Data Analytics

Course Taught at IIFT

Session 2: Descriptive Statistics and Probability Distributions

Dr. Tanujit Chakraborty

www.ctanujit.org

Today's Topics…

- Data summarization
- Graphical summarization
- Probability vs. Statistics
- Concept of random variable
- Probability distribution concept
- Discrete probability distribution
- Continuous probability distribution

TRP: An example

- Television rating point (TRP) is a tool provided to judge which programs are viewed the most.
  - This gives us an index of the choice of the people and also the popularity of a particular channel.
- For calculation purpose, a device is attached to the TV sets in few thousand viewers' houses in different geographic and demographic sectors.
  - The device is called as **People's Meter**. It reads the time and the programme that a viewer watches on a particular day for a certain period.
- An average is taken, for example, for a 30-days period.
- The above further can be augmented with a personal interview survey (PIS), which becomes the basis for many studies/decision making.
- Essentially, we are to analyze **data** for TRP estimation.

Data

Definition : **Data**

A set of data is a collection of observed values representing one or more characteristics of some objects or units.

Example: For TRP, data collection consist of the following attributes.

- **Age:** A viewer's age in years
- **Sex:** A viewer's gender coded 1 for male and 0 for female
- **Happy:** A viewer's general happiness
  - NH for not too happy
  - PH for pretty happy
  - VH for very happy
- **TVHours:** The average number of hours a respondent watched TV during a day

Data : Example

Viewer#	Age	Sex	Happy	TVHours
…	…	…	…	…
…	…	…	…	…
55	34	F	VH	5
…	…	…	…	…

Note:

- A data set is composed of information from a set of units.
- Information from a unit is known as an observation.
- An observation consists of one or more pieces of information about a unit; these are called variables.

Population

Definition : **Population**

A population is a data set representing the entire entities of interest.

Example: All TV Viewers in the country/world.

Note:

1. All people in the country/world is not a population.
2. For different survey, the population set may be completely different.
3. For statistical learning, it is important to define the population that we intend to study very carefully.

Sample

Definition : **Sample**

A sample is a data set consisting of a population.

Example: All students studying in MBA (IB) 2020-2022 is a sample, whereas those students belong to IIFT is population.

Note:

- Normally a sample is obtained in such a way as to be representative of the population.

Statistic

Definition : Statistic

A statistic is a quantity calculated from data that describes a particular characteristics of a sample.

Example: The sample mean (denoted by $\bar{y}$) is the arithmetic mean of a variable of all the observations of a sample.

Statistical Inference

Definition : **Statistical inference**

Statistical inference is the process of using sample statistic to make decisions about population.

Example: In the context of TRP

- Overall frequency of the various levels of happiness.
- Is there a relationship between the age of a viewers and his/her general happiness?
- Is there a relationship between the age of the viewer and the number of TV hours watched?

Data Summarization

- To identify the typical characteristics of data (i.e., to have an overall picture).
- To identify which data should be treated as noise or outliers.
- The data summarization techniques can be classified into two broad categories:
  - Measures of **location**
  - Measures of **dispersion**

Measurement of location

- It is also alternatively called as measuring the central tendency.
  - A function of the sample values that summarizes the location information into a single number is known as a measure of location.
- The most popular measures of location are
  - Mean
  - Median
  - Mode
  - Midrange
- These can be measured in three ways
  - Distributive measure
  - Algebraic measure
  - Holistic measure

Distributive measure

- It is a measure (*i.e. function*) that can be computed for a given data set by partitioning the data into smaller subsets, computing the measure for each subset, and then merging the results in order to arrive at the measure's value for the original (*i.e. entire*) data set.

Example

- ✓ sum(), count()

Algebraic measure

- It is a measure that can be computed by applying an algebraic function to one or more distributive measures.

- Example

$$\text{average} = \frac{\text{sum}()}{\text{count}()}$$

Holistic measure

- It is a measure that must be computed on the entire data set as a whole.

- Example

  Calculating median

  What about *mode?*

Mean of a sample

- The mean of a sample data is denoted as $\bar{x}$. Different mean measurements known are:
  - Simple mean
  - Weighted mean
  - Trimmed mean

- In the next few slides, we shall learn how to calculate the mean of a sample.

- We assume that given x_1, x_2, x_3,….., x_n are the sample values.

Simple mean of a sample

- **Simple mean**

  It is also called simply arithmetic mean or average and is abbreviated as (AM).

Definition : **Simple mean**

If $x_1, x_2, x_3, \ldots\ldots, x_n$ are the sample values, the simple mean is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Weighted mean of a sample

- **Weighted mean**

  It is also called weighted arithmetic mean or weighted average.

Definition : **Weighted mean**

When each sample value x_i is associated with a weight w_i, for $i = 1,2,\ldots,n$, then it is defined as

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Note: *When all weights are equal, the weighted mean reduces to simple mean.*

Trimmed mean of a sample

- **Trimmed Mean**

  If there are extreme values (*also called outlier*) in a sample, then the mean is influenced greatly by those values. To offset the effect caused by those extreme values, we can use the concept of trimmed mean

Definition : **Trimmed mean**

Trimmed mean is defined as the mean obtained after chopping off values at the high and low extremes.

Properties of mean

- **Lemma 1:**

  If $\bar{x}_i$, $i = 1,2,\ldots,m$ are the means of m samples of sizes $n_1, n_2, \ldots, n_m$ respectively, then the mean of the combined sample is given by:-

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{\sum_{i=1}^{m} n_i}$$

(Distributive Measure)

- **Lemma 2:**

  If a new observation x_k is added to a sample of size n with mean $\bar{x}$, the new mean is given by

$$\bar{x}' = \frac{n\,\bar{x} + x_k}{n+1}$$

Properties of mean

- **Lemma 3:**

  If an existing observation $\boldsymbol{x_k}$ is removed from a sample of size $\boldsymbol{n}$ with mean $\boldsymbol{\bar{x}}$, the new mean is given by

$$\boldsymbol{\bar{x}' = \frac{n\,\bar{x} - x_k}{n-1}}$$

- **Lemma 4:**

  If $\boldsymbol{m}$ observations with mean $\boldsymbol{\bar{x}_m}$, are added (*removed*) from a sample of size $\boldsymbol{n}$ with mean $\boldsymbol{\bar{x}_n}$, then the new mean is given by

$$\boldsymbol{\bar{x} = \frac{n\,\bar{x}_n \pm m\,\bar{x}_m}{n \pm m}}$$

Properties of mean

- **Lemma 5:**

  If a constant ***c*** is subtracted (*or added*) from each sample value, then the mean of the transformed variable is linearly displaced by *c*. That is,

$$\bar{x}' = \bar{x} \mp c$$

- **Lemma 6:**

  If each observation is called by multiplying (*dividing*) by a non-zero constant, then the altered mean is given by

$$\bar{x}' = \bar{x} * c$$

  *where, * is* ***x*** *(multiplication) or ÷ (division) operator.*

Mean with grouped data

Sometimes data is given in the form of classes and frequency for each class.

Class →	x_1- x_2	x_2- x_3	.....	x_i- x_{i+1}	.....	x_{n-1}- x_n
Frequency →	f_1	f_2	.....	f_i	.....	f_n

There three methods to calculate the mean of such a grouped data.

- Direct method
- Assumed mean method
- Step deviation method

Direct method

- **Direct Method**

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Where, $\mathbf{x_i} = \frac{1}{2}$ **(lower limit + upper limit)** of the i^{th} class, i.e., $x_i = \frac{x_i + x_{i+1}}{2}$ (also called class size), and f_i is the frequency of the i^{th} class.

Note: $\sum f_i (x_i - \bar{x}) = 0$

Assumed mean method

- **Assumed Mean Method**

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}$$

where, A is the assumed mean (it is usually a value $x_i = \frac{x_i + x_{i+1}}{2}$ chosen in the middle of the groups $d_i = (A - x_i)$ for each i)

Step deviation method

- **Step deviation method**

$$\bar{x} = A + \left\{ \frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} \; h \right\}$$

where,

A = assumed mean

h = class size (*i.e.*, $\mathbf{x_{i+1} - x_i}$ *for the* i^{th} *class*)

$u_i = \frac{x_i - A}{h}$

Mean for a group of data

- For the above methods, we can assume that…
  - All classes are equal sized
  - Groups are with inclusive classes, i.e., $\mathbf{x_i = x_{i-1}}$(*linear limit of a class is same as the upper limit of the previous class*)

10 - 19	20 - 29	30 - 39	40 – 49

Data with **exclusive** *classes*

9.5 – 19.5	19.5 – 29.5	29.5 – 39.5	39.5 – 49.5

Data with ***inclusive*** *classes*

Ogive: Graphical method to find mean

- **Ogive** (pronounced as O-Jive) is a cumulative frequency polygon graph.
  - When cumulative frequencies are plotted against the upper (lower) class limit, the plot resembles one side of an Arabesque or **ogival** architecture, hence the name.
  - There are two types of Ogive plots
    - Less-than (upper class vs. cumulative frequency)
    - More than (lower class vs. cumulative frequency)

Example:

Suppose, there is a data relating the marks obtained by 200 students in an examination

444, 412, 478, 467, 432, 450, 410, 465, 435, 454, 479, ...

(Further, suppose it is observed that the minimum and maximum marks are 410, 479, respectively.)

Ogive: Cumulative frequency table

444, 412, 478, 467, 432, 450, 410, 465, 435, 454, 479, ...

Step 1: Draw a cumulative frequency table

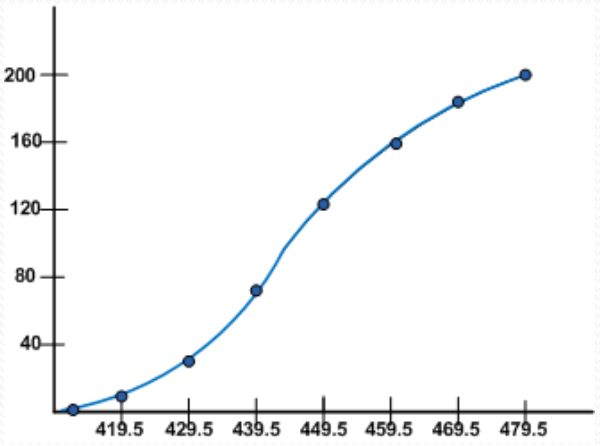
Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

Ogive: Graphical method to find mean

Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

Step 2: Less-than Ogive graph

Upper class	Cumulative Frequency
Less than 419.5	14
Less than 429.5	34
Less than 439.5	76
Less than 449.5	130
Less than 459.5	175
Less than 469.5	193
Less than 479.5	200

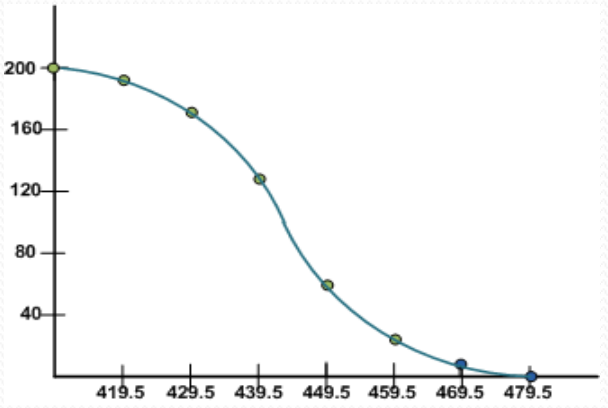


Ogive: Graphical method to find mean

Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

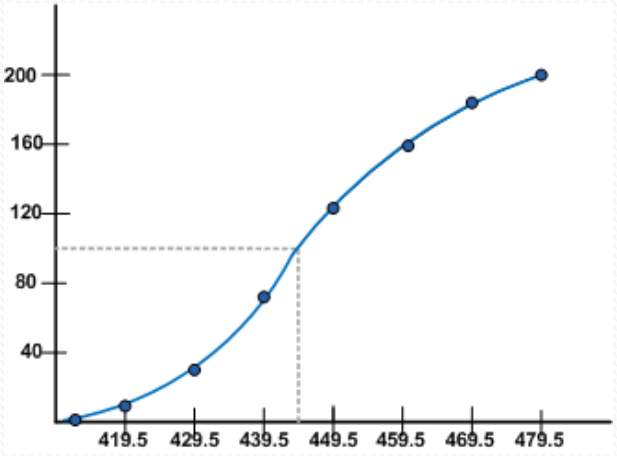
Step 3: More-than Ogive graph

Upper class	Cumulative Frequency
More than 409.5	200
More than 419.5	186
More than 429.5	166
More than 439.5	124
More than 449.5	70
More than 459.5	25
More than 469.5	7

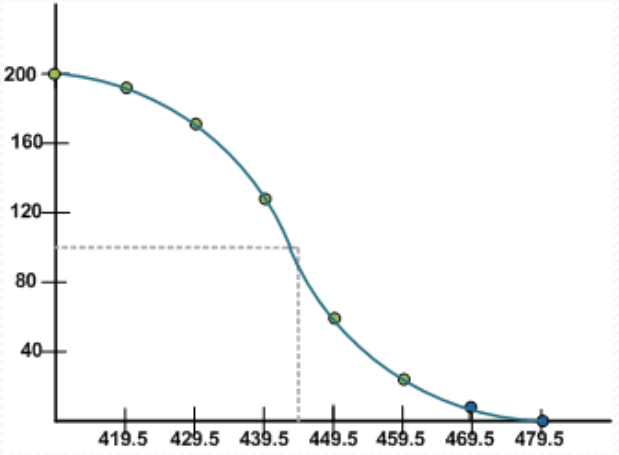


Information from Ogive

- Mean from Less-than Ogive



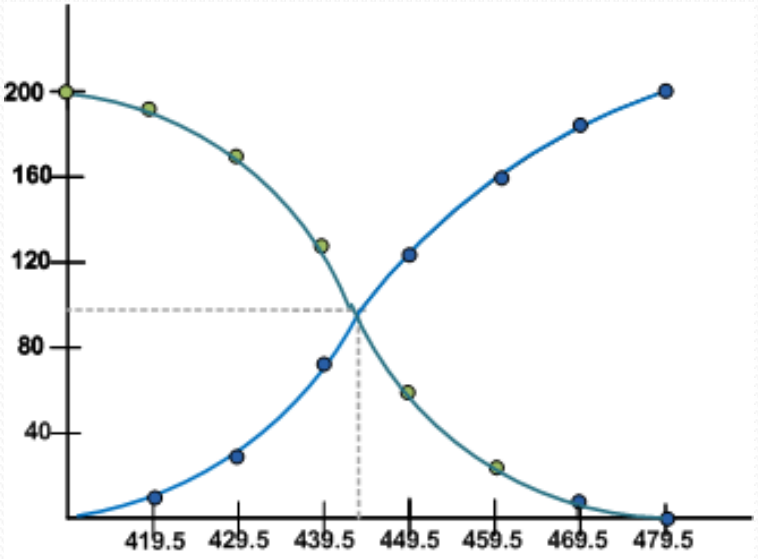
- Mean from More-than Ogive



- A % C frequency of .65 for the third class 439.5.....449.5 means that 65% of all scores are found in this class or below.

Information from Ogive

- Less-than and more-than Ogive approach



A cross point of two Ogive plots gives the mean of the sample.

Some other measures of mean

- Arithmetic Mean (**AM**)
  - $S: \{x_1, x_2\}$
  - $\bar{x} = \frac{x_1 + x_2}{2}$
  - $\bar{x} - x_1 = x_2 - \bar{x}$
- Geometric mean (**GM**)
  - $S: \{x_1, x_2\}$
  - $\tilde{x} = \sqrt{x_1 . x_2}$
  - $\frac{x_1}{\tilde{x}} = \frac{\tilde{x}}{x_2}$
- Harmonic Mean (**HM**)
  - $S: \{x_1, x_2\}$
  - $\hat{x} = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$
  - $\frac{2}{\hat{x}} = \frac{1}{x_1} + \frac{1}{x_2}$

Geometric mean

Definition : **Geometric mean**

Geometric mean of $\boldsymbol{n}$ observations (*none of which are zero*) is defined as:

$$\tilde{x} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

where, $n \neq 0$

Note

- GM is the arithmetic mean in "log space". This is because, alternatively,

$$log\tilde{x} = \frac{1}{n} \sum_{i=1}^{n} log\, x_i$$

- This summary of measurement is meaningful only when all observations are > 0.
- If at least one observation is zero, the product will itself be zero! For a negative value, root is not real

Harmonic mean

Definition : **Harmonic mean**

If all observations are non zero, the reciprocal of the arithmetic mean of the reciprocals of observations is known as harmonic mean.

For ungrouped data

$$\hat{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

For grouped data

$$\hat{x} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \left(\frac{f_i}{x_i}\right)}$$

where, f_i is the frequency of the i^{th} class with x_i as the center value of the i^{th} class.

Significant of different mean calculations

- There are two things involved when we consider a sample
  - Observation
  - Range

Example: Rainfall data

Rainfall (in mm)	r_1	r_2	...	r_n
Days (in number)	d_1	d_2	...	d_n

- Here, rainfall is the observation and day is the range for each element in the sample
- Here, we are to measure the mean "**rate of rainfall**" as the measure of location

Significant of different mean calculations

- **Case 1: Range remains same for each observation**

Example: Having data about amount of rainfall per week, say.

Rainfall (in mm)	35	18	…	22
Days (in number)	7	7	…	7

Significant of different mean calculations

- **Case 2: Ranges are different, but observation remains same**

Example: Same amount of rainfall in different number of days, say.

Rainfall (in mm)	**50**	**50**	**…**	**50**
Days (in number)	**1**	**2**	**…**	**7**

Significant of different mean calculations

- **Case 3: Ranges are different, as well as the observations**

Example: Different amount of rainfall in different number of days, say.

Rainfall (in mm)	21	34	…	18
Days (in number)	5	3	…	7

Rule of thumbs for means

- **AM:** When the range remains same for each observation

  Example: Case 1

Rainfall (in mm)	**35**	**18**	**…**	**22**
Days (in number)	7	7	…	7

$$\bar{r} = \frac{1}{n}\sum_{1}^{n} r_i$$

Rule of thumbs for means

- **HM:** When the range is different but each observation is same
  - Example: Case 2

Rainfall (in mm)	**50**	**50**	**…**	**50**
Days (in number)	1	2	…	7

$$\tilde{r} = \frac{n}{\sum_1^n \frac{1}{r_i}}$$

Rule of thumbs for means

- **GM:** When the ranges are different as well as the observations
  - Example: Case 3

Rainfall (in mm)	**21**	**34**	**…**	**18**
Days (in number)	5	3	…	7

$$\hat{r} = \left(\prod_{1}^{n} r_i \right)^{\frac{1}{n}}$$

Rule of thumbs for means

- The important things to recognize is that all three means are simply the arithmetic means in disguise!
- Each mean follows the "additive structure".
  - Suppose, we are given some abstract quantities $\{x_1, x_2, \ldots, x_n\}$
  - Each of the three means can be obtained with the following steps
    1. Transform each x_i into some y_i
    2. Taking the arithmetic mean of all y_i's
    3. Transforming back the to the original scale of measurement

Rule of thumbs for means

- For arithmetic mean
  - Use the **transformation** $y_i = x_i$
  - Take the arithmetic mean of all y_i *s* to get $\bar{y}$
  - Finally, $\bar{x} = \bar{y}$
- For geometric mean
  - Use the **transformation** $y_i = \log(x_i)$
  - Take the arithmetic mean of all y_i *s* to get $\bar{y}$
  - Finally, $\hat{x} = e^{\bar{y}}$
- For harmonic mean
  - Use the **transformation** $y_i = \frac{1}{x_i}$
  - Take the arithmetic mean of all y_i *s* to get $\bar{y}$
  - Finally, $\tilde{x} = \frac{1}{\bar{y}}$

AM ≥ GM ≥ HM

Median of a sample

Definition : **Median of a sample**

Median of a sample is the middle value when the data are arranged in increasing (*or decreasing*) order. Symbolically,

$$\hat{x} = \begin{cases} x_{(n+1)/2} & \textit{if } n \textit{ is odd} \\ \frac{1}{2}\left\{x_{n/2} + x_{(\frac{n}{2}+1)}\right\} & \textit{if } n \textit{ is even} \end{cases}$$

Median of a sample

Definition : **Median of a grouped data**

Median of a grouped data is given by

$$\hat{x} = l + \left\{ \frac{\frac{N}{2} - cf}{f} \; h \right\}$$

where h = width of the median class

$N = \sum_{i=1}^{n} f_i$

f_i is the frequency of the i^{th} class, and n is the total number of groups

cf = the cumulative frequency

N = the total number of samples

l = lower limit of the median class

Note

A class is called median class if its cumulative frequency is just greater than $N/2$

Mode of a sample

- Mode is defined as the observation which occurs most frequently.
- For example, number of wickets obtained by bowler in 10 test matches are as follows.

1 2 0 3 2 4 1 1 2 2

- In other words, the above data can be represented as:-

	0	1	2	3	4
# of matches	1	3	4	1	1

- Clearly, the mode here is “2”.

Mode of a grouped data

Definition : **Mode of a grouped data**

Select the modal class (it is the class with the highest frequency). Then the mode $\tilde{x}$ is given by:

$$\tilde{x} = l + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) h$$

where,

h is the class width

Δ_1 is the difference between the frequency of the modal class and the frequency of the class just after the modal class

Δ_2 is the difference between the frequency of the modal class and the class just before the modal class

l is the lower boundary of the modal class

Note

If each data value occurs only once, then there is no mode!

Relation between mean, median and mode

- There is an empirical relation, valid for moderately skewed data

$$\boldsymbol{Mean - Mode = 3 * (Mean - Median)}$$

- A given set of data can be categorized into three categories:-
  - Symmetric data
  - Positively skewed data
  - Negatively skewed data

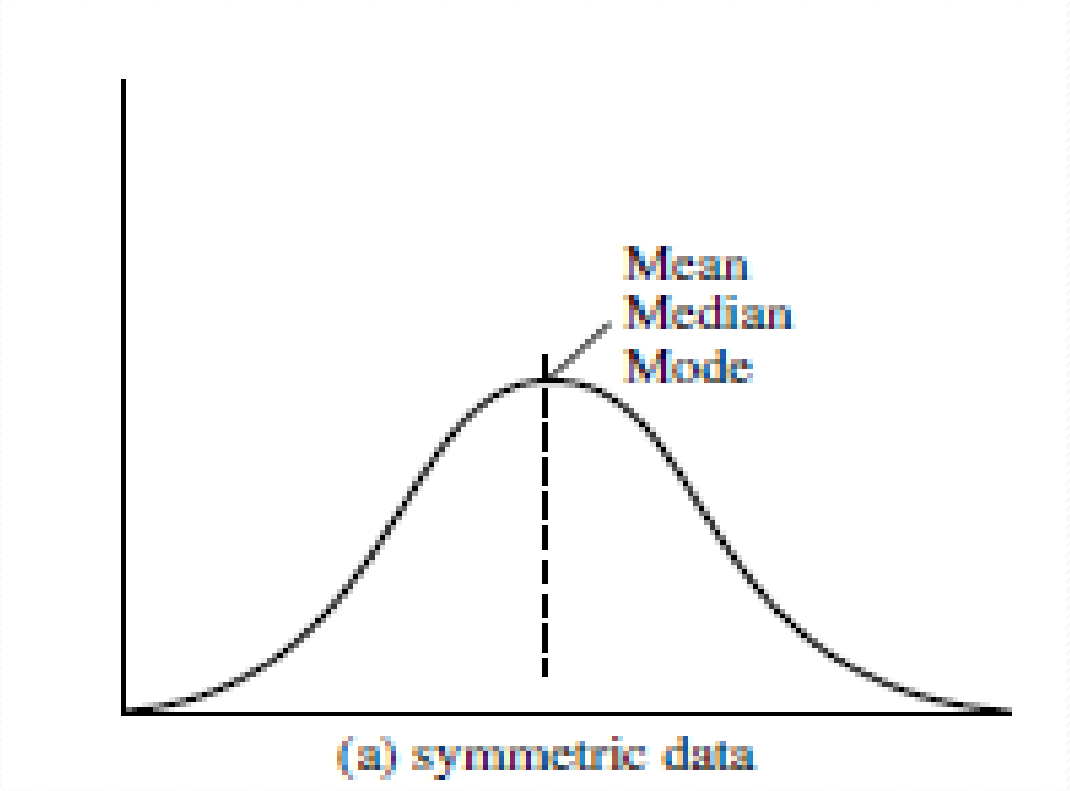
- To understand the above three categories, let us consider the following
- Given a set of $\boldsymbol{m}$ objects, where any object can take values $\boldsymbol{v_1, v_2, \ldots, v_k}$. Then, the frequency of a value $\boldsymbol{v_i}$ is defined as

$$\boldsymbol{Frequency(v_i) = \frac{Number\ of\ objects\ with\ value\ v_i}{n}}$$

$$for\ i = 1, 2, \ldots, k$$

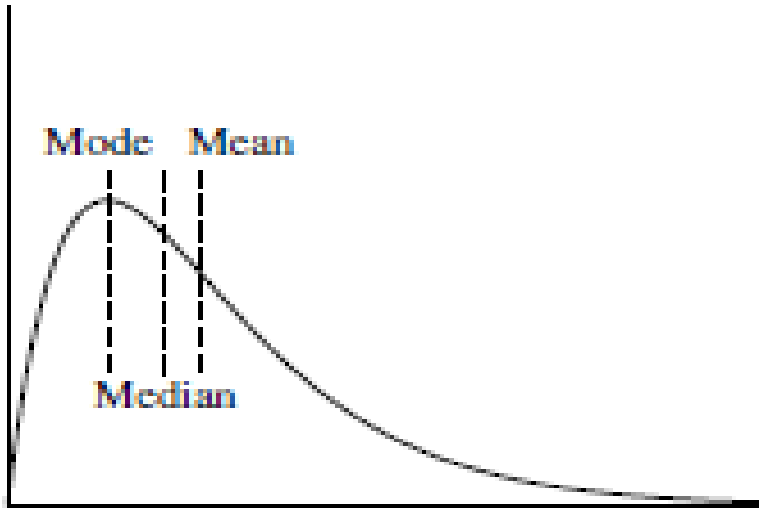
Symmetric data

- For symmetric data, all mean, median and mode lie at the same point

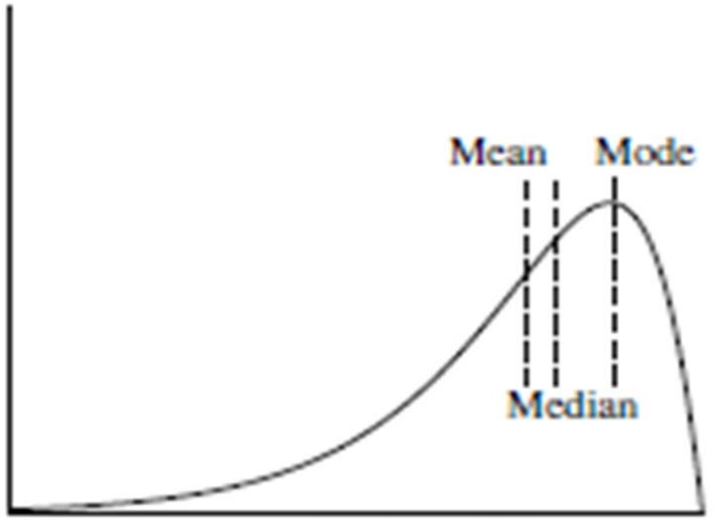


Positively & Negatively Skewed data

- Positively Skewed Data: Mode occurs at a value smaller than the median.
- Negatively Skewed Data: Mode occurs at a value greater than the median.



(b) positively skewed data



(c) negatively skewed data

Midrange

- It is the average of the largest and smallest values in the set.

- Steps

1. A percentage ‘p’ between 0 and 100 is specified.
2. The top and bottom of (p/2)% of the data is thrown out
3. The mean is then calculated in the normal way

- Thus, the median is trimmed mean with p = 100% while the traditional mean corresponds to p = 0%

Note

- Trimmed mean is a special case of Midrange.

Measures of dispersion

- Location measure are far too insufficient to understand data.
- Another set of commonly used summary statistics for continuous data are those that measure the dispersion.
- A dispersion measures the extent of spread of observations in a sample.

- Some important measure of dispersion are:
  - Range
  - Variance and Standard Deviation
  - Mean Absolute Deviation (MAD)
  - Absolute Average Deviation (AAD)
  - Interquartile Range (IQR)

Measures of dispersion

Example

- Suppose, two samples of fruit juice bottles from two companies ***A*** and ***B***. The unit in each bottle is measured in litre.

Sample A	0.97	1.00	0.94	1.03	1.06
Sample B	1.06	1.01	0.88	0.91	1.14

- Both samples have same mean. However, the bottles from company A with more uniform content than company B.
- We say that the dispersion (or variability) of the observation from the average is less for A than sample B.
  - The variability in a sample should display how the observation spread out from the average
  - In buying juice, customer should feel more confident to buy it from A than B

Range of a sample

Definition : **Range of a sample**

Let $\mathbf{X = x_1, \ldots, x_n}$ be **n** sample values that are arranged in increasing order.

The range **R** of these samples are then defined as:

$$\mathbf{R = max(X) - min(X) = x_n - x_1}$$

- Range identifies the maximum spread, it can be misleading if most of the values are concentrated in a narrow band of values, but there are also a relatively small number of more extreme values.
- The variance is another measure of dispersion to deal with such a situation.

Variance and Standard Deviation

Definition : **Variance and Standard Deviation**

Let $\mathbf{X = \{x_1, \ldots, x_n\}}$ are sample values of **n** samples. Then, variance denoted as $\boldsymbol{\sigma^2}$ is defined as :-

$$\boldsymbol{\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where, $\bar{x}$ denotes the mean of the sample

The standard deviation, **σ**, of the samples is the square root of the variance $\boldsymbol{\sigma^2}$

Coefficient of Variation

- **Basic properties**
  - **σ** measures spread about mean and should be chosen only when the mean is chosen as the measure of central tendency
  - **σ = 0** only when there is no spread, that is, when all observations have the same value, otherwise **σ > 0**

Definition : **Coefficient of variation**

A related measure is the coefficient of variation **CV**, which is defined as follows

$$\mathbf{CV} = \frac{\boldsymbol{\sigma}}{\bar{\mathbf{x}}} \times \mathbf{100}$$

This gives a ratio measure to spread.

Mean Absolute Deviation (MAD)

- Since, the mean can be distorted by outlier, and as the variance is computed using the mean, it is thus sensitive to outlier. To avoid the effect of outlier, there are two more robust measures of dispersion known. These are:

  - Mean Absolute Deviation (MAD)

$$\textbf{MAD (X) = median } (\{|x_1 - \bar{x}|, \ldots\ldots, |x_n - \bar{x}|\})$$

  - Absolute Average Deviation (AAD)

$$\textbf{AAD(X)} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

where, X = $\{x_1, x_2, \ldots\ldots, x_n\}$ *is the sample values of n observations*

Interquartile Range

- Like MAD and AAD, there is another robust measure of dispersion known, called as Interquartile range, denoted as IQR
- To understand IQR, let us first define *percentile* and *quartile*
- **Percentile**
  - The percentile of a set of ordered data can be defined as follows:
    - Given an ordinal or continuous attribute **x** and a number **p** between 0 and 100, the $\mathbf{p^{th}}$ percentile $\mathbf{x_p}$ is a value of **x** such that **p**% of the observed values of **x** are less than $\mathbf{x_p}$
    - Example: The $\mathbf{50^{th}}$ percentile is that value $\mathbf{x_{50\%}}$ such that **50%** of all values of **x** are less than $\mathbf{x_{50\%}}$.
  - **Note:** The median is the $\mathbf{50^{th}}$ percentile.

Interquartile Range

- **Quartile**
  - The most commonly used percentiles are quartiles.
    - The first quartile, denoted by $\mathbf{Q_1}$ is the $\mathbf{25^{th}}$ percentile.
    - The third quartile, denoted by $\mathbf{Q_3}$ is the $\mathbf{75^{th}}$ percentile
    - The median, $\mathbf{Q_2}$ is the $\mathbf{50^{th}}$ percentile.

- The quartiles including median, give some indication of the center, spread and shape of a distribution.

- The distance between $\mathbf{Q_1}$ and $\mathbf{Q_3}$ is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (**IQR**) and is defined as

$$\mathbf{IQR = Q_3 - Q_1}$$

Application of IQR

- **Outlier detection using five-number summary**
  - A common rule of the thumb for identifying suspected outliers is to single out values falling at least **1.5 × IQR** above $\mathbf{Q_3}$ and below $\mathbf{Q_1}$.
  - In other words, extreme observations occurring within **1.5 × IQR** of the quartiles

Application of IQR

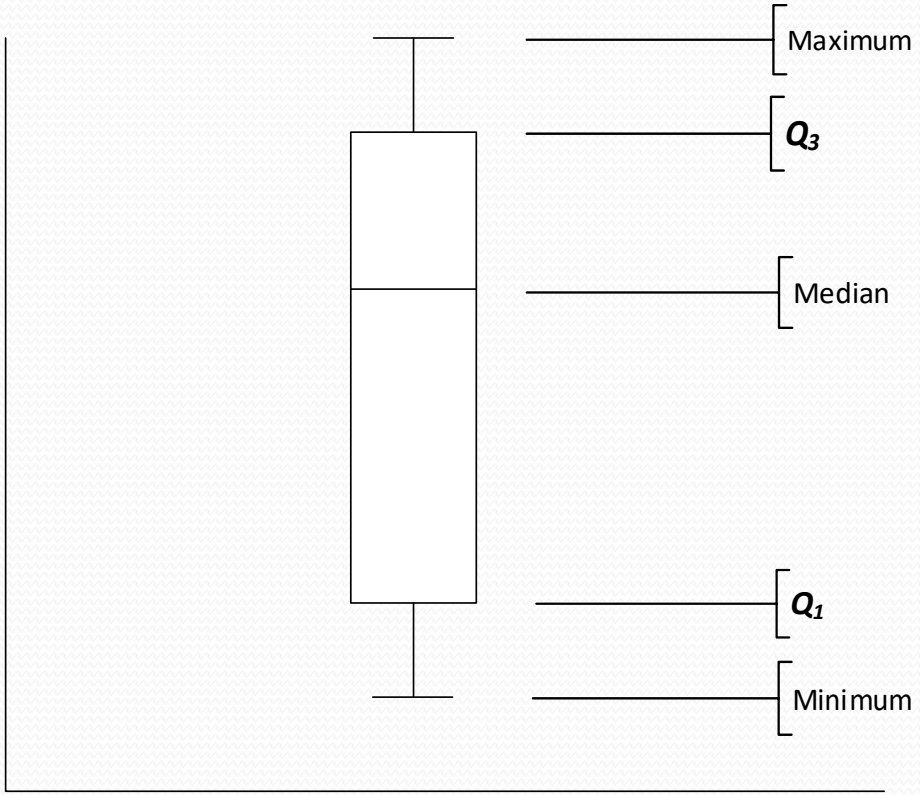
- **Five Number Summary**
  - Since, $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ together contain no information about the endpoints of the data, a complete summary of the shape of a distribution can be obtained by providing the lowest and highest data value as well. This is known as the five-number summary
  - The five-number summary of a distribution consists of :
    - The Median $\mathbf{Q_2}$
    - The first quartile $\mathbf{Q_1}$
    - The third quartile $\mathbf{Q_3}$
    - The smallest observation
    - The largest observation

These are, when written in order gives the **five-number summary**:

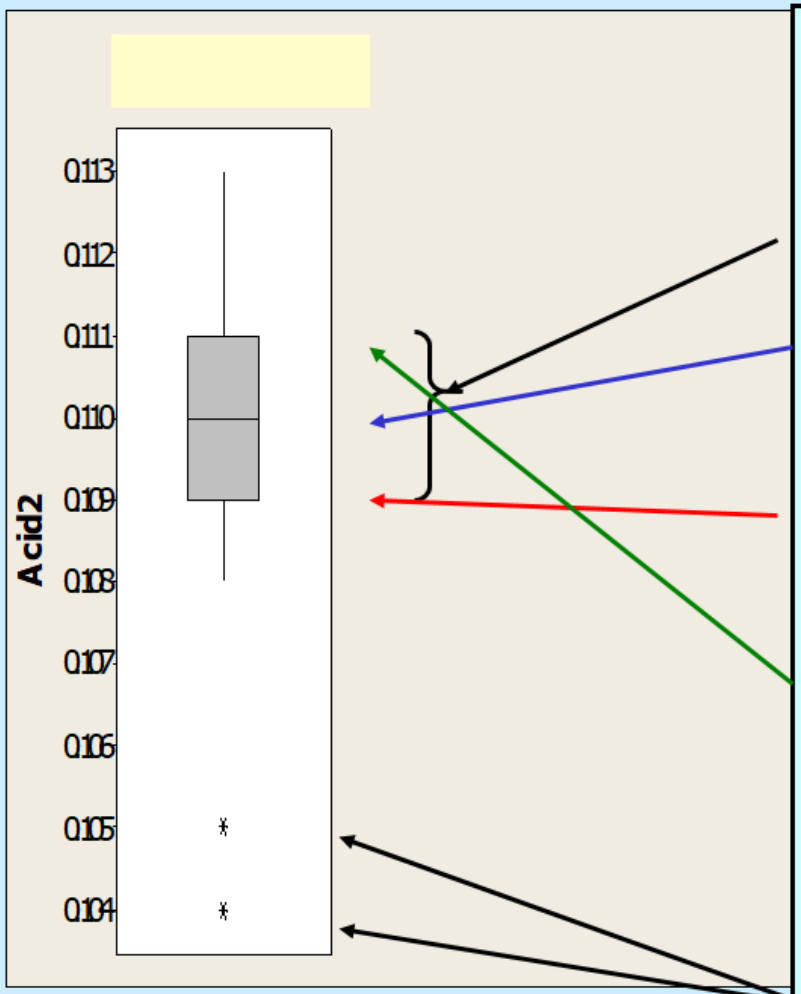
Minimum, $\mathbf{Q_1}$, Median ($\mathbf{Q_2}$), $\mathbf{Q_3}$, Maximum

Box plot

- **Graphical view of Five number summary**



Box plot



A box and whisker plot provides a 5 point summary of the data.

1)The box represents the middle 50% of the data.

2) The median is the point where 50% of the data is above it and 50% below it.

3) The 1st quartile is where, 25% of the data fall below it.

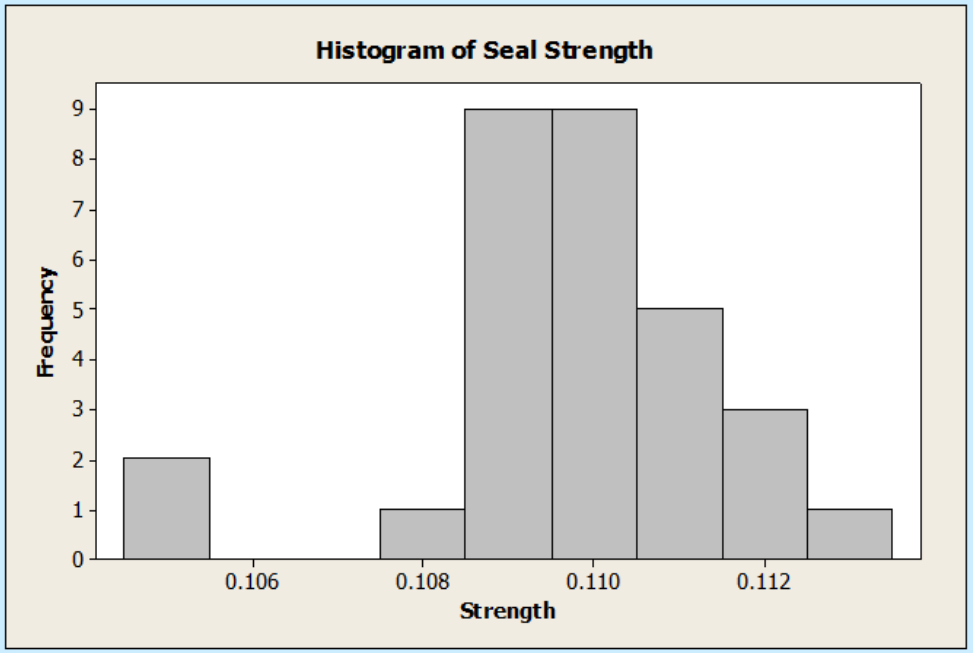
4) The 3rd quartile is where, 75% of the data is below it.

5) The whiskers cannot extend any further than 1.5 times the length of the inner quartiles.

If you have data points outside this, they will show up as outliers.

Histogram

Histogram is a basic graphing tool that displays the relative frequency or occurrence of continuous data values showing which values occur most and least frequently.



A histogram illustrates the

Shape,
Centering, and
Spread

of data distribution

and indicates whether there are any outliers.

Probability and Statistics

Probability is the chance of an outcome in an experiment (also called event).

Event: Tossing a fair coin
Outcome: Head, Tail

Probability deals with **predicting** the likelihood of **future** events.	Statistics involves the **analysis** of the **frequency** of **past** events

Example: Consider there is a drawer containing 100 socks: 30 red, 20 blue and 50 black socks.

We can use probability to answer questions about the selection of a random sample of these socks.

- **PQ1.** What is the probability that we draw two blue socks or two red socks from the drawer?
- **PQ2.** What is the probability that we pull out three socks or have matching pair?
- **PQ3.** What is the probability that we draw five socks and they are all black?

Statistics

Instead, if we have no knowledge about the type of socks in the drawers, then we enter into the realm of statistics. Statistics helps us to infer properties about the population on the basis of the random sample.

Questions that would be statistical in nature are:

- **Q1**: A random sample of 10 socks from the drawer produced one blue, four red, five black socks. What is the total population of black, blue or red socks in the drawer?
- **Q2**: We randomly sample 10 socks, and write down the number of black socks and then return the socks to the drawer. The process is done for five times. The mean number of socks for each of these trial is 7. What is the true number of black socks in the drawer?
- etc.

Probability vs. Statistics

In other words:

- In probability, we are given a model and asked what kind of data we are likely to see.
- In statistics, we are given data and asked what kind of model is likely to have generated it.

Example: Measles Study

- A study on health is concerned with the incidence of childhood measles in parents of childbearing age in a city. For each couple, we would like to know how likely, it is that either the mother or father or both have had childhood measles.
- The current census data indicates that 20% adults between the ages 17 and 35 (regardless of sex) have had childhood measles.
  - This give us the probability that an individual in the city has had childhood measles.

Defining Random Variable

Definition: **Random Variable**

A random variable is a rule that assigns a numerical value to an outcome of interest.

Example : In "measles Study", we define a random variable X as the number of parents in a married couple who have had childhood measles.

This random variable can take values of $0, 1$ *and* 2.

Note:

- Random variable is not exactly the same as the variable defining a data.
- The probability that the random variable takes a given value can be computed using the rules governing probability.
  - For example, the probability that $X = 1$ means either mother or father but not both has had measles is 0.32. Symbolically, it is denoted as **P(X=1) = 0.32.**

Probability Distribution

Definition : **Probability distribution**

A probability distribution is a definition of probabilities of the values of random variable.

Example : Given that 0.2 is the probability that a person (in the ages between 17 and 35) has had childhood measles. Then the probability distribution is given by

X	Probability
0	0.64
1	0.32
2	0.04

?

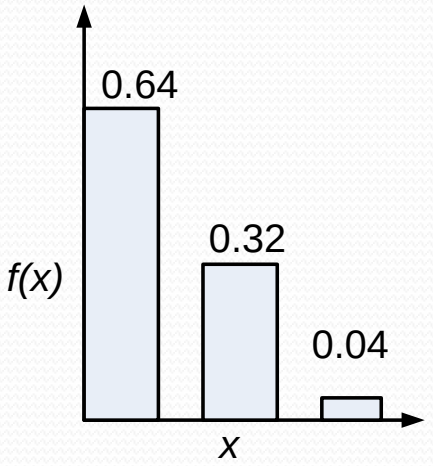
Probability Distribution

- In data analytics, the probability distribution is important with which many statistics making inferences about population can be derived .
  - In general, a probability distribution function takes the following form

x	x_1	$x_2 \dots \dots \dots \dots \dots x_n$
$f(x) = P(X = x)$	$f(x_1)$	$f(x_2) \dots \dots \dots f(x_n)$

Example: Measles Study

x	0	1	2
$f(x)$	0.64	0.32	0.04



Usage of Probability Distribution

- Distribution (discrete/continuous) function is widely used in simulation studies.
  - A simulation study uses a computer to simulate a real phenomenon or process as closely as possible.
  - The use of simulation studies can often eliminate the need of costly experiments and is also often used to study problems where actual experimentation is impossible.

Examples :

1) A study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use such a drug approximately follows a binomial distribution.

2) Operation of ticketing system in a busy public establishment (e.g., airport), the arrival of passengers can be simulated using Poisson distribution.

Binomial Distribution

- In many situations, an outcome has only two outcomes: success and failure.
  - Such outcome is called dichotomous outcome.
- An experiment when consists of repeated trials, each with dichotomous outcome is called Bernoulli process. Each trial in it is called a Bernoulli trial.

Example : Firing bullets to hit a target.

- Suppose, in a Bernoulli process, we define a random variable $X \equiv$ the number of successes in trials.
- Such a random variable obeys the binomial probability distribution, if the experiment satisfies the following conditions:

1) The experiment consists of n trials.
2) Each trial results in one of two mutually exclusive outcomes, one labelled a *"success"* and the other a *"failure"*.
3) The probability of a success on a single trial is equal to $\boldsymbol{p}$. The value of p remains constant throughout the experiment.
4) The trials are independent.

Defining Binomial Distribution

Definition: **Binomial distribution**

The function for computing the probability for the binomial probability distribution is given by

$$f(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

$$for\ x = 0, 1, 2, \ldots., n$$

Here, $f(x) = P(X = x)$, where X denotes "the number of success" and $X = x$ denotes the number of success in x trials.

Binomial Distribution

Example : Measles study

X = having had childhood measles a success

p = 0.2, the probability that a parent had childhood measles

n = 2, here a couple is an experiment and an individual a trial, and the number of trials is two.

Thus,

$$P(x = 0) = \frac{2!}{0!(2-0)!}(0.2)^0(0.8)^{2-0} \quad = \mathbf{0.64}$$

$$P(x = 1) = \frac{2!}{1!\,(2-1)!}(0.2)^1(0.8)^{2-1} = \mathbf{0.32}$$

$$P(x = 2) = \frac{2!}{2!\,(2-2)!}(0.2)^2(0.8)^{2-2} = \mathbf{0.04}$$

Binomial Distribution

Example : Verify with real-life experiment

Suppose, 10 pairs of random numbers are generated by a computer (Monte-Carlo method)

15 38 68 39 49 54 19 79 38 14

If the value of the digit is 0 or 1, the outcome is “had childhood measles”, otherwise, (digits 2 to 9), the outcome is “did not”.

For example, in the first pair (i.e., 15), representing a couple and for this couple, $x = 1$. The frequency distribution, for this sample is

x	0	1	2
$f(x)=P(X=x)$	0.7	0.3	0.0

Note: This has close similarity with binomial probability distribution!

The Multinomial Distribution

The binomial experiment becomes a multinomial experiment, if we let each trial has more than two possible outcome.

Definition: **Multinomial distribution**

If a given trial can result in the ***k*** outcomes $E_1, E_2, \dots\dots, E_k$ with probabilities $p_1, p_2, \dots\dots, p_k$, then the probability distribution of the random variables $X_1, X_2, \dots\dots, X_k$ representing the number of occurrences for $E_1, E_2, \dots\dots, E_k$ in n independent trials is

$$f(x_1, x_2, \dots\dots, x_k) = \binom{n}{x_1, x_2, \dots\dots, x_k} {p_1}^{x_1} {p_2}^{x_2} \dots\dots {p_k}^{x_k}$$

where $\binom{n}{x_1, x_2, \dots\dots, x_k} = \frac{n!}{x_1! x_2! \dots\dots x_k!}$

$$\sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} p_i = 1$$

The Poisson Distribution

There are some experiments, which involve the occurring of the number of outcomes during a given time interval (or in a region of space).

Such a process is called Poisson process.

Example :

Number of clients visiting a ticket selling counter in a metro station.

The Poisson Distribution

Properties of Poisson process

- The number of outcomes in one time interval is independent of the number that occurs in any other disjoint interval [Poisson process has no memory]
- The probability that a single outcome will occur during a very short interval is proportional to the length of the time interval and does not depend on the number of outcomes occurring outside this time interval.
- The probability that more than one outcome will occur in such a short time interval is negligible.

Definition : **Poisson distribution**

The probability distribution of the Poisson random variable X, representing the number of outcomes occurring in a given time interval t, is

$$f(x, \lambda t) = P(X = x) = \frac{e^{-\lambda t}.(\lambda t)^x}{x!}, x = 0, 1, \dots \dots$$

where λ is the average number of outcomes per unit time and $e = 2.71828 \dots$

Descriptive measures

Given a random variable X in an experiment, we have denoted $f(x) = P(X = x)$, the probability that $X = x$. For discrete events $f(x) = 0$ for all values of x except $x = 0, 1, 2, \ldots\ldots$

Properties of discrete probability distribution

1. $0 \leq f(x) \leq 1$
2. $\sum f(x) = 1$
3. $\mu = \sum x.f(x)$ [is the mean]
4. $\sigma^2 = \sum(x - \mu)^2 . f(x)$ [is the variance]

In $2, 3$ *and* 4, summation is extended for all possible discrete values of x.

Note: For discrete uniform distribution, $f(x) = \frac{1}{n}$ with $x = 1, 2, \ldots\ldots, n$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

Descriptive measures

Binomial distribution

The binomial probability distribution is characterized with p (the probability of success) and n (is the number of trials). Then

$$\mu = n.p$$

$$\sigma^2 = np(1-p)$$

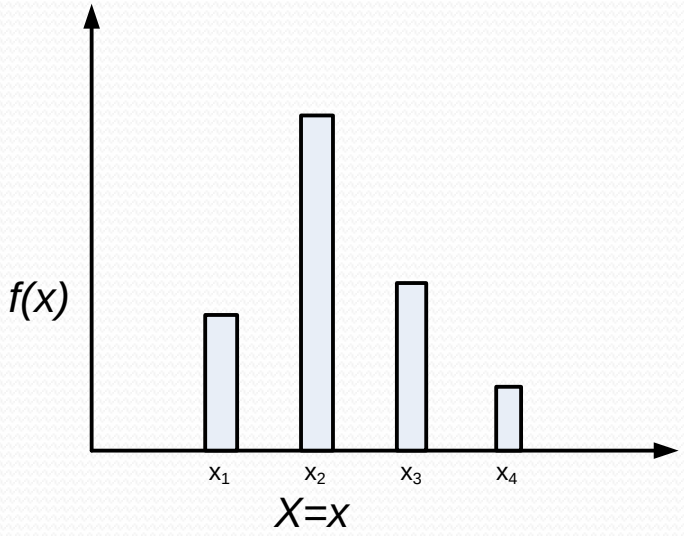
Poisson Distribution

The Poisson distribution is characterized with λ where $\lambda = the\ mean\ of\ outcomes$ and $t = time\ interval.$

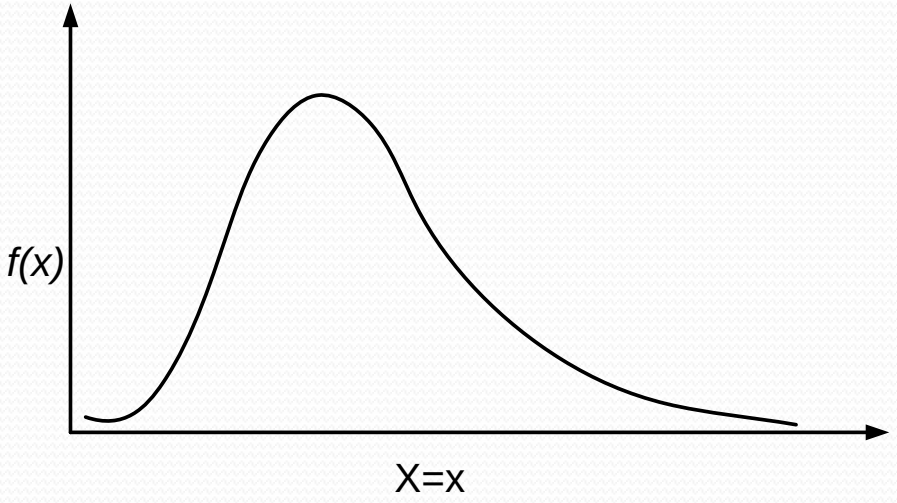
$$\mu = \lambda t$$

$$\sigma^2 = \lambda t$$

Discrete Vs. Continuous Probability Distributions



Discrete Probability distribution



Continuous Probability Distribution

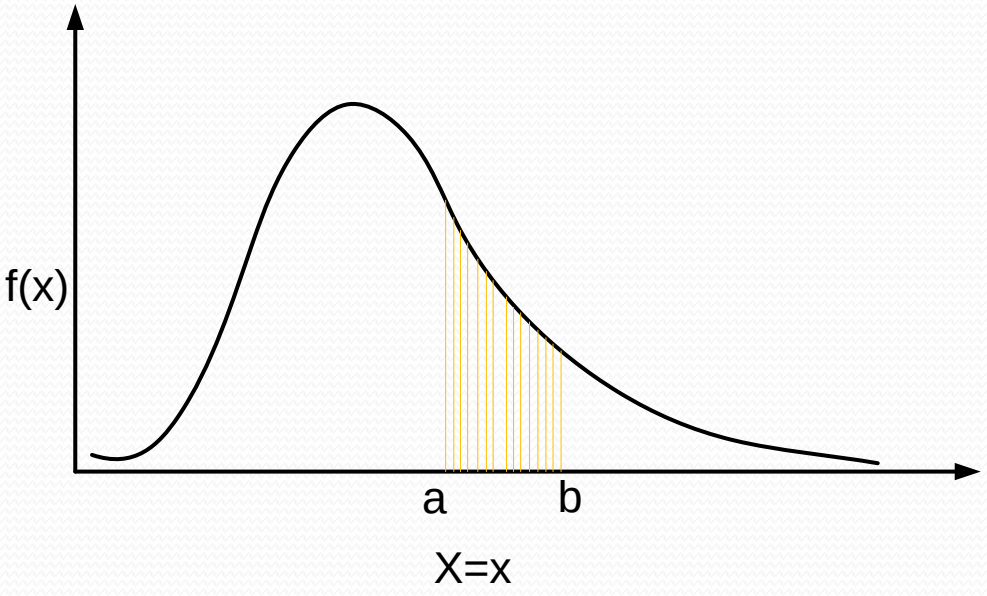
Continuous Probability Distributions

- When the random variable of interest can take any value in an interval, it is called continuous random variable.
  - Every continuous random variable has an infinite, uncountable number of possible values (i.e., any value in an interval)
- Consequently, continuous random variable differs from discrete random variable.

Properties of Probability Density Function

The function $f(x)$ is a probability density function for the continuous random variable X, defined over the set of real numbers R, if

1. $f(x) \geq 0$, for all $x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$
4. $\mu = \int_{-\infty}^{\infty} xf(x)\,dx$
5. $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$



Continuous Uniform Distribution

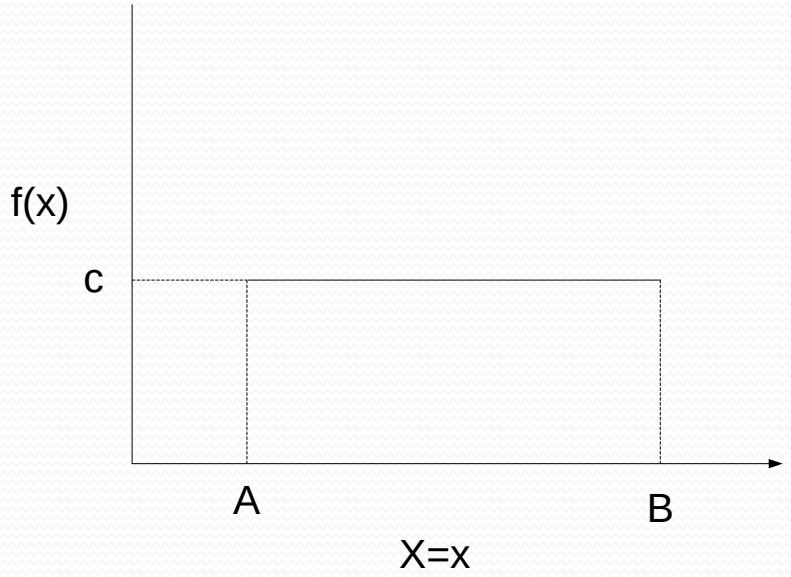
- One of the simplest continuous distribution in all of statistics is the continuous uniform distribution.

Definition : **Continuous Uniform Distribution**

The density function of the continuous uniform random variable X on the interval $[A, B]$ is:

$$f(x: A, B) = \begin{cases} \dfrac{1}{B - A} & A \leq x \leq B \\ 0 & Otherwise \end{cases}$$

Continuous Uniform Distribution



Note:

a) $\int_{\infty}^{-\infty} f(x)dx = \frac{1}{B-A} \times (B - A) = 1$

b) $P(c < x < d) = \frac{d-c}{B-A}$ where both c and d are in the interval (A,B)

c) $\mu = \frac{A+B}{2}$

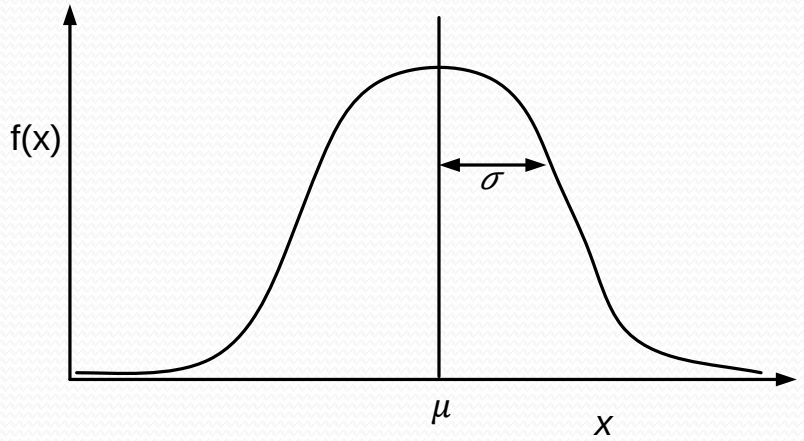
d) $\sigma^2 = \frac{(B-A)^2}{12}$

Normal Distribution

- The most often used continuous probability distribution is the normal distribution; it is also known as **Gaussian** distribution.
- Its graph called the normal curve is the bell-shaped curve.
- Such a curve approximately describes many phenomenon occur in nature, industry and research.
  - Physical measurement in areas such as meteorological experiments, rainfall studies and measurement of manufacturing parts are often more than adequately explained with normal distribution.
- A continuous random variable X having the bell-shaped distribution is called a normal random variable.

Normal Distribution

- The mathematical equation for the probability distribution of the normal variable depends upon the two parameters μ and σ, its mean and standard deviation.



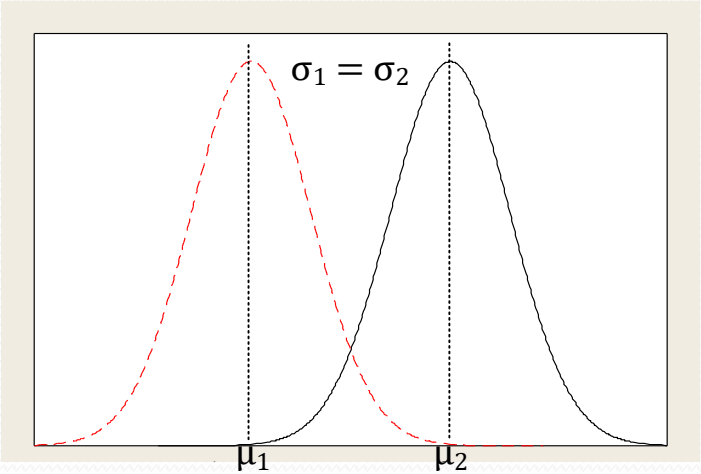
Definition : **Normal distribution**

The density of the normal variable x with mean μ and variance σ^2 is

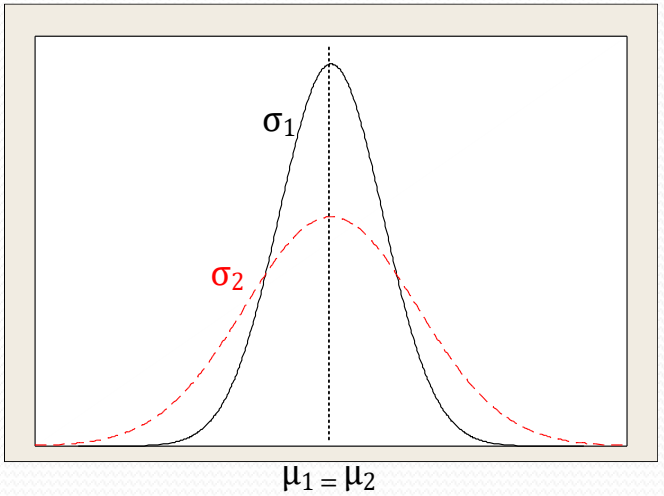
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

where $\pi = 3.14159$ ... and $e = 2.71828$ ….., the Naperian constant

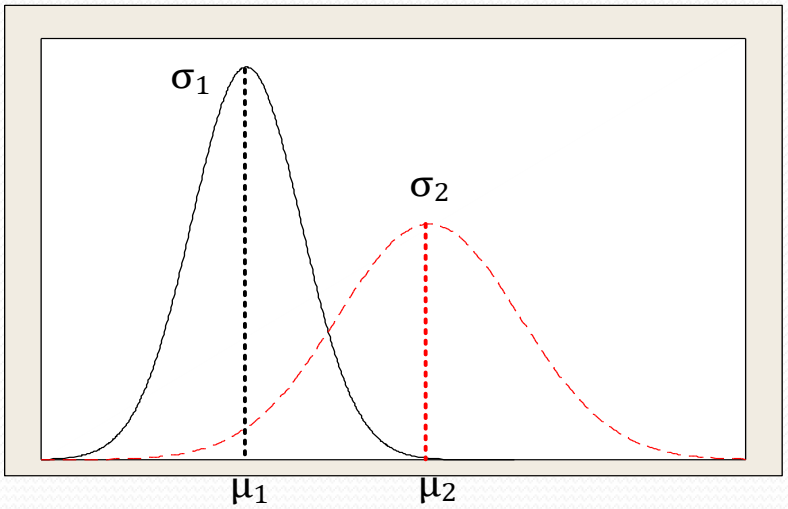
Normal Distribution



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$

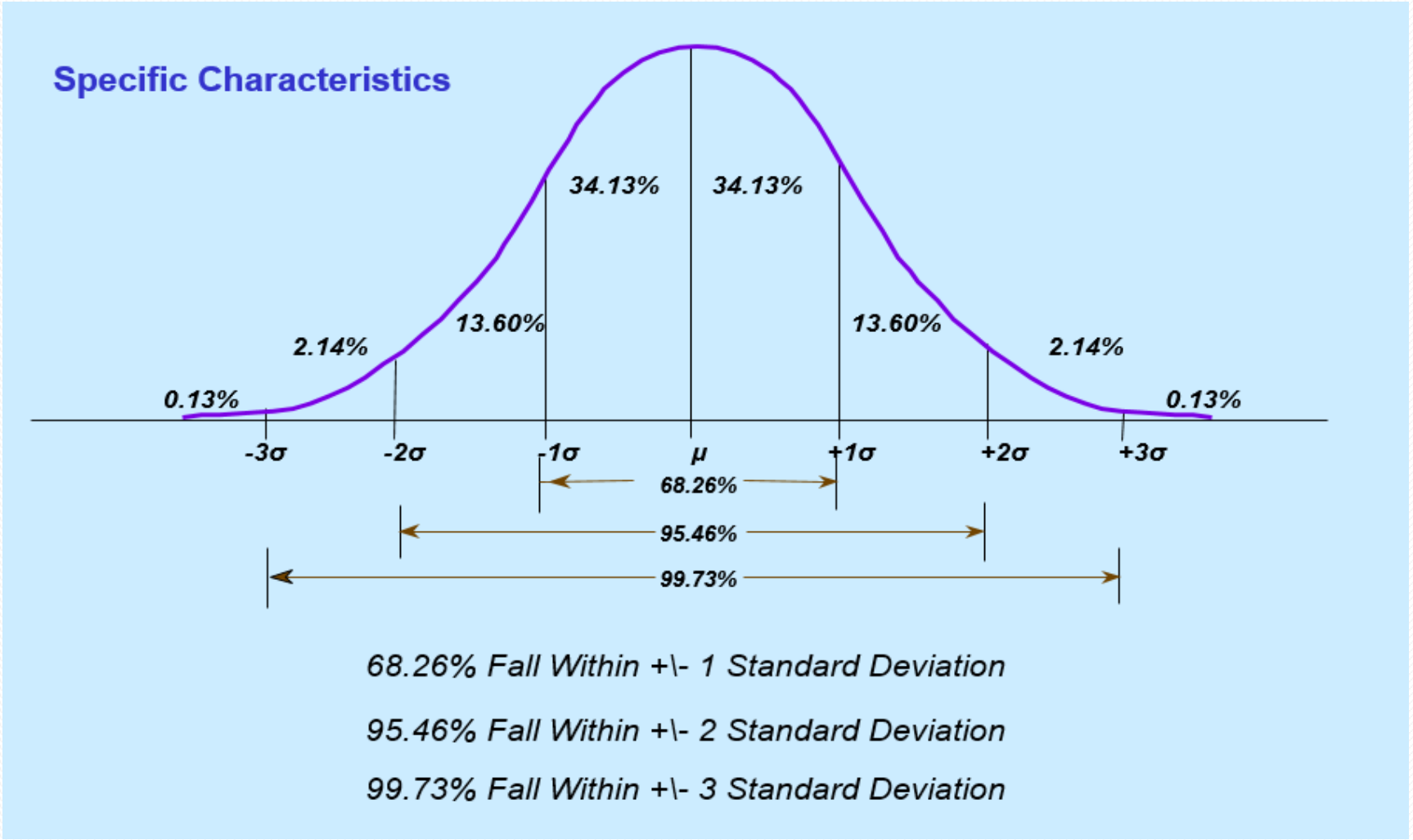


Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$

Normal Curve (6-sigma)



68.26% Fall Within +\- 1 Standard Deviation

95.46% Fall Within +\- 2 Standard Deviation

99.73% Fall Within +\- 3 Standard Deviation

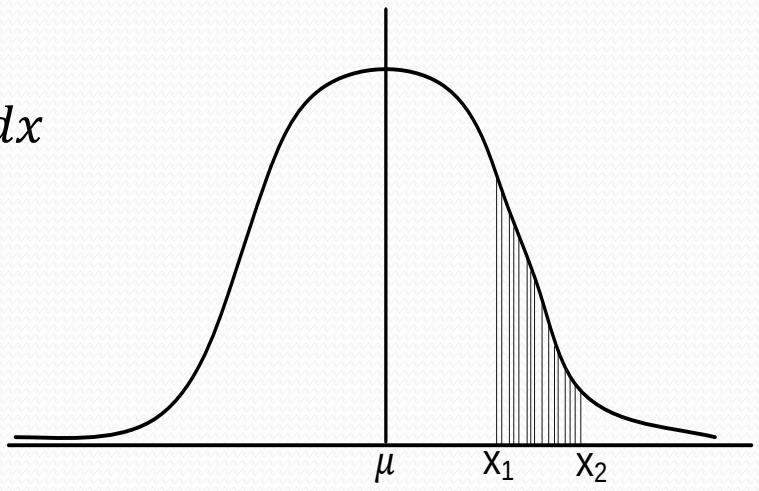
Properties of Normal Distribution

- The curve is symmetric about a vertical axis through the mean μ.
- The random variable x can take any value from $-\infty \; to \; \infty$.
- The most frequently used descriptive parameter s define the curve itself.
- The mode, which is the point on the horizontal axis where the curve is a maximum occurs at $x = \mu$.
- The total area under the curve and above the horizontal axis is equal to 1.

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 1$$

- $\mu = \int_{-\infty}^{\infty} x.f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x.e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$
- $\sigma^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2.e^{-\frac{1}{2}[(x-\mu)/\sigma^2]} dx$
- $P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

denotes the probability of x in the interval (x_1, x_2).



Standard Normal Distribution

- The normal distribution has computational complexity to calculate $P(x_1 < x < x_2)$ for any two (x_1, x_2) and given μ and σ
- To avoid this difficulty, the concept of z-transformation is followed.

$$z = \frac{x-\mu}{\sigma} \qquad \text{[Z-transformation]}$$

- X: Normal distribution with mean μ and variance σ^2.
- Z: Standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.
- Therefore, if *f(x)* assumes a value, then the corresponding value of $f(z)$ is given by

$$f(x:\mu,\sigma): \; P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

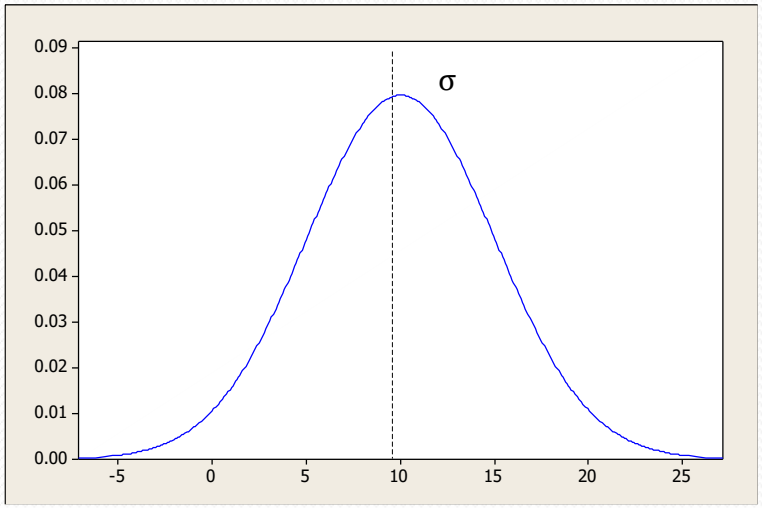
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz$$

$$= f(z:0,\sigma)$$

Standard Normal Distribution

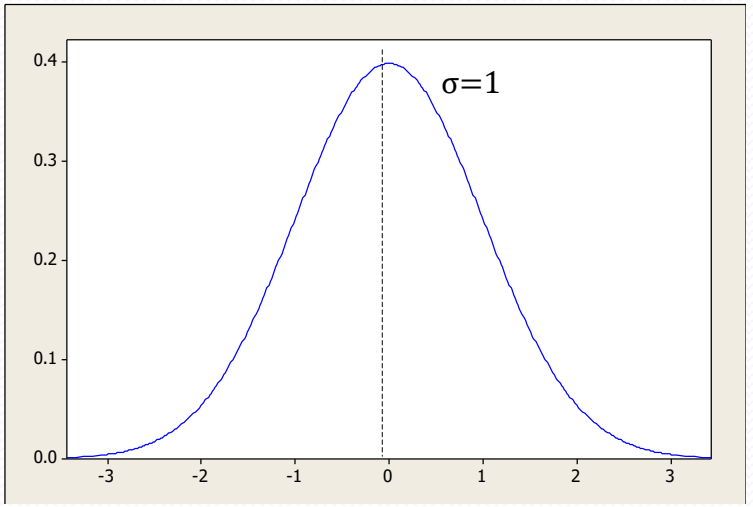
Definition : **Standard normal distribution**

The distribution of a normal random variable with mean 0 and variance 1 is called a standard normal distribution.



x=μ

f(x: μ, σ)



μ=0

f(z: 0, 1)

Reference Book

- Probability and Statistics for Engineers and Scientists (8th Ed.) by Ronald E. Walpole, Sharon L. Myers, Keying Ye (Pearson), 2013.

Any question?

You may also send your question(s) at ctanujit@gmail.com