STATISTICS FOR DATA SCIENCE

Central Limit Theorem

D. Uma

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

- To make accurate decisions from statistics about the parameters.
- Accuracy can be measured using probability by identifying the distribution.
- **Shape and the sample of the statistic.**
- **How the statistic is distributed? – Sampling Distribution.**
- Example **Unknown Parameters** – Proportion (p)
  -- Mean (μ) and standard deviation (σ)

$\bar{X} \sim Dist(\ )$ $N(\mu, \sigma^2)$

$\bar{X}$:

$\hat{p} \sim ?(\ )$

$s \sim ?(\ )$

Statistic: $\bar{X}$, $\hat{p}$, s

Parameter: μ, p, σ

Probability Distribution

X :	1	2	3	4	5	6
$P(X=x)$:	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

Uniform

Sample Mean: Recap

Let $X_1, X_2, \dots X_n$ be n independent r.v.s ~~uniform~~ a normal distribution with mean μ and variance σ^2.

normally distribution

$\mu_{X_1} = \mu$ $\quad X_1 \sim N(\mu, \sigma^2)$

$\sigma_{X} = \sigma$ $\quad X_2 \sim N(\mu, \sigma^2)$

$\mu_{X_2} = \mu$ $\quad X_3 \sim N(\mu,$

$\sigma_{X_2} = \sigma$

Population

Normally Distributed

μ, σ

Sample

$X_1, \dots, X_n$

$n > 30$

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n$$

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

$$\mu_{\bar{X}} = \mu_{\frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n}$$

$$= \frac{1}{n}\mu_{X_1} + \frac{1}{n}\mu_{X_2} + \dots + \frac{1}{n}\mu_{X_n}$$

$$= \frac{1}{n}\mu + \frac{1}{n}\mu + \dots + \frac{1}{n}\mu = \frac{1}{n}(\mu + \mu + \dots + \mu)$$

$$= \frac{1}{n}(n\mu) = \mu$$

$$\boxed{n > 30}$$

$$\sigma^2_{\bar{X}} = \sigma^2_{\frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n}$$

$$= \frac{1}{n^2}\sigma^2_{X_1} + \frac{1}{n^2}\sigma^2_{X_2} + \dots + \frac{1}{n^2}\sigma^2_{X_n}$$

$$= \frac{1}{n^2}\left[\sigma^2 + \sigma^2 + \dots + \sigma^2\right]$$

$$= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

Example – Sampling Distribution

population values = {1,3,5,7} on slips of paper and put them in a box.

List all possible samples of size n = 2 and calculate the mean of each.

Find the mean, variance, and standard deviation of the sample means.

Compare your results with the mean (μ = 4) variance (σ^2 = 5) and standard deviation (σ =2.236) of the population.

Predict the sampling distributions.

Example – Sampling Distribution

List of **all 16 samples** of **size 2** from the **population {1,3,5,7}** and the mean of each sample.

$\mu = 4$; $\sigma^2 = 5$; $\sigma = 2.236$

Sample	Sample Mean ($\bar{x}$)	Sample	Sample Mean ($\bar{x}$)
1, 1	1	5, 1	3
1, 3	2	5, 3	4
1, 5	3	5, 5	5
1, 7	4	5, 7	6
3, 1	2	7, 1	4
3, 3	3	7, 3	5
3, 5	4	7, 5	6
3, 7	5	7, 7	7

Population N=4
1, 3, 5, 7

Sample n=2

$\bar{x}$:

$P(\bar{x} = x)$:

sample

Population

Prob Distribution

$\bar{x}$: μ

σ

?? Static

Sampling Distribution

Probability Distribution of a Static

$\bar{x}$

$\bar{x}$ or $\hat{p}$ or s

Example – Sampling Distribution

List of **all 16 samples** of **size 2** from the **population {1,3,5,7}** 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	4

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

STATISTICS FOR DATA SCIENCE

Example – Sampling Distribution & Probability Histogram

Probability Distribution of all sample means

$\bar{x}$	frequency	Probability
1	1	1/16
2	2	2/16
3	3	3/16
4		
5		
6		
7		

16

Example – Sampling Distribution & Probability Histogram

Probability Distribution of all sample means

$\bar{x}$	frequency	Probability
1	1	1/16
2	2	2/16
3	3	3/16
4	4	4/16
5	3	3/16
6	2	2/16
7	1	1/16

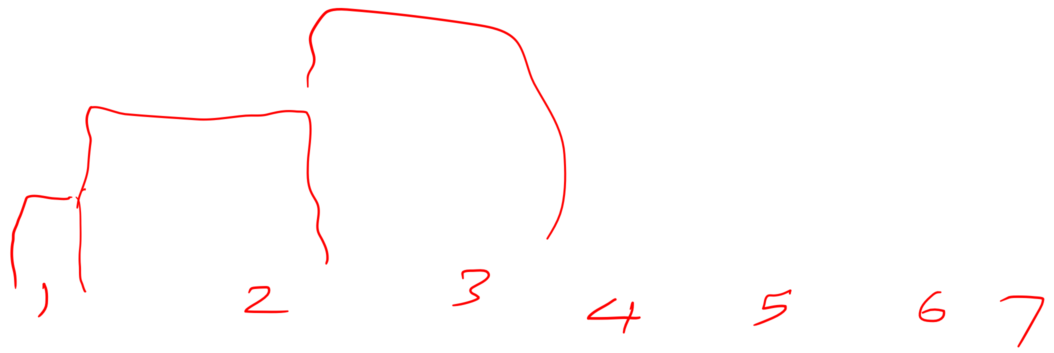
$\Sigma f = 16$

Example – Sampling Distribution & Probability Histogram

Probability Distribution of all sample means &
Probability Histogram.

$\bar{x}$	**frequency**	**Probability**
1	1	1/16
2	2	2/16
3	3	3/16
4	4	4/16
5	3	3/16
6	2	2/16
7	1	1/16

sample mean $\bar{x}$

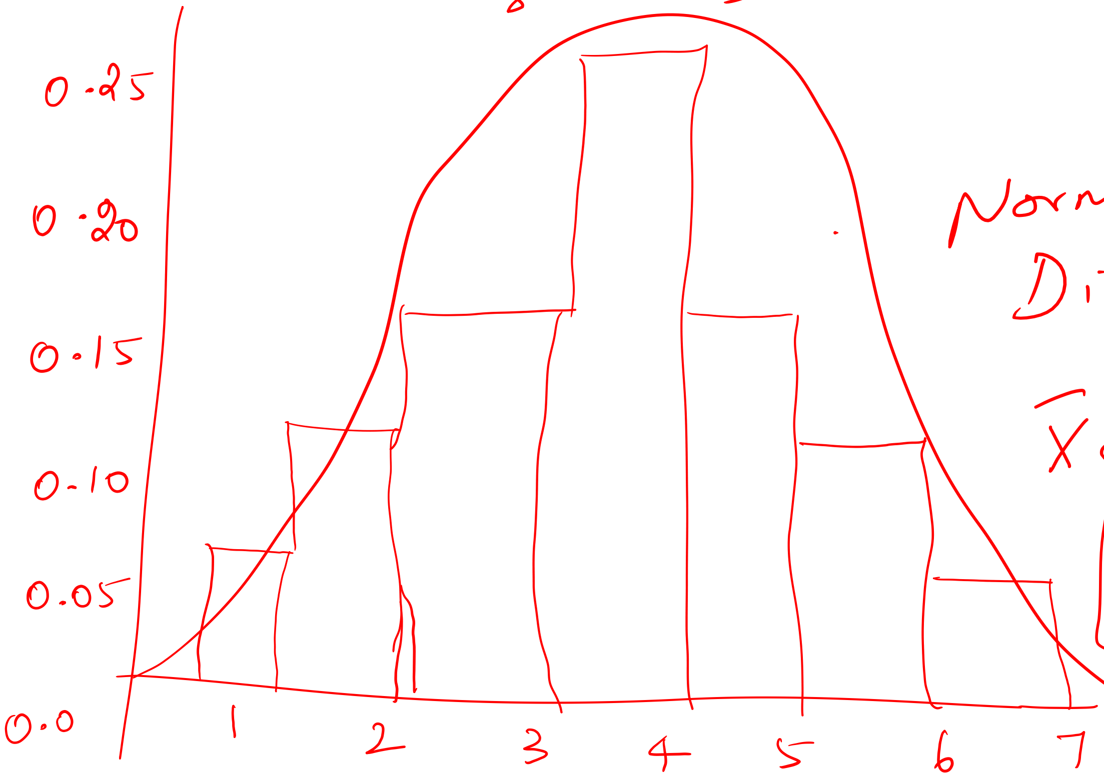


Example – Sampling Distribution & Probability Histogram

Probability Distribution of all sample means & **Probability Histogram.**

several random samples } $\bar{X}$: ---- $P(\bar{X}=x)$: ----

$\bar{x}$	frequency	Probability
1	1	1/16
2	2	2/16
3	3	3/16
4	4	4/16
5	3	3/16
6	2	2/16
7	1	1/16



Normal Distribution

$\bar{X} \sim N(\mu, \sigma^2/n)$

$\boxed{\mu_{\bar{X}} = \mu}$

$\sigma^2_{\bar{X}} = \sigma^2/n$

$\mu_{\bar{X}} = 4$ $\sigma_{\bar{X}} =$

$\mu = 4$

$P = \{1, 3, 5, 7\} \Rightarrow \mu = 4$

Probability Distribution of $\bar{X}$

$\sigma = \sqrt{n}\,\sigma_{\bar{X}}$ $\leftarrow$ $\boxed{\sigma_{\bar{X}} = \sigma/\sqrt{n}}$

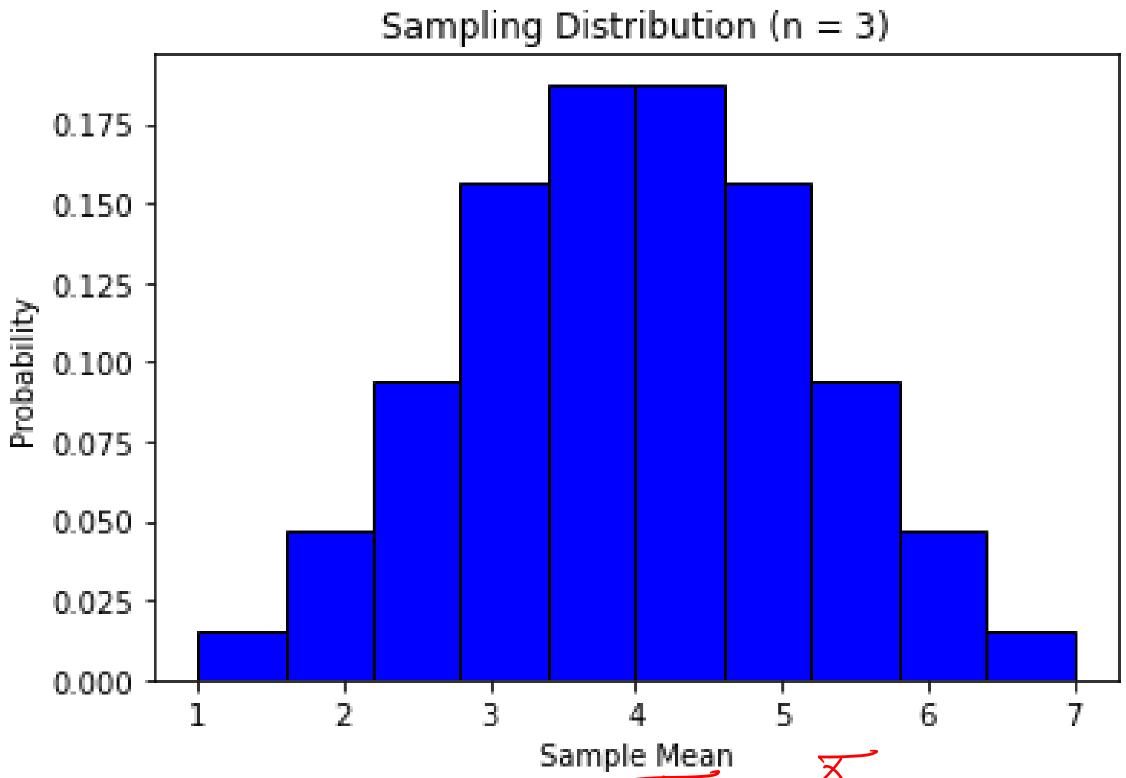
Example – Sampling Distribution & Probability Histogram

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	1/16 = 0.0625

Example – Sampling Distribution

Sampling distribution when n = 3.



$\bar{x}$

$N=4$

$n=3$

$n=4$

$n \uparrow$

$n=5$

$n=10$

$n=20$

$n=50$

Pop Dist

Sampling Distribution of 1 statistic

$\sim N(,)$

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

Mean and Standard Deviation of the Sampling Distribution

If $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$

$$\mu_{\bar{x}} = \mu$$

$$\sigma^2_{\bar{x}} = \sigma^2/n$$

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

Sampling Distributions

- The **probability distribution of statistic** is called as **sampling distribution**.
- Trials are repeated by taking sample size 'n' from a population.
- The distribution is sampling distribution of sample means.
- **Every sample statistic** has a **sampling distribution**.

SD of statistic

$\hat{p}$

$\hat{p}$:

$P(\hat{p} = \ )$

s : - - - - -

$P(s = \ )$:

SD of s

Data in Sample Distribution should be ....

$n > 30$

- "Law of Large numbers" - The **mean of the sample distribution** will be **same as** the **mean of the population** distribution when the size of the sample increases.

- Random Selection of samples and independent of each other.

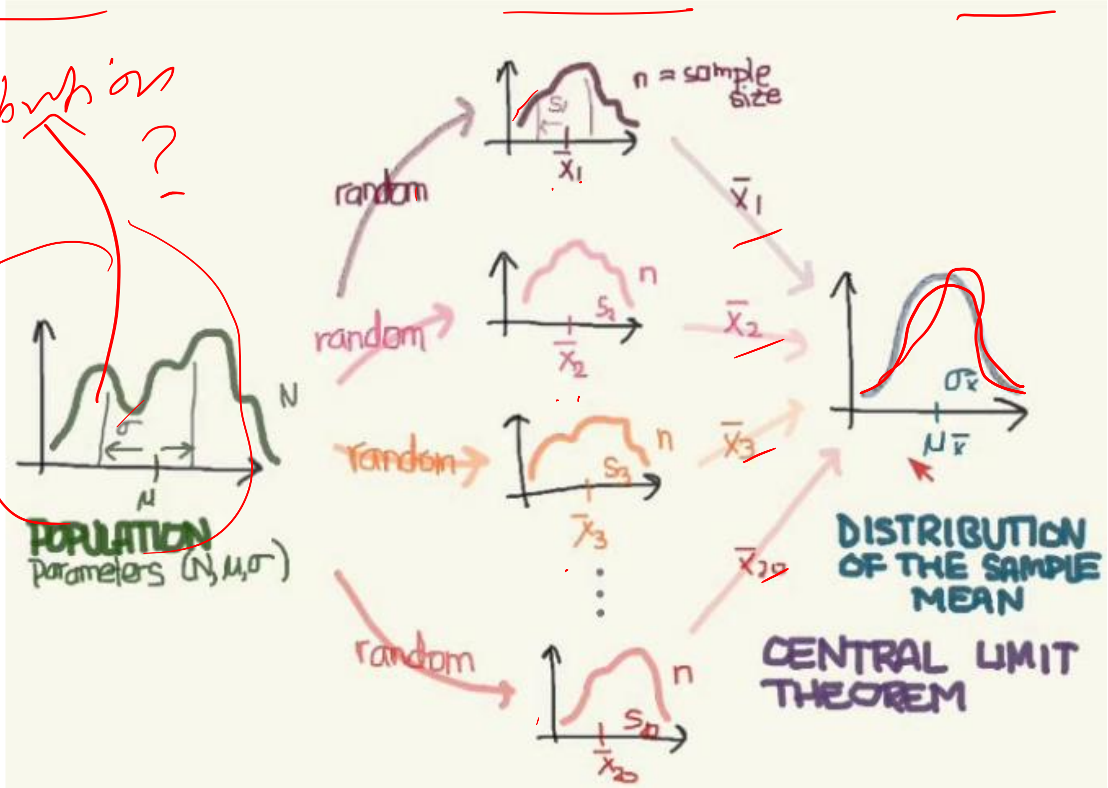
- **Sample size of 30** is mandatory.

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

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.

Distribution ?



The Central Limit Theorem

The Central Limit Theorem says that if we draw a large enough sample from a population, then the distribution of the sample mean is approximately normal, no matter what population the sample was drawn.

The Central Limit Theorem

Let $X_1, X_2 \ldots X_n$ be a random sample from a population with mean μ and variance σ^2.

Let $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ be the sample mean.

Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of the sample observations.

Then, if n is sufficiently large

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ approximately.

$S_n \sim N(n\mu, n\sigma^2)$ approximately.

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \qquad \bar{X} \sim N\left(\mu, \sigma^2/n\right)$$

$$S_n = n\bar{X}$$

$$\mu_{S_n} = \mu_{n\bar{X}} = n\mu_{\bar{X}} = n(\mu) = n\mu$$

$$\sigma^2_{S_n} = \sigma^2_{n\bar{X}} = n^2 \sigma^2_{\bar{X}} = n^2\left(\frac{\sigma^2}{n}\right) = n\sigma^2$$

$$\therefore S_n \sim N\left(n\mu, n\sigma^2\right)$$

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.

Numerical Data Source: https://medium.com/

Real Life Case Study

PES UNIVERSITY ONLINE

Given : Mean of μ= 32.66 kg

standard deviation of σ = 1.36 kg.

The plane you have can carry the max cargo weight up to 1193 kg.

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

$n = 36$

$n > 30$

CLT

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

$\bar{X} \sim N(32.66, 0.227^2)$

$\mu_{\bar{x}} = \mu$

$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \left(\cancel{1.36}\right)$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.36}{\sqrt{36}}$

$\sigma_{\bar{x}} = 0.227$

36 Boxes

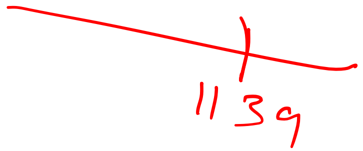
$\frac{1193}{36} = 33.14$

$P($

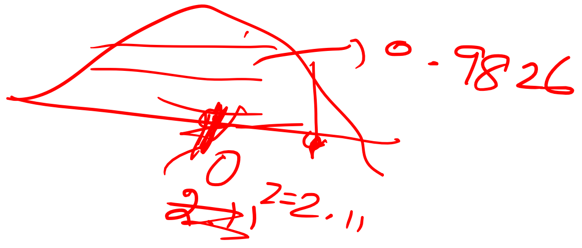
Numerical Data Source: https://medium.com/

The Central Limit Theorem: Case Study

$$\bar{X} = \frac{1193}{36} = 33.14$$



$$z = \frac{\bar{X} - \mu}{s} = \frac{33.14 - 33.66}{0.227} = 2.11$$



$P(T < 1139)$

$$P(z < 2.11) = 0.9826$$

can take off 98.3 %

1.7 % chance that the plane can't take off

Real Life Case Study

We know that $\bar{X} \sim N(\mu_{\bar{x}}, \sigma^2_{\bar{x}})$ where $\mu_{\bar{x}} = \mu = 32.66$ kg

$\therefore \bar{X} \sim N(32.66, 0.227^2)$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.36}{\sqrt{36}} = 0.227$

Allowable capacity for the plane is 1193 kg

X = critical point $= \frac{1193}{36} = 33.14$ kg/box

The corresponding Z-score is $Z = \frac{X - \mu}{\sigma_{\bar{x}}} = \frac{33.14 - 32.66}{0.227} = 2.11$

Area to the left of $Z = 2.11$ is 0.9826

$Z = 2.11$

$\therefore$ Probability that plane can take-off safely is 98.26 % or $\simeq$ 98.3 %.

$\therefore$ Probability that plane can't take off safely is 1.7 %.

Example 2

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.

$n = 30$

$\bar{x} = 30.01, \quad s = 0.1$

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?

Problem

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

$\mu_{\bar{x}} = 30.01$ L

$s_{\bar{x}} = 0.1$

T = Total amt. of soln in 50 drums

T or $S = n\bar{X}$ $\quad S \sim N(n\mu, n\sigma^2)$

$P(S > 1500) = ?$

$\mu_S = \mu_{n\bar{x}} = n\mu_{\bar{x}} = 50(30.01)$

$= 1500.5$

$\sigma_S = \sqrt{n}\, s = \sqrt{50} * 0.1 = 0.7071$

$Z = \frac{\bar{x} - \mu_S}{s_S} = \frac{1500 - 1500.5}{0.7071}$

$= -0.71$

$\therefore S \sim N(1500.5, 0.7071)$

$P(Z > -0.71) = 1 - 0.2389 = 0.7611$

$P(Z < -0.71) = 0.2389$

$\therefore P(S > 1500) = 0.7611$

Problem

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?

Let T: Total amount of solution in 80 drums.

$T = X_1 + X_2 + \cdots + X_{80}$ where X_i : Amt. of soln in drum X_i

As per CLT, $T \sim N(\mu_T, \sigma_T^2)$ where $\mu_T = n\mu = 80(30.01) = 2400.8$

$\sigma_T = s\sqrt{n} = 0.1 * \sqrt{80} = 0.8944$

To find: $P(T < 2401)$

$= P\left(\frac{T-\mu}{\sigma_T} < \frac{2401-2400.8}{0.8944}\right)$

$= P(z < 0.22)$

$P(T<2401) = \underline{0.5871}$

Problem

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

T : Total Amt. of solution in 80 drums.

We know that $T \sim N(\mu_T, \sigma_T^2) \Rightarrow T \sim N\left(80(30.01), (0.1\sqrt{80})^2\right)$

$n\mu$, $(\sigma\sqrt{n})^2$

$T \sim N(2400.8, 0.8944^2)$

Given probability is 0.9

The corresponding z-score is $z = 1.28$

$\therefore \quad z = \frac{x - \mu_T}{\sigma_T} \Rightarrow 1.28 = \frac{x - 2400.8}{0.8944}$

$\Rightarrow x = 2401.9 \, l$

$\therefore$ Vat should contain 2401.9 l solution.

Do It Yourself!!!

A simple random sample of 100 men is chosen from a population with mean height 70 in. & standard deviation 2.5 in.
What is the probability that the average height of the sample men is greater than 69.5 in?

The Central Limit Theorem: Problems

The manufacture of a certain part requires two different machine operations.

- The time on machine 1 has mean 0.5 hours and standard deviation 0.4 hours.
- The time on machine 2 has mean 0.6 hours and standard deviation 0.5 hours.
- The times needed on the machines are independent.
- Suppose 100 parts are manufactured.

1) What is the probability that the total time used by machine 1 is greater than 55 hours?
2) What is the probability that the total time used by machine 2 is less than 55 hours?
3) What is the probability that the total time used by both the machines together is greater than 115 hours.
4) What is the probability that the total time used by machine 1 is greater than the total time used by machine 2?

The Central Limit Theorem: Problems

The manufacture of a certain part requires two different machine operations.

- The time on machine 1 has mean 0.5 hours and standard deviation 0.4 hours.
- The time on machine 2 has mean 0.6 hours and standard deviation 0.5 hours.
- The times needed on the machines are independent.
- Suppose 100 parts are manufactured.

1) What is the probability that the total time used by machine 1 is greater than 55 hours?

The Central Limit Theorem: Problems

The manufacture of a certain part requires two different machine operations.

- The time on machine 1 has mean 0.5 hours and standard deviation 0.4 hours.
- The time on machine 2 has mean 0.6 hours and standard deviation 0.5 hours.
- The times needed on the machines are independent.
- Suppose 100 parts are manufactured.

2) What is the probability that the total time used by machine 2 is less than 55 hours?

The Central Limit Theorem: Problems

The manufacture of a certain part requires two different machine operations.

- The time on machine 1 has mean 0.5 hours and standard deviation 0.4 hours.
- The time on machine 2 has mean 0.6 hours and standard deviation 0.5 hours.
- The times needed on the machines are independent.
- Suppose 100 parts are manufactured.

3) What is the probability that the total time used by both the machines together is greater than 115 hours.

The Central Limit Theorem: Problems

The manufacture of a certain part requires two different machine operations.

- The time on machine 1 has mean 0.5 hours and standard deviation 0.4 hours.
- The time on machine 2 has mean 0.6 hours and standard deviation 0.5 hours.
- The times needed on the machines are independent.
- Suppose 100 parts are manufactured.

4) What is the probability that the total time used by machine 1 is greater than the total time used by machine 2?

THANK YOU

D. Uma

Computer Science and Engineering

umaprabha@pes.edu
+91 99 7251 5335