(*) Seaborn, KDE plot: Kernal Density Estimation plot.
→ combination of histogram and distribution.
→ sns.kdeplot(df['sepal length'])

(*) Distplot: Histogram + KDE Plot.

(*) KDE:

```
for col in ['sepal length', 'sepal width', 'petal length', 'petal width']:
    sns.kdeplot(df[col], shade=True)
```

(*) pairplot: plots pairwise relationships in a dataset.

→ sns.pairplot(df, hue='species', size=2.5);

(*) Heat Map:

→ flights_long = sns.load_dataset('flights')
flights = flights_long.pivot("month", "year", "passengers")

→ f, ax = plt.subplots(figsize=(9,6))
sns.heatmap(flights, annot=True, fmt="d", linewidth=.5, ax=ax)

(*) boxplot: sns.boxplot(y=df["Exchange Rate"])

Business statistics

→ Statistical Analysis is meant to collect and study the information available in large quantities
→ Branch of Mathematics where computation is done over a bulk of data.
→ Data collected for analysis here is called Measurements
→ if we need to measure the data, a sample is taken out a Population
→

(*) Types of Data:

① Descriptive stats describes characteristics of dataset.
② three basic categories of measure: ① Measures of central Tendency
② Measures of Variability
③ frequency distribution
③ central Tendency: center of dataset (Mean, Median, Mode)
④ Variability: Variance, (Standard deviation)

→ Measures of frequency distribution describe the occurence of data within a dataset (count)

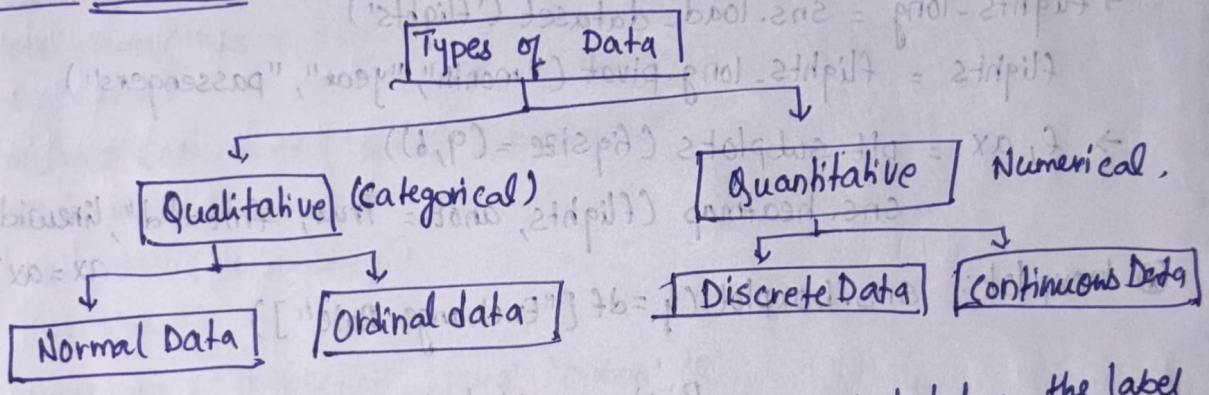
① Mean = Avg = $\frac{\Sigma()}{N} = \frac{25+30+15+25+35}{5} = \frac{26}{}$

② Median : 15, 25, 25, 30, 35 →
Median.

③ Mode: 25 → Max no. of occurences.

→ Standard Deviation (σ) = $\sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$ ↑ variance.

Inferential statistics:



→ Nominal Data: type of qualitative information which helps the label variables without providing numerical value.

→ Ordinal Data: They can be ordered → High, Mid, low
→ Used Mostly in Surveys, finance → These data are called ordinal data

Quantitative: Numerical data represent numerical value

① Discrete Data: Takes only discrete values.
Ex: No. of workers in a Company.
Ex: 40, 45, 50, ...

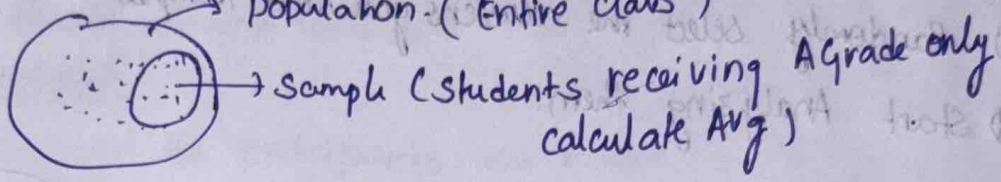
② Continuous Data: Data that can be counted
Ex: Temperature Range (97.3, 103.2, ...)

Sampling Techniques

① population: The Entire group you want to draw conclusions about.

② Sample: Subset of population. The specific group of individuals that you will collect data from.

→ In real world scenarios, we never work on population Data



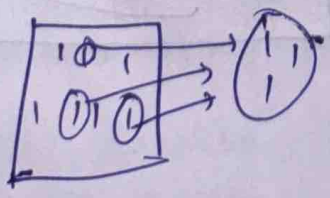
→ why sampling is important?

① Make information faster.
② Easily Analyse the data
③ Lesser Data collection Error.

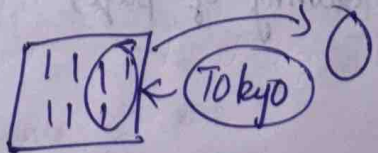
→ Types of sampling:

① probability sampling
→ Every member of population has an equal chance of being selected.

Examples:
- Simple random sampling
- Stratified random sampling
- cluster sampling
- Systematic sampling

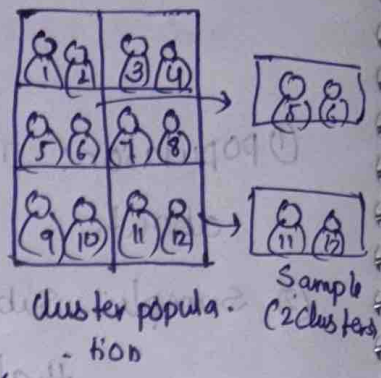


② Non probability sampling
→ selection basis of judgement (or) the convenience of accessing data.
→ Largely depends on a researchers sample selection skills
→ Examples:
- convenience sampling
- purposive sampling
- voluntary response
- Snow ball sampling



⑤ Cluster random Sampling:

① Divide population into groups
② Then randomly select group from all the groups



⑥ Steps to form clusters:

① Population → diverse, No repetition in clusters
→ Cluster should cover the entire population Members

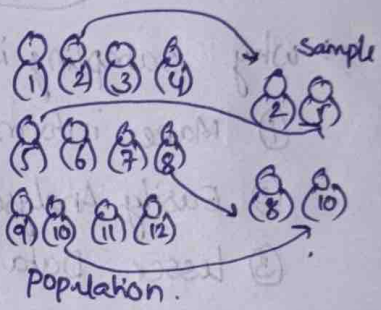
② Cluster it

③ Randomly select the clusters you need /samples out of it.

④ Start Analyzing them.

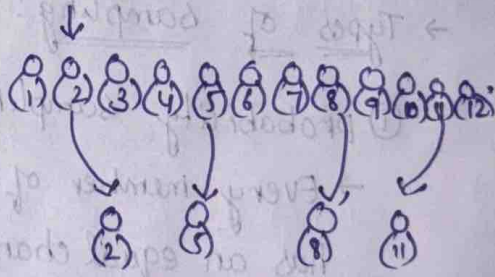
⑦ Random Sampling:

→ Randomly chose a member from the population
→ Every member and set of member has an equal chance of being selected



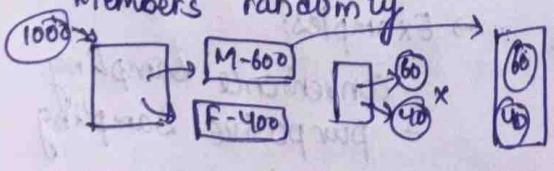
⑧ Systematic Sampling:

→ put a member of the population in some order and a starting point is choose as random the every "nth" member is selected to be in a sample.



⑨ Stratified Sampling

→ first divide population into groups
→ Then from each group we select Members randomly



⑩ Cluster Random sampling (explained starting of page)

Non probabilistic Sampling

1. convenience Sampling: Sampling based on the survey form. Include respondants/member who are easy to reach for researcher.
2. purposive Sampling: Select sample based on purpose of research
   → Researcher select sample by using their expertise and knowledge.
3. Voluntary Response sampling:
   → Based on ease of access
   → people volunteer to it, ex: LinkedIn polls.
4. Snowball Sampling:
   → Recruit the participants via research participants for test (or) study.
   → Used where its hard to find potential population for research.
   → like a subcontract.

population Sampling

→ Analysing or testing entire population is impossible and also a cost & time taking. we use sample
→ Subset of population which represent entire population.
→ Errors can lead to inaccurate and misleading result

Standard Deviation: Measure of how spread the numbers are. $\sqrt{v}$ (σ)

Variance: $\frac{\Sigma(x_i-\mu)^2}{N}$ $\mu \rightarrow$ Avg/Mean

for Sampling: it is going to be (N-1) instead of N.

But why N-1 and Not N?

→ Bessels correction.
→ corrects bias in estimation of population variance
→ partially corrects bias in estimation of population SD

Example:

x_i	$x_i - \mu$	$(x_i-\mu)^2$
5	5-16 = -11	121
10	-6	36
15	-1	1
20	4	16
30	14	196
		$\Sigma(x_i-\mu)^2 = 370$

$\mu = \frac{\Sigma x_i}{N} = \frac{80}{5} = 16$

$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}} = \sqrt{\frac{370}{5}} = \sqrt{74}$ ← Var