ST 228 : DATA ANALYTICS, ML and AI ASSIGNMENT-1

K. BHIKSHAPATHI - 24489.

Problem - 1

Assumptions

3 distributions were generated with following parameters:

1. Uniform distribution
   Lower bound (a) : 0
   Upper bound (b) : 10
   Population size (N) : 10,000

2. Normal distribution
   Mean (μ) : 0
   Standard deviation (σ) : 1
   Population size (N) : 10,000

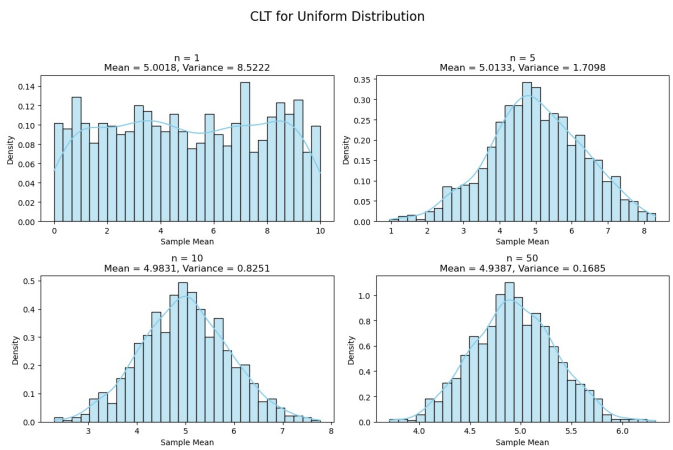
3. Exponential Distribution
   Rate parameter (λ) : 1 population size (N) : 10,000

(a) Central Limit Theorem (CLT)

Uniform Distribution:-

→ For smaller sample size (n=5), the distribution shows asymmetry

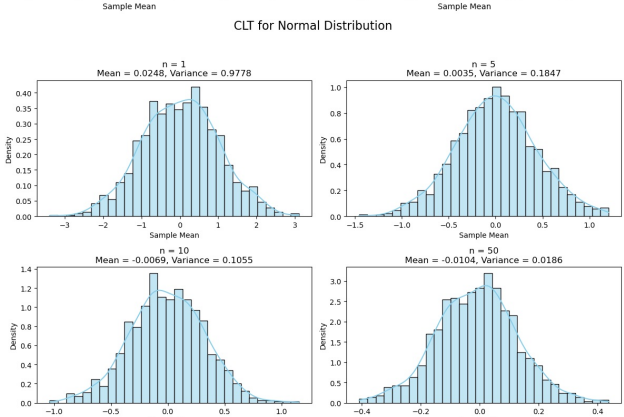
→ But as n increases the distribution approaches to normality with reduced varience



Normal distribution:-

→ Even for smaller sample size we see the distribution exhibited normality

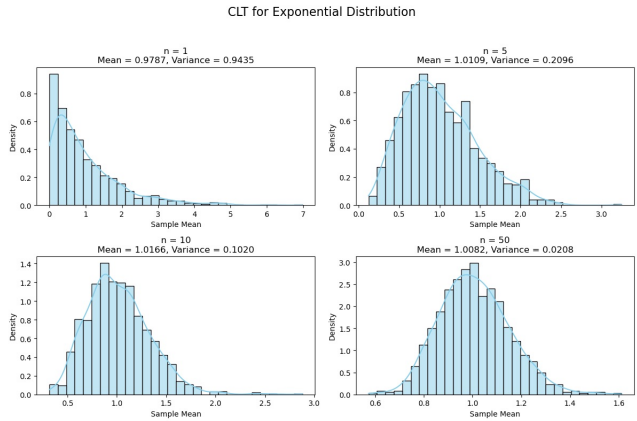
→ with increasing sample size (n=30), the distribution further increased precision towards normality



Exponential distribution

→ For smaller n values we see skewness bcz of inherited behaviour

→ But as n increase, the distribution more symmetrical and bell shaped



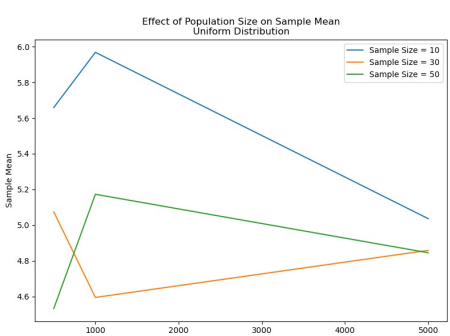
∴ It is confirmed that for larger sample size uniform, exponential and normal distribution exhibits central Limit theorm (CLT).

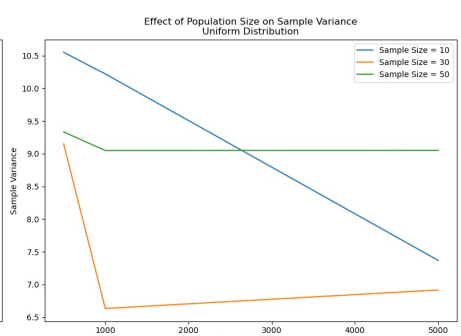
(b) Sample Mean:- (Uniform Distribution)

→ Sample mean (n=10) showed high variability as population size increased.

→ Large sample size (n=50) yielded stable means, converging towards population mean(5).

→ Increasing population size stabilized the sample mean for larger sample.



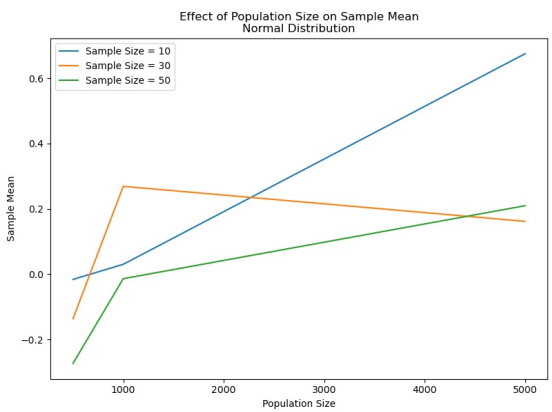


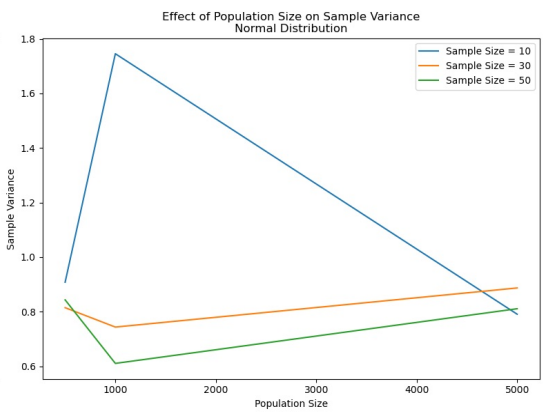
Sample varience :-

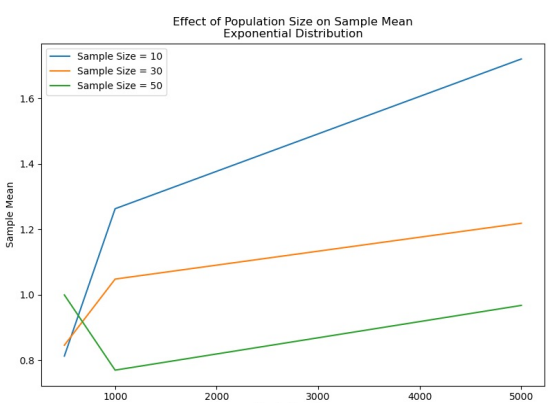
- Smaller sample size exhibited higher sensitivity to population size, with varience decreasing as population size increased.
- Larger sample size had stable varience regardless of population size.

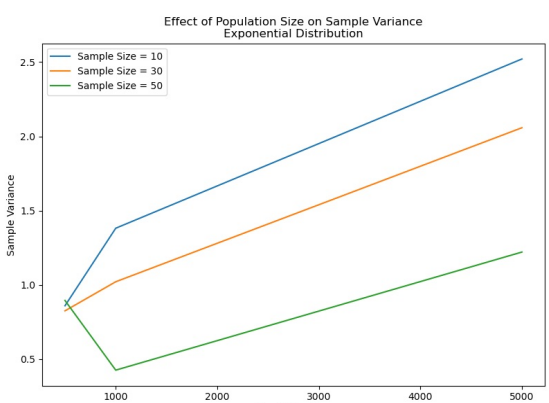
Conclusion :-

- Sample size has a stronger effect on stability than population size
- Larger sample size leads more stable estimates of population parameters.









(C)

Effect of skewness of CLT

Key insights:-

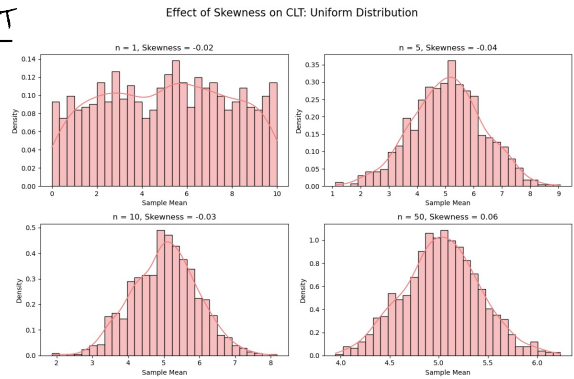
- Symmetric distributions (e.g normal, uniform) show rapid convergence to normality for sampling distribution of the mean even with small sizes
- Highly skewed distribution (eg exponential) requires larger sample size to achieve normality

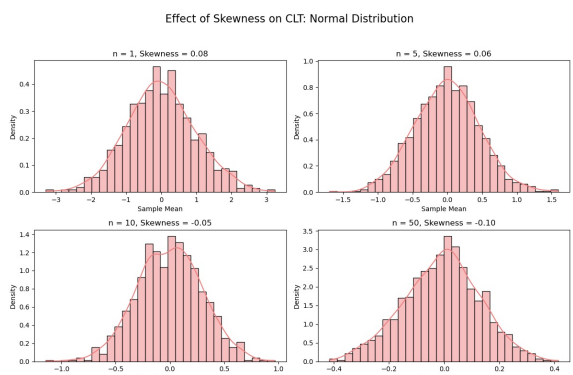
Behaviour of sample mean and varience :-

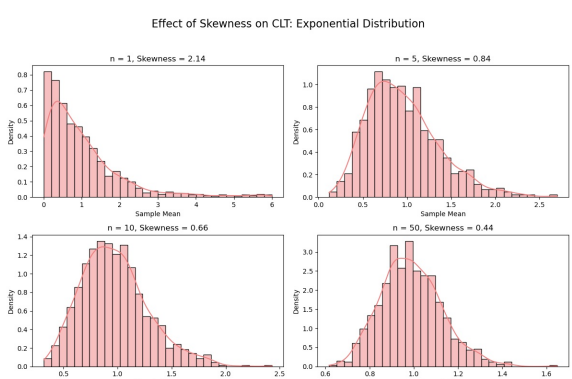
- Symmetric distribution sample mean and varience stabilize quickly as sample size increases.
- Skewed distribution, it exhibits greater variability and takes longer to stabilize

Conclusion:-

CLT hold across all types of distribution regardless of skewness but rate of convergence heavily depend on original population characteristics.







Problem-2

Given : Ten bottles tested data.

Days : 108, 124, 124, 106, 115, 138, 163, 159, 134, 139.

(a) Define the population mean shelf as μ. The hypothesis as follow

$H_0 = \mu \leq 120$ (Null hypothesis : The mean shelf life is less than or equal to 120)

$H_a = \mu > 120$ (Alternative hypothesis : shelf life exceeds 120 days.)

(b) Sample size $(n) = 10$

Null hypothesis mean $(\mu_0) = 120$

Sample mean $(\bar{x})$

$$\bar{x} = \frac{108 + 138 + 124 + 163 + 124 + 159 + 106 + 134 + 115 + 139}{10} = \frac{1310}{10}$$

$$\boxed{\bar{x} = 131.0}$$

Sample standard deviation (σ)

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{3438}{9} = \boxed{382.0}$$

$$\boxed{\sigma = 19.54}$$

Calculate t-statistic

Formula :- $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{131 - 120}{19.54/\sqrt{10}} = \frac{11}{6.18} \approx 1.78$$

$$\boxed{t = 1.78}$$

Critical value :-

Degree of freedom $= n-1 = 9$ (df)

From t distribution table for $\alpha = 0.05$ and $df = 0.05$

$$\boxed{t_{critical} = 2.821}$$

Decision :-

$t_{stat} = 1.78$ and $t_{crit} = 2.821$

Since $1.78 < 1.833$, we fail to reject null hypothesis

Conclusion :-

There is insufficient evidence to support the claim that shelf life of beverage exceeds 120 days at 1% significance level.

(c) find P value for $\alpha = 0.01$

→ using t distribution to find the p-value

For one tailed test the p value is calculated as

$$P = 1 - CDF(t, df)$$

$$P = 1 - CDF(1.78, 9)$$

$CDF(1.78, 9) \approx 0.946$

From t distribution table

$$P = 1 - 0.946 = \boxed{0.054}$$

(d) confidence interval of 99% on mean shelf life

$$\boxed{CI = \bar{x} \pm t_{\alpha/2} \cdot s/\sqrt{n}}$$

$t_{\alpha/2} = 3.25$ for 99% confidence $(\alpha/2 = 0.005)$

$$CI = 131.0 \pm 3.25 \cdot \frac{19.54}{\sqrt{10}}$$

$CI = 131.0 \pm 20.07$

$$\boxed{CI = (110.93, 151.07)}$$

The 99% confidence interval for mean life is b/w 110.91 and 151.09. Since the interval also include 120 days, it supports the hypothesis test's conclusion that mean shelf life might not exceed 120 days.

(e) Explain type I and type II errors

Type I Error (α) :- Rejecting the null hypothesis (H_0) when its true

- Example : concluding a drug when it doesn't (false positive)
- Risk : Controlled by significance level (α, e.g, 0.05)

Type - II error (β) :- Failing to reject the null hypothesis (H_0) when it is false.

- Example :- Missing that a drug works when it actually does (false negative)
- Risk :- Reduced by increasing sample size or test power.

Error	Reality	Decision
Type I	H_0 true	Incorrect rejection (α)
Type II	H_0 false	Incorrect acceptance (β)

Trade off :- Lowering Type I error increases Type II error and vice-versa. Adjust sample size to reduce both.