Theory SI Introduction

Friday, 27 June 2021 8:17 PM

Motivating examples:

1. Who will win the next general election in India
2. Investigate people's belief:
   - (a) Do you believe in life after death?
   - (b) Would you be willing to pay higher prices to protect environment?
   - (c) How much TV do you watch per day?

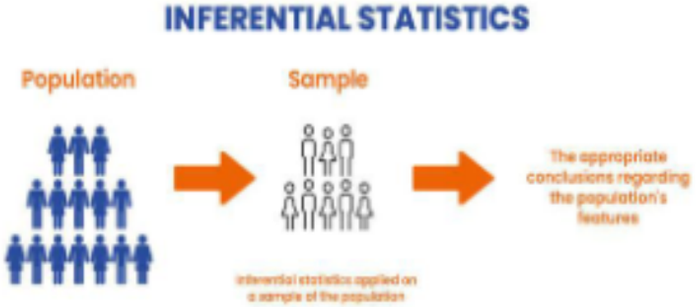
Inferential statistics:— The word infer means to arrive at a decision or prediction by reasoning from know evidence. Statistical inference does this using data as evidence.

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

Sample versus population:— population is a total set of similar items or events which is of interest for some statistical question or experiment.

Example: All voter in India

A sample is a subset of the population for whom we have (or plan to have) data, often randomly selected.



How does inferential statistics answer a statistical question?

Let us consider a question

How likely is that 50% or more population think that Indian economy is getting worse.

Population: All voters in India, All graduates in economics in India etc.

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 Will we be able to make 100% correct estimate 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, taking 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.

Theory SI Probability

Friday, 27 June 2025 8:23 PM

Probability :- With any random phenomenon, the probability of a particular outcome is the proportion of times that outcome would occur in long run of observations.

Example : A weather forecaster might say that probability of rain today is 0.70. This means that in a large number of days with atmospheric condition like those today, the proportion of days in which rain occurs is 0.70.

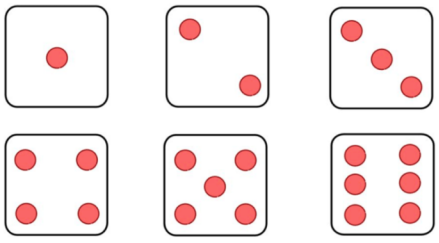
Sometimes probabilities are expressed as percentage, such as the weather forecaster might say that the probability of rain is 70 %.

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 $S = \{1, 2, 3, 4, 5, 6\}$ — denoted by S.

Sample Space for Rolling a Die:



6 outcomes

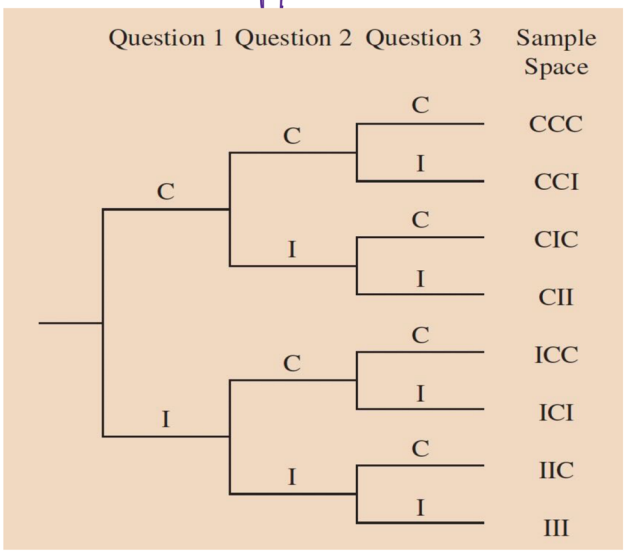
The integers 1 --- 6 represent the number of dots on the six faces of the die.

Real-life example for sample space :- Let a statistics instructor gives a pop quiz with three multiple choice questions.

The student's answer is either correct (C) or incorrect (I). For example, if a student answered the first two questions correctly & the last question incorrectly, the student's outcome on the quiz can be symbolized by CCI.

What is the sample space of the pop quiz?

Solution :- one technique for listing the possible outcomes for finding the sample space is to draw a tree-diagram, with branches showing what can happen on subsequent trials.



From the diagram, a student's performance has eight possible outcomes. Thus, the sample space is

{ CCC, CCI, CIC, CII, ICC, ICI, IIC, III }

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

Events:- An event is a subset of 'S' and may consider any number of sample points.

For the pop quiz example, an event may be

A = student answer all three question correctly = {ccc}

B = student passes (at least two questions are correct)

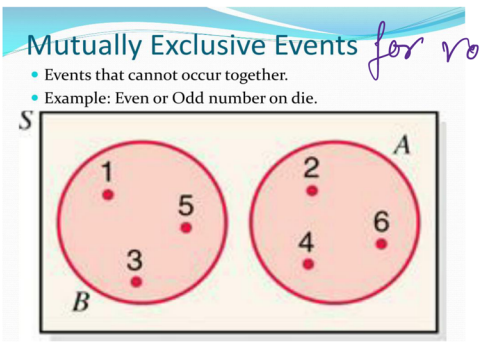
= {ccI, cIc, Icc, ccc}

Complement of an event:-

The compliment of an event A, denoted by $\bar{A}$, consists of all the sample points in 'S' that are not in A. If $A = \{2, 4\} \Rightarrow \bar{A} = \{1, 3, 5, 6\}$

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

Example;



for rolling a die example.

The above two events are mutually exclusive or disjoint events.

Union of events:- The union (sum) of two events is an event that consists all the sample points in the two events. Let $A = \{1, 2\}$, $B = \{5, 6\}$

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

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 $A = \{2,4\}$ and $B = \{1,2,3\} \Rightarrow A \cap B = \{2\}$

For mutually exclusive events (see fig)

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

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

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

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

① The probability of an event A satisfies the condition

$$P(A) \geq 0$$

② The probability of the sample space 'S'

$$P(S) = 1$$

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

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

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

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

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

Addition rule of probability

If two events A and B are not disjoint or mutually exclusive

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example: Consider a family with two children. Let F and M denote child gender for female and male respectively (a) find the sample-space

(b) find the prob of the event A = first child is a girl

(c) find the prob of the event B = second child is a girl

(d) find the prob of the event $A \cup B$

Solution:- (a) Sample space $S = \{FF, FM, MF, MM\}$

(b) $A = \{FF, FM\} \Rightarrow P(A) = 2/4 = 0.5$

(c) $B = \{FF, MF\} \Rightarrow P(B) = 2/4 = 0.5$

(d) $A \cup B = \{FM, MF, FF\} \Rightarrow P(A \cup B) = 3/4 = 0.75$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$A \cap B = \{FF\} \Rightarrow P(A \cap B) = 1/4 = 0.25$

$$P(A \cup B) = 0.5 + 0.5 - 0.25 = 0.75$$

Theory SI Random Variable

Friday, 27 June 2025 8:33 PM

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.

In other words, A random variable is a numerical measurement of the outcome of a random phenomenon.

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

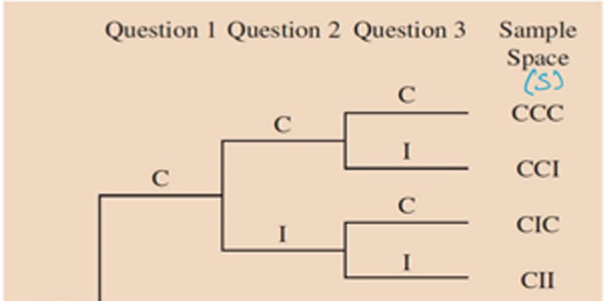
$$X(s) = \begin{cases} 1 & s = H \\ -1 & s = T \end{cases}$$

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

Types of random variable
- Discrete random variable
- continuous random variable

Discrete random variable :- If a random variable $X(s)$ [Typically denoted by X] takes finite many values, it is known as discrete random variable.

Example 1 :- Recall the pop quiz example having sample space given below.



$S = \{CCC, CCI, CIC, CII, ICC, ICI, IIC, III\}$

$X = \{7, 6, 5, 4, 3, 2, 1, 0\}$ → discrete random variable.

Example 2: X corresponding to tossing a coin and rolling a die.

Probability distribution of discrete random variable

Recall that a discrete RV takes finite many distinct values. Its probability distribution assigns a probability to each possible value, and the sum of probabilities of all possible values equals 1.

It is denoted by $p_X(x)$, which gives the probability that the random variable X takes a value x.

Example:- plot the probability distribution of the RV X that count the number of correct questions in the pop quiz exam. Assume each outcome to be equally likely.

Sample space (S)	X = # of correct questions
CCC	3
CCI	2
CIC	2
CII	1
ICC	2
ICI	1
IIC	1
III	0

The values taken by the random variable $X = 0, 1, 2, 3$

$$p_X(0) = P(X=0) = \frac{1}{8} = 0.125$$

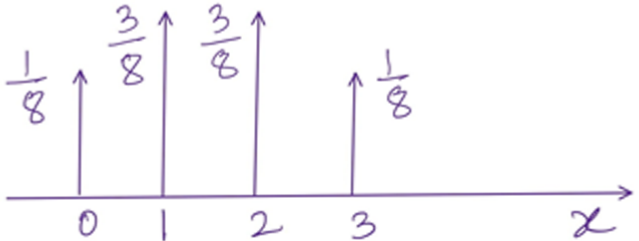
$$p_X(1) = P(X=1) = \frac{3}{8} = 0.375$$

$$p_X(2) = P(X=2) = 3/8 = 0.375$$

$$p_X(3) = P(X=3) = \frac{1}{8} = 0.125$$

Note that $p_X(0) + p_X(1) + p_X(2) + p_X(3) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$

Plot of prob distribution



once you have prob distribution, you can answer the questions like

(a) Prob that a student passes (at least two questions are correct)

$$= P(X \geq 2) = p_X(2) + p_X(3) = \frac{3}{8} + \frac{1}{8} = \frac{4}{8}$$

Continuous Random Variable: Many experiments have continuous outcomes. For example,

(a) Time that people take to commute to work in Bengaluru

(b) Salary of an engineer with one year of experience

The outcome in this can take any value in a range. Thus 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. If a random variable can take any value in a range, it is know as continuous random variable.

Probability distribution of continuous RV:-

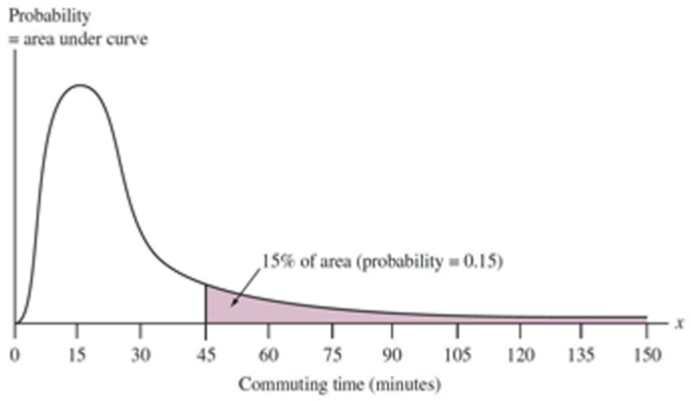
The probability distribution of a continuous random variable describes how probabilities are distributed over the range of possible values the variable can take.

It is represented by a probability density function (pdf), $f_x(x)$, which satisfies two key properties.

1. $f_x(x) \geq 0$ for all x, ensuring probabilities are non-negative
2. The total area under the curve of $f_x(x)$ over the variable's range equals 1.

Fig shows the curve for the probability distribution
X = Commuting time to work in the united state.

Fig shows the curve for the probability distribution
x = Commuting time to work in the united state.



The area of the shaded portion gives the prob that commute time is more than 45 minutes. Thus

$$P(X > 45 \text{ minutes}) = \int_{45}^{150} f_X(x)\,dx = 0.15$$

Statistical averages of random variable:

Mean or expected value of a random variable X is

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

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

It provides the long-term average of the variable.

$E[X] = M$ also known as the first moment of X.

The variance of a rv X

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

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

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.

Cummulative Distribution Function (CDF):

The CDF, denoted by $F_x(x)$, gives the probability that the random variable X takes value less than or equal to x.

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_x(t)\,dt \quad \text{for CRV}$$

$$F_X(x) = P(X \le x) = \sum_{x_i \le x} P(X = x_i) \quad \text{for DRV}$$

x_i : ith value taken by the DRV X.

Example 1: Find the probability that the commute time in the USA is less than equal to 45 minutes.

We need to find $P[X \le 45] = \int_{-\infty}^{45} f_x(x)\,dx$

$$= 1 - P[X > 45] = 1 - \int_{45}^{150} f_x(x)\,dx$$

$$P[X \le 45] = 1 - 0.15 = 0.85.$$

Example 2: In the pop quiz example, find the prob that a student fails in the exam.

Recall, condition of passing the exam = responses to atleast two questions are correct.

Prob that a student fails in the exam = $P[X \le 1]$

Example2: In the pop quiz example, find the prob that a student fails in the exam.

Recall, condition of passing the exam = responses to atleast two questions are correct.

Prob that a student fails in the exam = $P[X \leq 1]$

$$P[X \leq 1] = \sum_{x_i \leq 2} P[X = x_i] = P[X=0] + P[X=1]$$

$$P[X \leq 1] = \frac{1}{8} + \frac{3}{8} = \frac{4}{8}$$

or $P[X \leq 1] = 1 - P[X > 1.2] = 1 - \frac{4}{8} = \frac{4}{8}$.

Properties of a CDF:

1. $F_X(x)$ is monotonically non-decreasing function
   $F_X(x_1) \geq F_X(x_2)$ if $x_1 > x_2$
2. $0 \leq F_X(x) \leq 1$ for all x.
3. $F_X(\infty) = P(X \leq \infty) = \int_{-\infty}^{\infty} f_X(t)\, dt = 1$
4. $F_X(-\infty) = \int_{-\infty}^{\infty} f_X(t)\, dt = 0$
5. If x is continuous, the pdf can be obtained $f_X(x) = \frac{d}{dx} F_X(x)$

Some useful Probability distributions:

1. **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

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

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-M)^2}{2\sigma^2}},$$

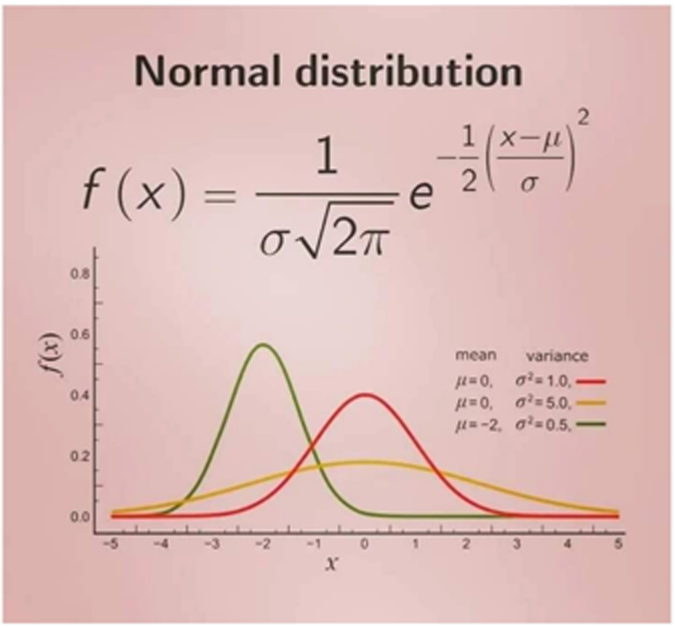
where M is the mean and σ is the standard deviation

$$E[X] = M, \quad E[(X-M)^2] = \sigma_x^2 = \sigma^2, \text{ and}$$

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

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

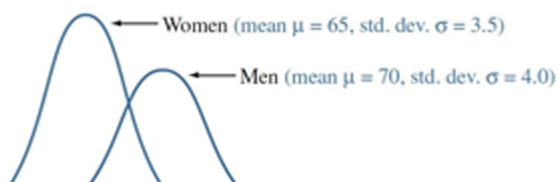
The PDF is symmetric with respect to $x = M$. It has a bell-shaped curve.



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

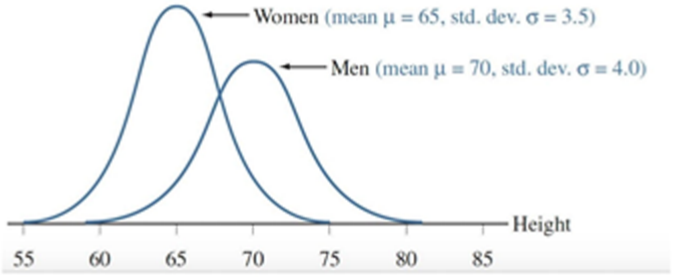
some real-life example for normal distribution :-

(a) Heights of female and male students in the US follow a bell-shaped distribution. This can be approximated to a normal distribution as shown below.



Some real-life example for normal distribution :-

(a) Heights of female and male students in the US follow a bell-shaped distribution. This can be approximated to a normal distribution as shown below.



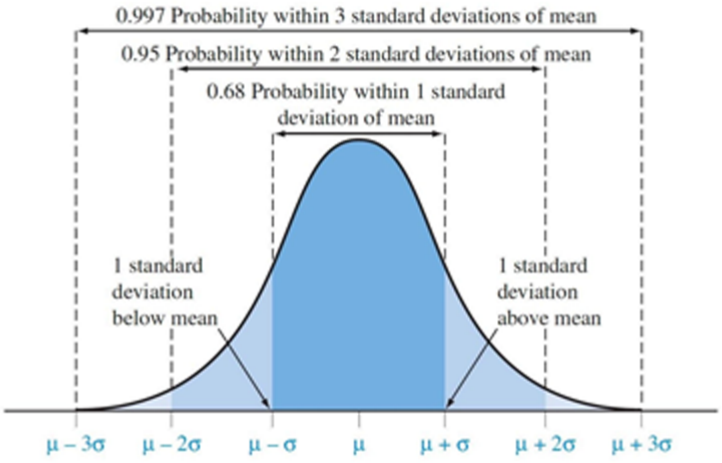
Important properties of normal distribution :-

(1) In practical work with the normal distribution, it is good to remember the probabilities within 1, 2, and 3 standard deviation away from the mean.

(a) $P(M-\sigma < X \leq M+\sigma) \approx 68\%$

(b) $P(M-2\sigma < X \leq M+2\sigma) \approx 95.5\%$

(c) $P(M-3\sigma < X \leq M+3\sigma) \approx 99.7\%$



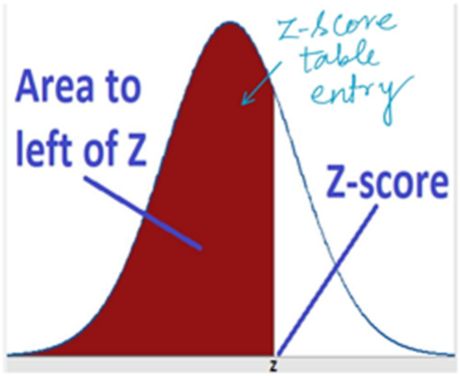
The multiple of standard deviation from the mean is denoted by a symbol z. For example, $z=2$ is 2 standard deviation away from the mean.

For a given z, the prob within z standard deviations

For a given z, the prob within z standard deviations away from the mean is the area under the normal curve between $\mu - z\sigma$ to $\mu + z\sigma$. This prob. is 0.68 for $z = 1$, as shown in Fig.

How to calculate z for a certain probability?
It is calculated using z-score table. A part of this table is shown below. Table entry for z is the area under the standard normal curve to the left of z.

Second Decimal Place of z										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
...										
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9139	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441



For example, let us find the value of z for cumulative prob of 0.923. From the table, this corresponds to $z = 1.4 + 0.03 = 1.43$. Thus, to get a cummulative prob of 0.923, we have to move 1.43σ away from the mean.

$$\int_{\mu - 1.43\sigma}^{\mu + 1.43\sigma} f_x(x)\, dx = 0.923$$

where $f_x(x)$ is the normal distribution.

For a given cummulative probability, one can find z-score from the table!

(2) Binomial distribution: (Probabilities when each observation has two possible outcomes): In many applications, each observation is binary (has two possible outcomes).
For example, a person may
(a) accept or decline an offer from a bank for a credit card
(b) have or not have health insurance
(c) vote yes or no in a referendum, such as whether to provide additional funds to school.

(2) Binomial distribution: (Probabilities when each observation has two possible outcomes): In many applications, each observation is binary (has two possible outcomes).
For example, a person may
(a) accept or decline an offer from a bank for a credit card
(b) have or not have health insurance
(c) vote yes or no in a referendum, such as whether to provide additional funds to school.

Binomial distribution is a prob distribution for a discrete random variable that takes two values.

Consider n cases, called trials, in which we observe a binary random variable. Let 'k' denote the number outcomes of interest in n trial.

The prob of getting 'k' success in 'n' independent trials is given by the Binomial distribution as

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

where p is the prob of success and

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient

$$E[X] = np \quad , \quad \sigma_x^2 = np(1-p)$$

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.

Central limit theorem :-

Let X_i, $i = 1, 2, \cdots n$, are independent and identically distributed (iid) random variables, each having

Let X_i, $i = 1, 2, \cdots n$, 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 $i = 1, 2 \cdots n$.

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

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

$$\boxed{\mu_y = \mu \;, \quad \sigma_y^2 = \frac{\sigma^2}{n}}$$

When y is viewed as point estimate for the mean μ, we note that its expected value is μ and its variance is 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 $n \to \infty$.

$$\boxed{y \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$$

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.

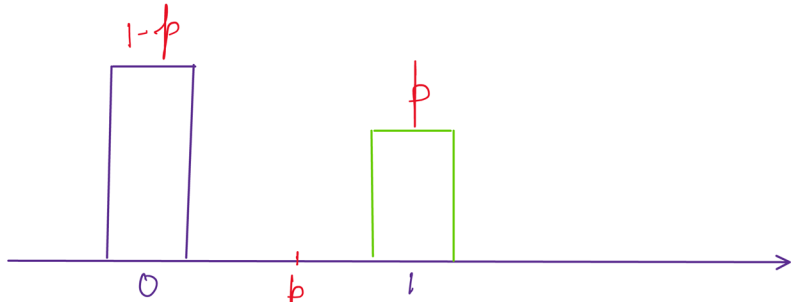
Theory SI Problem Solving

Friday, 27 June 2025 8:45 PM

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



economy is not getting worse

economy is getting worse

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 $\bar{x} = \frac{1(108) + 0(142)}{250}$

$\bar{x} = \frac{108}{250} = 0.432$ (point estimate of the population who think that the economy

250 population who think that the econ is getting worse)

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

$$\sigma_1^2 = \frac{34.8434 + 26.5}{249} = \frac{61.344}{249} = 0.246$$

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

We want a 99.1. confidence interval.

Confidence interval :- It is a range of values that likely contains the true population parameter, with a spacified level of confidence (like 95.1. or 9

Formula for a confidence interval (for population me

$$CI = \bar{x} \pm z\left(\frac{\sigma}{\sqrt{N}}\right)$$

where $\bar{x}$ = sample mean
z = Z-score for the desired confidence le
σ = population standard deviation (or Sam standard deviation if σ is unknown)
N = sample size

In our case, where $N = 1$ and σ is the standard deviation of the population.

Using CLT, we can say that $\bar{x}$ is comming from a Normal distribution with mean $\mu = \bar{x}$ and $\sigma_{\bar{x}} = \sigma/\sqrt{25}$

Recall that the variance of the sample mean

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

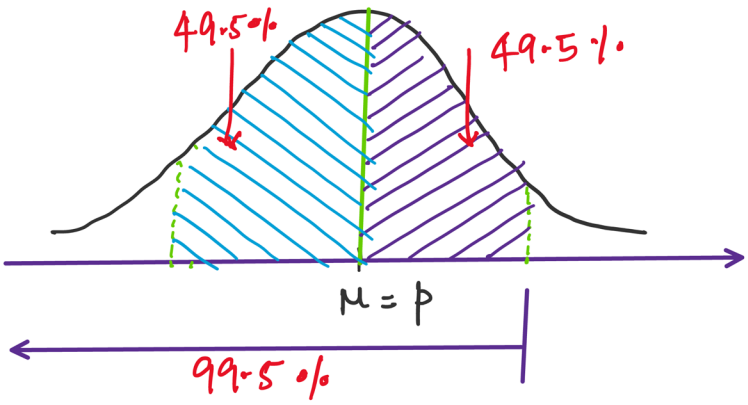
Recall that the variance of the sample mean

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

We do not know σ, so

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

For calculating the CI, we need to find Z.



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.

Thus the 99% CI $= \bar{x} \pm 2.58\,\sigma_{\bar{x}}$

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

CI upper $0.432 + 0.08 = 0.513$

CI lower $0.432 - 0.08 = 0.352$

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.