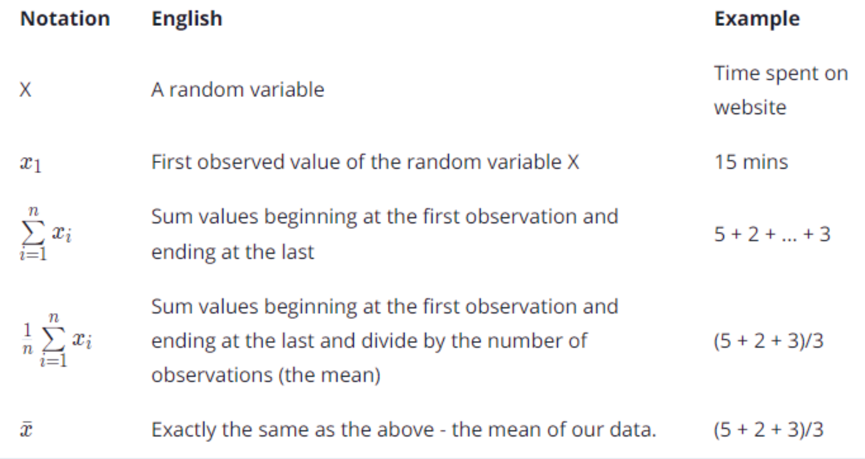
capital letters signify ***random variables***. When we look at individual instances of a particular random variable, we identify these as ***lowercase*** letters with subscripts attach themselves to each specific observation.

X= years of experience,Y= Department,Z= Part/Full-Time

X_1=5,Y_2-Finance ,Z_3=Full-Time,n=4



Years Experience	Department	Part/Full-Time
5	IT	Part-Time
10	Finance	Full-Time
8	HR	Full-Time
1	Finance	Part-Time



Notation	English	Example
X	A random variable	Time spent on website
x_1	First observed value of the random variable X	15 mins
$\sum_{i=1}^{n} x_i$	Sum values beginning at the first observation and ending at the last	5 + 2 + ... + 3
$\frac{1}{n}\sum_{i=1}^{n} x_i$	Sum values beginning at the first observation and ending at the last and divide by the number of observations (the mean)	(5 + 2 + 3)/3
$\bar{x}$	Exactly the same as the above - the mean of our data.	(5 + 2 + 3)/3

Summary

- Evaluate data types and variable types
- Analyze measures of center
- Implement notation

Histograms

Histograms are the most common visual used for quantitative data. help you understand the four aspects regarding a quantitative variable:center,spread,shape,outliers . Each bin represents a range of values in a dataset. The number of values that fall in the range of each bin determines the height of each histogram bar.

Five Number Summary

- **Minimum**: The smallest number in the dataset.
- **Q1**: The value such that 25% of the data fall below.
- **Q2:** The value such that 50% of the data fall below.
- **Q3**: The value such that 75% of the data fall below.
- **Maximum**: The largest value in the dataset.

Range

The range is then calculated as the difference between the maximum and the minimum.

IQR

The interquartile range is calculated as the difference between Q3 and Q1.

EX1: 1, 5, 10, 3, 8, 12, 4, 1, 2, 8

Sort>> 1,1,2,3, 4,5, 8,8,10,12 >> median=(4+5)/2

-range =12-1=11 , (1,1,2,3, 4 , 5, 8,8,10,12) TO find Q1,Q2

Q1=MID(1,1,2,3)=2, Q2= median ,Q3=MID(5, 8,8,10,12)= 8

EX2: 5, 10, 3, 8, 12, 4, 1, 2, 8

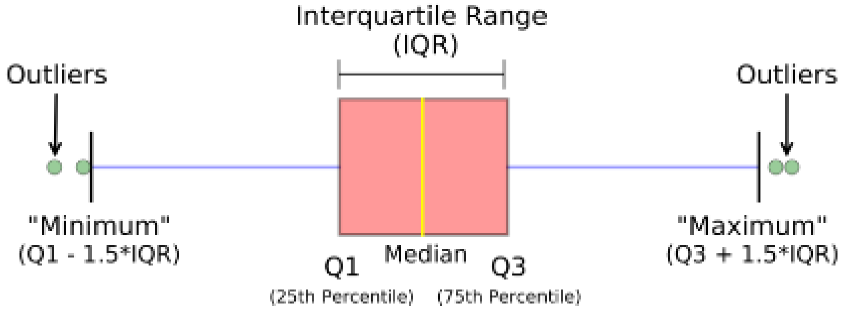
Sort>> 1,2,3,4,5,8,8,10,12>> median=5

-range =12-1=11 ,(1,2,3,4, 5, 8,8,10,12)

Q1=MID(1,2,3,4)=2.5 , Q2= median ,Q3=MID(8,8,10,12)= 9

Box plots

are useful for quickly comparing the spread of two data sets across some key metrics, like quartiles, maximum, and minimum.



Standard Deviation and Variance:

STD :It is defined as the average distance of each observation from the mean.

The variance is the average squared difference of each observation from the mean.

-Standard deviation is a common metric ***used to compare the spread of two datasets***. The benefits of using a single metric instead of the 5 number summary are:

-It simplifies the amount of information needed to give a measure of spread

-It is useful for inferential statistics

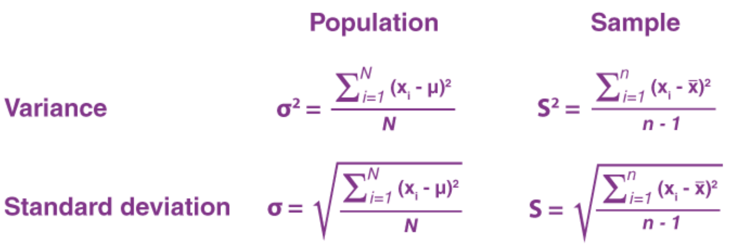
Ex: 10, 14, 10, 6

1- Calculate the mean = 10

2- $(X_i- X)^2$ = 0,16,0,16

3- Variance= (0+16+0+16)/4=8

4- STD= Sqrt(8)=2.83



	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

Ex: 1, 5, 10, 3, 8, 12, 4

Variance=13.55, STD =3.68

Ex: 15, 4, 3, 8, 15, 22, 7, 9, 2, 3, 3, 12, 6 Sort>> 2,3,3,3,4,6,7,8,9,12,15,15,22

Range=20,IQR =10.5 , variance=33.9 ,std=5.8 ,min, max =22

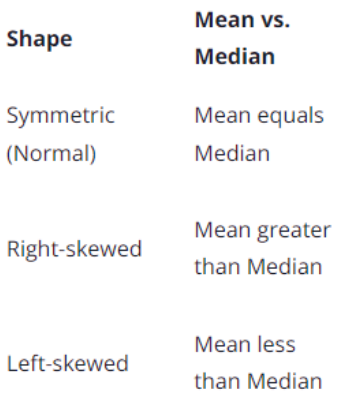
Shape

From a histogram, we can quickly identify the shape of our data

1. Right-skewed ex: Time between phone calls at a call center, Time until light bulb dies

2. Left-skewed ex: Grades as a percentage in many universities, Age of death, Asset price changes

3. Symmetric (frequently normally distributed) ex: Height, Weight, Errors, Precipitation



Shape	Mean vs. Median
Symmetric (Normal)	Mean equals Median
Right-skewed	Mean greater than Median
Left-skewed	Mean less than Median

Outliers

are points that fall very far from the rest of our data points.

Descriptive vs. Inferential Statistics

Descriptive statistics is about ***describing*** our collected data.

Inferential statistics is about using our collected data to draw conclusions about a larger population.

-Population - our entire group of interest.

-Parameter - numeric summary about a population

-Sample - a subset of the population

-Statistic numeric summary about a sample

-inference :Drawing conclusions regarding a population using information from a sample.

Recap

-Variable Types: categorical or quantitative. identify quantitative variables as either ***continuous or discrete***. identify ***categorical*** variables as either **ordinal or nominal**.

- Categorical Variables

-***Quantitative Variables***: ***four main*** aspects used to describe quantitative variables:

1- **Measures of Center 2 - Measures of Spread 3-Shape of the Distribution 4-Outliers**

-calculating measures of Center 1-Means 2 -Medians 3-Modes

-measures of Spread 1-Range 2- Interquartile Range 3-Standard Deviation 4- Variance

-Depending on the shape associated with our dataset, certain measures of center or spread may be better for summarizing our dataset.

-When we have data that follows a normal distribution, we can completely understand our dataset using the mean and standard deviation.

-However, if our dataset is skewed, the 5 number summary (and measures of center associated with it) might be better to summarize our dataset.

- When ***outliers*** are present we should consider the following points.

1. Noting they exist and the impact on summary statistics.

2. If typo - remove or fix

3. Understanding why they exist, and the impact on questions we are trying to answer about our data.

4. Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

5. Be careful in reporting. Know how to ask the right questions.

- Descriptive statistics is about describing our collected data, using measures of center, measures of spread, the shape of our distribution, and outliers. We can also use plots of our data to gain a better understanding.

- Inferential Statistics is about using our collected data to draw conclusions to a larger population. Performing inferential statistics well requires that we take a sample that accurately represents our population of interest.

<u>Summary</u>

-Evaluate measures of spread

1.Range 2.Interquartile Range (IQR) 3.Standard Deviation 4.Variance

-Analyze outliers

-Evaluate descriptive and inferential statistics