Lec IV. Sampling Of Variables

Sampling Of Variables!

- What is statistics?

Statistics is a branch of mathematics that involves collecting, analyzing interpreting and presenting data. It provides tools and method to understand and make sense of large amount of data and to draw conclusion and make decision based on data.

ie ① business data analyst — identifying customer behaviour and demand forecasting

② Government & politics — Conveying surveys & polling

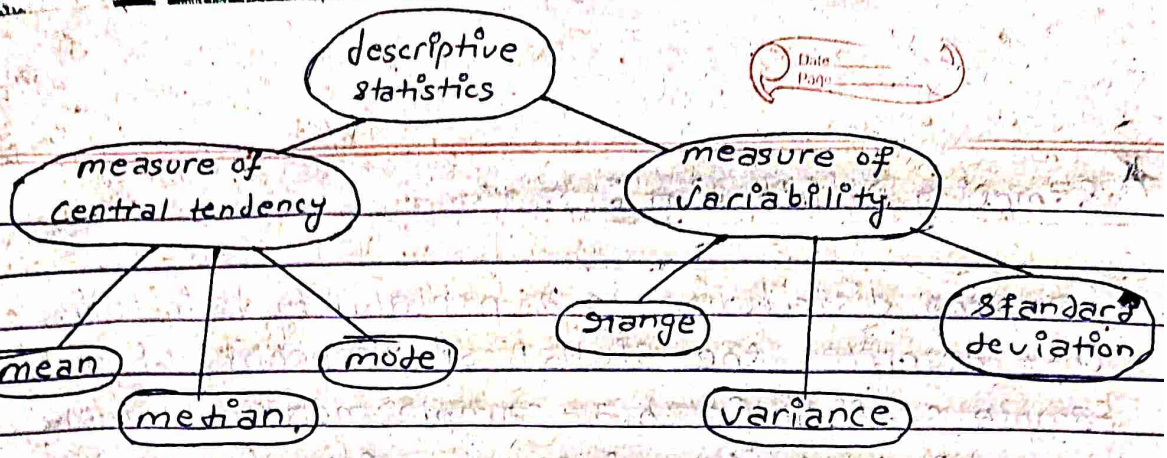
③ Medical — identify efficiency of new medicines (clinical trials)
Identifying risk factor for disease (Epidemiology)

- Types of statistics ?

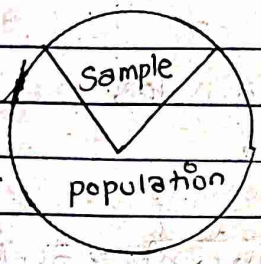
- Descriptive statistics → is a term given to the analysis of data that helps to describe, shows and summarize data in a meaningful way

It's a simple way to describe our data

Very Impo to present our raw data ineffective / meaningful way using numerical calculations or graphs or tables.



- Inferential statistics — prediction are made by taking any group of data in which you are interested



It can be defined as a random sample of data taken from a population to describe and make inferences about population

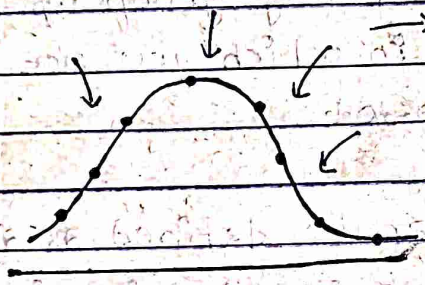
eg, hypothesis testing
= ANOVA (Analysis of Variance)
Chi- Square tests
Bayesian statistics
Sampling techniques.

Measure of Central tendency →

A statistical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is more representative of dataset as whole.

★ Sampling Distribution :

- Sampling distribution is a probability distribution that describes the statistical properties of a sample statistic (such as sample mean or sample proportion) computed from multiple independent samples of the same size from a population.



→ sample → 50 people

100 samples

population distribution (140 cr)

$x_1, x_2, x_3, \ldots, x_{50}$ → 50 people

↳ mean → $\bar{x}_1$

$x_1, x_2, x_3, \ldots, x_{50}$ → 50 people

↳ mean → $\bar{x}_2$

$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_{100}$ → 100 sample mean of population

[Sampling distribution of sample mean.]

- Why Sampling distribution is important ?

Sampling distribution is important in statistics and machine learning because it allows us to estimate the variability of a sample statistic, which is useful for making inferences about the population. By analyzing the properties of the sampling distribution, we can compute confidence intervals, perform hypothesis tests & make predictions about population based on sample data.

Central Limit Theorem →

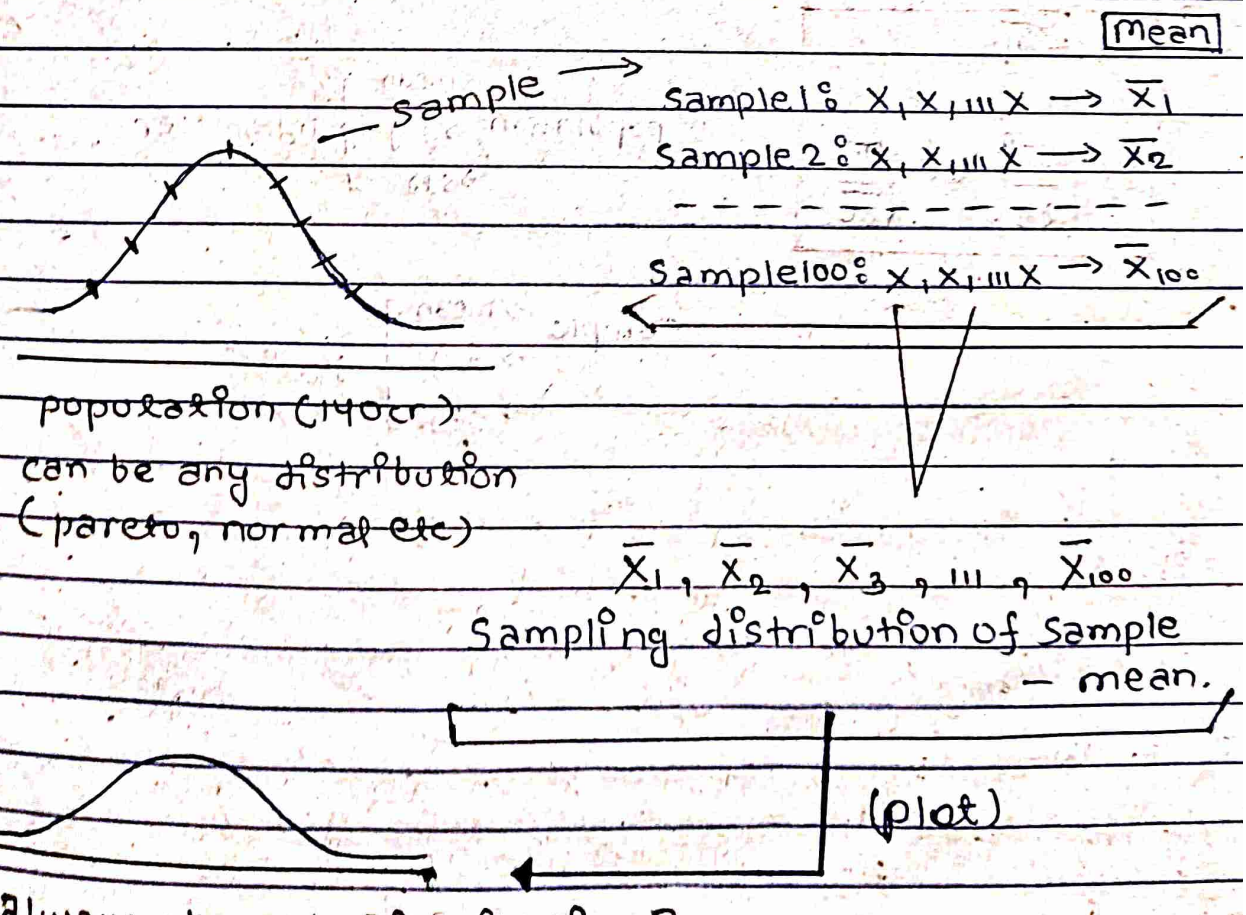
The CLT states that the distribution of the sample means of a large nos of independent and identically distributed random variables will approach a normal distribution, regardless of the underlying distribution of variables.

The conditions require for CLT to hold are –

(i) The sample size is large enough, typically greater than or equal to 30.

(ii) The sample is drawn from a finite population or an infinite population with a finite variance

(iii) The random variables in the sample are independent and identically distributed.



parameter & statistics —

characteristics	population (parameter)	sample (statistic)
Symbols	population size = N	sample size = n
	population mean = μ	sample mean = $\bar{x}$
	population SoD = σ	sample SoD = S
	population proportion = p	sample prop = $\hat{p}$

by applying CLT,

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

$$S = \frac{\sigma^2}{n} \text{ or } \frac{\sigma}{\sqrt{n}}$$

population → mean, avg, std → parameter

sample → mean, avg, std → statistic

Q) The records of weight of the male population follows the normal distribution, It's mean & std. dev. are 70kg & 15kg respectively. If a reasearchers consider a records of 50 males, then what would be the mean & std. dev. of chosen sample?

Mean of population = μ = 70 kg
Std. dev. of the population = σ = 15 kg
Sample size = n = 50

Mean of sample $\Rightarrow \bar{x}$
$\mu = \bar{x}$ = 70 kg

Std. dev. of sample $\Rightarrow s$

$$s = \frac{\sigma}{\sqrt{n}} \Rightarrow \frac{15}{\sqrt{50}} \Rightarrow 2.122 \text{ or } 2.12 \text{ kg (approx).}$$

Q) There are 250 dogs at a dog show who weigh an average of 12 pounds, with an std. dev. of 8 pounds If 4 dogs are chosen at random, what is the probability they have an average weight of greater than 8 pound and less than 25 pound?

$P(8 < \bar{X} < 25)$ $\quad Z = \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

$$\Rightarrow \frac{25-12}{\frac{8}{\sqrt{4}}} \Rightarrow \frac{13}{4} \Rightarrow 3.25$$

A z-score of 3.25 has an area of roughly 0.4.07

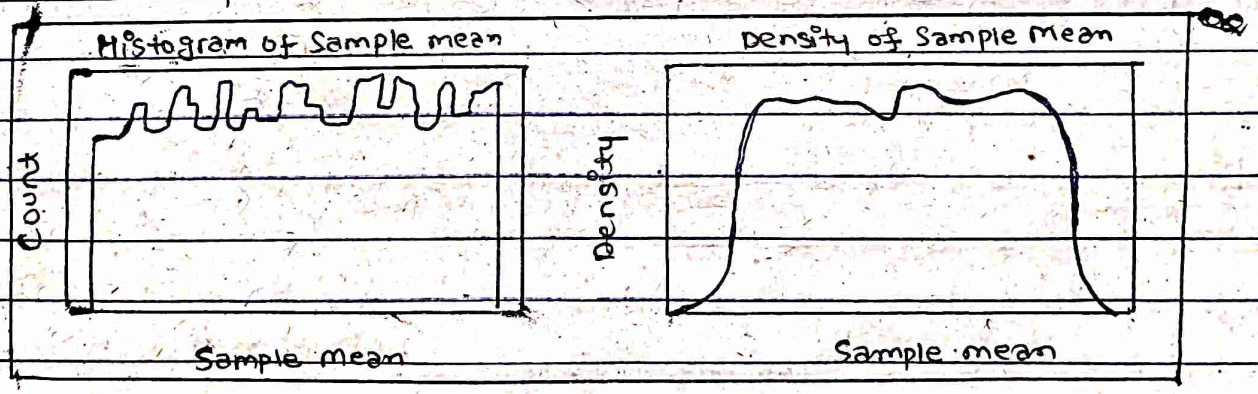
Code of Central Limit Theorem :

```
import numpy.random as np
import seaborn as sns
import matplotlib.pyplot as plt

population_size = 1000000
population = np.rand(1000000)

# population of size 1000000 consisting of
# random numbers.

nos_of_samples = 10000
sample_means = np.rand(nos_of_samples)
sample_size = 1

# as of now sample mean is randomly initialized
# later, it'll be used to store the means of sample
# drawn from population.

c = np.rand(nos_of_samples)
for i in range(0, nos_of samples):
    c = np.randint(1, population_size,
                   sample_size)

    sample_means[i] = population[c].mean()

# we run a for loop 10000 times. Each time
c take up value btw 1 & population_size
and size of c is same as 'sample size'

# The sample is drawn from population and
its mean is stored in 'sample_mean'.
```

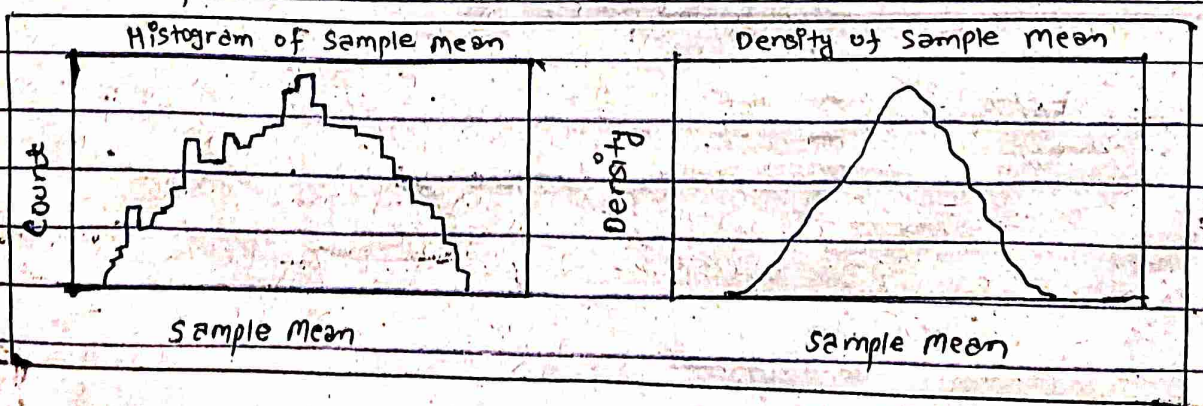
```
plt.subplot(1,2,1)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
sns.distplot(sample_means, bins=int(180/5),
             hist=True, kde=False)
plt.title('Histogram of sample mean', fontsize=20)
plt.xlabel('Sample mean', fontsize=20)
plt.ylabel('Count', fontsize=20)
plt.subplot(1,2,2)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)

sns.distplot(sample_means, hist=False, kde=True)
plt.title('Density of sample mean', fontsize=20)
plt.xlabel('Sample mean', fontsize=20)
plt.ylabel('Density', fontsize=20)

plt.subplots_adjust(bottom=0.1, right=2, top=0.9)
```

Op→

sample size = 1



sample size = 2



Sample size = 30