5/1/2021

* Random Experiment:

→ An experiment in which all possible outcomes are known in advance but we don't know which possible outcome occurs.

Eg 1: Tossing a coin

$\{H, T\}$

2: Throwing a die

$\{1, 2, 3, 4, 5, 6\}$

3: Drawing a card from pack of cards

52 cards

* Sample Space

→ Set of all possible outcomes of a Random Experiment.

Eg: 1 Tossing a coin

$S = \{H, T\}$

2: Tossing a pair of coins

$S = \{HH, HT, TH, TT\}$

note:

→ If we toss 'n' coins number of possible outcomes $= 2^n$

3. Tossing 3 coins

$2^3 = 8$ Possibilities

$2^3 = 8$

$\frac{8}{2} = 4$

$\frac{4}{2} = 2$

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

H H H ✓ 1st Sample point 2
H H T
H T H
H T T
T H H
T H T
T T H
T T T — 8th

∴ Total 8 possibilities.

8 possible outcomes.

4) 4 coins

$2^4 = \frac{16}{2} = 8H$ $8T$

$\frac{8}{2} = 4H$ $4T$

$\frac{4}{2} = 2H$ $2T$

$\frac{2}{2} = 1H$ $1T$

H H H H ✓ 1st Possibility
H H H T
H H T H
H H T T
H T H H
H T H T
H T T H
H T T T
T H H H
T H H T
T H T H
T H T T
T T H H
T T H T
T T T H
T T T T ✓ 16th

$S = \{HHHH, \ldots\ldots, TTTT\}$

II) Throwing a dice → Random experiment

$S = \{1, 2, 3, 4, 5, 6\}$

Throwing a pair of dice

$S = \{(1,1), (1,2), (1,3) \ldots\ldots (1,6)$

$(2,1), (2,2) \ldots\ldots\ldots\ldots$

$\ldots\ldots (6,6)\} \rightarrow$ 36 Possibilities.

note:

→ If I throw n dice then number of possible outcomes 6^n

(1) A class consists of 6 girls and 10 boys. If a committee of 3 is choosen at random from the class. find the probability that

(i) 3 boys are Selected

(ii) Exactly 2 girls are Selected.

Sol: Total number of Students = 16

6 G, 10 B

(i) Prob: → $\frac{\text{Number of favourable cases}}{\text{Total number of cases}}$

$\frac{{}^{10}C_3}{{}^{16}C_3}$ → ${}^nC_n = \frac{n!}{(n-n)!\,n!}$

$\frac{10 \times 9^3 \times 8^4}{1 \times 2 \times 3} = 120$

Selection

↓

Combination → Eg: Cricket

arrangement

↓

permutation → Batting order

$^{16}C_3 \rightarrow \frac{16 \times 15 \times 14}{1 \times 2 \times 3}$

(ii) Exactly 2 girls

$$\frac{^6C_2 \times {}^{10}C_1}{^{16}C_3}$$

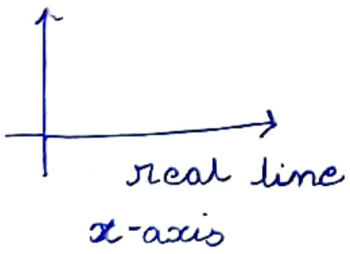
3 → Committee

$[\because {}^nC_1 \rightarrow n]$

7/1/2021

* Random Variable:

✓ A real Variable X whose value is determined by the output of a random Experiment is called a random Variable.



Eg:
0, 10, 10.5, 9.6 real number

Types of Random Variable

a) Discrete
b) Continous

} Random Variable

* Discrete RV

Eg: 1, 2, 3, 4, ...

* number of Students in a class

* Continous Random Variable:

Eg: 1.1, 1.2, ...

✓ weight, height of a person.

(1) Let X denote the number of heads in a single toss of 4 fair coins. Determine

(i) $P(X < 2)$

(ii) $P(1 < x \leq 3)$

Sol: Possible values of X is 4 → {0, 1, 2, 3, 4

↓

Discrete Random Variable

(∵ Given Single toss of 4 fair coins)

(RV → capital letters)

Probability Distribution

X	0	1	2	3	4
P(x)	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Discrete dist → table

0 heads → All tails

H H H H
H H H T
H H T H
H H T T
H T H H
H T H T
H T T H
H T T T
T H H H
T H H T
T H T H
T H T T
T T H H
T T H T
T T T H
T T T T

1> $\sum P(x) = 1$

I> $P(X < 2) = P(x = 0) + P(x = 1)$

$= \frac{1}{16} + \frac{4}{16}$

$= \frac{5}{16}$

(ii) $P(1 < x \leq 3)$

$P(x=2) + P(x=3)$

$\frac{6}{16} + \frac{4}{16} = \frac{10}{16} = \frac{5}{8}$

(2) Two dice are thrown. Let X assign to each point (a, b) in S. The maximum of its numbers i.e $X(a,b) = \max(a,b)$. Find the probability distribution. X is a random variable with $X(S) = \{1, 2, 3, 4, 5, 6\}$. Also find mean, variance.

Sol:

$$\begin{Bmatrix} (1,1)\ (1,2)\ (1,3)\ (1,4)\ (1,5)\ (1,6) \\ (2,1)\ (2,2)\ (2,3)\ (2,4)\ (2,5)\ (2,6) \\ (3,1)\ (3,2)\ (3,3)\ (3,4)\ (3,5)\ (3,6) \\ (4,1)\ (4,2)\ (4,3)\ (4,4)\ (4,5)\ (4,6) \\ (5,1)\ (5,2)\ (5,3)\ (5,4)\ (5,5)\ (5,6) \\ (6,1)\ (6,2)\ (6,3)\ (6,4)\ (6,5)\ (6,6) \end{Bmatrix}$$

$X[1,1] = \max(1,1) = 1$

$P(x=1) = P(1) = \frac{1}{36}$

$P[x=2] \quad x = \{(1,2)\ (2,1)\ (2,2)\}$

$\frac{3}{36}$

$P[x=3] = \{(1,3)(3,2)(3,3)(2,3)(3,1)\}$

$= \frac{5}{36}$

$X = x$	1	2	3	4	5	6
$P(X)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

(3) A Random variable X has following probability distribution

X	0	1	2	3	4	5	6	7
P(x)	0	K	2K	2K	3K	K^2	$2K^2$	$7K^2+K$

(i) Determine K

$\sum P(X=x) = 1$

P(x)

$\sum P(X=x) = 1$

↓

0 to any value

$0 + K + 2K + 2K + 3K + K^2 + 2K^2 + 7K^2 + K = 1$

$10K^2 + 9K - 1 = 0$

$K = \frac{1}{10}, -1$

-1 ✗

Rule 2:- $0 \leq P(E) \leq 1$

$K = \frac{1}{10}$

(ii) find $P(x < 6)$

$= P(x=0) + P(x=1) + P(x=2) + \cdots + P(x=5)$

$// \{P(x<6) = P(x=5)\}$

$= 0 + \frac{1}{10} + 2 \cdot \frac{1}{10} + 2 \cdot \frac{1}{10} + 3 \cdot \frac{1}{10}$

$= 0.81$

(iii) $P(x \geq 6) = P(X=6) + P(X=7)$

$P(x \geq 6) = 1 - P(x<6)$

(iv) $P(0 \leq x \leq 4)$

$P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)$

$0 +$

(v) $P(x \leq \textcircled{k}) > \frac{1}{2}$ $\longrightarrow$ find the minimum value of $\textcircled{k}$ x

$P(x \leq 0) > \frac{1}{2}$

$0 > 0.5$ ×

$x=1$

$P(x \leq \textcircled{x}_1) > \frac{1}{2}$

$0.1 > 0.5$ × false

$x=2$

$P(x \leq 2) > \frac{1}{2}$

$\frac{1}{10} + 2 \cdot \frac{1}{10} > \frac{1}{2}$

$\frac{3}{10} > \frac{1}{2}$

$0.3 > 0.5$ × false

x Satisfies 4, 5, 6, 7

{∵ min value ∴ x value is 4} ✓

$x=3$

$P(x \leq 3) > \frac{1}{2}$

$0 + K + 2K + 2K$

$5K > \frac{1}{2}$

$5 \cdot \frac{1}{10} > \frac{1}{2}$

$\frac{5}{10} > \frac{1}{2}$

$0.5 > 0.5$ × false

$x=4$

$P(X \leq 4) > \frac{1}{2}$

$8K > \frac{1}{2}$

$0.8 > 0.5$ ✓ True

* Distribution function

$F(x) = P(X \leq x)$

Denoted by $F(x)$

X	F(x)
0	0
1	1/10
2	3/10
3	5/10
4	4/10
5	0.81
6	0.83
7	1

$x = 5$

$P(X \leq 5) > \frac{1}{2}$

$8k^2 + k^2$

$8(1/10) + (1/10)^2 = 0.81$

$x = 6$

$P(x \leq 6) > 1/2$

$8\left(\frac{1}{10}\right)^2 + 3\left(\frac{1}{10}\right)^2 = 0.83$

$x = 7$

$P(x \leq 7) > 1/2$

$8K + 10K^2 + K = 9K + 10K^2 = \frac{10}{10} = 1$

* Cumulative Distribution function (CDF)

* Probability mass function (PMF) → discrete

* Probability density function (PDF) → Continous.

Source: towards Data Science

$x = 7$ [contd]

$\because 9K + 10K^2$

$\frac{9}{10} + 10\left(\frac{1}{10}\right)$

$= \frac{10}{10}$

$= 1$

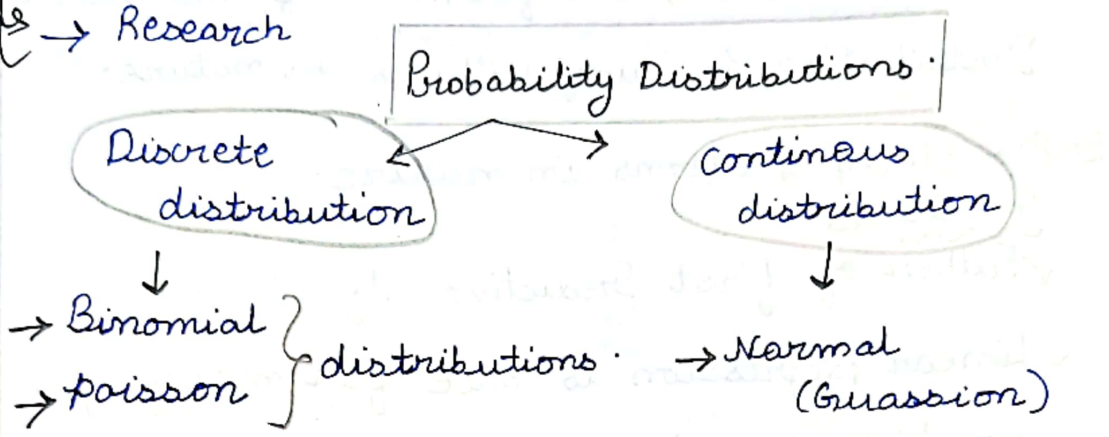
12/1/2021

Probability distributions:

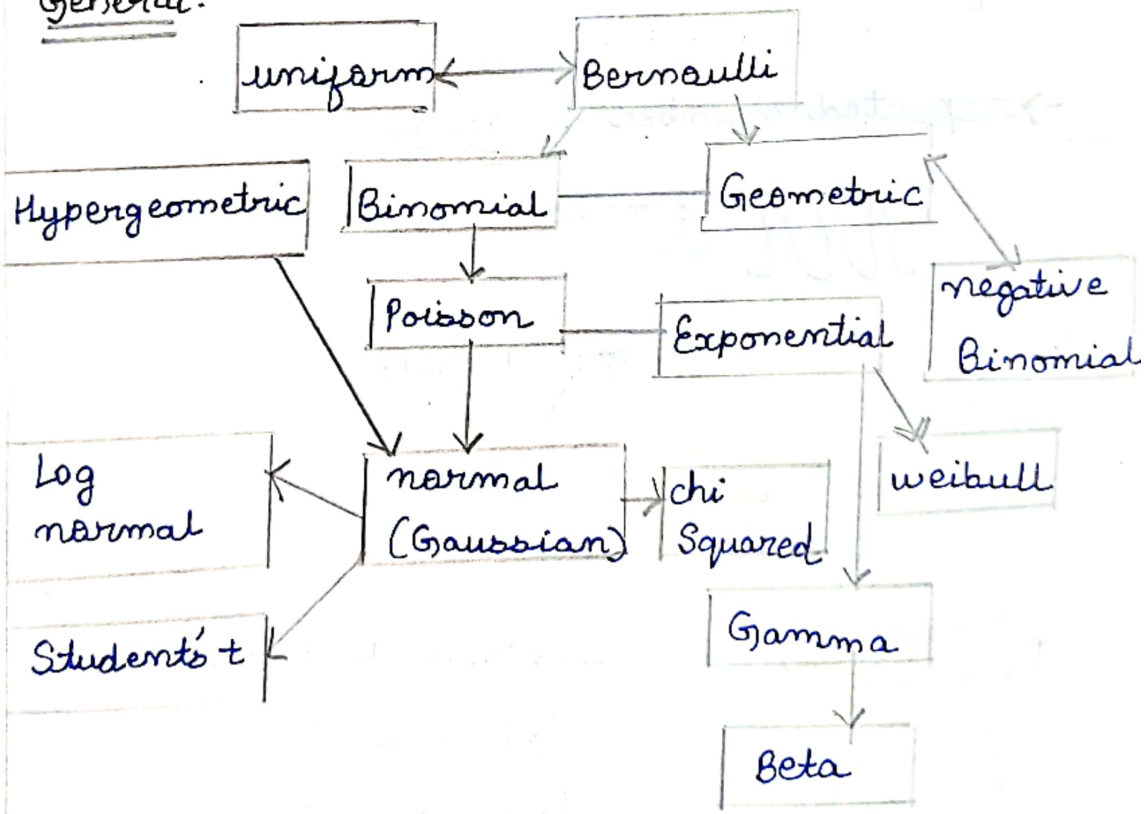
★★★★★ v.imp.

v.imp

areas
- → machine learning
- → Data Science
- → Research



General:



★★★★★ Imp. Normal Distributions:

✓Mother of all distributions is normal distribution.

✓Guassian distribution [Normal]

✓Biologists use this information of the normal Distribution to Study patterns in nature.

★ To Study patterns in nature.

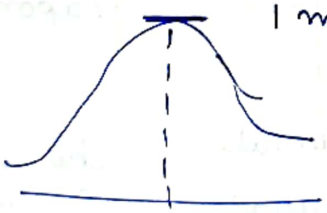
Gauss is

✓Father of first Predictive algorithm.

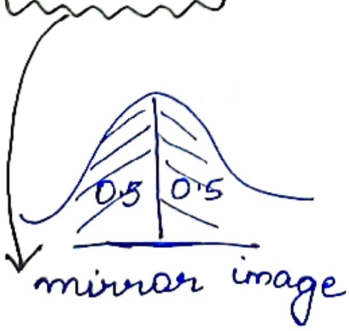
✓Linear Regression is base for most of machine learning and data Science algorithms.

→repeated number is mode

3 modes



1 mode



→ Normal distribution

↳ 1. Symmetric

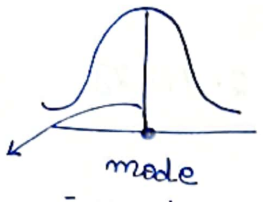
↳ 2. unimodal curve

↳ 3. Asymptotic

Asymptote: Tangent at ∞

↳ Touches / meets at Single Point.

x axis is tangent for normal curve

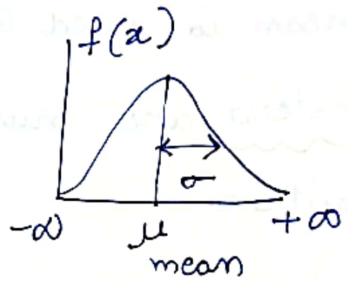


freq > = median = mean

*** mean = median = mode [In normal distribution]

* Properties

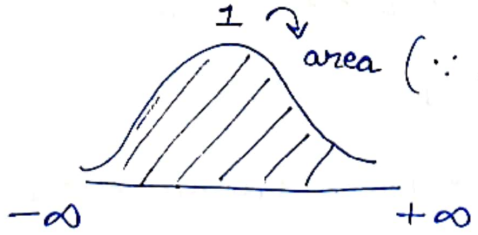
1. Bell Shaped
2. Symmetrical



3. mean, median, mode are equal
4. Location is characterized by mean μ
5. Spread is characterized by Standard deviation σ

** Imp information / Variance σ^2 = Variance

Random Variable has an infinite theoritical range $-\infty$ to $+\infty$



(∵ Sum of probabilities is 1)

V. Imp $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$; μ, σ are parameters

$f(x)$ = Probability density function of normal distribution

x values $-\infty$ to $+\infty$ x is variable

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

$e = 2.71828$

$\pi =$

Values $-\infty$ to $+\infty$

x is a Variable

μ, σ are parameters

[∵ mean = median = mode in normal distribution
mean is used Generally]

parameters are numerical descriptors of a distribution.

* μ, σ decides Shape, Structure in normal distribution.

15/1/2021:

General Lecture

Maths
- ✓ Linear Algebra
- ✓ Calculus

Statistics
- ✓ Descriptive
- ✓ Inferential

} Statistics

MNIST data Set

→ 64K images of {0, 1, ... 9}

Eg: 

Images are represented by matrices

Each image to represent we require 784 Pixels

$28 \times 28 = 784$

✓Each image is feature

$$D = \begin{bmatrix} & f_1 & f_2 & f_3 & \cdots & f_{784} \\ 1 & & & & & \\ 2 & & & & & \\ 3 & & & & & \\ \vdots & & & & & \\ 64K & & & & & \end{bmatrix}$$

64000×784

↓
Data

→dimensions → columns

784 $\xrightarrow{\text{Convert}}$ 2D or 3D to Visualize data

(Linear Algebra) plays crucial role

* Dimensionality Reduction
  - ✓Data Visualization [2D or 3D] [∵ ease]
    - ↓ even 10D ✓
  - To apply machine learning algorithm

→To convert to 2D

we need Eigen values; Consider top two

Eigen values λ_1 λ_2 Each Eigen value has Eigen vector

↓ ↓

v_1 v_2

with principle component analysis (1933)

$$D = \begin{bmatrix} & f_1 & f_2 \\ 1 & & \\ 2 & & \\ \vdots & & \\ 64K & & \end{bmatrix}$$

2D Visualization.

Eg: To identify dog, cat in an image we need Linear algebra

Calculus:

✓ GeoHrey Hinton is father of modern deep Learning

→ Deep learning —based on→ chain rule

↓

** Calculus

Back propagation algorithm → Calculus chain rule

Statistics:

* Decision making

- fact based
- not by Emotions
- Backed up data [information]

→ Branch of Mathematics is Statistics

Collect, organize, interpret data

Descriptive Statistics

- uni variate
- Bi variate
- multi variate

To Solve inferential ↓ depends on descriptive Statistics

→ Summarize data

*** vImp

✓ Inferential Statistics:

- Research
- academics
- Industry
  - Eg: Covid
- medical.
- Engineering
- ⋮

Inferential Statistics:

Hypothesis testing

imp ✓

model fitting → confirms which model fits best.

↳ regression
Logistic
multi linear
} model

→ Model is relation between dependent variable and independent variable

Descriptive Statistics.

- Mean
- median
- mode
- Standard deviation
- Sample variance
- Kurtosis
- Skewness
- Range
- maximum
- minimum
- Sum
- Count
- Geometric mean
- Harmonic mean
- AAD
- MAD
- IQR

$X = [\quad]_{64K \times 784}$

$X^T = [\quad]_{784 \times 64K}$

$C = X^T \times [\quad]_{784 \times 784}$

↓

Covariance matrix

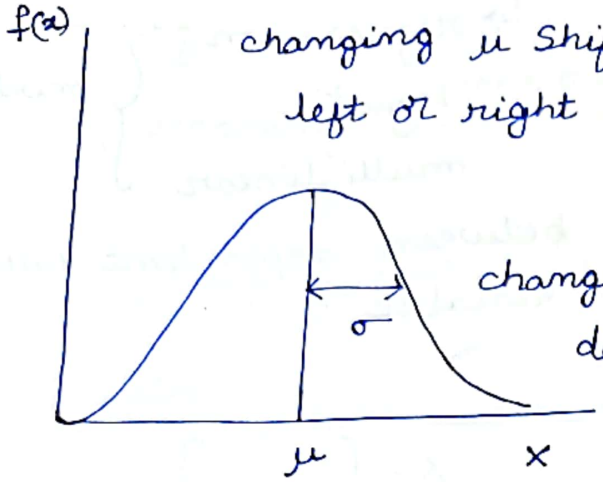
↓ under

multivariate descriptive matrix

16/1/2021:

Normal distribution ....[Contd]

Shape



changing μ Shifts the distribution left or right

changing σ increases or decreases the Spread.

→ μ → we can Shift mean

→ σ → spread of location

☆ Converting X → Z

↳ Normal distribution (X)

Imp ↳ Standardized Normal distribution (Z)

① Needs to be transform X units into Z units

② mean is 0 and Standard deviation is 1

③ * μ = 0 σ = 1 (in Z)

* Standardized Normal distribution:

$$Z = \frac{X - \mu}{\sigma}$$

Z → Standard normal Variate/variable

X → normal distribution

μ → mean

σ → Standard deviation

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$ → [Pdf of Standard normal deviation distribution]

$$Z = \frac{x-\mu}{\sigma}$$ [$\because \mu = 0, \sigma = 1$] Conditions.

$$Z = \frac{x-0}{1}$$

Z is increasing; f(Z) is decreasing

$$\left[\because f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty \le x \le \infty\right.$$

$$= \frac{1}{1 \times \sqrt{2\pi}} e^{-\frac{1}{2}\cdot\left(\frac{x-0}{1}\right)^0} \quad -\infty \le x \le \infty$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Z)^2} \quad -\infty \le x \le \infty$$

$$\left.\left(\frac{x-0}{1} = Z\right)\right]$$

mean is 0 and Standard deviation is 1

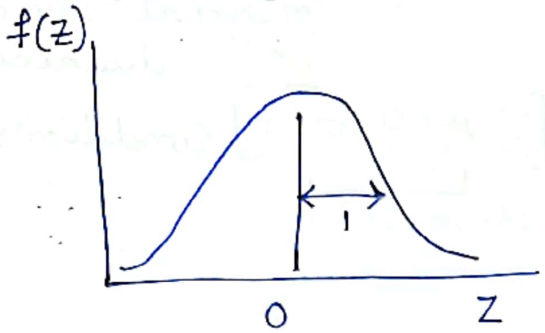
$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$ → Probability density function of Standardized normal distribution.

where e = the mathematical constant approximated by 2.71828

π = the mathematical constant approximated by 3.14159

Z = any value of the Standardized normal distribution

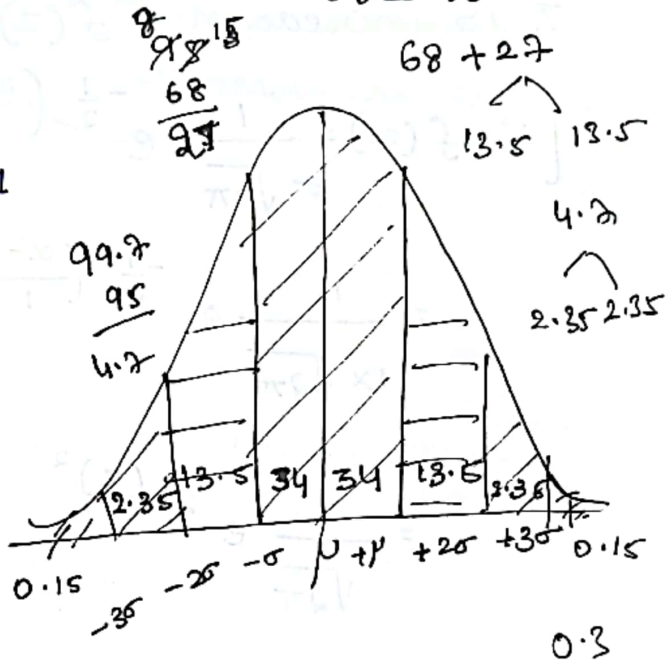
Standardized Normal distribution:



68
34 34
95% $\mu+2\sigma$ $\mu-2\sigma$

68 < 95

→ Z distribution
→ mean 0
→ Standard deviation 1
→ positive Z values
→ Negative Z values.



0.3
0.15 0.15

Eg:
Supposely

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

$$f(z) = e^{-z^2}$$

1SD 34 — $z = 0$, $f(0) = 1$

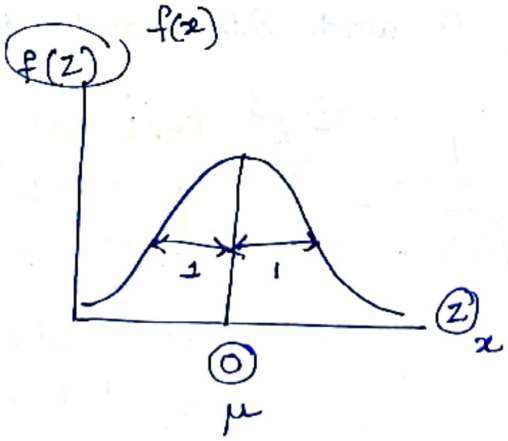
2SD 2.35 — $z = 1$, $f(1) = e^{-1}$

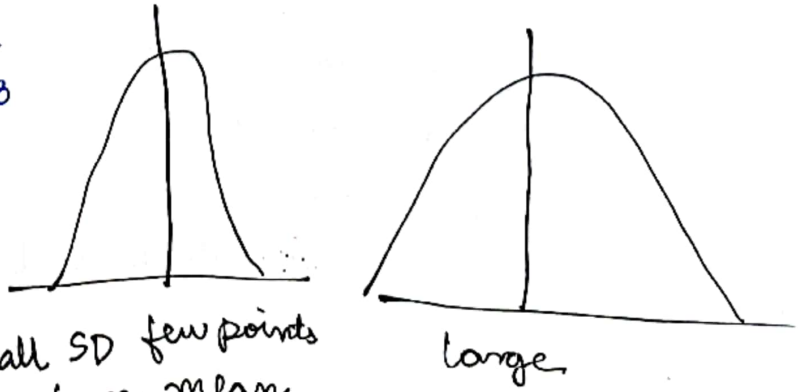
3SD 0.15 — $= \frac{1}{e}$

$= \frac{1}{2.713}$

34.00
02.35
00.15
36.50

Role of σ





small SD few points i.e. from mean

large

* Calculate the height of all individuals in world. 68% of people will be with 1 SD. avg in terms of ht.

① If X is distributed normally with mean of 100 and σ of 50, the Z value for X = 200 is

Sol:

$$Z = \frac{X - \mu}{\sigma}$$; σ is Standard deviation.

$$\Rightarrow \frac{200 - 100}{50} \Rightarrow \frac{100}{50} = 2$$

$$= 2.0$$

☆ applicable for forecasting when actual data pred. is ☹

** Empirical Rule for Normal Distribution:

↓ also known as (68-95-99.7 rule)

* Rule holds good only for (applicable) Normal distribution

Symmetrical curve. Normal Bell shaped.

Empirical rule is also called as

↳ 68-95-99.7 rule

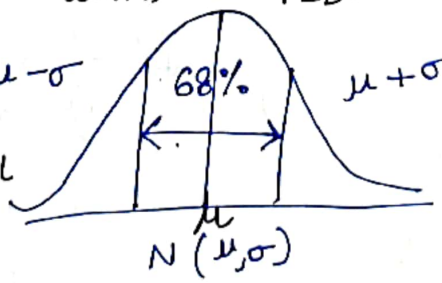
↳ Three Sigma rule

↳ 3 Standard deviations of a mean

* One Standard deviation distance:

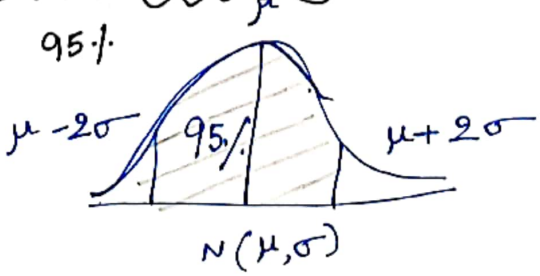
within μ 1 SD.

Supposely data points → 68% of worlds height avg will be in 1σ



many points closer to average.

* Two Standard deviation



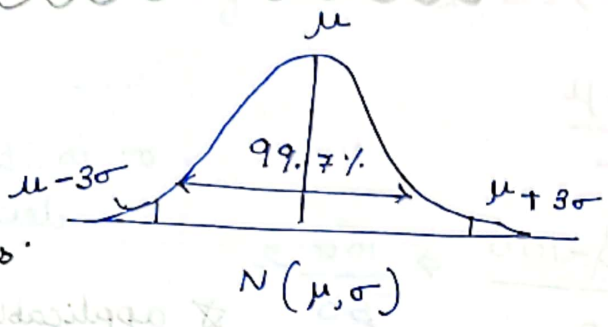
extreme values exists.

↑ ↓ } outliers.

Empirical Rule

* Three Standard deviation

99.7



Extremities.

$N(\mu, \sigma)$

99.7% within 3 Standard deviations of mean

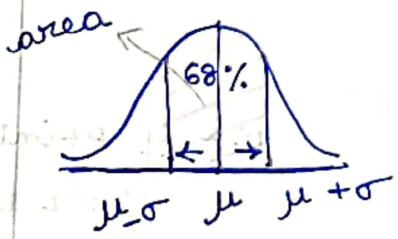
② Assume that the mean of 1 year-old girls in the US is normally distributed with mean of about μ 9.5 kg with Standard deviation of $\sigma \simeq$ 1.1 kg. without using calculator, estimate the % of 1 year-old girl in the US that meet the following conditions

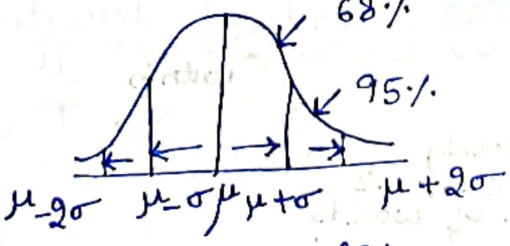
1. Less than 8.4 kg
2. Between 7.3kg and 11.7kg
3. more than 12.8kg

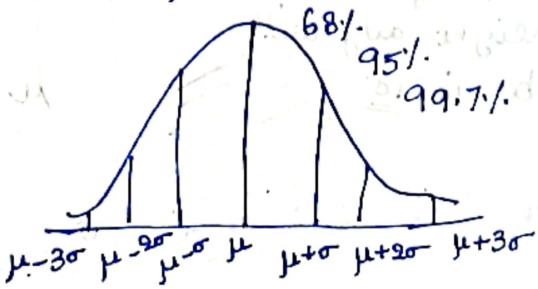
sol: Given $\mu = 9.5$ kg, $\sigma = 1.1$ kg

rule * 68-95-99.7

68% ⊂ 95%







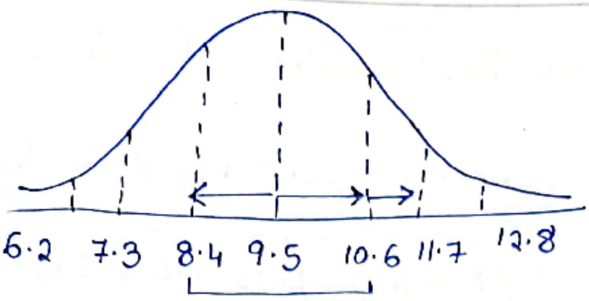
68% is area between

$\mu - \sigma$ and $\mu + \sigma$

$\mu - \sigma$ $\mu + \sigma$

$\mu = 9.5$ kg

$\sigma = 1.1$ kg

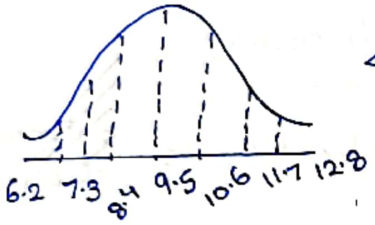


$\mu - 1\sigma = 9.5 - 1.1 = 8.4$

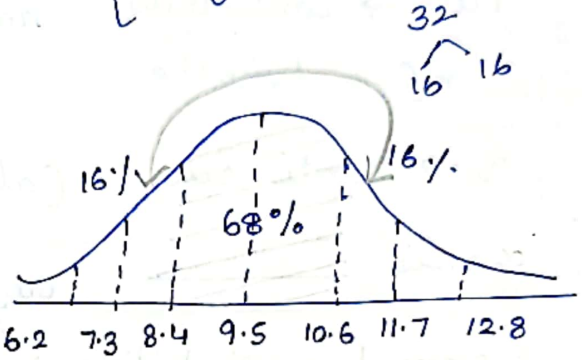
$\mu + 1\sigma = 9.5 + 1.1 = 10.6$

[∵ Symmetric distribution

(i) Less than 8.4 kg



< 8.4 kg



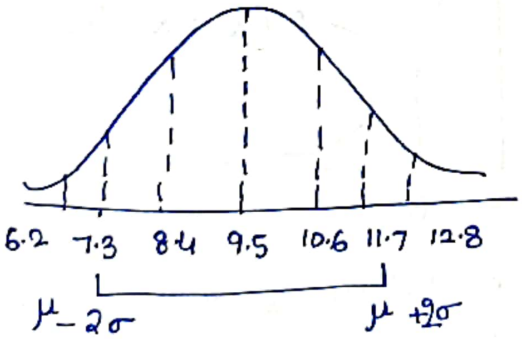
100% → 100 − 68

= 32

∴ Less than 8.4 kg = 16%.

$\frac{16}{100}$ is Probability.

(ii) Between 7.3 kg and 11.7 kg = $\frac{95}{100}$ = 0.

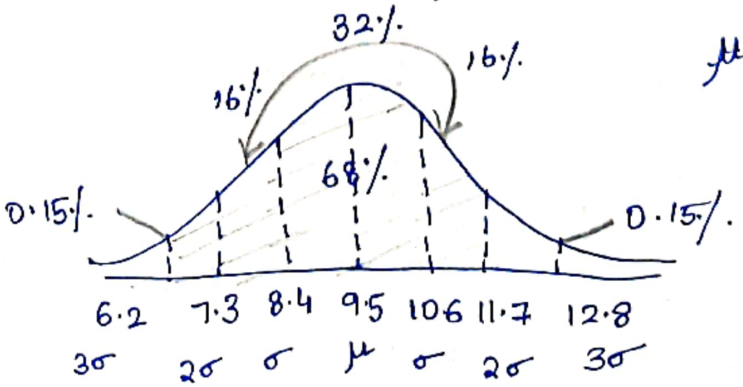


area between $\mu - 2\sigma$ and $\mu + 2\sigma$ is 95%

100 − 95 = 5%

$\frac{95}{100}$ probability

(iii) more than 12.8 kg = 0.15%. (iv) below than 6.2 is 0.15%



$\mu - 3\sigma$ and $\mu + 3\sigma$

99.7%

100 − 99.7

= 0.3%

0.15 0.15

Cumulative Distribution Function

CDF of random variable X is defined as

$$F_X(x) = P(X \leq x), \text{ for all } x \in R$$

Cdf → describes distribution of a random variable

Pdf → continous

Pmf → discrete

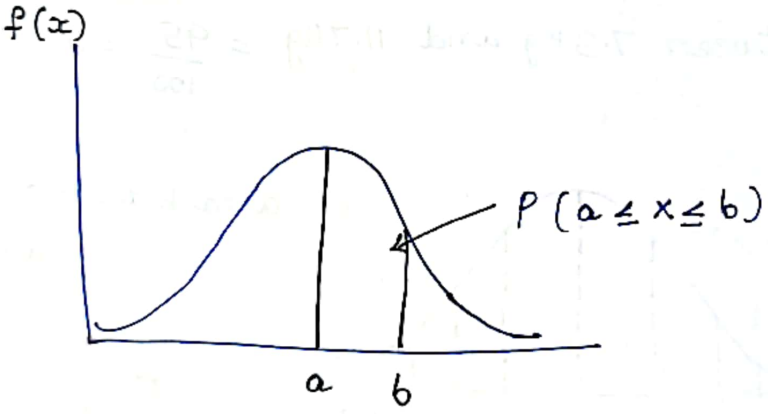
R → continous (also for discrete)

x axis ———— x axis - - - -

Cdf is applicable for continous / discrete

Normal probabilities:

→ Probability is measured by the area under the curve



Note that the probability of any zero individual value is zero.

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (\because \text{area})$$

[∵ In X Normal Distribution ∫ integration is complex]

↓

Convert to Z

no integration required.

Cumulative Distribution Function

CDF of random variable X is defined as

$$F_X(x) = P(X \leq x), \text{ for all } x \in R$$

Cdf → describes distribution of a random variable

Pdf → continous

Pmf → discrete

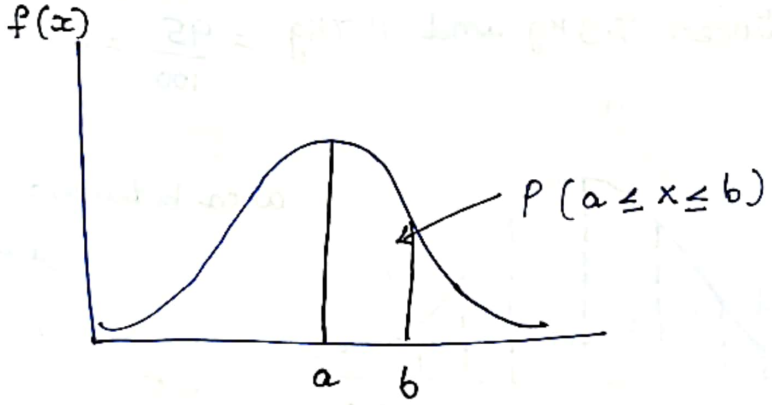
R → continous (also for discrete)

x axis ———— x axis - - - -

Cdf is applicable for Continous / discrete

Normal probabilities:

→ Probability is measured by the area under the curve



Note that the probability of any zero individual value is zero.

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (\because \text{area})$$

[∵ In X Normal Distribution ∫ integration is complex]

↓

Convert to Z

no integration required.

Applications of Cumulative Distribution Function.

* Standard Normal table
* unit Normal Table } also called is applicable in field of data Science.
* Z table

[∵ we need not to find integration to find probability using above tables.

→ using PDF and CDF we can find probability of a Random Variable.

$$\begin{cases} f(x) = Pdf \\ F(x) = cdf \\ f(x) = \frac{d}{dx}[F(x)] \end{cases}$$

* Statistics in Understanding Data

* To understand data we require Statistics.

"There are two kinds of Statistics, the kind you look up and the kind you make up"
Rex Stout

What is Statistics?

*A branch of mathematics that deals with collecting, analyzing and interpreting data.

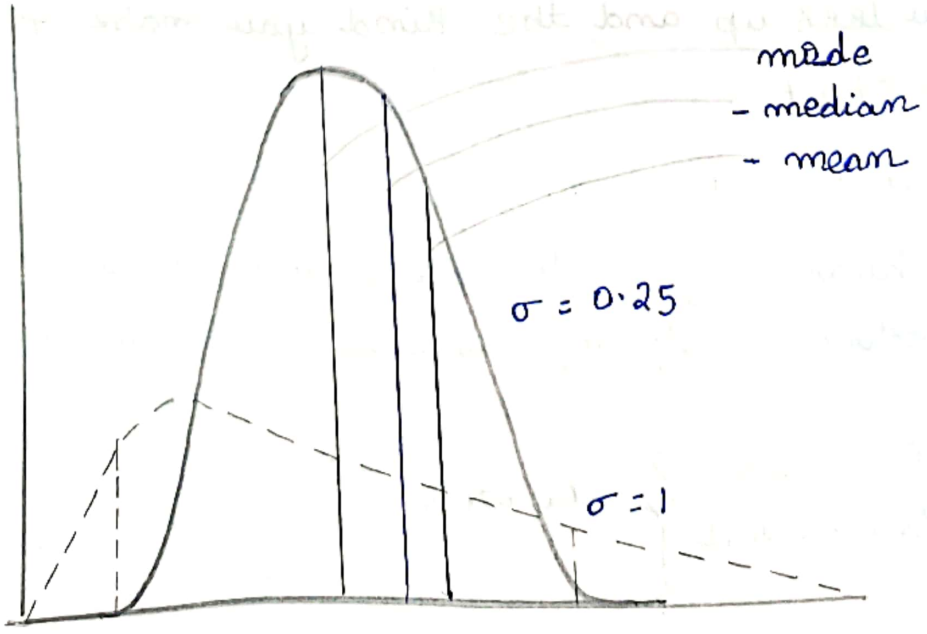
✓ Descriptive
✓ Inferential } Statistics

Descriptive Statistics:

- ✓ Summarize, describes the data
- ✓ Identifies important elements in dataset without seeking to find explanations of those elements
- ✓ Depending on how many variables are involved descriptive statistics are of 3 types

- ✓ Summarize Single Variable → univariate
- Relation b/w 2 Variables → Bivariate
- Describe → multiple variables multivariate

- ✓ Summarize data as it is
- ✓ Do not posit any hypothesis about data
- ✓ Detect outliers
- ✓ Plan how to prepare data
- ✓ precursor to feature engineering



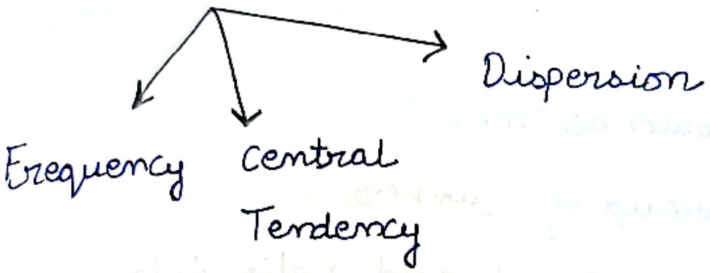
✓Feature engineering involves extracting (features) data to build mode

↓ we require features

we require feature engineering

✓ Precursor to feature engineering.

univariate (Single variable)



Measures of frequency:

↳ Frequency Tables

↳ Histograms

Summary of Statistics:

Measures of central Tendency:

Measures of Location

- Arithmetic mean
- ✓ weighted mean
- ✓ median
- ✓ mode
- ✓ Percentile
- ✓ Geometric
- ✓ Harmonic

} means ↓ infrequently used

Measures of Dispersion

- ✓ Skewness
- ✓ Kurtosis
- ✓ Range
- ✓ Interquartile range
- ✓ Variance
- ✓ Standard score
- ✓ Co-efficient of Variance

Arithmetic Mean

→ Data : {16, 17, 10, 13, 20, 18, 13, 14, 18}

$$\bar{x} = \frac{16+17+10+13+20+18+13+14+18}{9}$$

$$\bar{x} = \frac{139}{9}$$

$$\bar{x} = 15.444$$

☆ Commonly known as 'mean'

☆ Average of Group of numbers

☆ Applicable for interval and ratio data

Eg:

Interval data

[2000, 2020]

Heights, weights

↓

wts are allowed

☆ Not applicable for nomidal or ordinal data

Imp

Gender { male - 1, female - 0

1. Prof, AssoP, AssisP
2. Grades of Students

☆ Affected by each value in the dataset, including extreme values

* Computed by summing all values in the dataset and dividing the sum by the numbers of values in the dataset.

Population Mean:

Data: 60 20 10 40 50 30

$$\mu = \frac{\Sigma x_i}{n} = \frac{60+20+10+40+50+30}{6} = 35$$

* Impact of outliers:

Data: 60 20 10 40 50 30 1000

$$\mu = \frac{\Sigma x_i}{n} = \frac{60+20+10+40+50+30+1000}{7}$$

Mean = 172.85

Population mean $\rightarrow \mu$

Sample mean $\rightarrow \bar{x}$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

$$= \frac{57+86+42+38+90+66}{6}$$

$$= \frac{379}{6}$$

$$= 63.167$$

* Mean of Grouped Data

★ weighted average of class midpoints

★ class frequencies are the weights

$$\mu = \frac{\Sigma fM}{\Sigma f}$$ (fM: f → frequency, M → class midpoint)

$$= \frac{\Sigma fm}{N}$$

$$= \frac{f_1 M + f_2 M_2 + f_3 M + \cdots + f_i M_i}{f_1 + f_2 + f_3 + \cdots f_i}$$

f_1 is frequency of first class

M_1 is class midpoint of first class

Calculation of Grouped Mean:

class Interval	Frequency (f)	class (M)	fM
20 - under 30	6	25 (midpoint)	150
30 - under 40	18	35	630
40 - under 50	11	45	495
50 - under 60	11	55	605
60 - under 70	3	65	195
70 - under 80	1	75	75
	50		2150

(20+30)/2 = 50/2

$$\mu = \frac{\Sigma fM}{\Sigma f} = \frac{2150}{50}$$

$$= 43.0$$

20/1/2021

* Finding Normal probability procedure.

→ To find $P(a < X < b)$ when X is distributed normally:

→ Draw the normal curve for the Problem in terms of X.

✓ Translate X-values to Z-values
use the Standardized Normal Table

→ Accessing Normality:

CLT
↳ Very Imp
many Analytical and Statistical tools

Data → Data cleaning → verify ND

* Demo on Normal Distribution:

```
from Scipy.Stats import norm
Val, m, S = 68, 65.5, 2.5
print(norm.cdf(val, m, S))
```

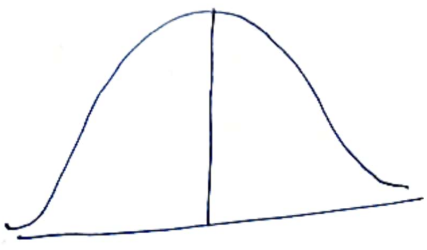
Val = 68, m = 65.5, S = 2.5
↓ ↓ ↓
x = 68 μ = 65.5 σ = 2.5

$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{68 - 65.5}{2.5}$$

$$= \frac{2.5}{2.5}$$

$$= 1$$



20/1/2021

Central Limit Theorem :

↳ Real time applications

↳ Super useful

↳ Statistics.

↳ Mathematics

Developing a Sampling distribution :

Assume there is a population ...

Population Size $N = 4$

Random variable x, is age of individuals

Values of x :

18, 20, 22, 24 (years)

$$\mu = \frac{\sum x_i}{N}$$

$$= \frac{18 + 20 + 22 + 24}{4} \Rightarrow 21$$

μ is Population mean

(by default)

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

$$= (x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + (x_4 - \mu)^2$$

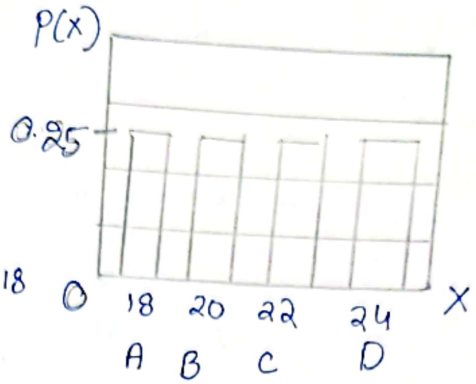
18 20 22 24

x_1 x_2 x_3 x_4

on Substitution

$\sigma = 2.236$

$\therefore \mu = 21$



$\therefore$ Probability for 18 $\frac{1}{4}$

uniform distribution

Sly 18, 20, 22, 24

$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ $\left\{ \frac{\text{favourable cases}}{\text{Total}} \right\}$

Now consider all possible Samples of Size

$n = 2$

1st obs	2nd observation 18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

note:

N^n → with repetition

$N_n C$ → without

N_C → without repetition

16 possible Samples (Sampling with replacement)

→ $N=4, n=2$ $N^n = 4^2 = 16$

$n = 2$ $= 16$

1st Obs	2nd observation 18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 Sample means

1st Obs	2nd Obs 18	19	20	21
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

above 2nd table is obtained by averages

avg:

$\frac{18+18}{2} = \frac{36}{2} = 18$

$\frac{18+18}{2} = 18$, $\frac{18+20}{2} = 19$, ... so on

mean is called first expectation

16 Sample means

	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

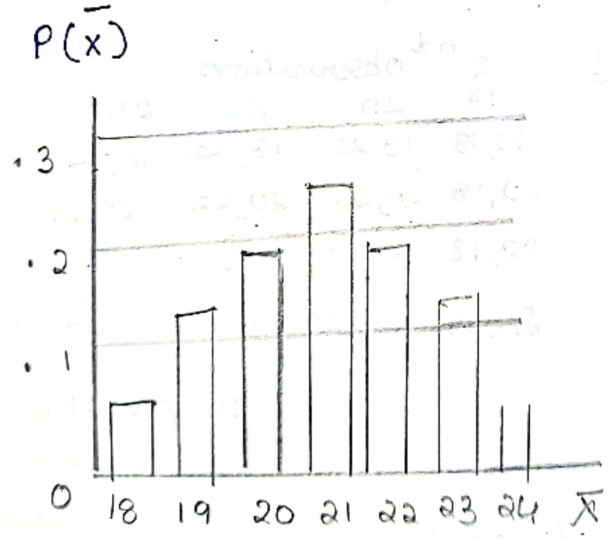
} means of Sample Size = 2

CLT / not applicable ✓

Pareto ✓

{Cauchy ✗ not use in CLT

Sample means distribution



no longer uniform

∴ μ is always population mean

$\bar{x}$ sample mean

Summary measures of this Sampling distribution

$$E(\bar{x}) = \frac{\Sigma \bar{x}_i}{N} = \frac{18+19+21+\cdots+24}{16} = 21 = \mu$$

μ and $E(\bar{x})$ are always same (values)

$$\sigma_{\bar{x}} = \sqrt{\frac{\Sigma (\bar{x}_i - \mu)^2}{N}}$$

$$= \sqrt{\frac{(18-21)^2 + (19-21)^2 + \cdots + (24-21)^2}{16}}$$

$$= 1.58$$

σ Value 2.236

always $\sigma_{\bar{x}} < \sigma$ values

Comparing the population with its Sampling distribution

Population

$N = 4$

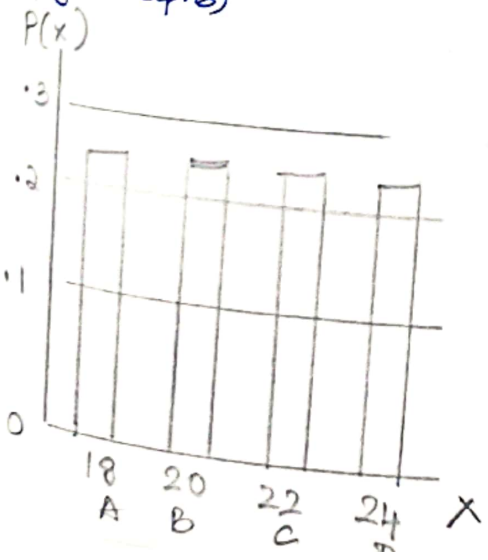
$\mu = 21$ $\sigma = 2.236$

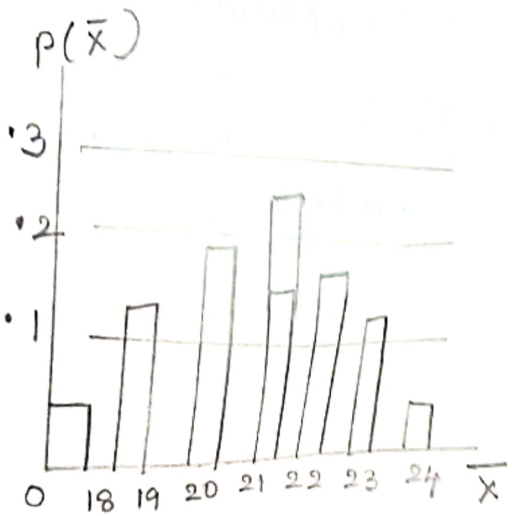
Sample means distribution

$n = 2$

$\mu_{\bar{x}} = 21$, $\sigma_{\bar{x}} = 1.58$

fig (Graphs)





Even if you're not
normal...
... the average...
... is normal!!!

* CLT [Central Limit Theorem]

Statement:

Let x be an independent identically distributed random Variable with finite Polulation mean μ and finite population Variance σ^2 then a random Variable converges in a distribution to the standard normal Variables as n goes to infinity with mean μ and Variance σ^2/n

Here n is the Size of Sample from given distribution

{ no matter what is the Shape of the population is distribution of Sampling distribution will be normal if n is very large.

Limitations:

not applicable for

* Pareto, Cauchy } Distributions $\because \mu = \infty$

$\because$ mean is ∞

Let us assume X is a random Variable with finite Population mean and finite Population Variance

$X(\mu, \sigma^2)$

I am Collecting m Samples

$S_1, S_2, S_3 \ldots S_m$

Size of each Sample is $n = 30$, my Sample Size

Let me assume I am Collecting 1000 Samples

$n = 30, m = 1000$

$m \times n = 1000 \times 30$

$= 30K$ Samples

For each of the Sample, let us Calculate mean

$S_1, S_2, S_3, \ldots S_m$

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

$\bar{x}_1 \quad \bar{x}_2 \quad \bar{x}_3 \quad \bar{x}_m$

Distribution of $\bar{x}_i$, $i = 1$ to m

Distribution of $\bar{x}_i$ = Sampling distribution of Sample mean

Now Central limit theorem Says

$\bar{x}_i$ is distributed with Gaussian Normal distribution with mean μ and Variance $\frac{\sigma^2}{n}$

as $n \to \infty$

i.e $\bar{x}_i \longrightarrow N\left(\mu, \frac{\sigma^2}{n}\right)$

as $n \to \infty$

* Sampling distribution Properties:

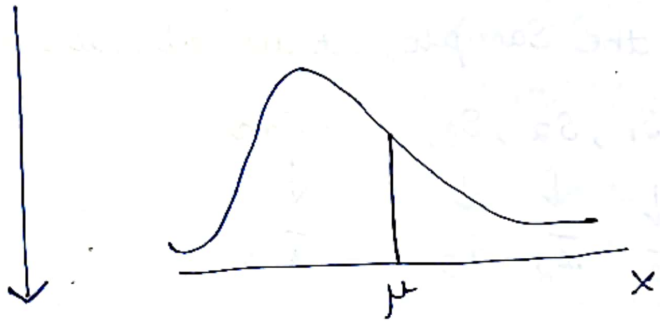
Central Tendency

$$\mu_{\bar{x}} = \mu$$

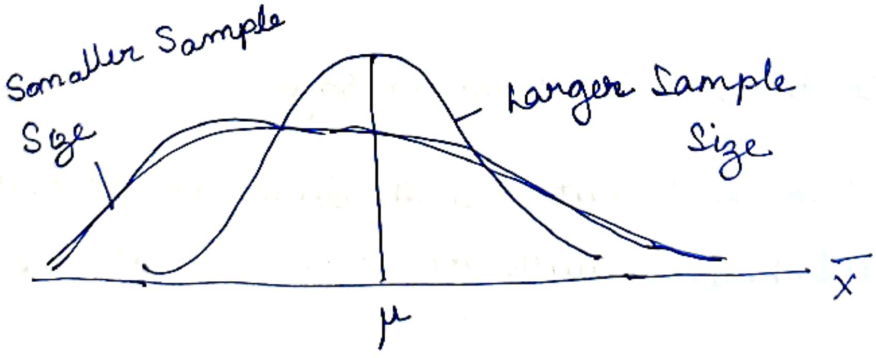
Variation

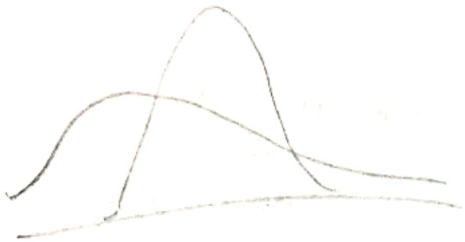
$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

* Population distribution:



* Sampling distribution (becomes normal as n increases)





21/1/2021

① Suppose a large population has mean $\mu = 8$ and Standard deviation $\sigma = 3$. Suppose a random Sample of Size $n = 36$ is Selected.
What is the probability that the Sample mean is between 7.8 and 8.2?

Sol: $\because n = 36$ ✓ large size → CLT is applicable $n \geq 25$ ✓

Even if the Population is not normal distributed, the Central theorem can be used

$\mu_{\bar{x}} = \mu$ $(n > 25)$

↓ Sample mean ↓ population mean

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

∴ So the Sampling distribution of $\bar{x}$ is approximately normal

$-0.5 < Z < +0.5$

$\therefore$ 7.8 8.2

with mean $\mu_x = 8$

and Standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$$

$$P(7.8 < \mu_x < 8.2) = P\left(\frac{7.8 - 8}{3/\sqrt{36}} < \frac{\mu_x - \mu}{\sigma/\sqrt{n}} < \frac{8.2 - 8}{3/\sqrt{36}}\right)$$

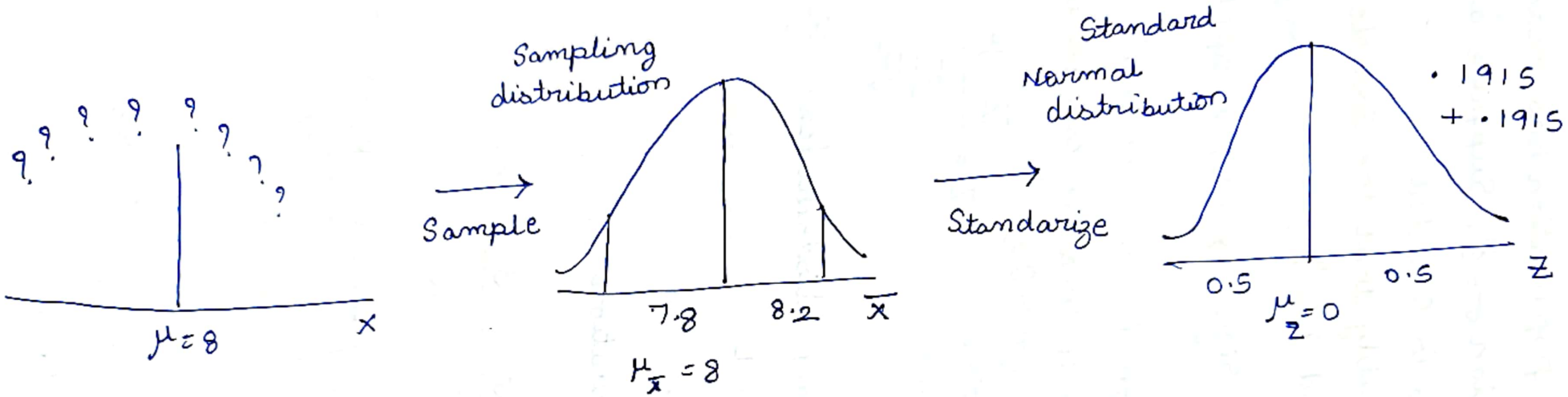
$$= P(-0.5 < Z < 0.5) = 0.3830$$

0.1915

$=$ ∵ from Z table

Value of 0.5 is 0.1915

$0 \quad 2 \times 0.1915$



parametric family of distributions:

① Binomial distribution:

Martin clothing Store problem.

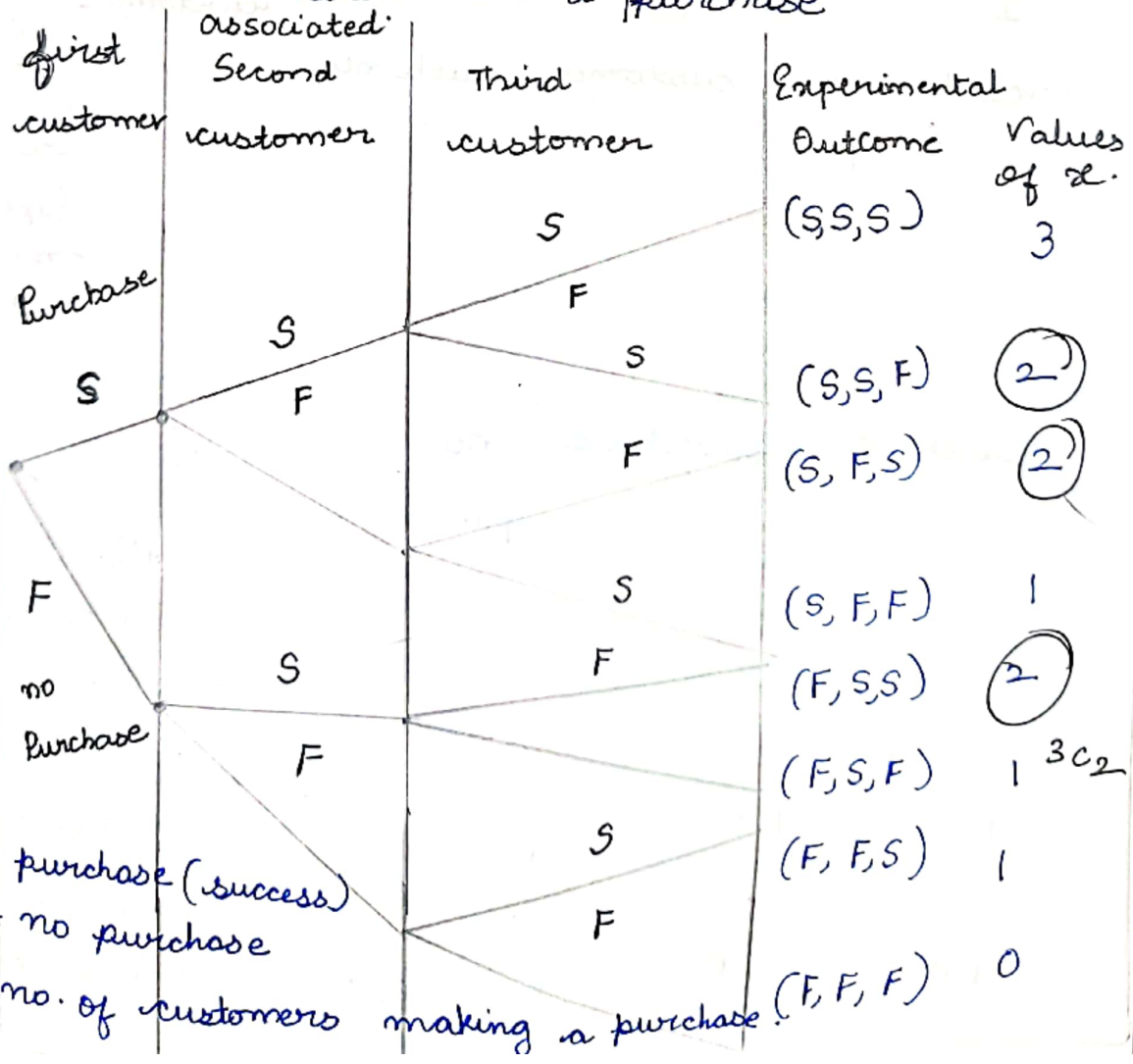
→ Let us Consider the purchase decisions of the next 3 customers who enter a Store.

→ on the basis of past experience, the store manager estimates the probability that any

→ one customer will purchase is 0.30

$$\star \begin{cases} p = 0.30 & \text{purchasing } p^* \\ 1-p = 0.70 & \text{not purchasing } p^* \end{cases}$$

→ what is the probability that two of the next 3 customers will make a purchase



S = purchase (success)
F = no purchase
x = no. of customers making a purchase.

$^3C_2$ is no. of Combinations to purchase 2 customers out of 3 customers

2 2 2

$^3C_2 = 3$

$^nC_r = {}^nC_{n-1}$

$^3C_2 = {}^3C_1 = 3$

Trial outcomes :-

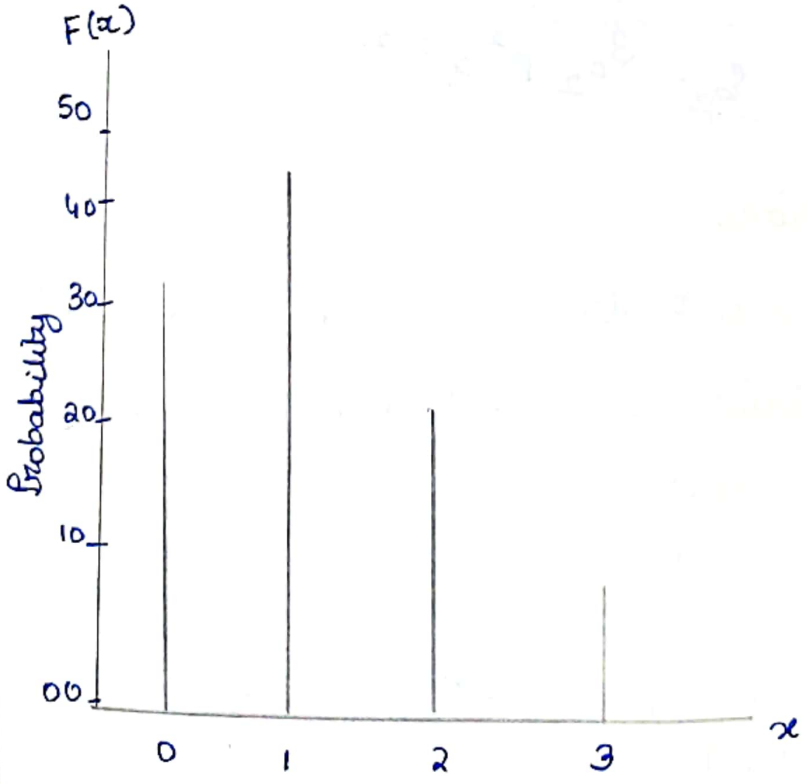
1st customer	2nd customer	3rd customer	Experimental customer	P
purchase	purchase	no purchase	(S,S,F)	PP(1-P) = $P^2(1-P)$ = $(0.30)^2(0.70)$ = 0.063
P	NP	P	(S,F,S)	P(1-P)P = 0.063
NP	P	P	(F,S,S)	(1-P)(PP) = 0.063

Probability of Experimental outcome.

Success P
Failure (1-P)

Graphical representation of probability dist. for number of customers making a purchase

x	$P(x)$
0	$0.7 \times 0.7 \times 0.7 = 0.343$
1	SFF $(0.3) \times 0.7 \times 0.7 +$ FSF $0.7 \times 0.3 \times 0.7 +$ $0.7 \times 0.7 \times 0.3 = 0.441$ ↳ FFS
2	0.189
3	0.027



number of customers making a purchase

from trial outcomes:

$$^3C_2\, p^2 (1-q)$$

$$p^2(1-p) + p^2(1-p) + p^2(1-p)$$

$$= 3p^2(1-p)$$

$$= {}^3C_2\, p^2 (1-p)^{3-2}$$

$${}^3C_2 = 3$$

$${}^nC_x\, p^x q^{n-x}$$

Binomial distribution

$$P(X=x) = {}^nC_x\, p^x q^{n-x}$$

Eg: $${}^3C_2\, p^2 q^{3-2}$$

Assumptions:

- ☆ Experiment involves n identical trials.
- ☆ Each trial has exactly two possible outcomes success or failure.
- ☆ Each trial is independent of the previous trial
- ☆ p is the probability of a success on any one trail
- ☆
- ☆

* $q = (1-p)$ is the probability of a failure on any one-trial.

* p and q are constant throughout the experiment.

probability function: $P(X) = \frac{n!}{X!(n-X)!} p^X \cdot q^{n-x}$ for $0 \leq x \leq n$

mean value $\mu = n \cdot p$

Variance and Standard deviation. $\sigma^2 = n \cdot p \cdot q$

$\sigma = \sqrt{\sigma^2} = \sqrt{n \cdot p \cdot q}$

$P(X = x) = {}^nC_x p^x q^{n-x}$

${}^nC_r = \frac{n!}{n!(n-r)!}$

Mean and Variance.

★ Suppose that for the next month the clothing store forecasts 1000 customers will enter the store.

★ what is the expected number of customers who will make a purchase?

★ The answer is $\mu = np$ $\quad n = 1000$

$= (1000)(0.3) = 300$

✓ for the next 1000 customer entering the store, the Variance and Standard deviation for the number of customers who will make a purchase are

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

$= 1000\ (0.3)(0.7)$

$= 210$

$\sigma = \sqrt{210}$

$= 14.49$

★ $\mu = np$

$\sigma^2 = npq$

```
import Scipy
import numpy as np
from Scipy.stats import binom
```

Poisson Distribution: Assumptions:

* Statistical Distribution → how many times event is occured within specified period of time.
* Discrete event { occuring / not occuring
* Describes rare events in large population
  mutation of cells in body: mutuation acquisition.
* Each occurence is independent any other occurences
* resources: [discrete occurences over interval]
* The expected number of occurences must hold constant throughout the experiment

Example:

(1) Measurements of number of occurences within a time period.
  arrival patterns

Eg1: Arrivals at queuing systems

* Airports - people, airplanes and baggage
* Banks - people, automatic and loan applications
* customers at a store.
* Computer file servers - read and write operations.

number of calls to a switch board.

The number of arrivals in any service facility like at an ATM, railway station, petrol pumps.

Eg 2: Defects in manufactured goods

* number of defects per 1,000 feet of extruded copper wire
* number of blemishes per Square foot of painted Surface.
* number of errors per typed page.

Eg 3: Number of Extreme weather events

Eg 4: The number of aircraft/road accidents in any time interval.

*poisson distributions model the number of Occurences within a fixed time period.

* can be used to approximate binomial distributions for very small Success probabilities.

Binomial:

$$P(X=x) = {}^{n}C_{x}\, p^{x} q^{n-x}$$

$\downarrow$ $n \to \infty$, $p \to 0$

poisson distribution

$$\frac{e^{-\lambda} \lambda^{x}}{x!} = P(X=x)$$

poisson distribution formula:

$$\frac{e^{-\lambda} \lambda^{x}}{x!}$$

e = Euler's constant ≈ 2.718

λ = mean or expected value of the variable

x = number of successes for the event

! = factorial.

mean	Variance	Standard deviation.
λ	λ	λ

$\lambda = 3.2$ customers / 4 minutes

$X = 10$ customers / 8 minutes

$$P(x) = \frac{\lambda^{x} e^{-\lambda}}{x!}$$

$$P(X=10) = \frac{6.4^{10} e^{-6.4}}{10!}$$

$$= 0.0528$$

$$P(X=x) = \frac{\lambda x e^{-\lambda}}{x!}$$

$\lambda = 3.2$ customers / 4 minutes

$X = 6$ customers / 8 minutes

Adjusted λ

6.4 customers / 8 minutes

$$P(x) = \frac{\lambda^{x} e^{-\lambda}}{x!}$$

$$P(x=6) = \frac{6.4^{6} e^{6.4}}{6 4} = 0.1856$$

=

* Sampling Theory:

* Population:

✓ The totality of observations with which we are concerned, whether this number be finite or infinite, constitutes what we call Population

(or)

Population (or universe) is the aggregate or totality of Statistical data forming a Subject of investigation.

Eg: 1. Population of Heights of Indians.

2. Population of Standarized banks.

* Sampling:

→ Study of entire Population may not be possible to carry out and hence a Part alone is selected from Given Population.

Part alone → Sample.

Sample → we need Sample to understand about Population.

☆ { Sample Size n / Population N Size }

Large Sample:

If Size of the Sample $(n) \geq 30$, the Sample is said to be large Sample.

Small Sample: $n < 30$

Correction factor:

* $\frac{N-n}{N-1}$

(i) what is the value of correction factor if $n=5$, and $N=200$

Sol: $\frac{200-5}{200-1} = \frac{195}{199} = \underline{0.98}$

(ii) How many different Samples of Size two can be choosen, from a finite Population of Size 25

Sol:

$n = 2$

$N = 25$

$N_{C_n} = {}^{25}C_n = \frac{25 \times \cancel{24}^{12}}{1 \times \cancel{2}}$

without replacement $= 300$ Samples

with replacement $N^n = 25^2$

$= 625$ Samples.

③ A Population consists of 5 numbers 2, 3, 6, 8, 11. Consider all possible samples of size two which can be drawn with replacement from this Population.

Sol: (i) find

6 (i) mean of the Population

3.29 (ii) Standard deviation of the Population

6 (iii) Mean of the Sampling distribution.

(iv) The Standard deviation of Sampling distribution of means

2.32.

Sol: (i) mean $= \frac{2+3+6+8+11}{5}$

$\mu = 6.0$

$\frac{\Sigma (x_i - \bar{x})^2}{n}$

(ii) $\sigma^2 = \frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5}$

$\sigma^2 = 10.8$

$\sigma = 3.29$

(iii) $n = 2, N = 5$

with replacement : $N^n = 5^2 = 25$ Samples

(2,2)	(2,3)	(2,6)	(2,8)	(2,11)
(3,2)	(3,3)	(3,6)	(3,8)	(3,11)
(6,2)	(6,3)	(6,6)	(6,8)	(6,11)
(8,2)	(8,2)	(8,6)	(8,8)	(8,11)
(11,12)	(11,3)	(11,6)	(11,8)	(11,11)

The corresponding Samples are

2.0	2.5	4.0	5.0	6.5
2.5	3.0	4.5	5.5	7.0
4.0	4.5	6.0	7.0	8.5
5.0	5.5	7.0	8.0	9.5
6.5	7.0	8.5	9.5	11.0

The mean distribution is

$$\mu_{\bar{x}} = \frac{\text{Sum of all Sample means (above)}}{25}$$

$$= \frac{150}{25}$$

$$= 6.0$$

$$\boxed{\mu_{\bar{x}} = \mu}$$

(iv)

$$\sigma_x = \frac{(2-6)^2 + \cdots\cdots + (11-6)^2}{25}$$

$$= \frac{135}{25}$$

$$= 5.40$$

$$\sigma_x = \sqrt{5.40}$$

$$= 2.32$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.29}{2} = 2.32$$

8/1/2021:

* Moments:

✓ measure of turning or rotating in physics

✓ In Statistics these moments are used to measure the characteristics like desperation, Skewness and Kurtosis of a frequency distribution.

Eg: find the first four moments for the set of numbers 2, 4, 6, 8

Sol: Here

$N = 4$

$$\text{mean} = \frac{2+4+6+8}{4} = 5$$

Numbers x_i	$d_i = x_i - 5$	d_i^2	d_i^3	d_i^4
x_1 2	-3	9	-27	81
x_2 4	-1	1	-1	1
x_3 6	1	1	-1	1
x_4 8	3	9	-27	81
$\Sigma x_i = 20$	$\Sigma d_i = 0$	$\Sigma d_i^2 = 20$	$\Sigma d_i^3 = 0$	$\Sigma d_i^4 = 164$

moments about mean

$$\mu_1 = \frac{\Sigma d_i}{N} = \frac{0}{4} = 0$$

$$\mu_2 = \frac{\sum d_i^2}{N} = \frac{20}{4} = 5 ; \quad \mu_3 = \frac{\sum d_i^3}{N} = \frac{0}{4} = 0$$

$$\frac{\sum d_i^4}{N} = \frac{164}{4} = 41$$

$$\beta_1 = \text{Skewness} = \frac{\mu_3^2}{\mu_2^3} = 0$$

[$\because$ Skewness $= 0 \rightarrow$ Normal distribution]

$$\beta_2 = \text{Kurtosis} = \frac{\mu_4^2}{\mu_2^2} = \frac{41}{25} = 1.64$$

* Properties of central moments:

(i) The first moment about mean is always zero
i.e $\mu_1 = 0$

(ii) The second moment about mean measures Variance. i.e $\mu_2 = \sigma^2$ or $\sigma = \sqrt{\mu_2}$

(iii) The third moment about mean measures Skewness of given distribution

(iv) (a) If $\mu_3 > 0$, the distribution is Positively Skewed.

(b) If $\mu_3 < 0$, the distribution is negatively Skewed

(c) If $\mu_3 = 0$, the distribution is Symmetrical