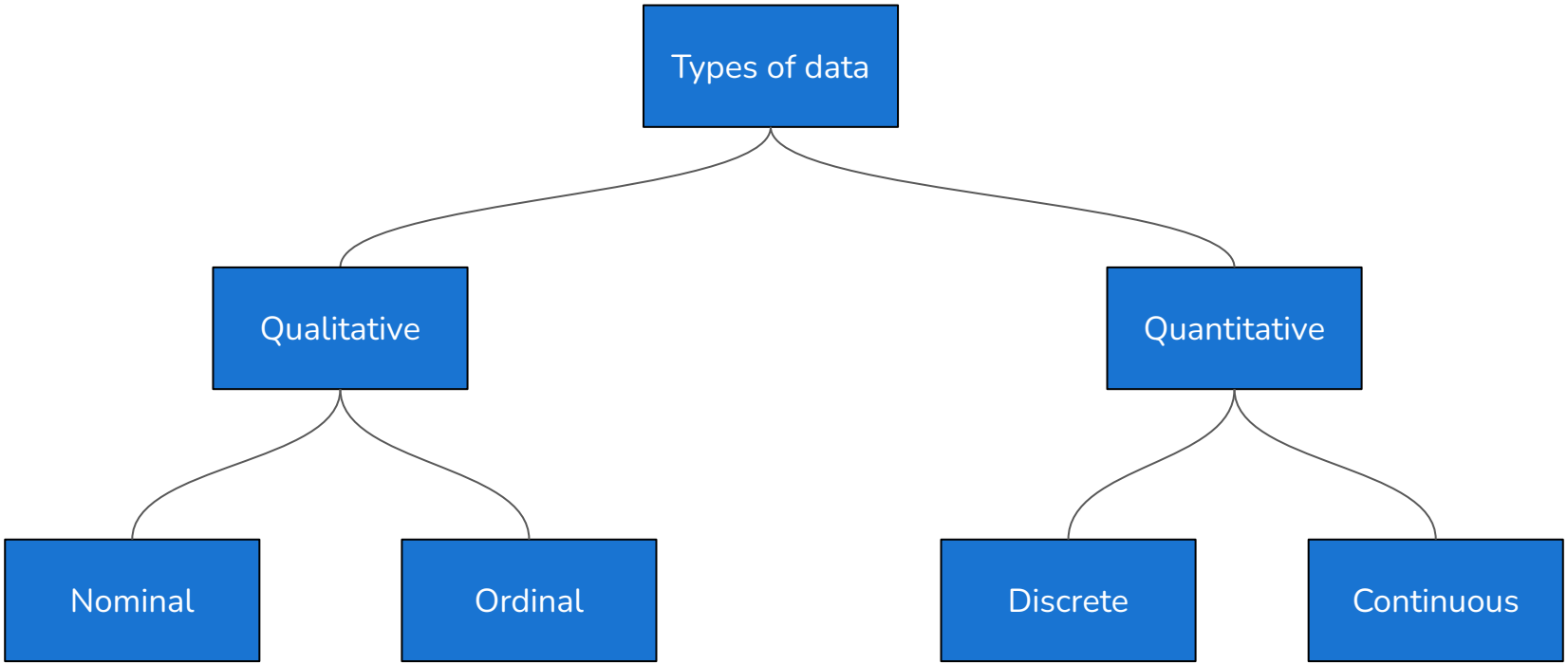
“Statistics for Data Science”
Week 2

Gauge your understanding

1. What are the different types of data?
2. What values help us identify the central tendencies in the data?
3. How do we identify the dispersion in data?
4. Probability theory and probability distributions
5. What is a random variable and how is it related to probability distribution?
6. What is Central Limit Theorem (CLT) and when is it used?
7. What do you mean by estimations?

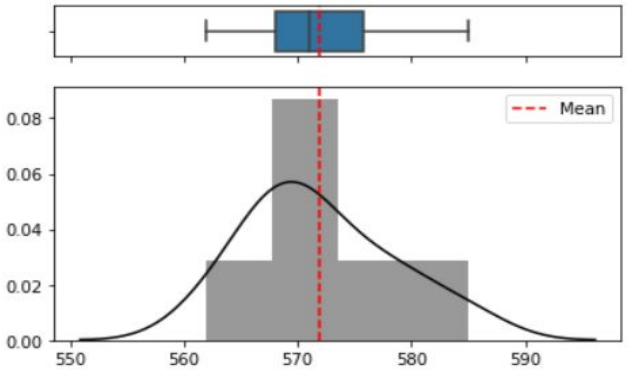
What are different types of Data?



Central Tendencies - Mean

Mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

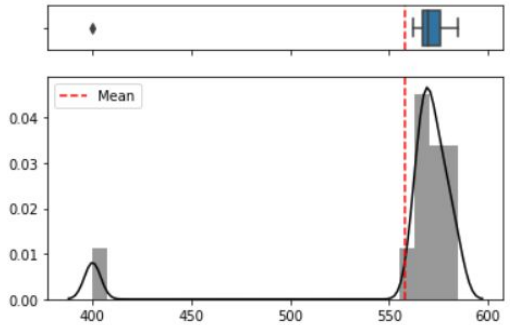
Example: Suppose over the last 12 days, a store sold 570, 568, 565, 572, 568, 585, 568, 578, 580, 575, 562, 572 litres of milk.



Mean = 571.92

But if store closed early on 1 day and sold only 400 litres of milk, the mean will be

Mean = 557.75

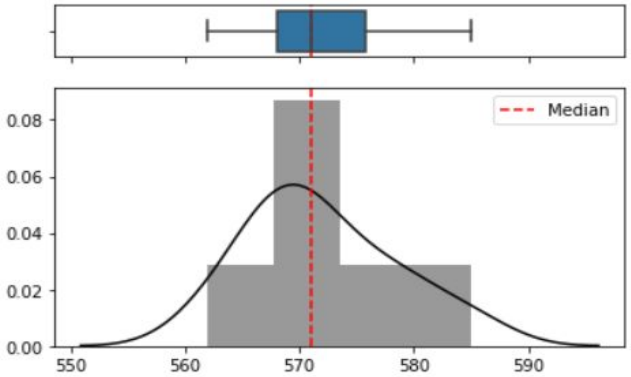


The mean has changed a lot. **Mean is affected by outliers.**

Central Tendencies - Median

The median is the middle score for a set of data that has been arranged in order of magnitude.

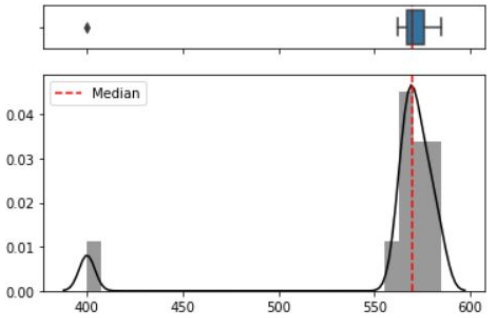
Example: Suppose over the last 12 days, a store sold 570, 568, 565, 572, 568, 585, 568, 578, 580, 575, 562, 572 litres of milk.



Median = 571

But if store closed early on 1 day and sold only 400 litres of milk, the median will be

Median = 570

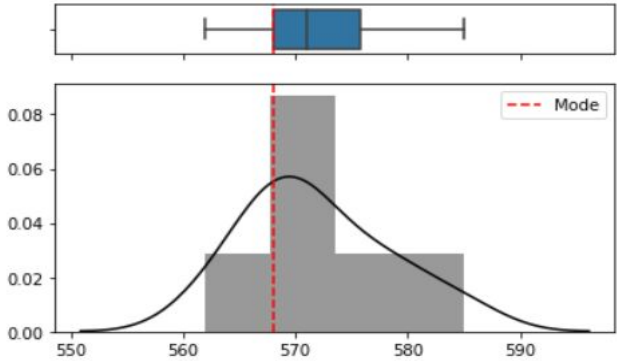


Hence median is less affected by outliers.

Central Tendencies - Mode

The mode is the most frequent score in our data set. This is the only central tendency measure that can be used with nominal data, which have purely qualitative category assignments.

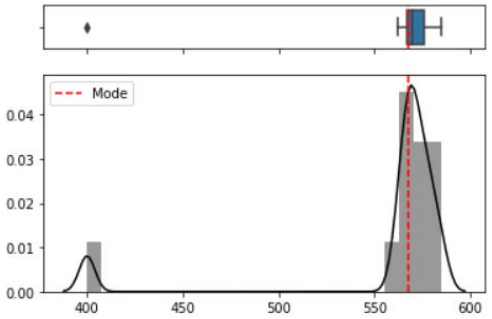
Example: Suppose over the last 12 days, a store sold 570, 568, 565, 572, 568, 585, 568, 578, 580, 575, 562, 572 litres of milk.



Mode = 568

But if store closed early on 1 day and sold only 400 litres of milk, the mode will be

Mode = 568



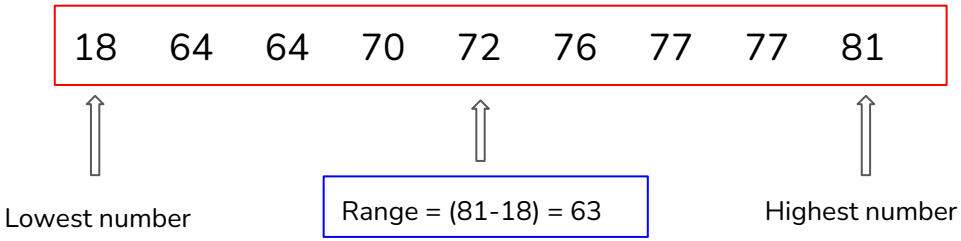
Hence mode is not affected by outliers.

Measure of dispersion, Range and IQR

Measures of dispersion: It indicates how large the spread of distribution in around the central tendency.

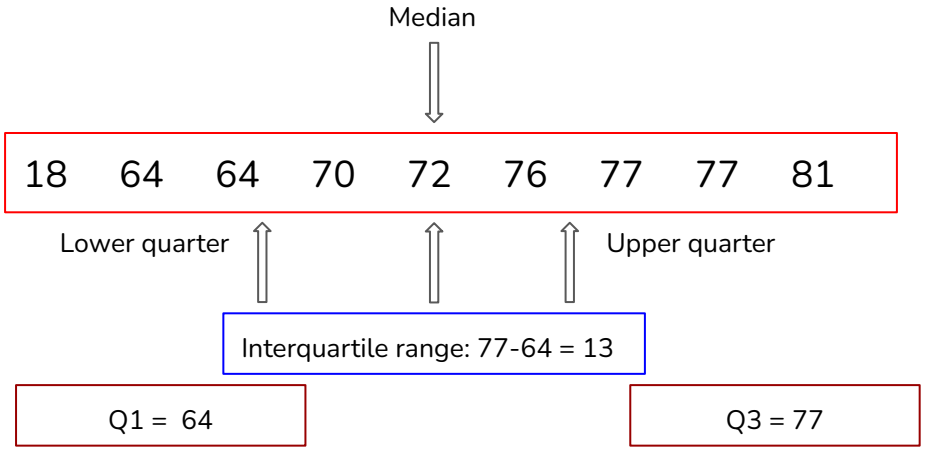
Range: Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in dataset.

range = X(maximum) - X(minimum)



Interquartile range (IQR): It is a measure of variability, based on dividing a data set into quartiles i.e. into four parts represented by Q1,Q2,Q3 and Q4.

IQR = Q3 - Q1



Standard Deviation

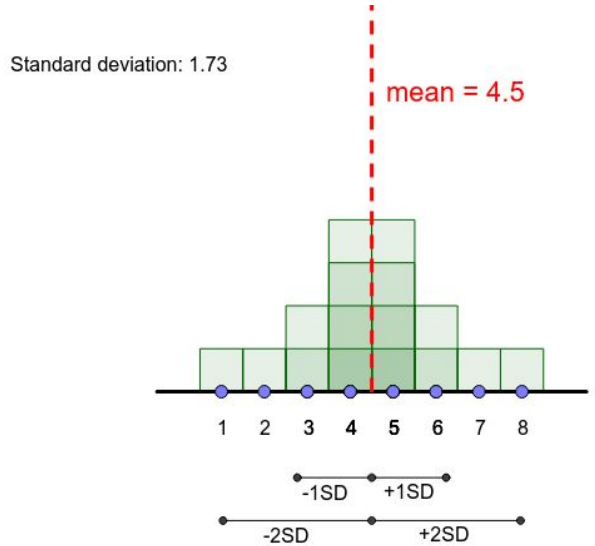
Standard Deviation: It is a measure of how spread out the numbers in a distribution are. It is the measure of dispersion of a data from its mean.

Formula for SD of a population:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

Formula for SD of a sample:

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



Why denominator term for Sample is n-1

Assume a population with “N” items. Suppose that we want to take samples of size “n” from that population . If we could list all possible samples of “n” items that could be selected from the population of “N” items, then we could find the SD for each possible sample.

If all the sample drawn as “Unbiased” , then the average of the sample SD for all possible samples would be equal the population SD. However practically it is not possible to get all possible samples from a population.

Hence when we divide by (n –1) when calculating the sample SD , then it turns out that the average of the sample SD for all possible samples is equal the population SD. So the sample SD is what we call an unbiased estimate of the population SD. (This is also known as Bessel’s Correction)

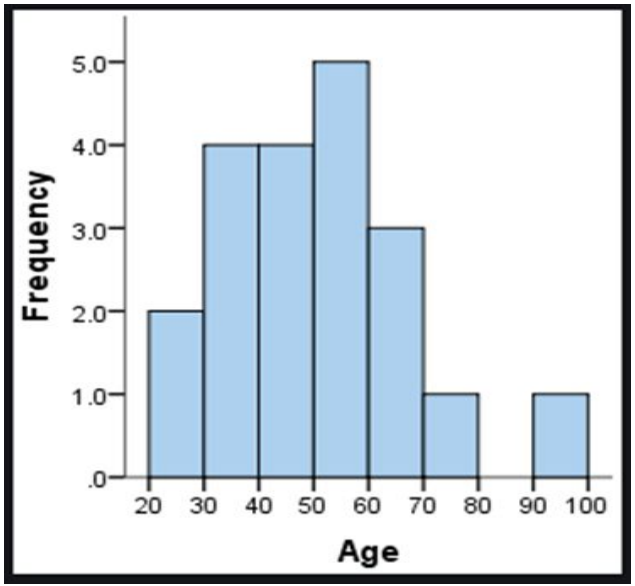
Histogram

Histogram is bar chart which represents a frequency distribution.

Horizontal axis of the histogram represents the data points within an interval called as “bin” and vertical axis represents the corresponding “frequency”

How to calculate “bin” from a numeric set of data points

- Count the number of data points.
- Calculate the number of bins by taking the square root of the number of data points and round up.
- Calculate the bin width by dividing the Range (i.e. Max-Min) by the # of bins.



Box and Whisker Plot

Box and Whisker Plot displays the Five-Point summary of the data.

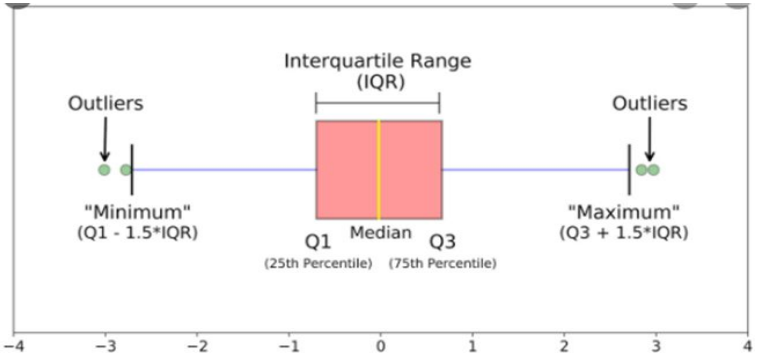
The five-number summary is the minimum, first quartile, median, third quartile, and maximum.

In a box plot, box is from the first quartile to the third quartile.

A vertical line goes through the box at the median.

Minimum value represented through a Whisker is (Q1-1.5* IQR)

Maximum value represented through a Whisker is (Q3 + 1.5* IQR)



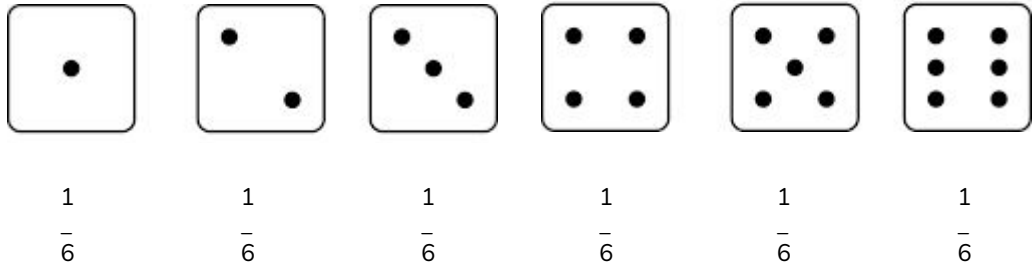
Any point which is below the Minimum Value and/or above the maximum value is an outlier

Probability

Probability refers to chance or likelihood of a particular event taking place. It ranges between 0 to 1

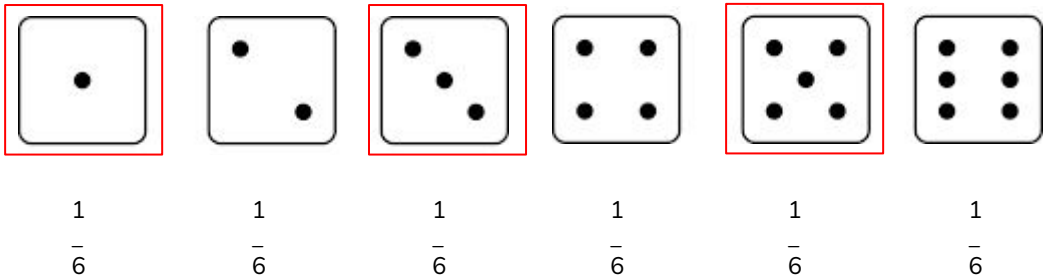
Example: If a dice is rolled, what is the probability of getting an odd number?

Possible Outcomes



$\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$

Getting an odd number



$\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$

$$\frac{3}{6} = \frac{1}{2}$$

Mutually Exclusive and Independent Events

Mutually exclusive events

Two events A and B are said to be mutually exclusive if the occurrence of A precludes the occurrence of B. The probability of both the events occurring together is zero.

$$P(A \text{ and } B) = 0$$

Examples:

1. Tossing a coin is a mutually exclusive event because either you will get a head or a tail. You can never get head and tail simultaneously while tossing a coin.
2. While turning to right or left, either you will turn left or right not both.

Independent Events

Two events A and B are said to be independent if the occurrence of A is in no way influenced by the occurrence of B. Likewise occurrence of B is in no way influenced by the occurrence of A.

Example: Getting a head by tossing a coin and getting 5 by rolling a die

Joint, Marginal and Conditional probability

A survey was conducted with 1000 people in New York to determine people's favourite sports. The options were Cricket, Football and others. The following table contains the response gathered:

	Male	Female	Total
Cricket	240	150	390
Football	200	50	250
Others	100	260	260
Total	540	460	1000

Probability proportions

	Male	Female	Total
Cricket	0.24	0.15	0.39
Football	0.2	0.05	0.25
Others	0.1	0.26	0.26
Total	0.54	0.46	1

Joint Probability: It is the probability of two events occurring together at the same time. **Ex.** What is the probability of someone being female and liking football?

P (female and football) = 0.05

Joint, Marginal and Conditional probability

Marginal Probability: It is the probability of an event irrespective of the outcome of another variable. **Ex.** Probability of being a male:

$$P(male) = 0.54$$

which completely ignores the sport.

Conditional Probability: It is the probability of an event occurring given that another event has occurred. **Ex.** What is the probability that a person would like to play cricket given that the person is male.

$$P\ (Cricket \mid Male) = P\ (Cricket, Male) / P\ (Male)$$

$$= 0.24 / 0.54$$

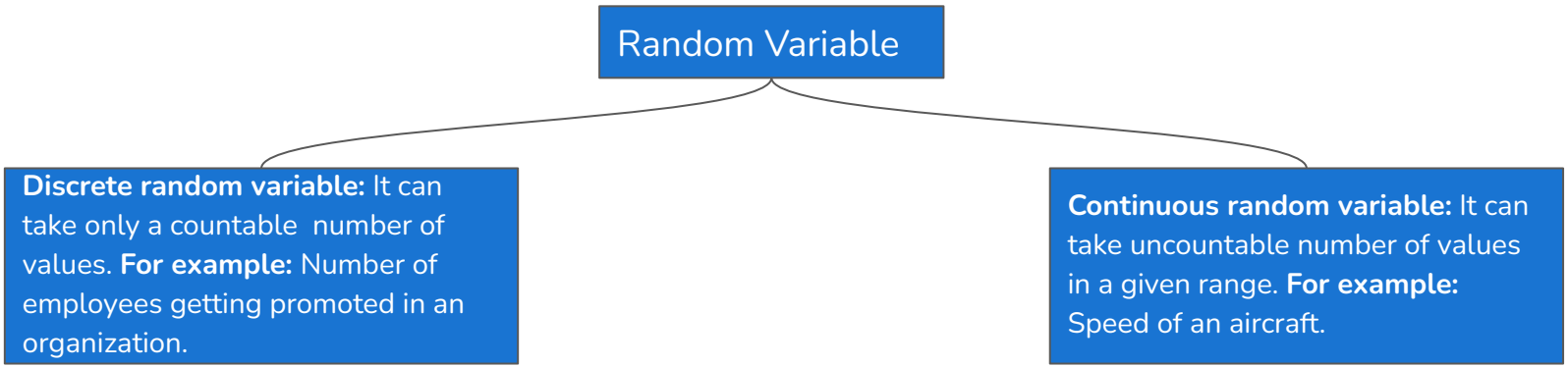
$$= 0.44$$

Case Study

What is a Random Variable?

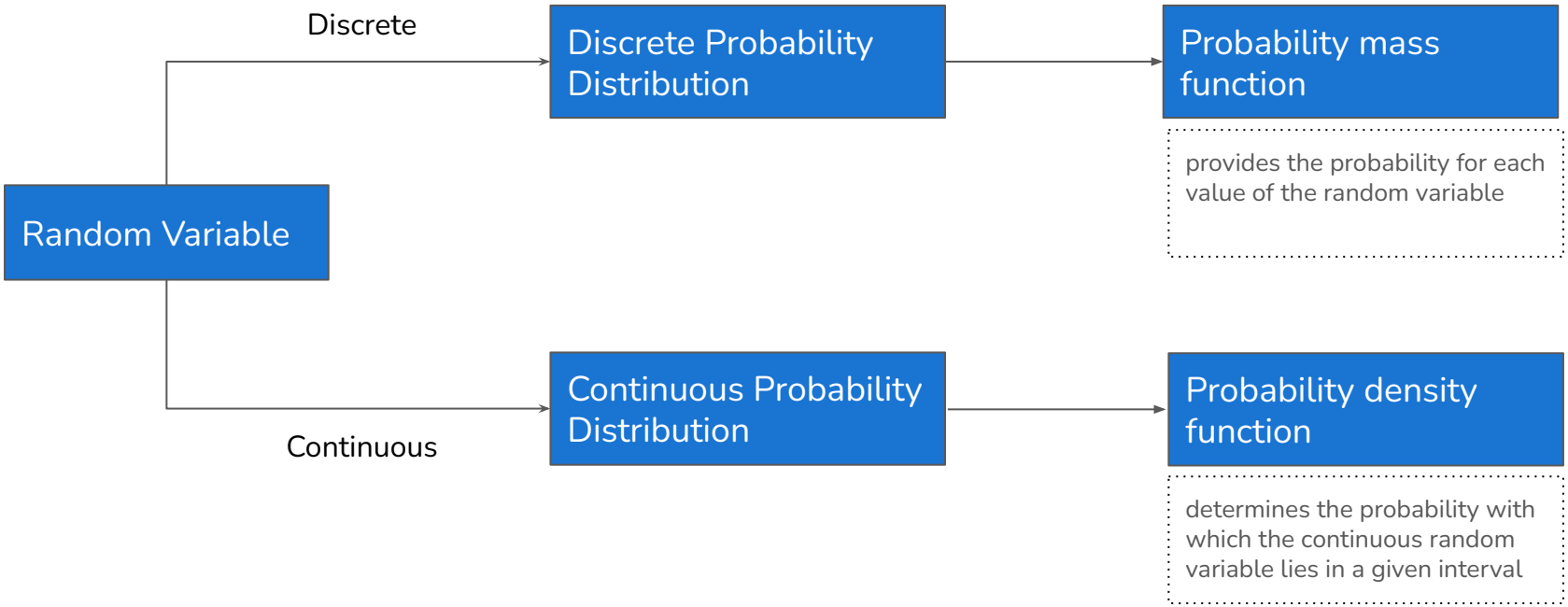
A random variable is a variable which its value is determined by a random process or by a random phenomenon. A random process is an event or experiment that has a random outcome. It is usually denoted by capital letter X and the probability associated with any particular value of X is denoted by P(X=value) = Probability of that value.

Example: Suppose that a fair coin is tossed twice and the possible outcome are {HH, HT, TH, TT}. Let X be the random variable representing the number of heads that can come up. So, X can take values from the set {2, 1, 0}.
The probability of two heads coming up is P(X=2) = ¼.



What is a Probability Distribution?

The probability distribution of a random variable describes the values that the random variable can take along with the probabilities of those values.



Binomial Distribution

The binomial distribution is the probability distribution of a success or failure outcome of an experiment that is conducted multiple times. **Ex.** Probability of getting a head after tossing a coin 10 times

Binomial distributions must also meet the following three criteria:

1. **The number of observations or trials is fixed.**
2. **Each observation or trial is** independent. In other words, none of your trials have an effect on the probability of the next trial.
3. The **probability of success** (tails, heads, fail or pass) is **exactly the same** from one trial to another.

The binomial distribution formula is:

$$b(x; n,p) = nCx * p^{x} * (1 - p)^{(n - x)}$$

Where: b = binomial probability

x = total number of successes

P = probability of a success on an individual trial

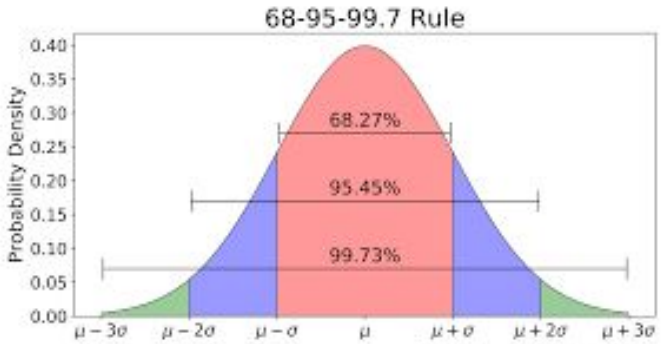
n = number of trials

Normal Distribution

The normal distribution is the probability distribution that is symmetric about the mean. It is also known as bell curve.

Properties:

- In a normal distribution, mean is zero and standard deviation is 1
- It has a zero skewness
- Mean = Median = Mode



Cumulative distribution function (cdf): It is the area under the curve for the given value.

Ex. What is the chance that a man is less than 165 cm tall?

Percent point function (ppf): It is the inverse of the cdf value.

Ex. Given that I am looking for a man who is smaller than 95% of all other men, what size does the subject have to be?

Sampling Distributions

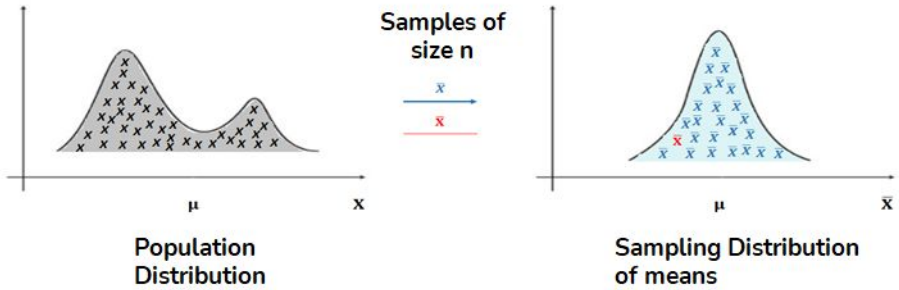
What is the need for sampling?

Given the limited resources and time, it is not always possible to study the population. That's why we choose a sample out of the population to make inference about the population.

Example: Suppose a new drug is manufactured and it needs is to be tested for the adverse side effects on a country's population. It is almost impossible to conduct a research study that involves everyone.

What are Sampling Distributions?

It is a distribution of a particular sample statistic obtained from all possible samples drawn from a specific population.



Central Limit Theorem

The sampling distribution of the sample means will approach normal distribution as the sample size gets bigger, no matter what the shape of the population distribution is.

Assumptions

- Data must be **randomly sampled**
- Sample values must be **independent** of each other
- Samples should come from the **same distribution**
- Sample size must be **sufficiently large (≥30)**

Let's see CLT in action by simulation - Link to external site

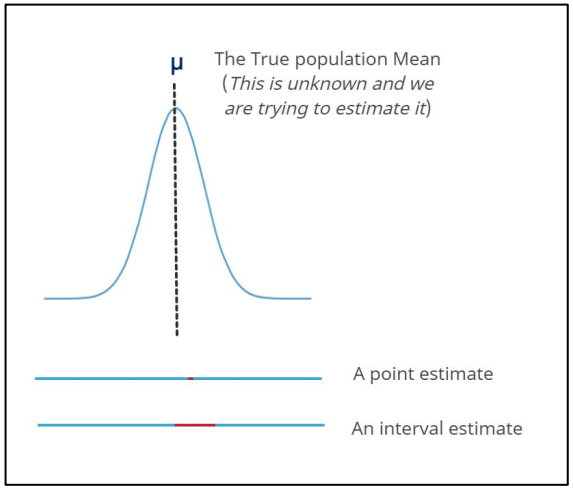
Estimations

Estimation

Make inferences about population parameter based on sample statistic

Point Estimation

Single value of a population parameter
Ex. Population mean as estimated from the sample mean is $20



Interval Estimation

The range of values within which the population parameter lies with some confidence
Ex. Population mean should lie in the range $15-$25, with 95% confidence

Case Study

Great Learning

Happy Learning !