Inferential Statistics

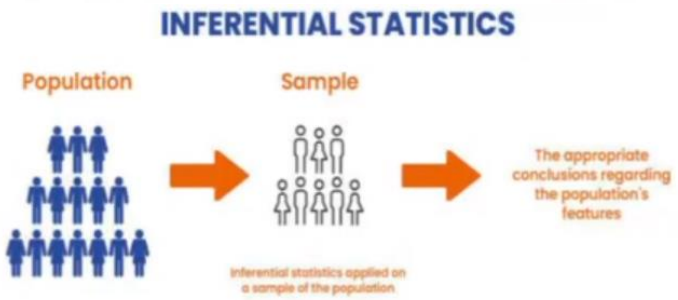
What is inferential statistics?

It is the process of generating conclusion about a population from sample(s) from the population.

What is population?

It is a set of similar items or events which is of interest for some question or experiment.

Example: ① All votters in India
② Part-time students in Bengaluru



Motivating examples:

① Who is going to win the next general election in India?
population: the votters who will vote for candidates. The votters can be in in a particular group - city, district, state.

② How likely is that 50% or more population think that Indian economy is getting worse.
population:- Indian citizens or the votters in India. Again they can be in a particular group.

To answer the above question, let us take a sample of 250 people. One can then take sample mean to arrive to a conclusion.

1. Can we directly use this sample mean as the population mean?

2. What other information do we need to know?

3. Will we be able to make 100% correct estimation? or we have to go with approximate value?

We can work this out in two ways

1. We draw whole lot of samples of size 250. Take the average of the sample mean of the each sample.

   Challenges :- Since the population involves people, takeing lots of samples may be difficult and costly. there is an element of uncertainty as how well the sample represents the population. the way the sample is taken matters.

2. Other way to work out this problem is to use probability theory.

The second approach is widely popular. For this we need to understand the fundamentals of probability and random variable.

Probability is the measure of how likely an event will occur.

Example:- We may model winning outcomes of an election for a party as if they are random when they are result of many components such as age, qualification, religion, gender etc. of the votters

Probability in this case would be the long run proportion of time winning outcome in repeated independent trial

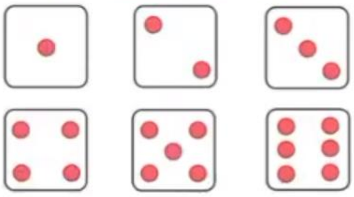
Probability Terminology:- Let us consider an example of rolling a die.

Sample space:- Sample space of an experiment consists of the set of all possible outcomes.

In the case of rolling a die, the sample space is

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

Sample Space for Rolling a Die:



6 outcomes

The integers 1 - - - 6 represent the number of dots on the six faces of the die.

Sample points:- The six possible outcomes are the sample points of the experiment.

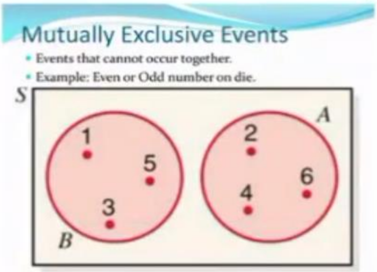
Event:- An event is a subset of 'S' and may consider any number of sample points.
For example, an event A is defined as

$$A = \{2, 4\}$$

The compliment of an event A, denoted by $\bar{A}$, Consists of all the sample points in 'S' that are not in A. Thus,

$$\bar{A} = \{1, 3, 5, 6\}$$

Mutually Exclusive events:- Two events are said to be mutually exclusive if they have no sample points in common.



Union of events:- The union (sum) of two events is an event that consists all the sample points in the two events. For the events A and B above

$$A \cup B = \{1, 2, 3, 4, 5, 6\}$$

The above implies $A \cup \bar{A} = S$ (sample space)

Intersection of events:- Intersection two events is an event that consists of the points that are common to the two events.

If $A = \{2, 4\}$, $B = \{1, 2, 3\}$

$$A \cap B = \{2\}$$

For mutually exclusive events (see fig)

$$A \cap B = \phi \text{ (Null event)}$$

Thus, $$A \cap \bar{A} = \phi$$

The concept of Union and Intersection can be extented to more than two events in a straightforward manner.

Probability rules:- Associated with each event A in S has probability of occurrence P(A), which is defined as

$$\boxed{P(A) = \frac{\text{Number of points in } A}{\text{Number of points in } S}}$$

(1) The probability of an event 'A' satisfies the condition

$$\boxed{P(A) \geq 0}$$

(2) The probability of the sample space 'S'

$$\boxed{P(S) = 1}$$

(1) and (2) imply that $\boxed{0 \leq P(A) \leq 1}$

(3) Let A_i for $i = 1, 2, \ldots$ are mutually exclusive events in 'S', that is

$$A_i \cap A_j = \phi \quad , \quad i \neq j = 1, 2, \ldots$$

Then, the probability of Union of these mutually exclusive events satisfies the condition

$$\boxed{P\left(\bigcup_i A_i\right) = \sum_i P(A_i)}$$

Example:- Let in the experiment of rolling a die each outcome is equiprobable (die is fair)

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

Let us consider an event $A = \{2, 4\}$. since the event A consists of two mutually exclusive subevents or outcomes

$$P(A) = P(2) + P(4) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Consider another event $B = \{1, 3, 6\}$

$$P(B) = P(1) + P(3) + P(4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

$$P(A \cup B) = P(A) + P(B) = \frac{1}{3} + \frac{1}{2} = \frac{5}{6}$$ because

A and B are mutually exclusive.

$$P(S) = P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1$$

Conditional probability:- It is often required to find the probability of an event B under the condition that an event A has occurred.

This prob is called the conditional probability of B given A and is denoted as

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad P(A) \neq 0$$

Similarly

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) \neq 0$$

Multiplication rule :- If A and B are events in a sample space S and $P(A) \neq 0$, $P(B) \neq 0$, then

$$\boxed{P(A \cap B) = P(A)\,P(B|A) = P(B)\,P(A|B)}$$

Independent Events :- If events A and B are such that $P(A \cap B) = P(A)\,P(B)$, they are called independent events.

Assuming $P(A) \neq 0$, $P(B) \neq 0$, we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\,P(B)}{P(B)} = P(A)$$

Similarly $P(B|A) = P(B)$

This implies that the prob of A does not depend on the occurrence or non-occurrence B. This justifies the term INDEPENDENT.

The concept can be extented for more than two events.

Example :- Consider an experiment of tossing a fair die. Let two events $A = \{1, 3, 6\}$, $B = \{1, 2, 3\}$. Find the prob $P(B|A)$

$$P(B|A) = \frac{P(A \cap B)}{P(A)},$$

The event $A \cap B = \{1, 3\} \Rightarrow P(A \cap B) = \frac{2}{6} = \frac{1}{3}$

$$P(A) = 3/6 = 1/2$$

Thus,

$$\boxed{P(B|A) = \frac{1/3}{1/2} = 2/3}$$

Random variable:- consider an experiment with a sample space 's' and sample points $s \in S$.
Let us define a function $X(s)$, which maps the outcome $s \in S$ of the experiment on the real line $(-\infty, \infty)$.
The function $X(s)$ is called a random variable.

Example ① flipping a coin
the possible outcomes are head (H) and Tail (T)

sample space $S = \{H, T\}$

let us define a function $X(s)$, $s \in S$, such that

$$X(s) = \begin{cases} 1 & s = H \\ -1 & s = T \end{cases}$$

$X(s)$ is a random variable that can take value ± 1

② Tossing a die : $S = \{1, 2, 3, 4, 5, 6\}$

A random variable defined on this sample space may be $\boxed{X(s) = s = \{1, 2, 3, 4, 5, 6\}}$

Another random variable may be
$$X(s) = s^2 = \{1, 4, 9, 16, 25, 36\}$$

Discrete random variable:- If a random variable $X(s)$ (typically denoted by X) can take finite many values, it is known as discrete random variable.
Example:- $X(s)$ corresponding to tossing a coin and rolling a die.

Continuous random variable:- many of the experiments have continuous outcomes. For example, measuring temparature of city, finding salary of an engineer with one year of experience.

The outcome in this can take any value in certain range.

In this, the sample space 's' is continuous so is the mapping $X(s) = s$.

The random variable x in this case can take any value in a certain range. Thus, x in this case know as continuous random variable.

Cumulative distribution function (CDF):- It shows distribution of the probabilities of events in an experiment. It is also know as probability distribution function.

For both descrete and continuous random variables, it is denoted by a distribution function

$$F_X(x) = P(X \leq x), \quad -\infty < x < \infty$$

The function $F_X(x)$ is called cummulative distribution function (CDF) of the random variable x.

Properties of a CDF:-

1. $0 \leq F_X(x) \leq 1$
2. $F_X(-\infty) = P(X \leq -\infty) = 0$
3. $F_X(\infty) = P(X \leq \infty) = 1$

Probability density/mass function (PDF/PMF):

The derivative of CDF function $F_X(x)$ is called probability density function (PDF) of a continuous random variable X. It is denoted by $f_X(x)$

$$f_X(x) = \frac{d}{dx} F_X(x), \quad -\infty < x < \infty$$

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$$

When a random variable is discrete, the probability mass function (PMF) is expressed as

$$p_X(x) = \sum_{i=1}^{n} P(X = x_i)\, \delta(x - x_i)$$

Here $x_1, x_2 \cdots x_n$ are the values taken by the discrete random variable X; $P(X = x_i)$, $i = 1, 2, \cdots n$ are the corresponding probabilities, and $\delta(x)$ denotes an impulse at $x=0$.

Example: (i) plot the PMF and CDF for a random variable which specifies the number of heads in the experiment of tossing a coin twice

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

The random variable $X(s)$ = number of heads

S	X = # of heads
HH	2
HT	1
TH	1
TT	0

X can take values 0, 1, 2

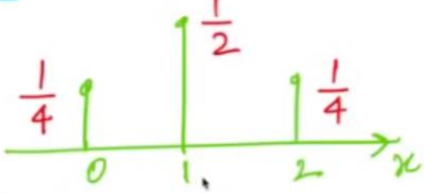
X is a discrete random variable

$$p_X(0) = P(X=0) = \frac{1}{4}$$

$$p_X(1) = P(X=1) = \frac{2}{4}$$

$$p_X(2) = P(X=2) = \frac{1}{4}$$

So, the PMF plot is



CDF $\quad F_X(x) = P(X \le x)$

$$F_X(-1) = P(X \le -1) = 0$$

$$F_X(0) = P(X \le 0) = \frac{1}{4}$$

$$F_X(0.5) = P(X \le 0.5) = \frac{1}{4}$$

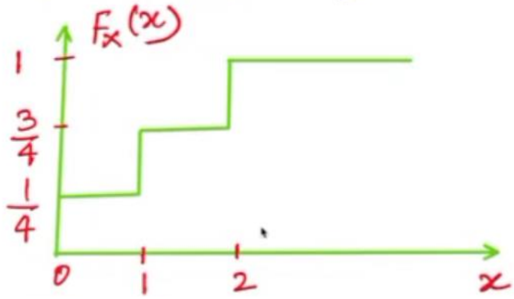
$$F_X(1) = P(X \le 1) = p_X(0) + p_X(1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

$$F_X(1.9) = P(X \le 1.9) = \frac{3}{4}$$

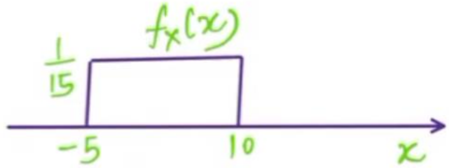
$$F_X(2) = P(X \le 2) = p_X(0) + p_X(1) + p_X(2)$$

$$= \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

$$F_X(10) = P(X \le 10) = 1$$



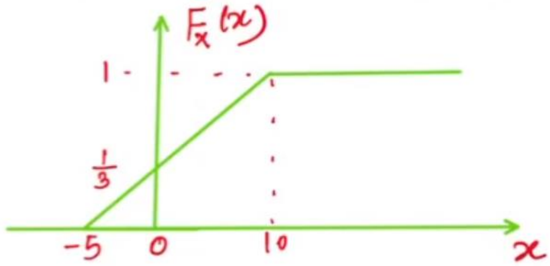
(2) For a continuous random variable X, the PDF is given below



Find the CDF?

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du = \int_{-5}^{x} \frac{1}{15}\,du$$

$$F_X(x) = \frac{1}{15}\left[u\right]_{-5}^{x} = \frac{x}{15} + \frac{5}{15} = \frac{x+5}{15}, \quad -5 \leq x \leq 10$$



Properties of PDF/PMF:

PMF: (1) $0 \leq p_X(x_i) = P(X=x_i) \leq 1$

(2) $\sum_i p_X(x_i) = 1$

CDF: (1) $f_X(x) \geq 0$ because $F_X(x)$ is a monotonically increasing function

(2) $$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)\,dx$$
$$= P(X \leq x_2) - P(X \leq x_1)$$
$$= F_X(x_2) - F_X(x_1)$$
$$= \int_{-\infty}^{x_2} f_X(x)\,dx - \int_{-\infty}^{x_1} f_X(x)\,dx$$

(3) $$P(X > x) = \int_{x}^{\infty} f_X(u)\,du$$

Statistical averages of random variable:

Mean or expected value of a random variable X is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx \quad \text{For continuous random variable}$$

$$E[X] = \sum_i x_i p_X(x_i) \quad \text{For discrete random variable}$$

It provides the long-term average of the variable.

Because of the law of large number, the value of the variable converges to the expected value as the number of repetitions approaches ∞.

$E[X]$ also known as the first moment of X.
In general, the nth moment of a random variable

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x)\, dx \quad \text{For CRV}$$

$$E[X^n] = \sum_i x_i^n p_X(x_i) \quad \text{For DRV}$$

let $E[X] = \mu$, then

$$E[(x-\mu)^n] = \int_{-\infty}^{\infty} (x-\mu)^n f_X(x)\, dx$$

is called the nth central moment of the rv X, because the moment is taken relative to the mean μ.

when $n = 2$, the central moment is called the variance of the rv X

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx \quad \text{For CRV}$$

$$\sigma_x^2 = \sum_i (x_i-\mu)^2 p_X(x_i) \quad \text{For DRV}$$

Variance σ_x^2 provides a measure of the dispersion of the random variable. In other words, it measures how much the random variable X varies from its mean.

Higher the value of σ_x^2, larger is the difference between values taken by X and its mean value.

standard deviation $\sigma = \sqrt{\sigma_x^2} = \sigma_x$

with the help of σ, we can find dispersion of values of X relative to its mean.

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx = \int_{-\infty}^{\infty} (x^2+\mu^2-2\mu x)\, f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx + \mu^2 \int_{-\infty}^{\infty} f_X(x)\, dx - 2\mu \int_{-\infty}^{\infty} x f_X(x)\, dx$$

$$\sigma_x^2 = E[X^2] + \mu^2 - 2\mu^2$$

$$\boxed{\sigma_x^2 = E[X^2] - \{E[X]\}^2}$$

Some useful Probability distributions:

① **Binomial distribution:** It is a discrete probability distribution for the number of success in a sequence of n independent experiments / trials. Outcome of each trial is success with prob p or failure with prob $1-p$.

A single success / failure experiment is called a Bernoulli trial. The prob of getting 'K' success in 'n' independent trials is given by the Binomial PMF as

$$\boxed{p_X(X=K) = \binom{n}{K} p^K (1-p)^{n-K}}$$

$\binom{n}{K} = \frac{n!}{K!(n-K)!}$ is the Binomial coefficient

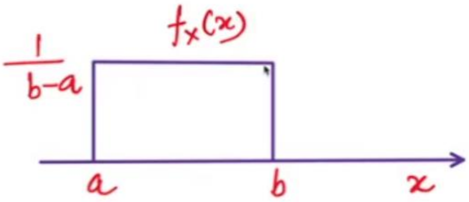
$$\boxed{E[x] = np \quad , \quad \sigma_x^2 = np(1-p)}$$

The probability distribution can be interpreted as: k success with prob p^k and $(n-k)$ failures occur with prob $(1-p)^{n-k}$. However, k success can occur anywhere among the n trials, hence there are $\binom{n}{k}$ different ways of distributing k-success in a sequence of n-trials.

② Uniform distribution: It is a distribution with equally likely outcomes. It can be for both continuous and discrete random variables.

PDF of uniform distribution is

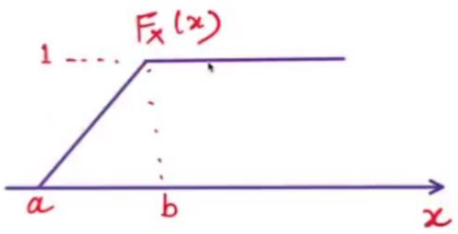
$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for} \quad a \le x \le b \\ 0 & \text{for} \quad x < a,\ x > b \end{cases}$$



The CDF

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = \frac{1}{b-a}\int_a^x du$$

$$F_X(x) = \frac{x-a}{b-a} \quad , \quad a \le x \le b$$



$$\boxed{E[x] = \frac{1}{2}(a+b) \quad , \quad \sigma_x^2 = \frac{1}{12}(b-a)^2}$$

Normal distribution :- It is the most important continuous distribution because in many applications random variables are normal random variables (they have a normal distribution) or they are approximately normal or can be transformed into normal random variable. Furthermore, it is very useful due to Central limit theorem (to be discussed later)

A random variable x is said to be Normal distributed or Gaussian distributed if it obeys the PDF

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

The Normal distribution has the following features

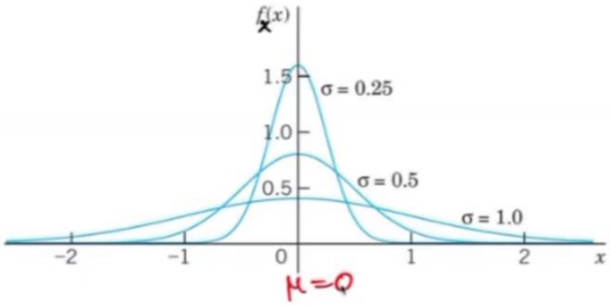
① μ is the mean and σ is the standard deviation

$$E[X] = \mu \ , \ E[(X-\mu)^2] = \sigma_x^2 = \sigma^2$$

② $\frac{1}{\sigma\sqrt{2\pi}}$ is a constant term that makes the area under the curve $f_X(x)$ from $-\infty$ to ∞ equal to 1.

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

③ The PDF is symmetric with respect to $x = \mu$. It has a bell-shaped curve.



④ The PDF goes to zero very fast as σ or σ^2 decreases.

⑤ The CDF

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}}\, du$$

Let $z = \frac{u-\mu}{\sigma} \Rightarrow dz = \frac{1}{\sigma} du$

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{z^2}{2}}\, dz$$] The result of this integral is given in Q-table.

$$\boxed{F_X(x) = Q\left(\frac{x-\mu}{\sigma}\right)}$$

The function $Q\left(\frac{x-\mu}{\sigma}\right)$ is frequeny used for finding the area under the tail of the normal distribution

6. For $\mu = 0$ and $\sigma = 1$, $f_X(x)$ is know as the standard normal distribution

$$f_X(x)\Big|_{\mu=0,\sigma=1} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty \le x \le \infty$$

The CDF of the standard normal random variable

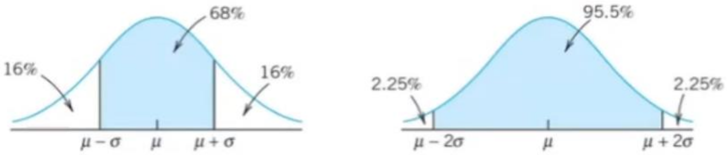
$$F_X(x) = Q(x)$$

7. In practical work with the normal distribution, it is good to remember that

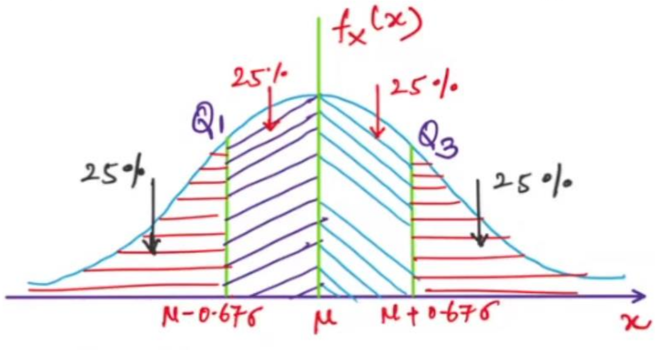
   a. $P(\mu - \sigma < X \le \mu + \sigma) \approx 68\%$

   b. $P(\mu - 2\sigma < X \le \mu + 2\sigma) \approx 95.5\%$

   c. $P(\mu - 3\sigma < X \le \mu + 3\sigma) \approx 99.7\%$



8.



Q_1: first quartile
Q_3: third quartile] — devide the entire normal distribution in 4 equal parts

Examples :- ① The bottom 25% of the distribution falls below what numer ?

Ans: $\mu - 0.67\sigma$

② compute the probability of obtaining at least two 'six' in rolling a fair die 4 times.

given fair die $\Rightarrow P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$

Prob of success $p = \frac{1}{6}$

Prof of failure $1-p = \frac{5}{6}$

The event 'at least two' six occurs if we obtain 2 or 3 or 4 'six'

Prob of getting k successes in 'n' trials is modelled using the Binomial distribution

$$p_X(x=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Prob of getting at least two successes

$$P = p_X(x=2) + p_X(x=3) + p_X(x=4)$$

$$= \binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 + \binom{4}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1 + \binom{4}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^0$$

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4\times3\times2}{2\times2} = 6$$

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4\times3\times2}{3\times2} = 4$$

$$\binom{4}{4} = \frac{4!}{4!(4-4)!} = \frac{4\times3\times2}{4\times3\times2} = 1$$

$$P = \frac{1}{6^2}\left(\frac{6\times25}{36} + 4\times\frac{5}{36} + \frac{1}{36}\right)$$

$$P = \frac{1}{6^4}(150 + 20 + 1) = \frac{171}{1296} = 0.1319$$

Central limit theorem :-

Let x_i, $i = 1, 2, \dots n$, are independent and identically distributed (iiD) random variables, each having a finite mean 'μ' and a finite variance σ^2

Let Y be defined as the normalized sum, called sample mean

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

The random variable Y is frequently encountered in estimating the mean of a random variable x from a number of observations x_i, for $i = 1, 2 \dots n$.

$$E[Y] = \mu_y = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$\boxed{E[Y] = \mu}$$

$$Var[Y] = \sigma_y^2 = E[Y^2] - \mu^2 = \frac{\sigma^2}{n}$$

$$\boxed{\sigma_y^2 = \frac{\sigma^2}{n}}$$

When Y is viewed as an estimate for the mean μ, we note that its expected value is μ and its variance is decreases as n increases. It means uncertainty in the estimate of μ decreases as n increases.

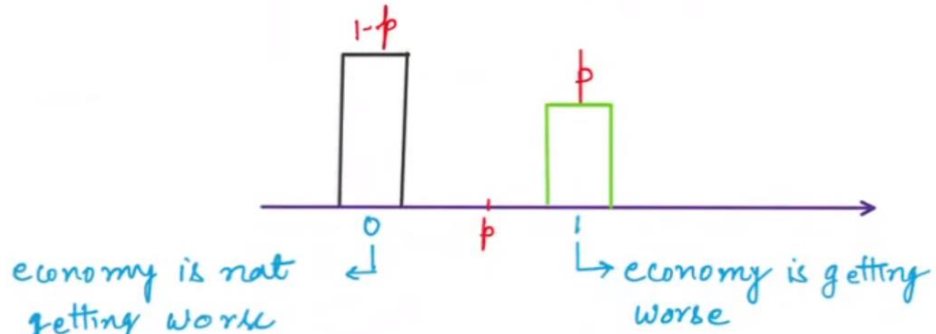
The central limit theorem states that the sum of statistically independent and identically distributed (IID) random variable approached to a Normal (Gaussian) distribution as $n \to \infty$.

In more simple form, the central limit theorem states that that if we take sufficiently large number of samples from a population, the sample mean is Normally distributed even if the population is not Normally distributed.

Since we have understood probability, random variable, central limit theorem, we can now solve the problem stated at the starting.

Problem: Assume that 40% population in India says that the economy is getting worse. If we take a sample of 250 people, how likely is that 50% or more of them will say they think the economy is getting worse. Calculate 99% confidence interval for proportion of the population who felt that the economy is getting worse.

Solution: We will not be able to survey the entire population to answer this question. But the entire population can be put in two buckets as follows



This is a binomial distribution.
Let we derive a sample of 250 people, in which 108 say the economy is getting worse.

We have 250 samples with in a sample

The sample mean $\bar{x} = \dfrac{1(108) + 0(142)}{250}$

$$\bar{x} = \frac{108}{250} = 0.432$$

sample variance $\sigma_1^2 = \dfrac{108(1-0.432)^2 + 142(0-0.432)^2}{250-1}$

$$\sigma_1 = \sqrt{0.246} = 0.4963$$

Recall that the variance of the samples

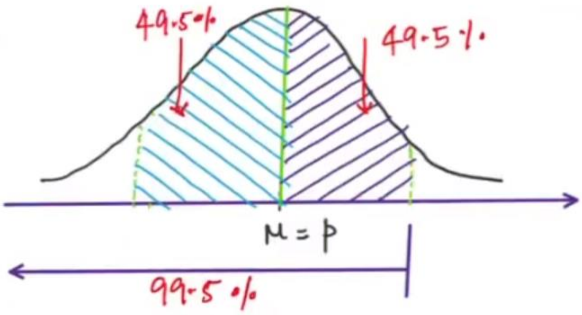
$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{\sigma^2}{250}$$

We do not know σ, so

$$\sigma_{\bar{x}} \approx \frac{\sigma_1}{\sqrt{250}} = \frac{0.4963}{\sqrt{250}} = 0.0314$$

We want a 99% confidence interval, which communicate how accurate our estimate is likely to be.

Using CLT, we can say that $\bar{x}$ is comming from a Normal distribution with mean $\mu = \bar{x}$ and $\sigma_1 = \sigma/\sqrt{n}$.



99% confidence interval means, how many standard deviation away from the mean we have to be so that we are 99% confident that any sample from this sampling distribution will be in that interval.

from the z-table, $0.5 + 0.495 = 0.995$ corresponds to $z = 2.58$. Thus, $2.58\,\sigma_{\bar{x}}$ away from the mean μ would give us the 99% confidence interval.

In other words, 99% chance that a random $\bar{x}$ is with $2.58\,\sigma_{\bar{x}}$ of $\mu = p$.

We are 99% confidence that p is with in

$2.58\,\sigma_{\bar{x}} = 2.58 \times 0.0314 = 0.0810$ of 0.432

i.e. 0.432 ± 0.08

upper $0.432 + 0.08 = 0.513$

$0.432 - 0.08 = 0.352$

We are 99% confident that true population proportion is within the range 35.2% to 51.3%.

or

the true %age of the people who think that the economy is getting worse is in the range 35.2% to 51.3%. There is 99% chance of this.