STATISTICS FOR DATA SCIENCE

Central Limit Theorem

D. Uma D

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Central Limit Theorem

D. Uma

Topics to be covered...

- ✓Statistical Inference
- ✓Sampling Distributions
- ✓Central Limit Theorem

Statistical Inference

- It allows us to make accurate decisions from the numerical descriptive measures of parameter from statistic.

- Accuracy can be measured using probability.

- It is entrusted by identifying the distribution of the random variables from the population from where they come from.

- With this the shape and the location of the sample mean, decision about the population parameters can be made.

- To understand the shape of the sample data, sampling distribution can be employed.

Examples of How do we deal with unknown parameters?

- A medical researcher surveys and the responses are based on agreement or disagreement which will follow Binomial distribution, but the proportion (p) of people who agree in the population may be unknown.

- The heights of the male in the city is assumed to be normally distributed. Estimate the mean (μ) and standard deviation (σ) which are unknown.

- So, to interpret the above two cases, and to get reliable information about the population, the sample should be chose accordingly.

- This can be done by choosing the appropriate sampling methods.

- The **probability distribution of statistic** is called as **sampling distribution**.

- This distribution based on how many trials are repeatedly taken of size ‘n’ from a population.

- If the sample statistic is the **sample mean**, then the distribution is sampling distribution of sample means.

- Every sample statistic has a sampling distribution.

Example – Sampling Distribution

Consider a population of 10 numbers from 1 to 10. You randomly choose two numbers with replacement. List all the possible samples of size n = 2 and calculate mean of each. Predict the sampling distributions.

Example – Sampling Distribution

List of all 16 samples of size 2 from the population and the mean of each sample.

Sample	Sample Mean ($\bar{x}$)
1,1	1
1,3	2
1,5	3
1,7	4
3,1	2
3,3	3
3,5	4
3,7	5

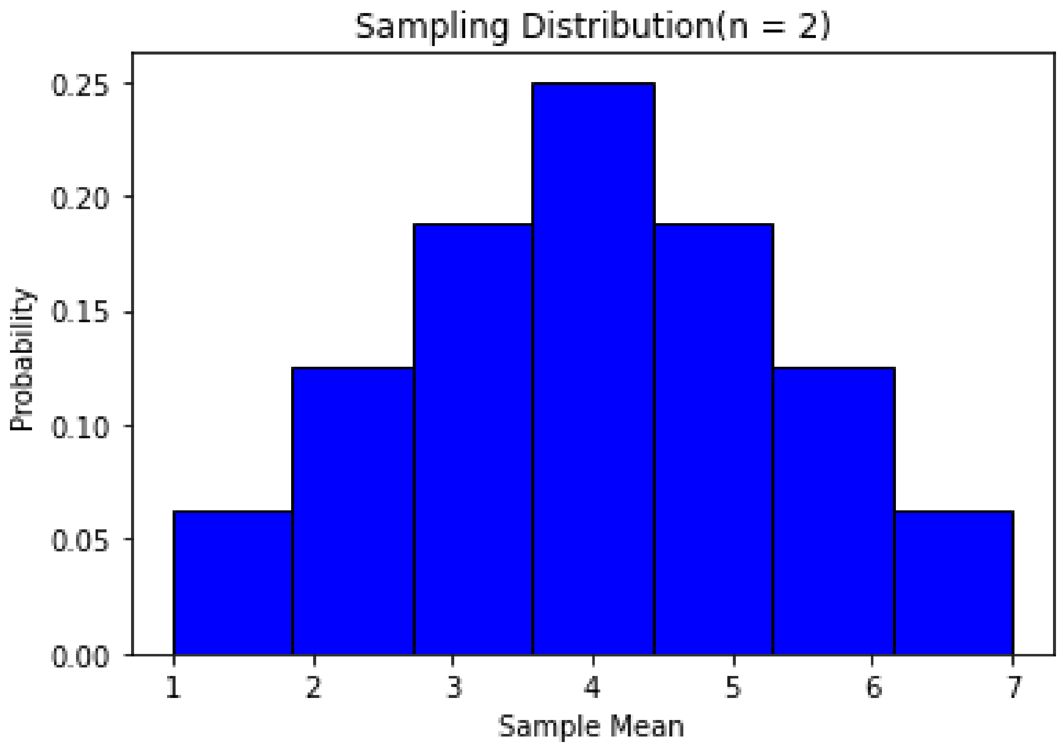
Sample	Sample Mean ($\bar{x}$)
5,1	3
5,3	4
5,5	5
5,7	6
7,1	4
7,3	5
7,5	6
7,7	7

Numerical Data Source: www.ruckdeschel.pbworks.com

Example – Sampling Distribution & Probability Histogram

Probability Distribution of all sample means & Probability Histogram.

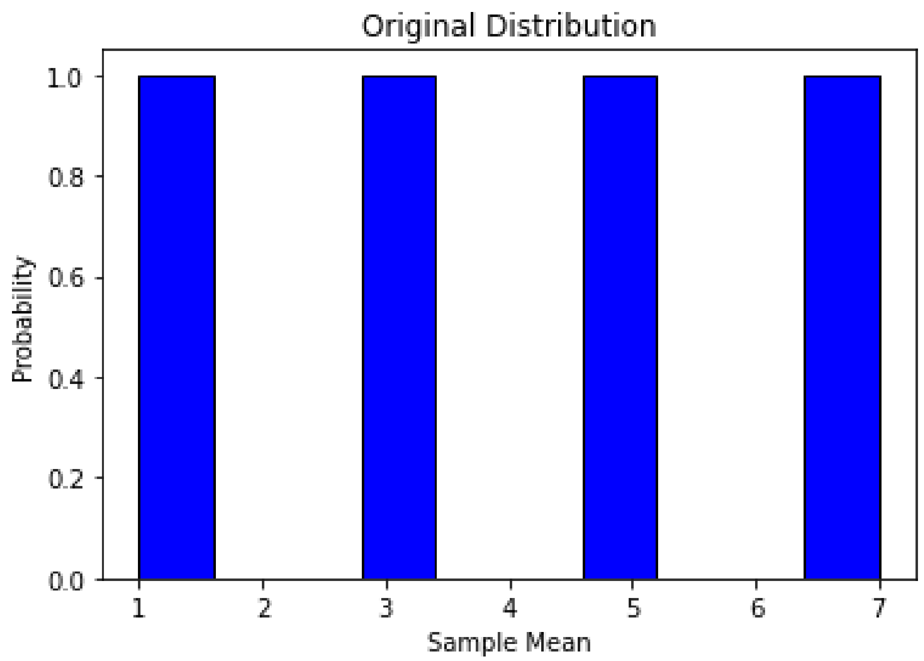
$\bar{x}$	frequency	Probability
1	1	1/16 = 0.0625
2	2	2/16 = 0.1250
3	3	3/16 = 0.1875
4	4	4/16 = 0.2500
5	3	3/16 = 0.1875
6	2	2/16 = 0.1250
7	1	2/16 = 0.0625

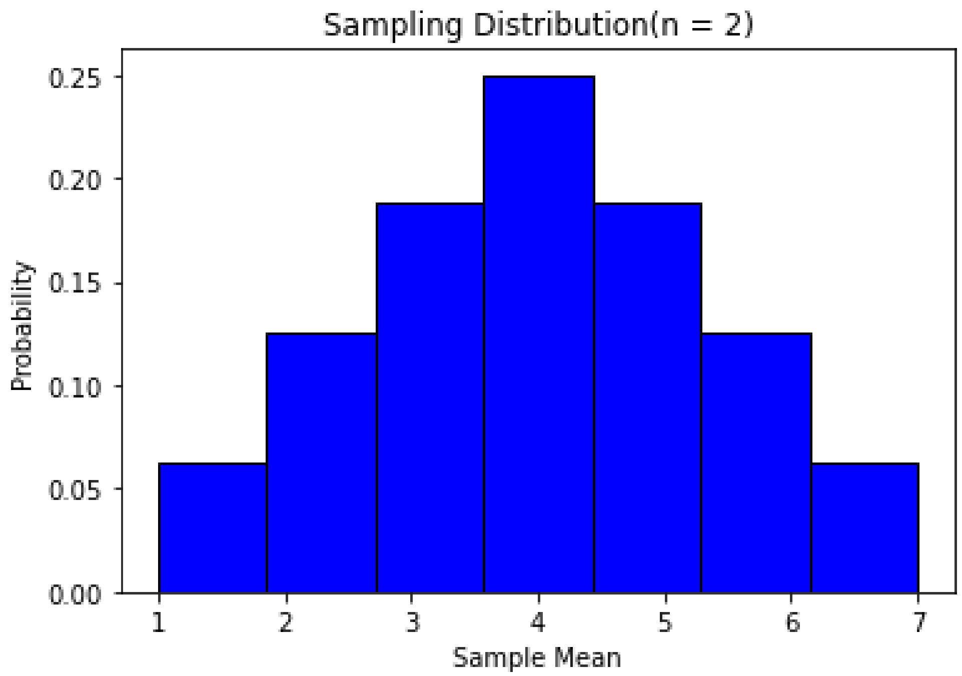


Numerical Data Source: www.ruckdeschel.pbworks.com

Example – Distribution

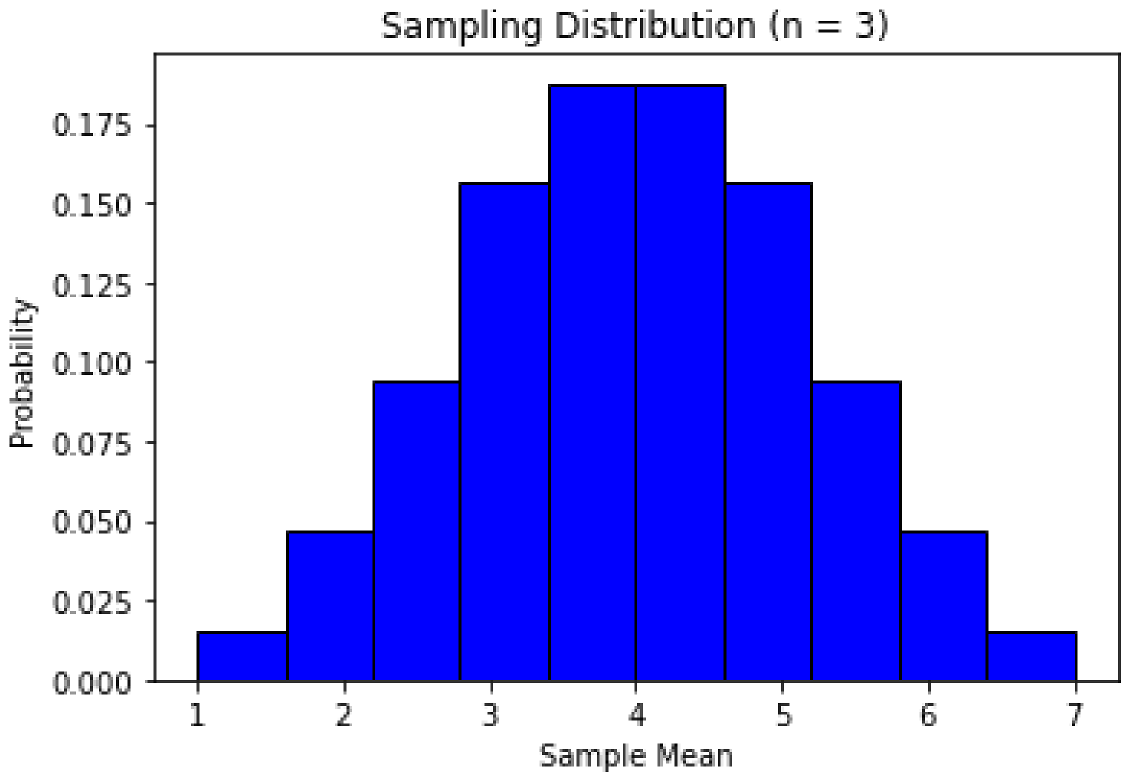
The original distribution and the sampling distribution when n = 2.





Example – Sampling Distribution

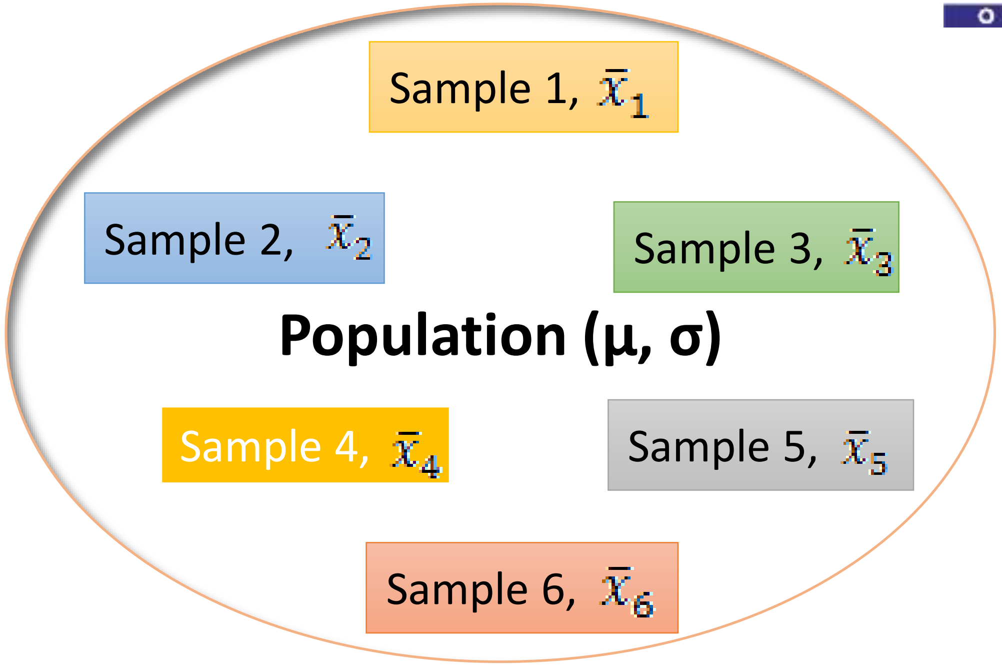
Sampling distribution when n = 3.



It is understood that when the sample size (n) increases, the shape is getting closer and closer to the normal distribution.

The idea behind Central Limit Theorem (CLT)

- This is where the idea of Central Limit theorem comes.
- When some *'x'* samples are collected each of size *'n'* and when the sample mean for each of them is computed then, it will probably follow Normal Distribution.
- Moreover the mean of the sample means would be almost equal to the true parameter of the population, which applies to the other parameters also.



What is Central Limit Theorem?

Central Limit Theorem states that the distribution of sample means that is calculated from sampling will follow normal distribution as the size of **'n'** increases regardless of the samples that may be drawn from any population distribution.

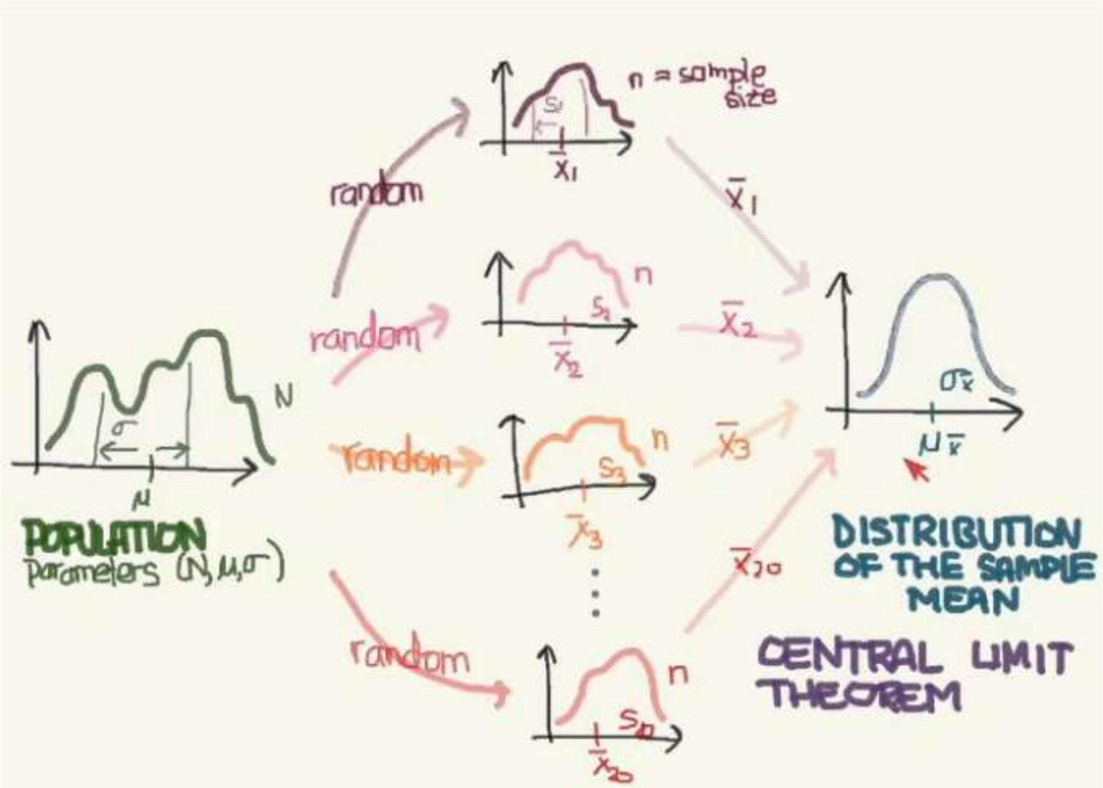


Image Source: https://medium.com/

Data in Sample Distribution should be ....

- As “Law of Large numbers” states that the mean of the sample distribution will be same as the mean of the population distribution when the size of the sample increases.

- The samples should be randomly selected.

- The samples must be independent of each other.

- The sample size should be large enough for the distribution to be normal, a sample size of 30 is mandatory for getting more representative sample.

- When the sampling is done without replacement, the sample size should not be more than 10% of the population.

The Central Limit Theorem

Let $X_1, \ldots, X_n$ be a simple random sample from the population with mean μ and variance σ^2

Let $\overline{X} = \dfrac{X_1 + \ldots . + X_n}{n}$ be the sample mean.

Let $S_n = X_1 + \ldots . + X_n$ be the sum of sample observations

Then if n is sufficiently large,

$$\overline{X} = \sim N(\mu, \frac{\sigma^2}{n})$$

$$S_n = N(n\mu, n\sigma^2)$$

The Central Limit Theorem

- The central limit theorem specifies that $\mu_{\bar{x}} = \mu$ and $\sigma^2_{\bar{x}} = \dfrac{\sigma^2}{n}$ which hold for any sample mean.

- The sum of the sample items is equal to the mean multiplied by the sample size, that is $S_n = n\bar{X}$

- It follows that $\mu_{S_n} = n\mu$ and $\sigma^2_{S_n} = \dfrac{n^2\sigma^2}{n} = n\sigma^2$

Real Life Case Study

A business client of FedEx wants to deliver urgently a large freight from Denver to Salt Lake City.

When asked about the weight of the cargo they could not supply the exact weight, however they have specified that there are total of 36 boxes.

You are working as a **Business analyst** for FedEx.

And you have been challenged to tell the executives quickly whether or not they can do certain delivery.

Real Life Case Study

Since, we have worked with them for so many years and have seen so many freights from them we can confidently say that the type of cargo they follow is a distribution with a

mean of μ= 72 lb (32.66 kg)

standard deviation of σ = 3 lb (1.36 kg).

The plane you have can carry the max cargo weight up to 2640 lb (1193 kg).

Based on this information what is the probability that all of the cargo can be safely loaded onto the planes and transported?

Real Life Case Study – How to Proceed?

- Using Central Limit Theorem, find the mean and standard deviation of the sample mean.
- Calculate the critical point of each box by dividing the allowable capacity of the plane to carry weight with total number of boxes.
- So, to safely takeoff the plane, the average weight of the each box should not exceed 73.06 lb/box.
- Finally, calculate the Z-score.

Real Life Case Study – Solution

$$\mu_{\bar{x}} = \mu_x = 72 \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$$

$$\text{Plane Capacity} = 2{,}640\text{lb}$$

$$x_{critPoint} = \frac{2640\ lb}{36\ boxes} = \frac{73.06 lb}{box}$$

$$z = \frac{x_{critPoint} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{73.06 - 72}{0.5} = 2.12$$

$$P(x < x_{critPoint}) = 0.9830 = 98.3\%$$

So, The plan can safely takeoff is 98.3% and 1.7 % chance it cannot takeoff.

Problem

Drums labeled 30 L are filled with a solution from a large vat. The amount of solution put into each drum is random with mean 30.01 L and standard deviation 0.1 L.

a) What is the probability that the total amount of solution contained in 50 drums is more than 1500 L?

b) If the total amount of solution in the vat is 2401 L, what is the probability that 80 drums can be filled without running out?

c) How much solution should the vat contain so that the probability is 0.9 that 80 drums can be filled without running out?

Solution for part (a)

Let $S = X_1, \dots . X_{50}$ be the amounts of solution in 50 drums.

$\mu_x = 30.01$ and $\sigma_x = 0.1$

Assuming that S is approximately normally distributed, then

$\mu_S = 50(30.01) = 1500.5$ and $\sigma_S = 0.1\sqrt{50} = 0.7071$

By calculating $z - \text{score}$, $z = \dfrac{1500 - 1500.5}{0.7071} = -0.71$

The area to the right of $z = -0.71$ *is* $1 - 0.2389 = 0.7611$

$P(S > 1500) = 0.7611$

Solution for part (b)

Let D = $X_1, \ldots . X_{80}$ be the amounts of solution in 80 drums.

$\mu_x = 30.01$ and $\sigma_x = 0.1$

Assuming that D is approximately normally distributed, then

$\mu_D = 80(30.01) = 2400.8$ and $\sigma_D = 0.1\sqrt{80} = 0.8944$

By calculating z – score, $z = \dfrac{2401 - 2400.8}{0.8944} = 0.22$

The area to the left of z = 0.22 *is* 0.5871

$P(S < 2401) = 0.5871$

Solution for part (c)

The z − score for 90th percentile is z = 1.28

From part (b), D is approximately normally distributed with

$\mu_D = 80(30.01) = 2400.8$ and $\sigma_D = 0.1\sqrt{80} = 0.8944$

$$1.28 = \frac{x - 2400.8}{0.8944}$$

$$x = 2401.9L$$

THANK YOU

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335