Inferential Statistics 15 Feb 25

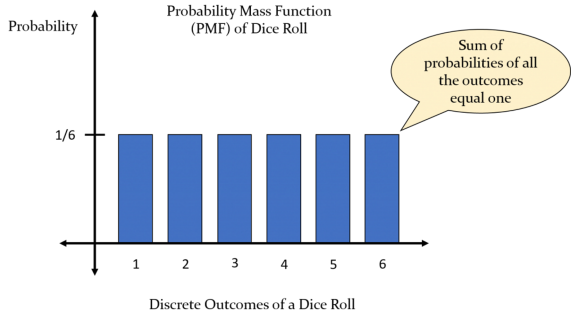
Friday, 14 June 2024 2:08 PM

Probability Distribution →

A mathematical function that gives the probabilities of occurence of different possible outcomes of an experiment. In simple words, it lists out all possible outcomes in the sample space with their probabilities.

Ex → **Throwing a dice** →

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



Discrete Probability Distribution →

Model the probabilities of random variables that can have discrete values as outcomes.

Ex → Let's take random variable X to be the sum of two die throws. X can take values (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12).

P(2) : (1,1) → 1/36

P(3) : (1,2), (2,1) → 2/36

P(4) : (2,2), (1,3), (3,1) → 3/36

P(5) → 4/36

P(6) → 5/36

P(7) → 6/36

⋮

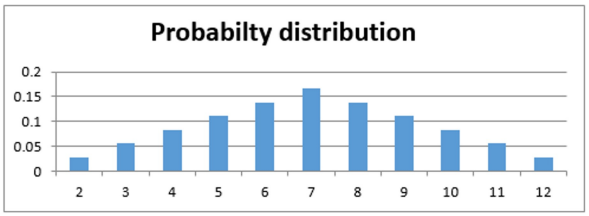
P(11) → 2/36

P(12) → 1/36

P(success) = 5/36

P(Failure) = 31/36

The probability function for a discrete random variable is the 'probability mass function'. It shows the exact probabilities for a particular value of the random variable.



Frequently used Discrete probability distribution in Data Science:

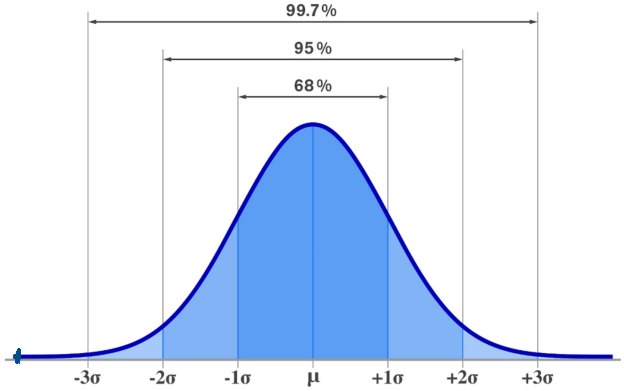
→ Bernoulli Distribution.
→ Binomial Distribution.
→ Poisson Distribution.
→ Multinomial Distribution.

Continuous Probability Distribution →

Describes the probabilities of the possible values of a continuous random variable.

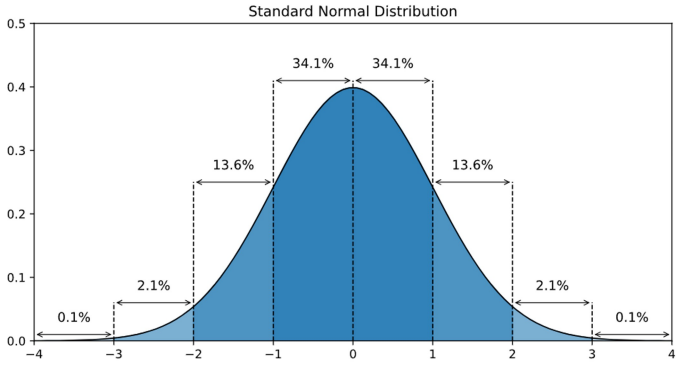
Normal Distribution →

One of the most used distribution characterized by its symmetric, bell-shaped curve.

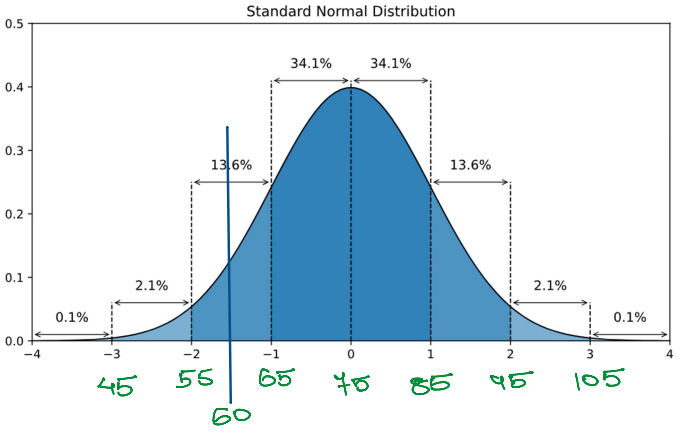


Many natural phenomenon also show normal distribution.

Standard Normal Distribution & Z-scores →



Normal Distribution can be represented as Standard Normal Distribution so that we can Calculate the percentage of data present within a range.



Suppose we have a data on salary of employees of a company where mean = 75k and standard deviation = 10k.

Here we want to find how many employees are earning upto 60k.

We will use z-score for the same:

$$z = \frac{x - \text{Mean}}{SD}$$

$$= \frac{60 - 75}{10}$$

$$= -1.5$$

Now we check for this value in Z-score

$= -1.5$

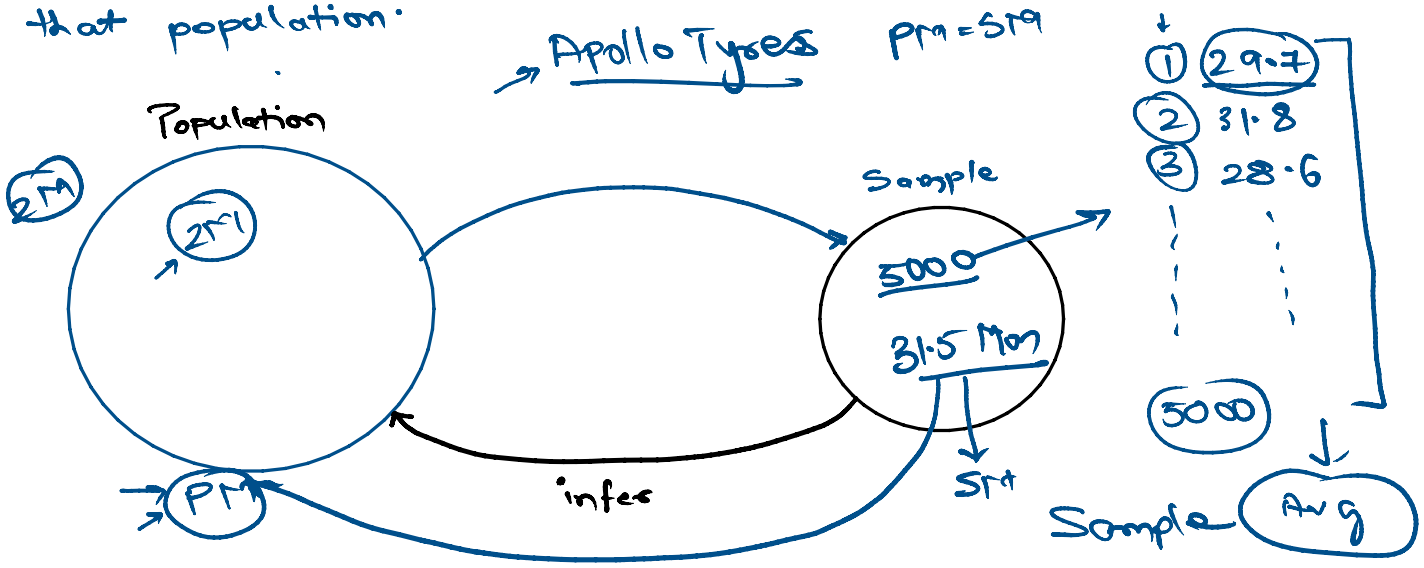
Now we check for this value in Z-score table.

$Z = -1.5 \Rightarrow 0.0668$ or 6.68%.

Z-score tells us about the percentage of data behind a certain value.

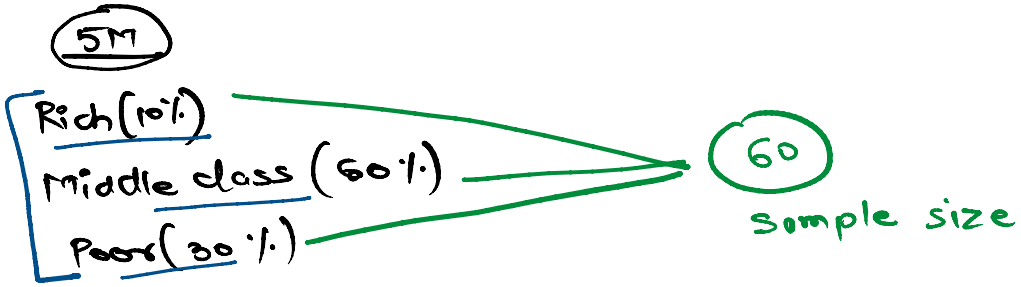
Inferential Statistics →

Involves making predictions or inferences about a popula-tion based on the sample data taken from that population.



→ We use sample mean to infer population mean.

→ The sample should be representative of the population.



Disproportionate sampling.

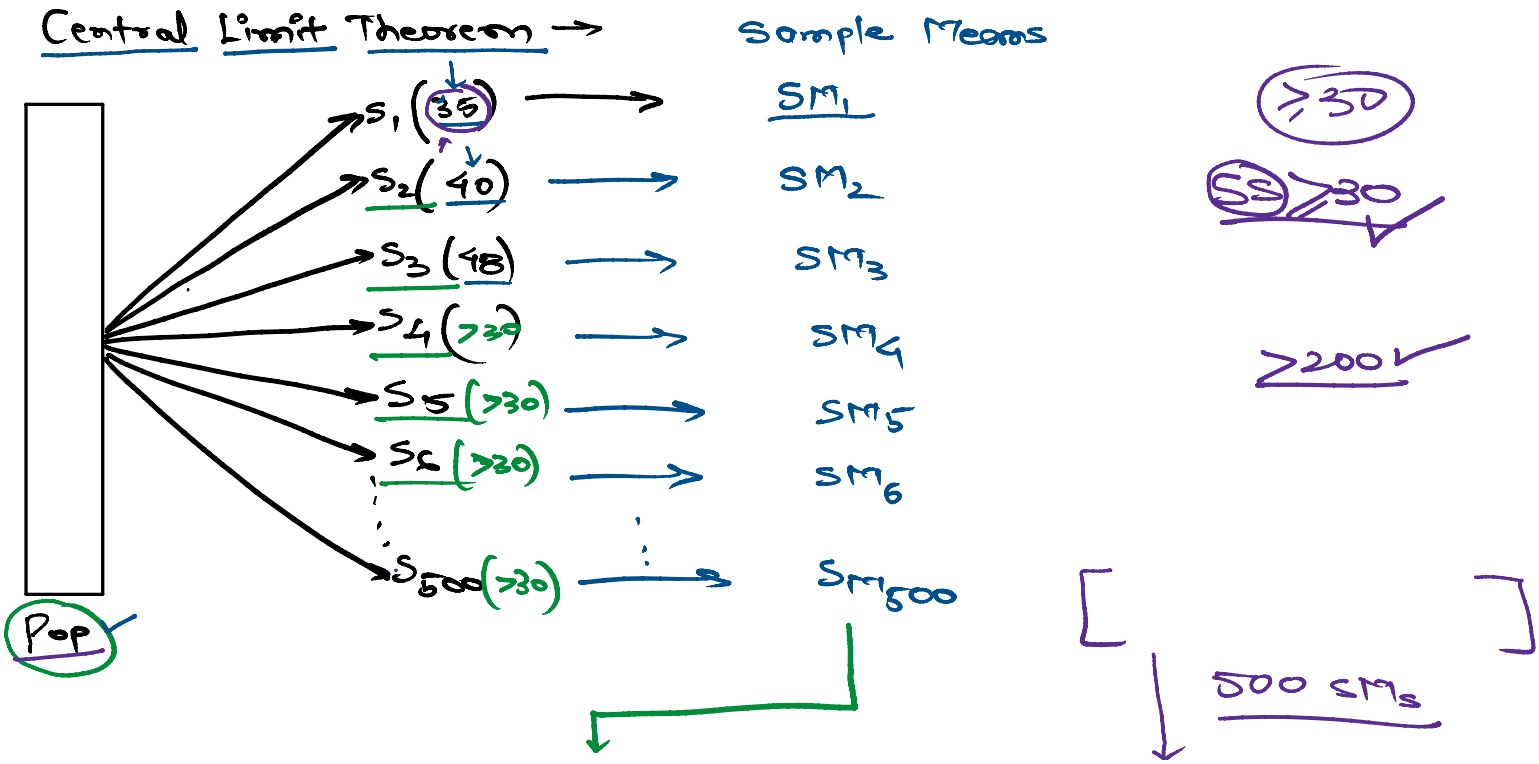
Disproportionate sampling.

5M

Rich (10%) → 6
Middle class (50%) → 36
Poor (30%) → 18

60 sample size

Proportionate Sampling.

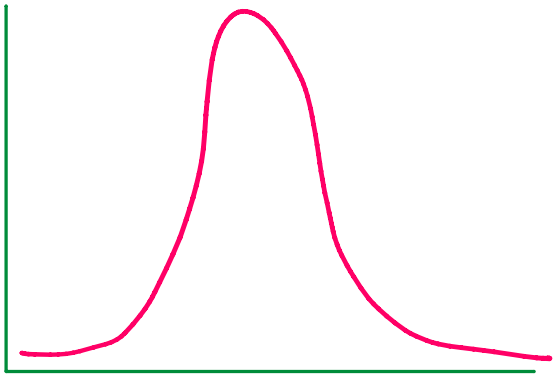


Take all the sample means & plot these value.

(sns.kdeplot or sns.displot)

↓

It will always come out to be a "Normal Distribution".



Note: Sample sizes > 30.

Takeaway: If we take one sample only from any population, we can safely assume that the sample will be 'Normally Distributed', given that the sample size > 30..

Sample = 50 Tyres

↓

Test these

1. 24.3 months.
2. 29.8 months
3. 30.5 months.
4. 25.9 months.

⋮

50. 31.9 months.

→ Plot these (kdeplot)

↓

Normal Distribution

↓

Population Mean = Sample Mean

↓

Point Estimate

Point Estimate

Mostly incorrect

With some adjustments, we can say →

Population Mean	=	Sample Mean	±	Margin

$= 24.7 \pm 1$

$= 24.7 + 1 \text{ to } 24.7 - 1$

$= 23.7 \text{ to } 25.7$

Interval / Range

90% confidence → 24.7k ± 1k

95% confidence → 24.7k ± 1.5k

99% confidence → 24.7k ± 2k

↓

Confidence Interval.

Formula to calculate the Confidence Interval →

$$CI = \underbrace{\bar{x}}_{\text{SM}} \pm \underbrace{z^* \times \frac{s}{\sqrt{n}}}_{\text{Margin}}$$

$\bar{x}$: Sample Mean

s: standard deviation of the sample.

$\bar{x}$: sample mean

S : standard deviation of the sample.

n : Sample size

z^* : z-score for a certain confidence level.

z^*	confidence level
1.65	90%. ✓
1.96	95%. ✓
✓ 2.58	99%. ✓

Ex: Estimate whether the mean lead content in maggi packets is within the allowed range or not?

≥30

Allowed range = 2.5 ppm.

$n = 100$

$\bar{x} = 2.3$ ppm

$S = 0.3$ ppm

①	2.20 ppm
②	2.43 ppm
③	2.37 ppm
④	2.21 ppm
⋮	
100	2.28 ppm

Sample data → Sample mean ($\bar{x}$)

50 pkts

100 pkts

1 pkt

SM = 2.3 ppm ✗

2.5 ppm

$$CI = \bar{x} \pm z^* \times \frac{S}{\sqrt{n}}$$

$$= 2.3 \pm 2.58 \times \frac{0.3}{\sqrt{100}}$$

$$= 2.3 \pm 2.58 \times 0.03$$

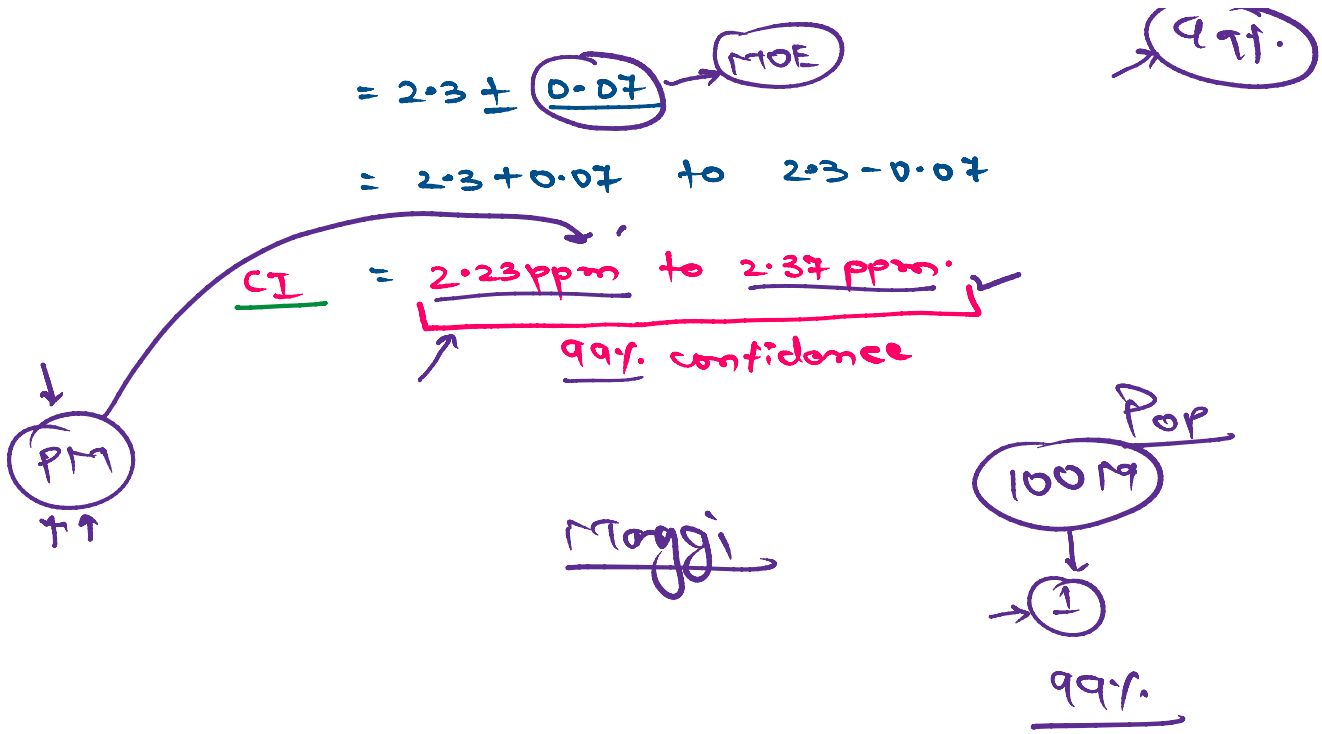
$$= 2.3 \pm 0.07$$ → MOE

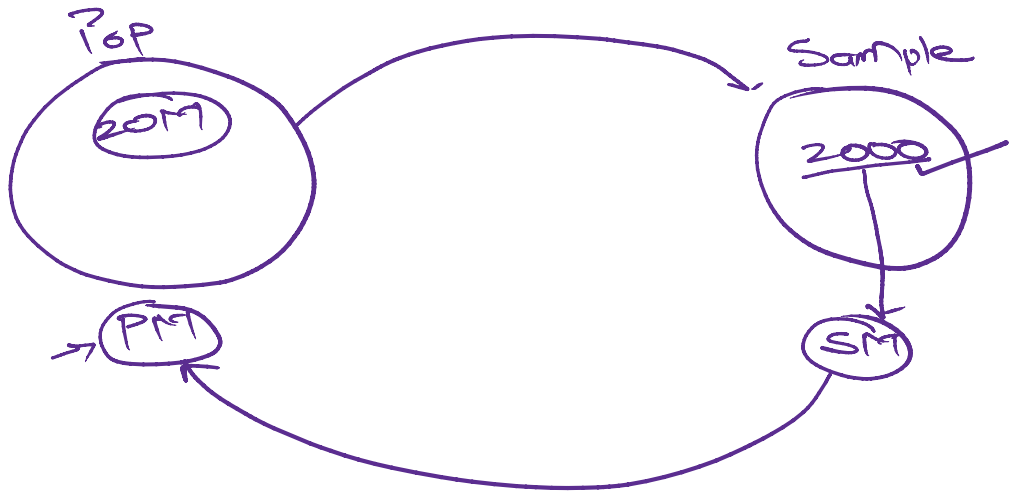
z^*

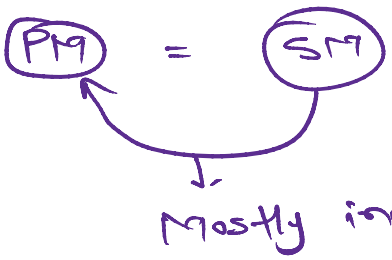
90%.

95%.

99%.







Mostly incorrect

PM = SM ± margin of error

one collection of data

Numerical

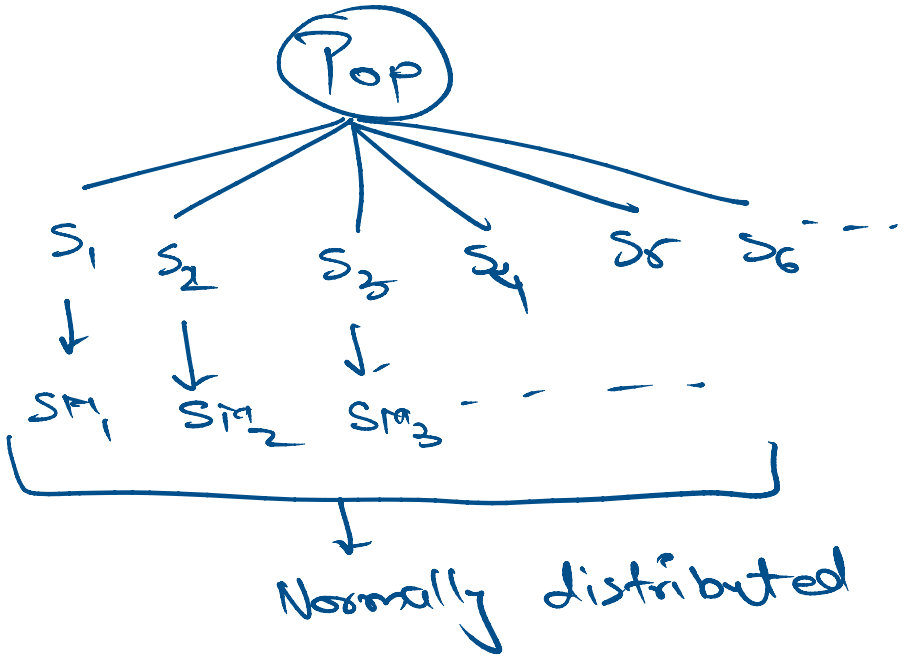
- Histogram
- kdeplot
- displot

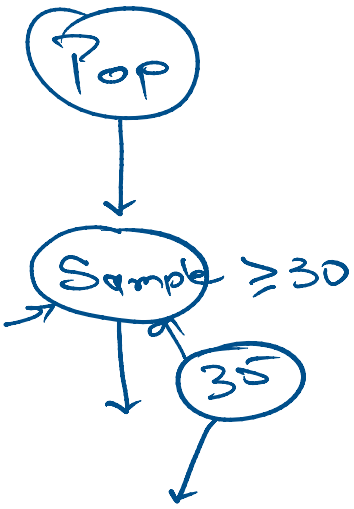
Categories

- Bar
- Pie

Gender

M F F M M F

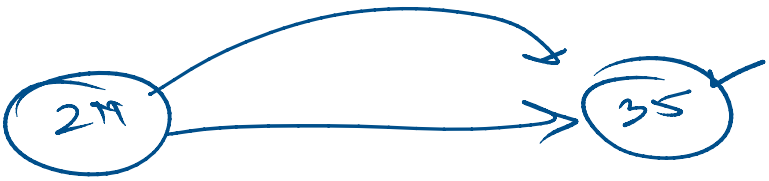




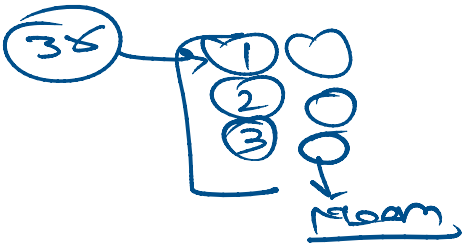
Normally Distributed.

≥30 → Through CLT

Proven



Average inc. of indian families



PM = SM ± Margin of Error

PM = SM

PM = 25K ± 1K

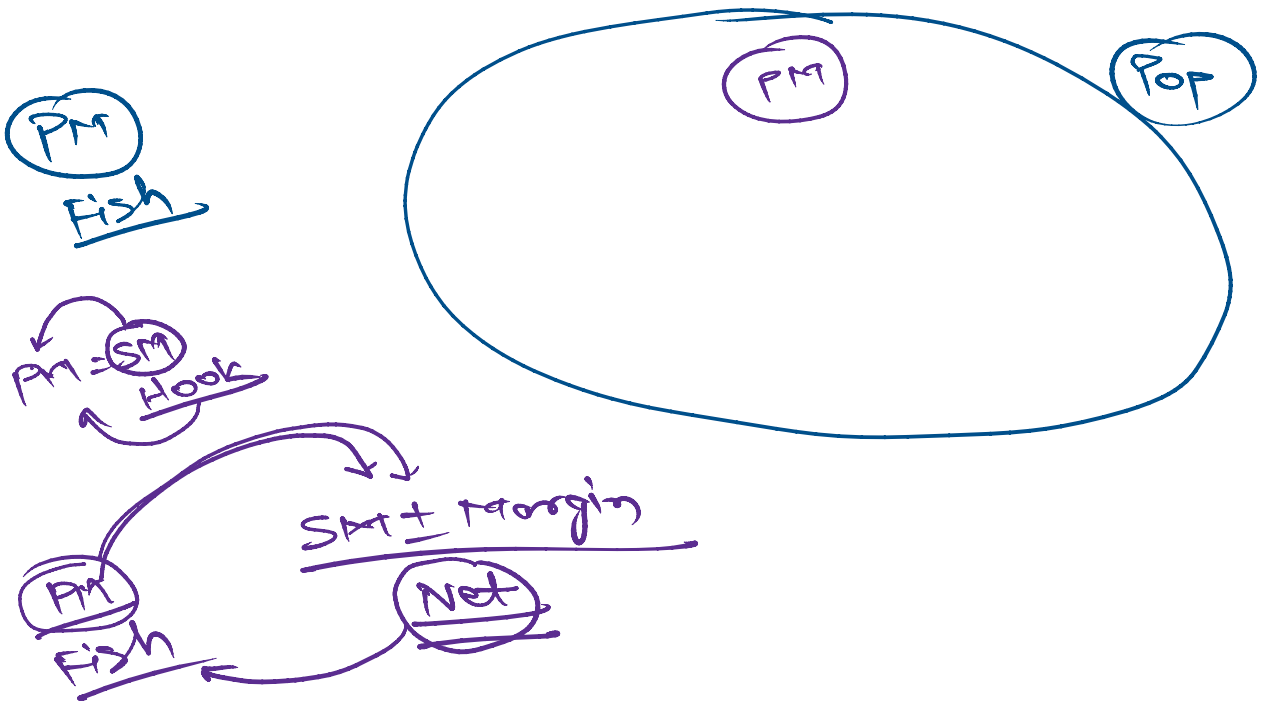
= 25-1 to 25+1

= 24K to 26K → Range/Interval

PM

PM = SM ± Margin

= 20K ± 1K



SM ± margin

90% confidence → 25K ± 1K → 24K to 26K

95% confidence → 25K ± 1.5K → 23.5K to 26.5K

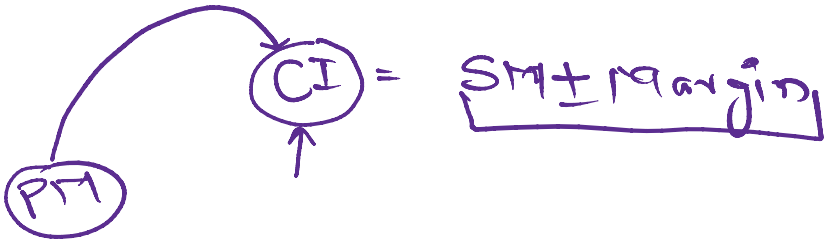
99% confidence → 25K ± 2.5K → 22.5K to 27.5K

Confidence Interval

90% + 24K to 26K

Interval

Interval



Margin / Confidence Interval:

$$CI = \bar{x} \pm z^* \times \frac{s}{\sqrt{n}}$$

→ Margin of error.

R