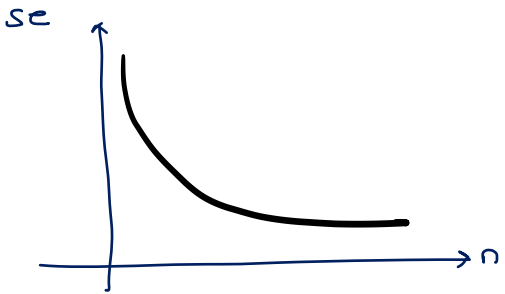


Central Limit Theorem

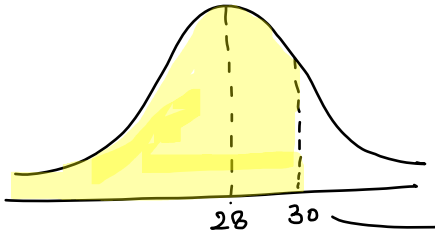
Note:
Sample doesn’t have any fixed size. It depends on feasibility and flexibility of the person performing sampling. For example a person doing survey on a population would not have this much time and money to reach out to 1 Lakh people but a brand like TATA can easily do survey on 1 M people.
According to statistics, sample size should be sufficiently large and generally sample size of 30 is considered statistically good enough but we should always try to get as large sample as possible to reduce error in the estimations.

Quiz Question

Sumit uses his mobile phone for X minutes each day to chat with his girlfriend Ankita. X is a random variable which may be modelled by a normal distribution with mean 28 minutes and standard deviation 8 minutes.

Find the probability that on a particular day Sumit chat's with Ankita using his mobile phone for less than 30mins:

A) 0.52
B) 0.63
C) 0.59
D) 0.55

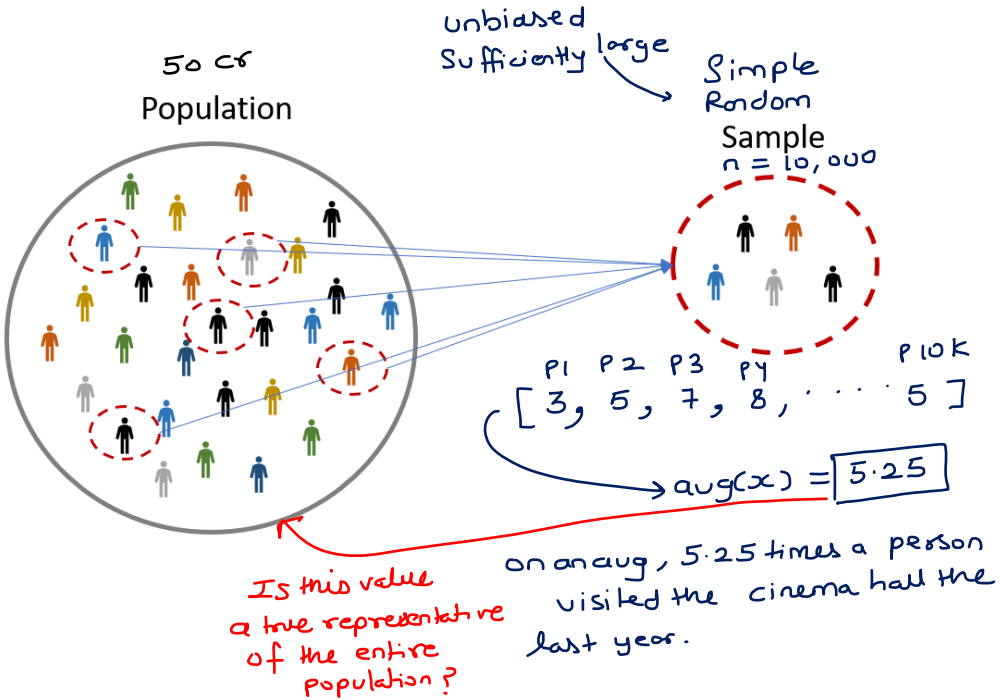


$N(28, (8)^2)$

$z = \frac{x-\mu}{\sigma} = \frac{30-28}{8}$

normcdf( )

=



You work for a small marketing firm and your manager wants you to find out on an average how many times a person in urban India visited the cinema hall the last year?

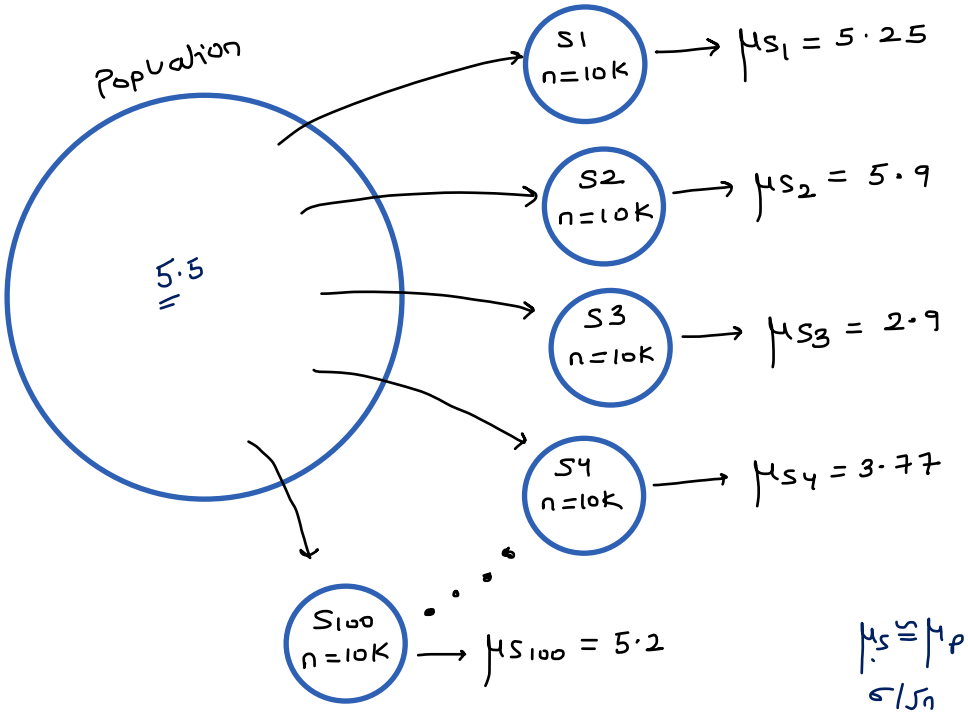
Problems:

1. It's impossible to survey on the entire population.
2. The company is not capable of putting this much time, money and resources.

I am being 100% confident ← (1) ✓ The avg no. of times a person visited the cinema hall the last year is 5.25 times.

Since we have performed survey on a sample, we can't be 100% confident and hence it's better to go with a range with some degree of confidence. ← (2) ✓ The avg no. of times a person visited the cinema hall the last year is b/w 4.5 times to 5.5 times and I am 90% confident about the same.

I am 90% confident that the actual population avg will lie somewhere between 4.5 to 5.5



$\mu_S \cong \mu_P$

$\sigma/\sqrt{n}$

list of all sample averages

μ_{S_1} μ_{S_2} μ_{S_3} μ_{S_4} . . . . $\mu_{S_{100}}$

[5.25, 5.9, 2.9, 3.77, ... , 5.2]

$n=10$

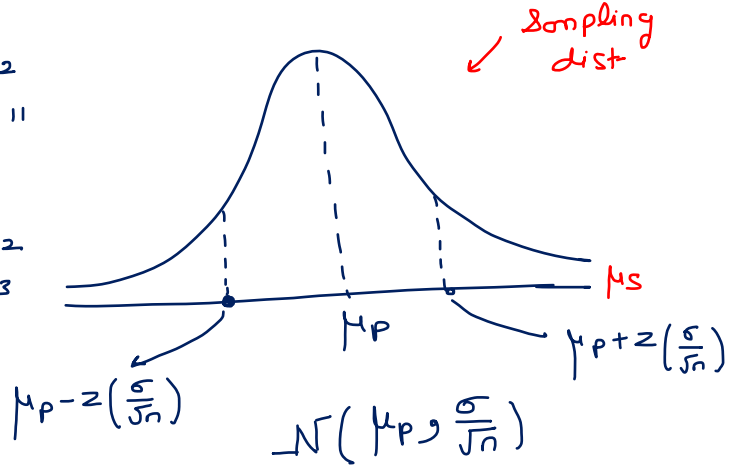
$\mu_{S_1} = 2$

$\mu_{S_2} = 11$

$n=100$

$\mu_{S_1} = 5.2$

$\mu_{S_2} = 5.3$



Central limit theorem (CLT)

① The avg of all sample averages is approx. equals to the population avg

$\mu_S \cong \mu_P$

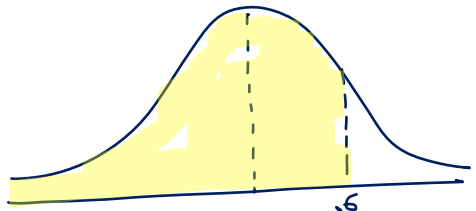
② The stdard deviation of the sampling dist $= \frac{\sigma}{\sqrt{n}}$

- σ → population std. deviation $\cong \sigma_S$
- n → # of observation in the sample.

The average time taken for a customer to complete a purchase is 4 minutes with a standard deviation of 1 minute. What is the probability that the average time for the next 5 customers is less than 6 minutes?

Population info. $N(4, (1)^2)$

Sample $n = 5$



$z = \frac{x - \mu}{\sigma/\sqrt{n}}$

$N(4, 1/\sqrt{5})$

0.99

$z = \frac{6 - 4}{1/\sqrt{5}}$

we don't need to actually perform the entire Sampling process

If n is large, and your sample mean is μ_s then

Sampling dist $= N(\mu_s, \frac{\sigma}{\sqrt{n}})$

Sample info $\rightarrow$ CLT

The sample mean recovery time of 100 patients after taking a drug was seen to be 10.5 days with a standard deviation of 2 days

Find the 95% confidence interval of the true mean.

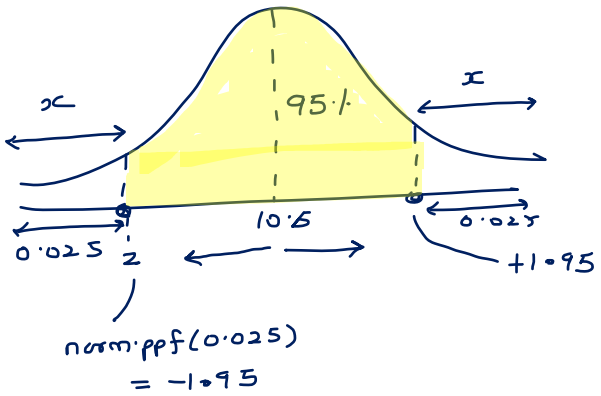
$n = 100$

$\mu_s = 10.5 \cong \mu_p$ (as per CLT)

$SE = \frac{\sigma}{\sqrt{n}} = \frac{\sigma_s}{\sqrt{n}} = \frac{2}{\sqrt{100}}$

That there are 95% chances that the actual population recovery time of the patients who will take this drug would be b/w 10.11 to 10.89 days

$= N\left(10.5, \left(\frac{2}{\sqrt{100}}\right)\right)$



norm.ppf(0.025)
$= -1.95$

$x + 0.95 + x = 1$

$2x = 1 - 0.95$

$2x = 0.05$

$x = 0.025$

$\mu - 1.95\left(\frac{\sigma}{\sqrt{n}}\right), \mu + 1.95\left(\frac{\sigma}{\sqrt{n}}\right)$

$10.5 - 1.95\left(\frac{2}{\sqrt{100}}\right), 10.5 + 1.95\left(\frac{2}{\sqrt{100}}\right)$

$(10.11, 10.89)$

The mean Youtube watch time of a sample of 100 students was found to be 3.5 hours, with a standard deviation of 1 hour.
Construct a 90% confidence interval for the true watch time.

$n = 100$

$\mu_S = 3.5 \approx \mu_P$

$\sigma_S = 1 \approx \sigma$

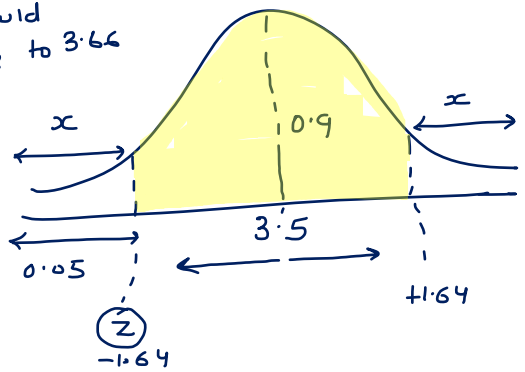
$N\left(\mu_P, \frac{\sigma}{\sqrt{n}}\right)$

$N\left(3.5, 1/\sqrt{100}\right)$

$(3.33 \text{ to } 3.66)$

① I am 90% confident that the actual pop. mean would lie somewhere b/w 3.33 to 3.66

② 90% of population of students are using youtube for 3.33 hrs to 3.66 hrs.



norm.ppf(0.05)

=

$x + 0.9 + x = 1$

$2x = 1 - 0.9$

$2x = 0.1$

$x = \frac{0.1}{2}$

$x = 0.05$

$\mu - 1.64(Se)$, $\mu + 1.64(Se)$

$3.5 - 1.64\left(\frac{1}{\sqrt{100}}\right)$, $3.5 + 1.64\left(\frac{1}{\sqrt{100}}\right)$

f_1 f_2 f_3 f_4 f_5 f_6 ... f_{100}

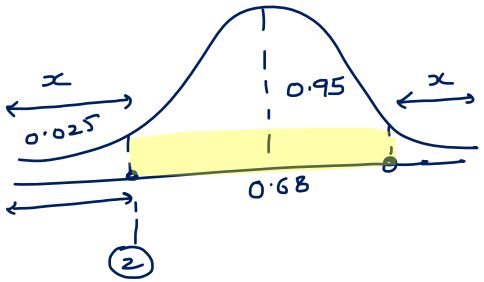
[1, 0, 1, 0, 1, 1 ..... 1]

Sumit was head over heels for Ankita, but he wasn't sure if she felt the same way. He decided to conduct a survey among their mutual friends to estimate the proportion of people who thought Ankita liked him back. Sumit followed through with the survey and found that 68 out of the 100 friends believed Ankita had feelings for him. Assume that the population standard deviation (σ) for the proportion of friends who think Ankita likes Sumit is known to be 0.2. Using the Central Limit Theorem, construct a 95% confidence interval for the true proportion of friends who think Ankita likes Sumit.

$\mu = \frac{68}{100} = 0.68$

$\sigma = 0.2$

64 to 71 friends believed ~~sumit~~ lik. ankit like sumit



norm.ppf(0.025)

-1.96

$N(0.68, 0.2/\sqrt{100})$

$x + 0.95 + x = 1$

$0.95 + 2x = 1$

$2x = 1 - 0.95$

$x = 0.025$

$\mu - 1.96(SE), \mu + 1.96(SE)$

CLT ⟶ Assumption ⟶ sample size sufficiently large

Confidence Interval Using Bootstrapping

Actual data

Person	1	2	3	4	5	6
Salary	20	37	17	50	53	33

(1) Each Bootstrapped sample is of same length as actual data

(2) observations may repeat

(3) not all observation from actual data are the part of bootstrapped sample

B.S-1

Person	2	2	6	5	3	6
Salary	37	37	33	53	17	33

B.S-2

Person	6	1	1	5	2	3
Salary	33	20	20	53	37	17

B.S-3

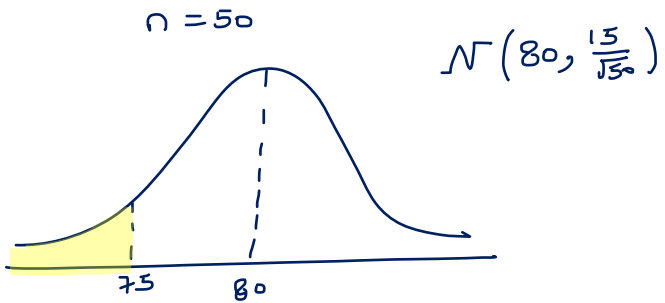
Person	2	3	3	4	5	6
Salary	37	17	17	50	53	33

35|30|34.5

Quiz-1: In an e-commerce website, the average purchase amount per customer is 80 with a standard deviation of 15. } pop. info

If we randomly select a sample of 50 customers,

what is the probability that the average purchase amount in the sample will be less than $75?



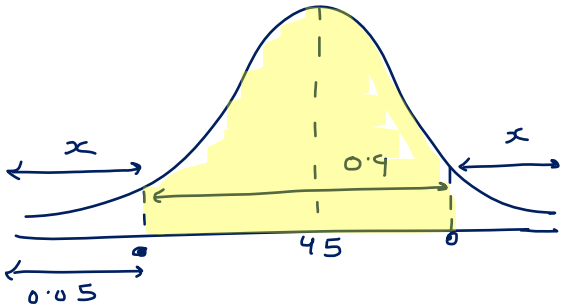
Quiz-2: From a sample of 80 endangered birds, the average wingspan was found to be 45 cm, with a population standard deviation of 10 cm. What is the correct confidence interval of the mean wingspan of the entire population with 90% confidence.

$n = 80$

$\mu_s = 45\,cm$

$\sigma = 10\,cm$

$N\left(45cm, \frac{10}{\sqrt{80}}\right)$



$x + 0.9 + x = 1$

$2x = 1 - 0.9$

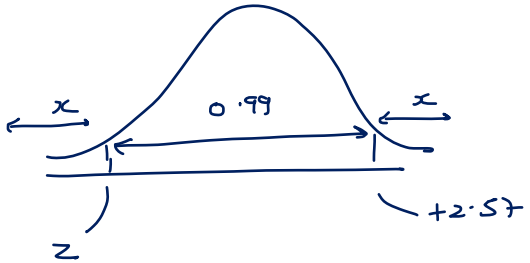
$2x = 0.1$

$x = 0.05$

norm.ppf(0.05) = −1.64

$(\mu - 1.64(SE), \mu + 1.64(SE))$

Quiz-2: In a software project, the team estimates bug resolution time at an average of 6 hours with a standard deviation of 2 hours. To estimate the mean resolution time with 99% confidence, the project manager samples 25 resolved bugs. What is the correct confidence interval?



$x + 0.99 + x = 1$

$2x = 1 - 0.99$

$x = \frac{0.01}{2}$

$x = 0.005$

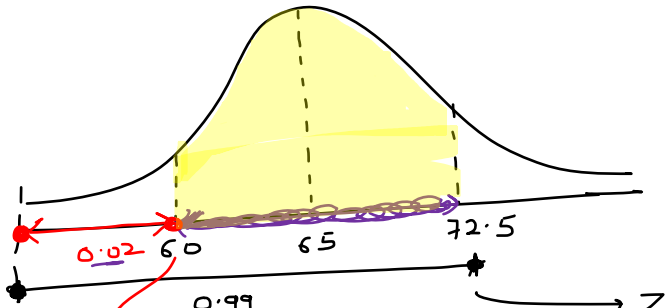
norm.ppf(0.005)

-2.57

$6 - 2.57\left(\frac{2}{\sqrt{25}}\right)$, $6 + 2.57\left(\frac{2}{\sqrt{25}}\right)$

the height of people in gaussian with mean 65 inch and std dev 2.5 what is the fraction of people whose heaight between 60 and 72.5

$$\mathcal{N}(65, (2.5)^2)$$



$$Z = \frac{(72.5 - 65)}{2.5}$$

norm.cdf(z)

$$Z = \frac{(60 - 65)}{2.5}$$

norm.cdf(z)

0.02

0.99 - 0.02

0.97

Sample very small ⟶ cannot apply CLT

[1, 5, 9, 10, 12]

BS1
BS2
BS3
⋮
BS1000

[m(BS1), m(BS2), m(BS3) . . . m(BS1000)]