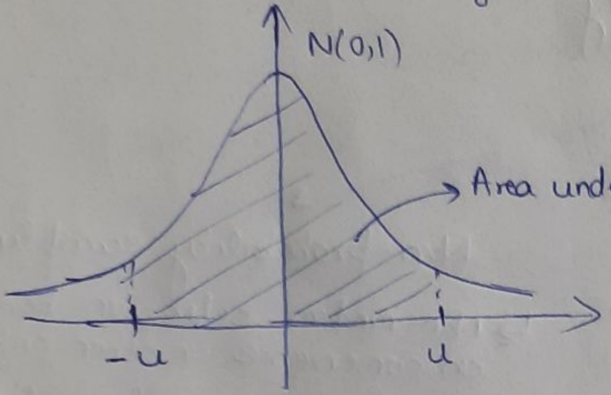
Lecture 10

* CDF and Error function:

Recall for $X \sim N(0,1)$ $\quad \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{x^2}{2}}\, dx$

Sometimes, tables are provided in terms of the error function:

$$\mathrm{erf}(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2}\, dx$$



Area under the curve $= \Phi(u) - \Phi(-u)$

$$= \Phi(u) - (1 - \Phi(u))$$

$$= 2\Phi(u) - 1$$

$$2\Phi(u) - 1 = \int_{-u}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$$

$$= 2\int_0^u \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$$

$x = \sqrt{2}y$

$dx = \sqrt{2}\, dy$

$$= 2\int_0^{u/\sqrt{2}} \frac{1}{\sqrt{2\pi}} e^{-y^2} \sqrt{2}\, dy$$

$$= \frac{2}{\sqrt{\pi}} \int_0^{u/\sqrt{2}} e^{-y^2}\, dy = \mathrm{erf}\left(\frac{u}{\sqrt{2}}\right)$$

$$\Rightarrow \boxed{2\Phi(u) - 1 = \mathrm{erf}\left(\frac{u}{\sqrt{2}}\right)}$$

* Central Limit Theorem:

Let $X_1, X_2, \ldots, X_n$ are random samples, i.e. X_i's are iid random variables with mean μ and variance σ^2.
↳ (independent and identically distributed)

$$\overline{X}_n \text{ (sample mean)} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

And, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, then ~~the~~

$$\lim_{n \to \infty} \left(\frac{\bar{X}_n - \mu}{\sigma_{\bar{x}}} \right) \rightarrow N(0,1)$$

Note: Convergence is independent of the distribution that X_i's follow.

—— x —— x —— x —— x —— x —— x ——

Types of Sampling

Probability Sampling
- ↳ involves random selection
- ↳ can make strong statistical inferences

Non-probability sampling
- ↳ Non-random selection based on convenience or other criteria
- ↳ allows you to easily collect data

* Probability Sampling

→ every member of the population has a chance of being selected

→ mainly used in quantitative research to produce results that are representative of the whole population.

Types of Probability Sampling:

① Simple Random Sampling

- Every member of population has equal chance of being selected
- Sampling frame includes entire population

Eg: Suppose you want to sample 100 employees of ~~a~~ company from 1000 employees.

~~Assign~~: Arrange employees in alphabetical order, assign them numbers from 1-1000 and use random number generator to select 100 numbers.

② Systematic Sampling

- Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Eg: List employees in alphabetical order. Select a random number from 1 to 10, say 6.
Now sample 6, 16, 26, 36, . . . ,

- Make sure that the dataset has no hidden patterns
- Suppose instead of alphabetical order, HR has them by groups and within a group, people are listed by seniority. Then sampling at periodic intervals selects ~~to~~ employees roughly at the same seniority level and not completely random.

③ Stratified sampling

- Involves dividing population into subgroups (strata) based on characteristics (like, gender, age, income, etc.)

Eg: Company has 600 female employees and 400 male employees
Select 60 women and 40 men using random/systematic sampling

④ Cluster sampling:

- Involves dividing population into subgroups, but each subgroup has similar characteristics. Randomly select an entire subgroup.
  ↳ Can do multi-stage sampling if subgroup is large.

Eg: A company has offices in 10 cities across India. Each office has a similar construct (organization structure). Then one can randomly select 3 locations and survey employees there.

* Non-probability Sampling Methods

↳ Used for exploratory or qualitative research

↳ Aim is not to test hypothesis about a broad population but to develop initial understanding

Types of Non-probability Sampling :

① Convenience Sampling

- Includes individuals/items who happen to be most accessible.
- Easy and Inexpensive, but there is no way to tell if the sample is representative of entire population.

Eg: Survey on a topic with your fellow students

② Purposive Sampling (also known as Judgement sampling)

- Involves the researcher using their expertise to select a sample that is most useful to the purposes of the research

Eg: Want to learn about the experiences of disabled students. Select students with different support needs.

③ Snowball sampling

- Recruit participants via other participants.

Eg: Suppose, we want to conduct research on homeless people. There is no database of homeless people. Suppose we manage to meet one homeless person, who then puts us in contact with others.

④ Volunteer Response Sampling.

- Instead of choosing participants, people volunteer themselves.
- One disadvantage of this type of sampling method is that people with strong views (+ve or −ve) are likely to volunteer more.

* Point and Interval Estimation

↓ Estimating a parameter (mean value)

↓ estimating its range or interval

Eg: Suppose VP of a bank needs to estimate the average balance in all the accounts in this bank.

↳ Randomly samples 500 accounts and computes the average balance of all accounts.

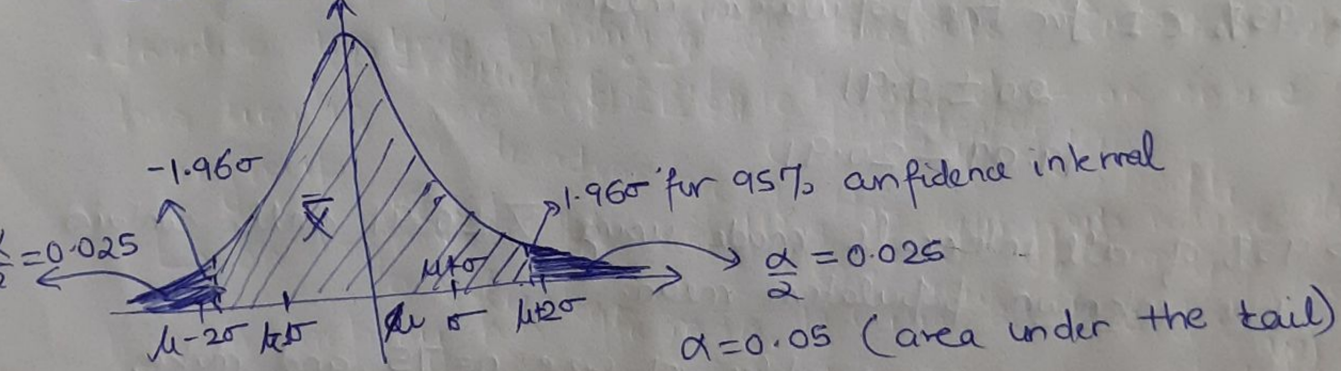
Since, this number goes in official bank records, it is important to know the "level of confidence" for the mean the VP is going to be satisfied with.

* Confidence Interval: Point Estimate ± Margin of Error
(Sample mean)

↳ affected by standard deviation, σ

- Sample size, n
- level of confidence we are satisfied with

- Standard Normal Curve:



↓
The distribution of many sample means of a given sample size.

- If the true population standard deviation σ is known, we use standard normal curve (z-curve).
- If σ is unknown, we use t-curve.

$$\underset{\substack{\downarrow \\ \text{sample mean}}}{\bar{x}} \pm \underbrace{1.96\,\sigma_{\bar{x}}}_{\text{Error Margin}} \rightarrow \text{(Confidence interval 95\%)}$$

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

- Samples of same size have the same standard error $\sigma_{\bar{x}}$.

Ex: 1 Restaurant sales:

To estimate the mean amount spent per customer at a restaurant chain, data was collected for 75 customers.

Assume $\sigma = \$4$.

① At 95% confidence interval (C.I.), what is the margin of error?

② If the sample mean is \$20, what is the 95% C.I. for the population mean?

A: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{75}} = 0.46$

$\bar{x} = \$20$

$\Big\} \quad \bar{x} \pm 1.96\,\sigma_{\bar{x}}$

- Margin of error $= 1.96\,\sigma_{\bar{x}} = 1.96 \times 0.46 = 0.91$
- 95% C.I. for the population mean

$$20 \pm 0.91$$

$$19.09 - 20.91$$

→ 95% of all intervals made using $\bar{x} \pm 0.91$ will contain the unknown population mean.

So, if we take 100 samples of size $n = 75$, and make intervals $\bar{x} \pm 0.91$, 95% of them will contain μ.

Eg: Customer Service drain

In Brick-and-Mortar stores, sales representatives end up engaging with them. Management determined that if < 15% of a salesperson's 8-hour day (4320 seconds) is spent with "false customers", then show-rooming is not a major problem.

Store samples. $n = 125$ salespersons. Assume $\sigma = 1958$ seconds and sample mean $\bar{X} = 3661.5$ seconds.

① Standard error (S.E.) of the mean:

$$\sigma_{\bar{X}} = \frac{1958}{\sqrt{125}} = 175.13 \text{ second}$$

At 95% C.I., S.E. $1.96\sigma_{\bar{X}} = 1.96(175.13)$

$= 343.25$ seconds

② C.I. at 95% confidence:

$\bar{X} \pm 343.25$ (3318.25 seconds — 4004.75 seconds)

We are 95% confident that salespersons spend b/w 3318.25 seconds and 4004.75 seconds interacting with false customers.

Policy intervention Not Needed!

* When population variance σ is unknown, we use t-curve instead of z-curve:

If s is the sample variance, then

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Student's t-distribution

↳ shape is not very different from Z-curve.

- It's shorter in the middle and fatter in the tails
  ↳ greater chance of extreme values
- t-distribution depends on the sample size
- Degrees of freedom $(n-1)$.
- Smaller the sample size, fatter the distribution
- As $n \to \infty$, t-distribution → z-distribution.