Sane's STATS Academy of Statistics

Statistical Inference

Essentials

Essential Terms

- Independence of Variables
- Independent Identically Distributed Variables
- Law of Large Numbers
- Types of Sampling
- Sampling Distribution
- Central Limit Theorem

Sane's STATS Academy of Statistics

Independence of Variables

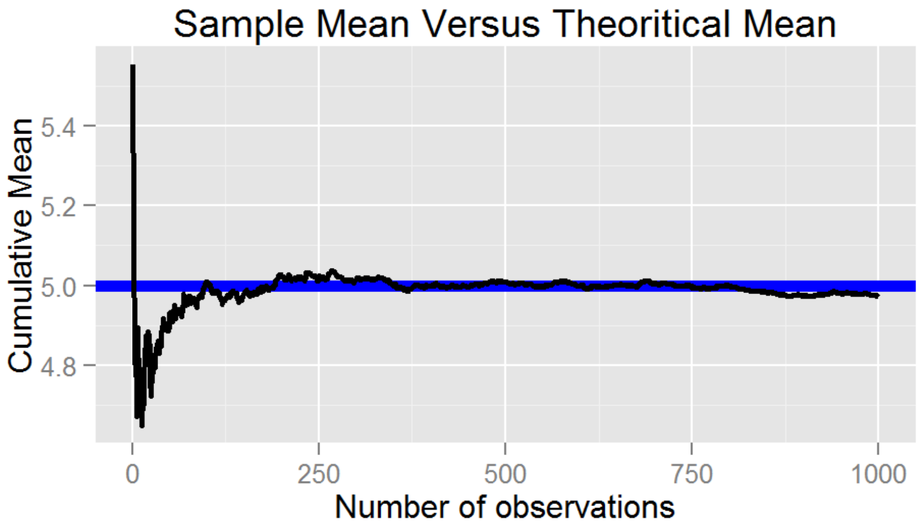
- The variables are said to be independent if any of the variable taking any value does not depend on the values of other variables
- e.g. In a table of employees, Variable ***Gender*** has values Male and Female, Variable ***City*** has values Pune, Bangalore and Hyderabad. Here ***Gender*** and ***City*** residing are independent of each other.

Identically Distributed Variables

- Identically Distributed variables are the variables with same probability distribution
- Independent Identically Distributed variables (iid) are the variables with same probability distribution and are mutually independent
- e.g. Variables X_1, X_2, X_3 all following same Normal Distribution with mean 90 and variance 140, where X_1, X_2, X_3 represent the scores of three batsmen

Law of large Numbers

- Law of large numbers states that, as the number of identically distributed, randomly generated observations increases, their sample mean (average) approaches their theoretical mean.
- The law of large numbers was first proved by the Swiss mathematician Jakob Bernoulli in 1713.



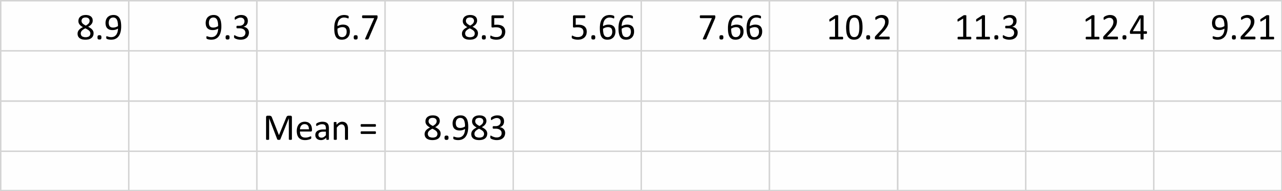
Sampling Distribution

A Sample

- When we draw a sample and study it, we find its characteristics by calculating its measures.
- We may calculate its mean and standard deviation.
- By calculating the measures, we intend to find the estimates of corresponding population parameters.

Example

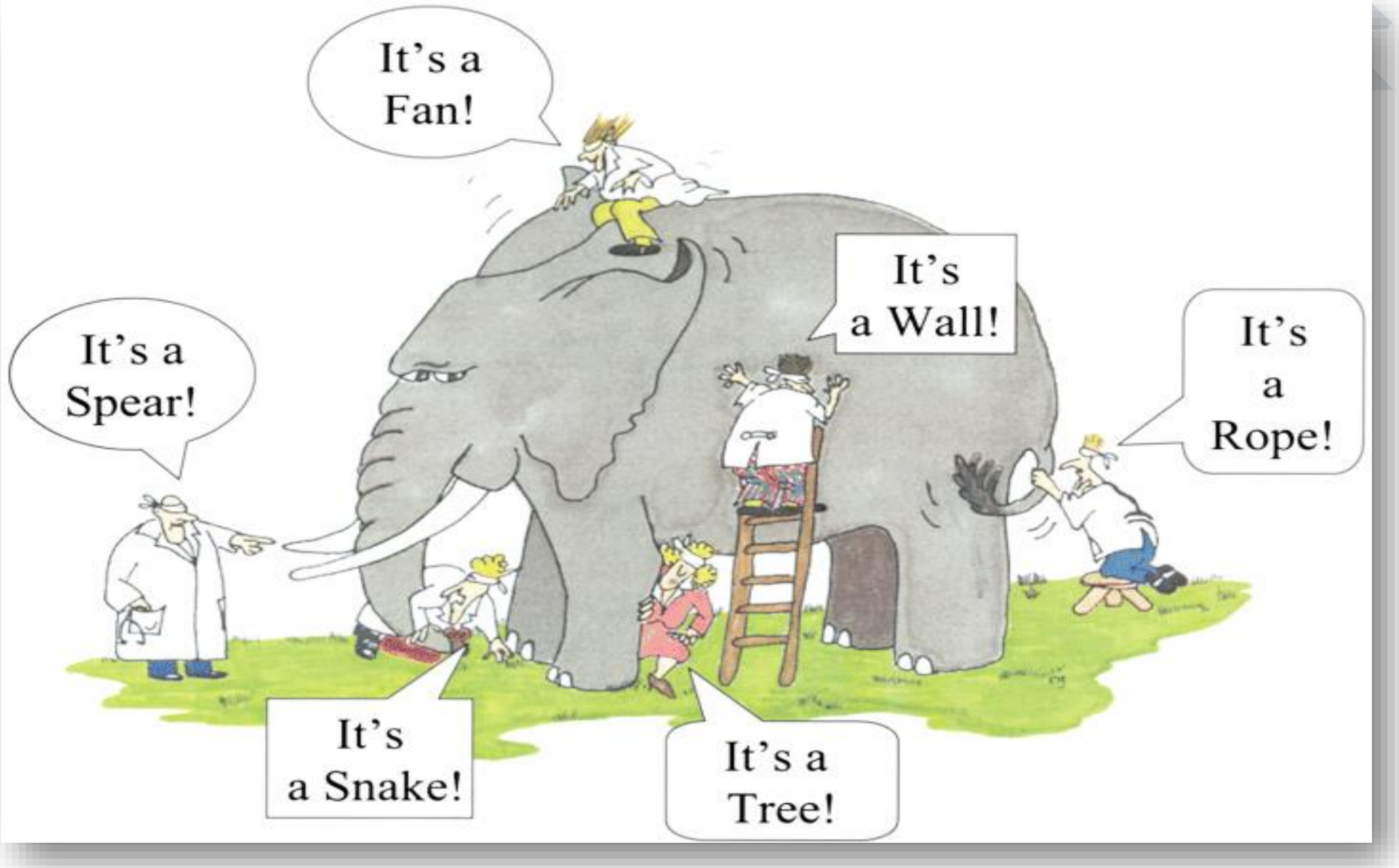
- Consider that we want to estimate the average salaries of Oracle Functional Consultants working in IT industry in India with 3-4 years experience.
- So we draw a random sample of size 10 and calculate the measures.



8.9	9.3	6.7	8.5	5.66	7.66	10.2	11.3	12.4	9.21
		Mean =	8.983						

- Hence we have the sample mean as 8.983 lakhs.

To what extent can we believe in the sample?



Our sample
Mean = 8.983

Sample from person B:
Mean = 9.683

Hence we see that if the experiment is performed by different people we get different values for the sample mean.

Sample from person C:
Mean = 10.09

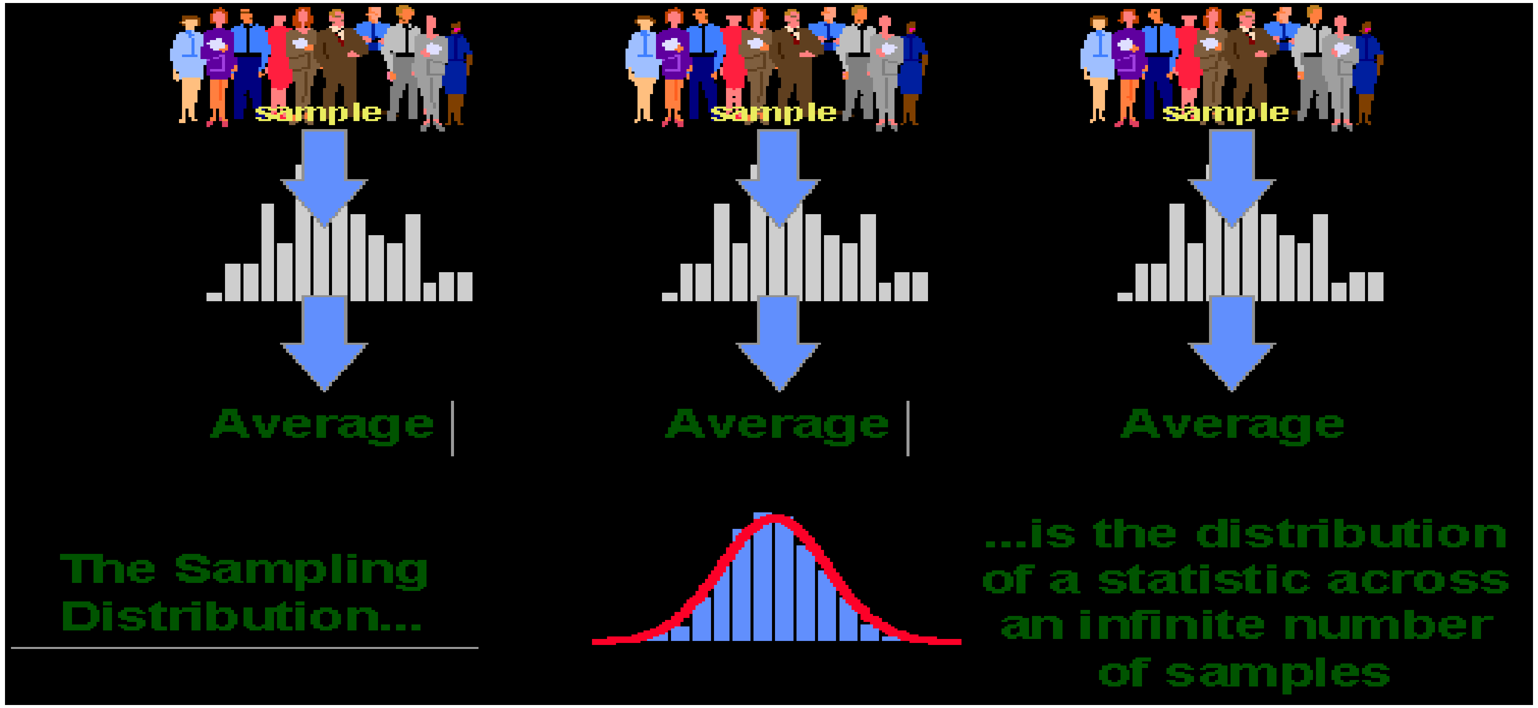
Sample from person D:
Mean = 8.77

Sample from person A:
Mean = 8.64

Each Sample consists of 10 observations

So we implies that…

- The values from different samples namely by persons A, B, C, D etc. also follow a random pattern.
- This pattern of randomness is the sampling distribution.


sample
sample
sample
Average
Average
Average
The Sampling Distribution…
…is the distribution of a statistic across an infinite number of samples

Population and Sample Notations

	Population	Sample
Mean	µ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s

Population Parameters

Statistics

A Statistic is said to be an estimator of a population parameter.

Central Limit Theorem

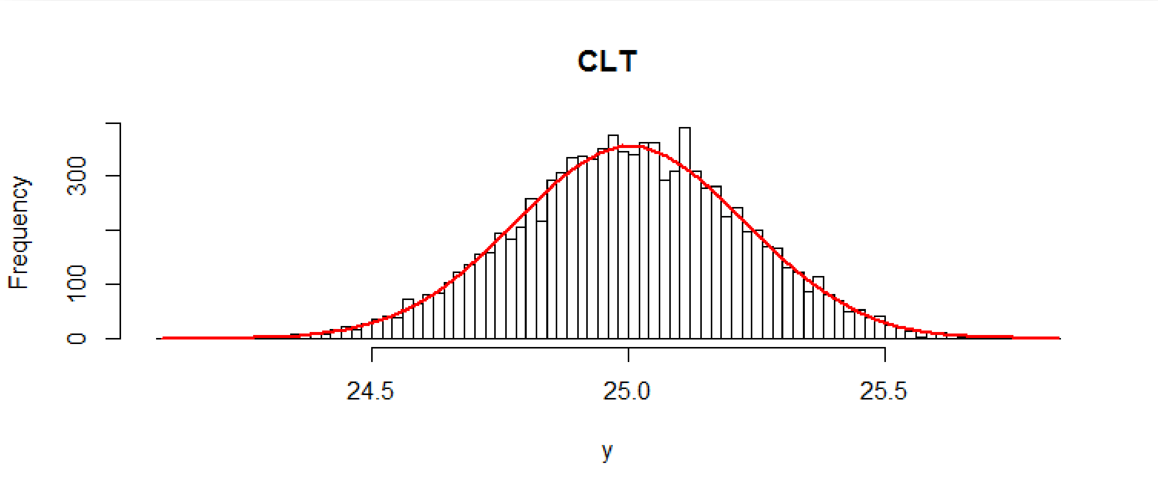
- The central limit theorem states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough.
- In practice, some statistics practitioners say that a sample size of 30 is large enough for the population distribution to be roughly bell-shaped.
- Hence, we can state population parameters for distribution of means as:

$$\boxed{E(\overline{X}) = \mu}$$

$$\boxed{Var(\overline{X}) = \frac{\sigma^2}{n}}$$

Illustration of CLT

- Suppose a sample of 1000 observations is taken 10,000 times.
- Each time the sum of the sample of 1000 is calculated and we have 10,000 such means
- Then we draw the following histogram for the distribution means
- Hence we see here that the shape of the histogram tends to be a bell shaped symmetric curve



Estimation Terminology

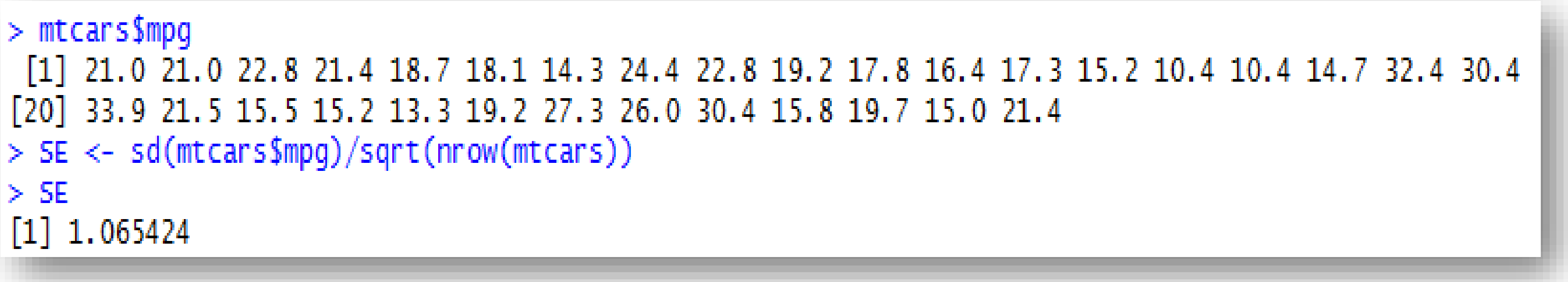
- Point Estimate
- Estimation Error
- Standard Error of the estimate
- Interval Estimate

Point Estimate

- Point estimate is calculated as being a “best guess” of the population parameter. e.g. Sample mean is point estimate of population mean.
- Points estimates can never be equal to the population parameter. The difference between the point estimate and the true value of the population parameter is called estimation error or sampling error.
- As the sample size increases point estimates come closer to their corresponding population parameters
- Also we have seen that distribution of point estimates is Sampling Distribution

Standard Error of the Estimate

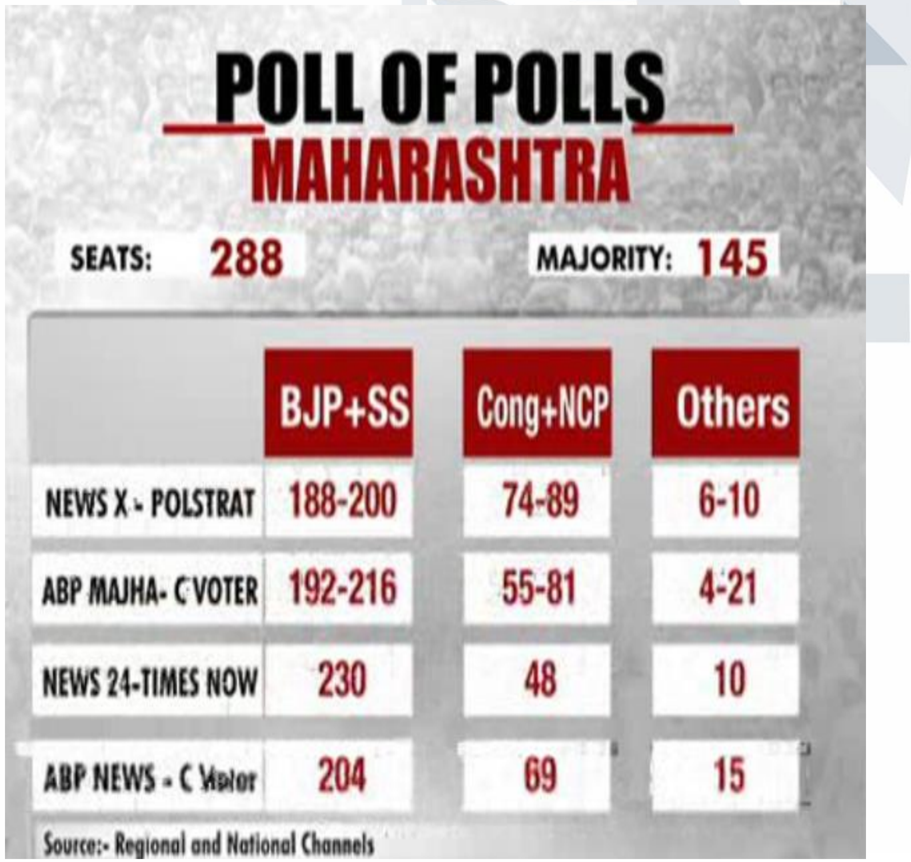
- It is a measure of uncertainty associated with the point estimate
- For Population, $SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- As σ is unknown for any population we use the following formula: $SE = \frac{s}{\sqrt{n}}$ where n is sample size



```
> mtcars$mpg
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7 32.4 30.4
[20] 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4
> SE <- sd(mtcars$mpg)/sqrt(nrow(mtcars))
> SE
[1] 1.065424
```

Interval Estimation

- A confidence interval is an interval around the point estimate calculated from the sample data, where it is strongly believed that the true value of the population parameter lies.
- Here we get the lower bound value and upper bound value
- e.g. [12.3, 41.5] C.I. indicates that the true population parameter value might be between 12.3 and 41.5



POLL OF POLLS
MAHARASHTRA

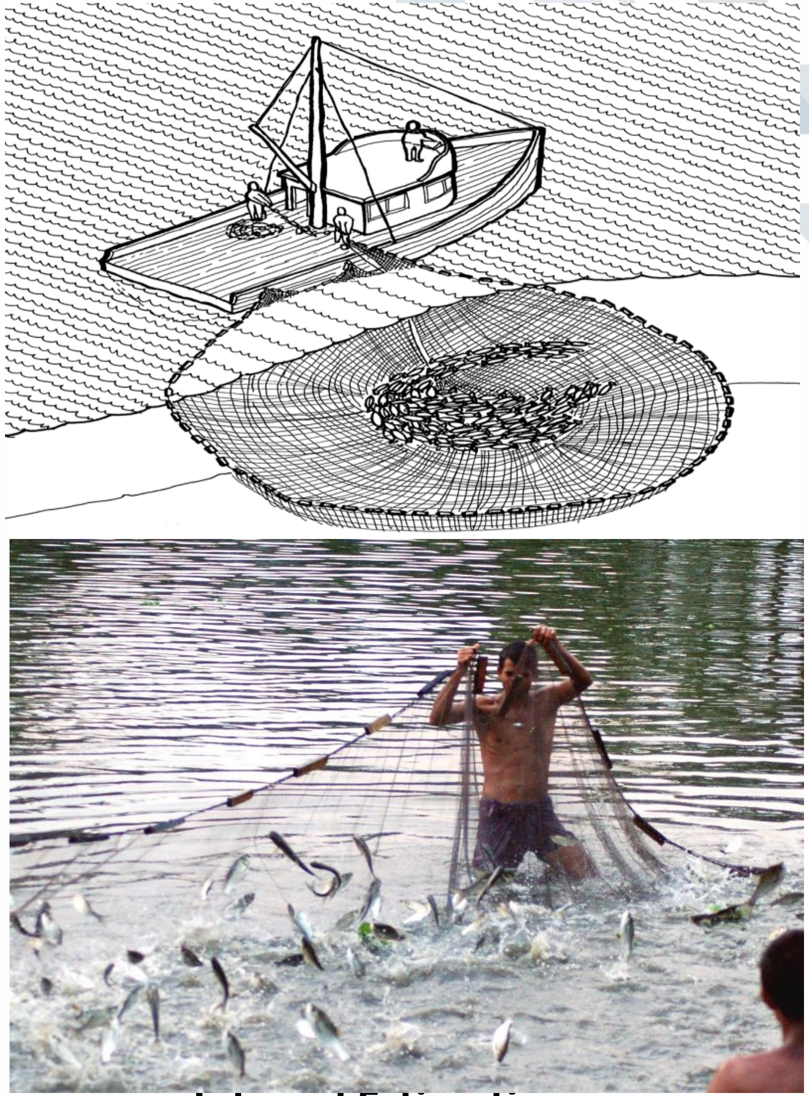
SEATS: 288 MAJORITY: 145

	BJP+SS	Cong+NCP	Others
NEWS X - POLSTRAT	188-200	74-89	6-10
ABP MAJHA- C VOTER	192-216	55-81	4-21
NEWS 24-TIMES NOW	230	48	10
ABP NEWS - C Votor	204	69	15

Source:- Regional and National Channels

Difference in Point and Interval Estimation

Point Estimation



Interval Estimation

Confidence Interval

- Interval (x1,x2) is said to be 95% confidence interval of population parameter μ,
  - If P(x1≤ μ ≤ x2) = 0.95
- Similarly,
- Interval (x1,x2) is said to be 99% confidence interval of population parameter μ,
  - If P(x1≤ μ ≤ x2) = 0.99

C.I. of µ

- Assuming Normal Distribution, C.I. of mean µ with known standard deviation is given by the formula(Sample size = n):

$$(\bar{x} - z.value\frac{\sigma}{\sqrt{n}}, \bar{x} + z.value\frac{\sigma}{\sqrt{n}})$$

- Assuming Normal Distribution, C.I. of mean µ with unknown standard deviation is given by the formula(Sample size = n):

$$(\bar{x} - t.value\frac{s}{\sqrt{n}}, \bar{x} + t.value\frac{s}{\sqrt{n}})$$

Margin of Error

- The quantity t.value* s / $\sqrt{n}$ is margin of error.
- More is the margin of error wider is the C.I.
- If its 95% C.I. then its confidence coefficient is 0.95.
- If its 99% C.I. then its confidence coefficient is 0.99.
- Confidence coefficient is denoted by (1- α)
- More the confidence coefficient wider is the C.I.
- Also greater is the sample size n, lesser would be the margin of error.