Micromasters “SDS” Module One

Introduction to Statistics and Data

-Dr. Fauwaz Parkar

Importance of Statistics in Data Science

- - Understand and describe data
- - Make inferences and predictions
- - Identify relationships and patterns
- - Drive decisions using data

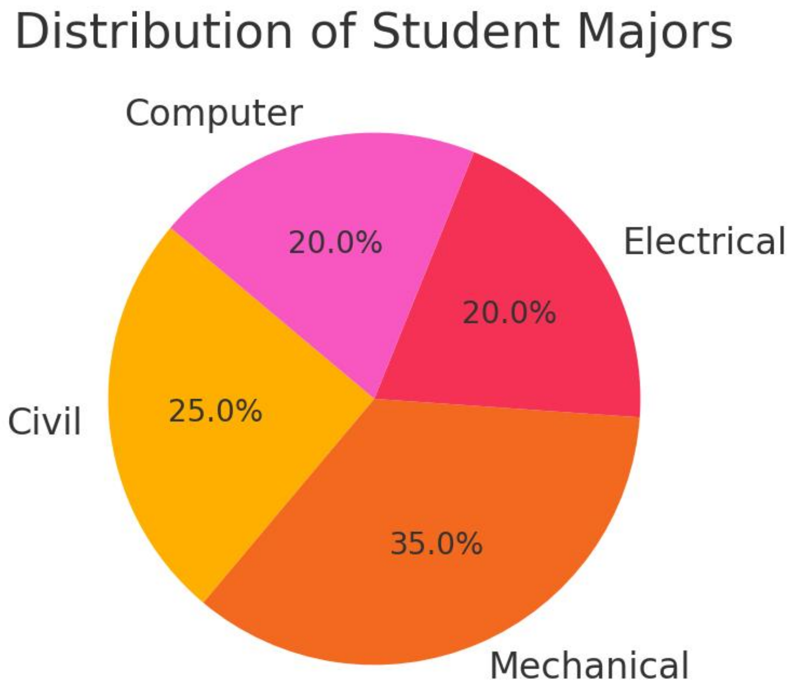
- Example:
- Average monthly sales = ₹45,000
- Prediction: Next month’s sales = ₹47,000 ± ₹2,000

Types of Data – Overview

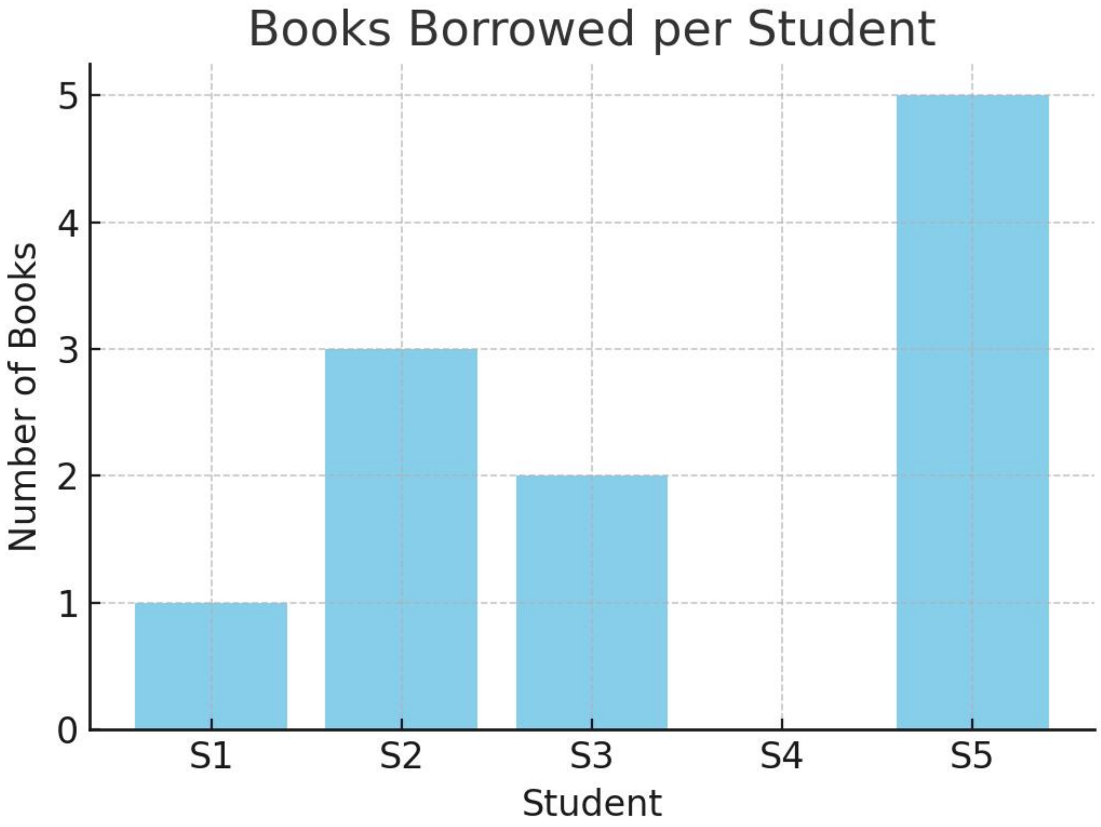
- - Categorical (Qualitative): Labels, e.g., City, Gender
- - Numerical (Quantitative): Discrete and Continuous

- Examples:
- - Discrete: Number of cars
- - Continuous: Distance in km

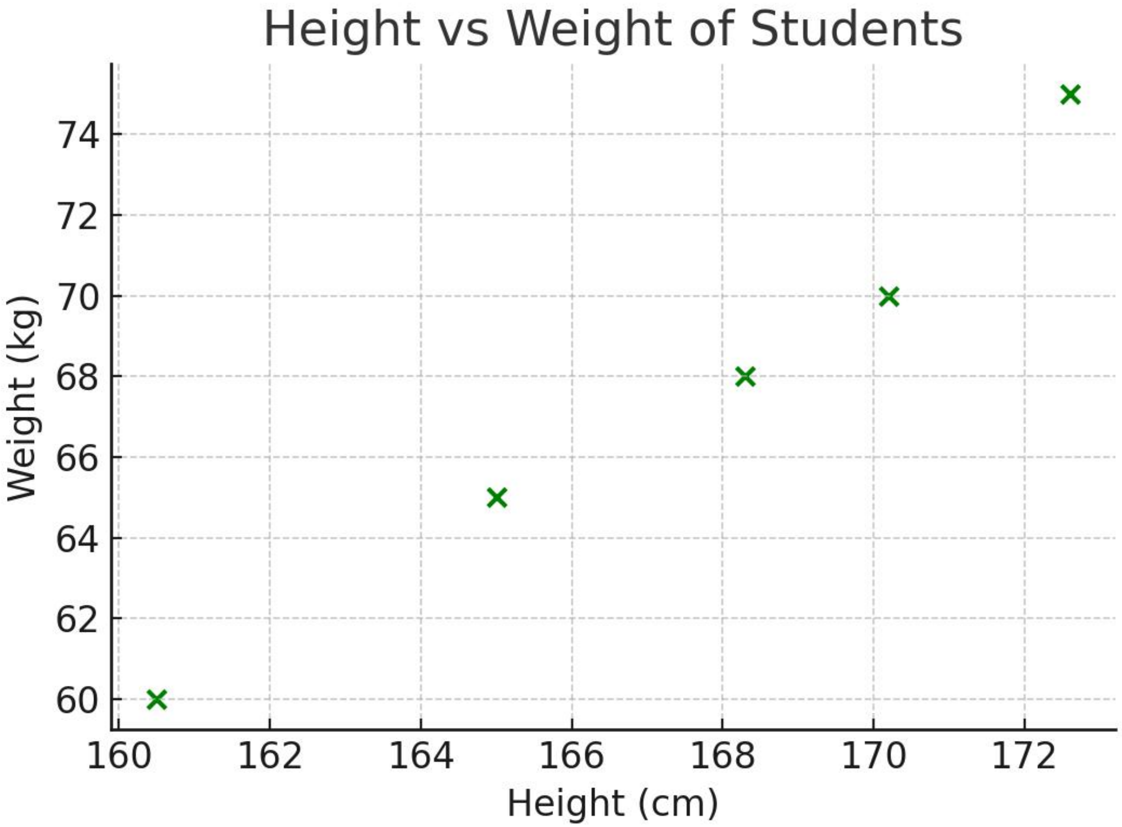
Categorical Data – Pie Chart



Discrete Data – Bar Chart



Continuous Data – Scatter Plot



Levels of Measurement

- 1. Nominal: Labels (e.g., Gender)
- 2. Ordinal: Ordered (e.g., Poor to Excellent)
- 3. Interval: Equal gaps (e.g., Temperature °C)
- 4. Ratio: True zero (e.g., Salary)

Data Collection Techniques

- - Surveys: Questionnaires and forms
- - Observations: Watching and recording
- - Experiments: Controlled tests
- - Secondary Data: Using existing datasets

Sampling Techniques

- 1. Simple Random: Equal chance for all
- 2. Stratified: Sample from subgroups
- 3. Systematic: Every kth element
- 4. Cluster: Randomly selected groups

Sampling – Numerical Example

- Population: 1,000 students
- Systematic Sampling with k=10
- Selected: 1st, 11th, 21st, ..., 991st

In short...

- - Statistics enables understanding and prediction
- - Data types and measurement levels are foundational
- - Visual tools (charts, plots) aid analysis
- - Sampling ensures reliable data insights

- Statistics is a tool for converting ***data*** into ***information***:

Statistics

Data

Information

But where then does ***data*** come from? How is it gathered? How do we ensure its accurate? Is the data reliable? Is it representative of the population from which it was drawn? This module explores some of these issues.

Methods of Collecting Data…

- There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:
- Direct Observation
- Experiments, and
- Surveys.

TYPES OF DATA

1) PRIMARY DATA : Are those which are collected a **fresh** and for the **first time** and thus happen to be **original in character** and known as Primary data.

2) SECONDARY DATA : Are those which have been **collected by someone else** and which have **already been passed** through the statistical process are known as Secondary data.

SOURCES OF DATA

The sources of data may be classified into

(a) Primary sources

(b) Secondary sources.

Primary Sources

Primary sources are original sources from which the researcher directly collects data that have not been previously collected, e.g., collection of data directly by the researcher on brand awareness, brand preference, brand loyalty and other aspects of consumer behaviour from a sample of consumers by interviewing them. Primary data are first-hand information collected through various methods such as observation, interviewing, mailing etc.

Secondary Sources

These are sources containing data that have been collected and compiled for another purpose. The secondary sources consist of readily available compendia and already compiled statistical statements and reports whose data may be used by researches for their studies, e.g., census reports, annual reports and financial statements of companies, Statistical statements, Reports of Government Departments, Annual Reports on currency and finance published by the National Bank for Ethiopia, Statistical Statements relating to Cooperatives, Federal Cooperative Commission, Commercial Banks and Micro Finance Credit Institutions published by the National Bank for Ethiopia, Reports of the National Sample Survey Organisation, Reports of trade associations, publications of international organisations such as UNO, IMF, World Bank, ILO, WHO, etc., Trade and Financial Journals, newspapers, etc.

Secondary sources consist of not only published records and reports, but also unpublished records. The latter category includes various records and registers maintained by firms and organisations, e.g., accounting and financial records, personnel records, register of members, minutes of meetings, inventory records, etc.

Features of Secondary Sources: Though secondary sources are diverse and consist of all sorts of materials, they have certain common charac-teristics.

First, they are readymade and readily available, and do not require the trouble of constructing tools and administering them.

COLLECTION OF PRIMARY DATA

- There are several methods of collecting primary data, particularly in surveys and descriptive researches. In descriptive research, we obtain primary data either through observation or through direct communication with respondents in one form or another or through personal interviews.

COLLECTION OF SECONDARY DATA

- These are already available i.e. they refer to the data which have **already been collected and analyzed by someone else.**
- Secondary data may either be published or unpublished data. Researcher must be very careful in using secondary data, because the data available may be sometimes unsuitable.

Methods of data Collection :Primary Data

- 1) OBSERVATION METHOD : Observation method is a method under which data from the field is collected with the help of observation by the observer or by personally going to the field.
- In the words of P.V. Young, "Observation may be defined as systematic viewing, coupled with consideration of seen phenomenon."

ADVANTAGES:

- Subjective bias eliminated (No bias info)
- Information researcher gets is Current information
- Independent to respondent's variable (as in interview and may be bias)
- DISADVANTAGES :
- It is expensive method (time requires more)
- Limited information
- Unforeseen factors may interfere with observational task
- Respondents opinion can not be recorded on certain subject

TYPES OF OBSERVATION

Structured and Unstructured Observation

- When observation is done by characterizing style of recording the observed information, standardized conditions of observation , definition of the units to be observed , selection of pertinent data of observation then it is structured observation
- When observation is done without any thought before observation then it is unstructured observation

Participant & Non Participant Observation

- When the Observer is member of the group which he is observing then it is Participant Observation
- In participant observation Researcher can record natural behavior of group , Researcher can verify the truth of statements given by informants in the context of questionnaire , Difficult to collect information can obtain through this method but in this researcher may loose objectivity of research due emotional feelings. Prob. of control in observation isn’t solved.

Non Participant Observation

- When observer is observing people without giving any information to them then it is non participant observation

Controlled & Uncontrolled Observation

- When the observation takes place in natural condition i.e. uncontrolled observation. It is done to get spontaneous picture of life and persons
- When observation takes place according to definite pre arranged plans , with experimental procedure then it is controlled observation generally done in laboratory under controlled condition.

INTERVIEW METHOD

- This method of collecting data involves presentation or oral-verbal stimuli and reply in terms of oral-verbal responses.
- Interview Method This is Oral Verbal communication . Where interviewer asks questions(which are aimed to get information required for study) to respondent

There are different type of interviews as follows :

PERSONAL INTERVIEWS :

The interviewer asks questions generally in a face to face contact to the other person or persons.

Types of Personal Interview

Personal Interview

- Predetermined questions
- Standardized techniques of recording
- Interviewer follows rigid procedure laid down i.e. asking questions in form & order prescribed
- Time required for such interview is less than non structured interview
- Not necessary of skill or specific knowledge
- Analysis of data becomes easier Bcoz information is collected in prescribed manner

Structured Interview

- Flexibility in asking questions
- No Predetermined questions
- No Standardized techniques of recording
- Interviewer has freedom to ask , omit , add questions in any manner
- Ask questions without following sequence
- Deep knowledge & skill required
- Analysis of data is difficult

Merits of Personal Interview

- Information at greater depth
- Flexibility of restructuring the Questionnaire
- Interviewer by his skill can come over resistance
- Non Response generally low
- Samples can controlled more effectively
- Personal information can be obtained

- Interviewer can collect supplementary information about respondent’s personal characteristics and environment which has value in interpreting results

De Merits Of Interview

- Expensive method
- Respondent may give bias information
- Some Executive people are not approachable so data collected may be inadequate
- Takes more time when samples are more
- Systematic errors may be occurred
- Supervisors has to do complex work of selecting ,training and supervising the field staff.

TELEPHONIC INTERVIEWS

- Contacting samples on telephone
- Uncommon method may be used in developed regions

MERITS

- Flexible compare to mailing method
- Faster than other methods
- Cheaper than personal interview method
- Callbacks are simple and economical also
- High response than mailing method.
- when it is not possible to contact the respondent directly, then interview is conducted through – Telephone.

- Replies can be recorded without embarrassment to respondents
- Interviewer can explain requirements more easily
- No field staff is required
- Wider distribution of sample is possible

DEMERITS

- Little time is given to respondents
- Survey is restricted to respondents who have telephones
- Not suitable for intensive survey where comprehensive answers are required
- Bias information may be more
- Very difficult to make questionnaire because it should short and to the point

- structured interviews : in this case, a set of pre-decided questions are there.

- unstructured interviews : in this case, we don’t follow a system of pre-determined questions.

- focused interviews : attention is focused on the given experience of the respondent and its possible effects.

- clinical interviews : concerned with broad underlying feelings or motivations or with the course of individual’s life experience, rather than with the effects of the specific experience, as in the case of focused interview.

- group interviews : a group of 6 to 8 individuals is interviewed.
- qualitative and quantitative interviews : divided on the basis of subject matter i.e. whether qualitative or quantitative.
- individual interviews : interviewer meets a single person and interviews him.
- selection interviews : done for the selection of people for certain jobs.
- depth interviews : it deliberately aims to elicit unconscious as well as other types of material relating especially to personality dynamics and motivations.

QUESTIONNAIRE METHOD

- This method of data collection is quite popular, particularly in case of big enquiries. The questionnaire is mailed to respondents who are expected to read and understand the questions and write down the reply in the space meant for the purpose in the questionnaire itself. The respondents have to answer the questions on their own.
- Questionnaire Method Questionnaire is sent to persons with request to answer the questions and return the questionnaire Questions are printed in definite order , mailed to samples who are expected to read that questions understand the questions and write the answers in provided space .

Merits of Questionnaire

- Merits of Questionnaire Low cost even the geographical area is large to cover Answers are in respondents word so free from bias Adequate time to think for answers Non approachable respondents may be conveniently contacted Large samples can be used so results are more reliable

Demerits of Questionnaire

- Demerits of Questionnaire Low rate of return of duly filled questionnaire Can be used when respondent is educated and co operative It is inflexible Omission of some questions Difficult to know the expected respondent have filled the form or it is filled by some one else Slowest method of data collection

Main Aspects of Questionnaire

- Main Aspects of Questionnaire General Form Structured Questionnaire Alternatives or yes no type questions are asked Easy to interpret the data but unuseful for the survey which is aimed to probe for attitudes, and reasons for certain actions Unstructured Questionnaire open ended questions

- Respondents gives answers in his own words On the basis of the pre test researcher can decide about which type of questionnaire should be used Question Sequence Question sequence should be clear and smoothly moving (relation of one question to another should readily apparent First question important for creating interest in respondents mind

- Question which gives stress on memory or of a personal character and wealth should be avoided as opening questions Easier question should be at the start of the questionnaire General to specific questions should be the sequence of questions Question Formulation and Wording Question should easily understood Question should be simple and concrete.

- Closed questions are easy to handle but this is like fixing the answers in people’s mouth. So depending upon problem for which survey is going on both close ended and open ended question may be asked in Questionnaire. Words having ambiguous meaning should be avoided, catch words ,words with emotional connotations , danger words should be avoided

Essentials of Good Questionnaire

- Essentials of Good Questionnaire Should Short & simple Questions should arranged in logical sequence (From Easy to difficult one) Technical terms should avoided Some control questions which indicate reliability of the respondent (To Know consumption first expenditure and then weight or qty of that material)
- Questions affecting the sentiments of the respondents should avoided Adequate space for answers should be provided in questionnaire Provision for uncertainty (do not know, No preference) Directions regarding the filling of questionnaire should be given Physical Appearance - - Quality of paper, color

HOW TO CONSTRUCT A QUESTIONNAIRE

Researcher should note the following with regard to these three main aspects of a questionnaire:

- General form
- Question Sequence
- Determine the type the Questions :
  - A) Direct Question
  - B) Indirect Question
  - C) Open Form Questionnaire
  - D) Closed Form Questionnaire
  - E) Dichotomous Questions
  - F) Multiple Choice Questions (MCQ)

SCHEDULE METHOD

- It is one of the important methods for the study of social problems.
- Schedules Like Questionnaires but it filled by enumerator . Enumerators are specially appointed for filling questionnaire Enumerators explain the aim and objective to respondent and fill the answers in provided space .
- In the words of Thomas Carson Macormic, “The schedule is nothing more than a list of questions which it seems necessary to test the hypothesis .”

Questionnaire V/S Schedule

Questionnaire	Schedule
• Q generally sent through mail and no further assistance from sender	• Schedule is filled by the enumerator or research worker
• Q is cheaper method	• Costly requires field workers
• Non Response is high	• Non response is low

Questionnaire

- In questionnaire it is not confirmed that expected respondent have filled the answers
- Very slow method
- No Personal contact
- Q can be used only when respondent is educated and co operative

Schedule

- In Schedule identity of person is known
- Information is collected well in time
- Direct personal contact
- Info can collected from illiterates also

- Wider distribution of sample is possible
- Incomplete and wrong information is more
- Depends on quality of questionnaire
- Physical appearance of questionnaire should attractive
- Observation method can not use

- Difficulty for wider area
- Relatively more correct and complete
- Depends on Honesty and competence of enumerator
- Not necessary in Schedule method
- It is possible to use observation at the time of filling schedule by enumerator

Other Methods Of Data Collection

- **Warranty Cards** Post card size cards sent to customers and feedback collected through asking questions on that card
- **Distributor or Store Audits** Audits are done by distributor or manufacturer's salesperson. Observation or copying information about inventory in retail shops. Useful method for knowing market share ,market size , effect of in store promotion.

- **Pantry Audits** From the observation of pantry of customer to know purchase habit of the people (which product , of what brand etc.) Questions may be asked at the time of audit
- **Consumer Panels** When pantry audit is done at regular basis, Daily record of consumption of certain customers. Or repeatedly interviewed at the specific periods to know their consumption.
- **Transitory consumer panels** – for limited time Continuing Consumer panel For indefinite period

- **Use of Mechanical Device** Eye Cameras to record eyes focus on certain sketch
- Psycho galvanometer to measure body excitement to visual stimulus
- Motion Picture camera to record movement of body at the time of purchase
- Audiometer concerned to TV . Useful to know Channel, program preference of people

- **Depth Interview** To discover the underlying motives or desires of samples . To explore needs , feelings of respondents. Skill is required , indirect question or projective techniques are used to know behavior of the respondent.
- **Content Analysis** analyzing contents of documentary material as news paper , books , magazines about certain characteristics to identify and count

CASE STUDY METHOD

- It is essentially an intensive investigation of the particular unit under consideration. Its important characteristics are as follows :

a) the researcher can take one single social unit or more of such units for his study purpose.

b) the selected unit is studied intensively i.e. it is studied in minute details.

SURVEY METHOD

- One of the common methods of diagnosing and solving of social problems is that of undertaking surveys.
- Festinger and Kat of the opinion that, "Many research problems require systematic collection of data from population through the use of personal interviews or other data gathering devices".

PANEL METHOD

In this method, data is collected from the same sample respondents at the some interval either by mail or by personal interview. This is used for studies on :

- 1) Expenditure Pattern
- 2) Consumer Behaviour
- 3) Effectiveness of Advertising
- 4) Voting Behaviour and so on

Secondary Data

Sources of data

- Publications of Central, state , local government
- Technical and trade journals
- Books, Magazines, Newspaper
- Reports & publications of industry ,bank, stock exchange
- Reports by research scholars, Universities, economist
- Public Records

Factors to be considered before using secondary data

- Reliability of data – Who, when , which methods, at what time etc.
- Suitability of data – Object ,scope, and nature of original inquiry should be studied, as if the study was with different objective then that data is not suitable for current study
- Adequacy of data– Level of accuracy,
- Area differences then data is not adequate for study

Selection of proper Method for collection of Data

- Nature ,Scope and object of inquiry
- Availability of Funds
- Time Factor
- Precision Required

Surveys…

- A ***survey*** solicits information from people; e.g. Gallup polls; pre-election polls; marketing surveys.

- The ***Response Rate*** (i.e. the proportion of all people selected who complete the survey) is a key survey parameter.

- Surveys may be administered in a variety of ways, e.g.
- Personal Interview,
- Telephone Interview,
- Self Administered Questionnaire, and
- Internet

Questionnaire Design…

- Over the years, a lot of thought has been put into the science of the design of survey questions. Key design principles:

1. Keep the questionnaire as short as possible.
2. Ask short, simple, and clearly worded questions.
3. Start with demographic questions to help respondents get started comfortably.
4. Use dichotomous (yes|no) and multiple choice questions.
5. Use open-ended questions cautiously.
6. Avoid using leading-questions.
7. Pretest a questionnaire on a small number of people.
8. Think about the way you intend to use the collected data when preparing the questionnaire.

Sampling…

- Recall that statistical inference permits us to draw conclusions about a population based on a sample.

- Sampling (i.e. selecting a sub-set of a whole population) is often done for reasons of ***cost*** (it's less expensive to sample 1,000 television viewers than 100 million TV viewers) and ***practicality*** (e.g. performing a crash test on every automobile produced is impractical).

- In any case, the ***sampled population*** and the ***target population*** should be **similar** to one another.

Sampling Plans…

- A ***sampling plan*** is just a method or procedure for specifying how a sample will be taken from a population.

- We will focus our attention on these three methods:

- Simple Random Sampling,
- Stratified Random Sampling, and
- Cluster Sampling.

- Random sampling,by far, is the most common one used.

Simple Random Sampling…

- A ***simple random sample*** is a sample selected in such a way that every possible sample of the same size is equally likely to be chosen.
- Drawing three names from a hat containing all the names of the students in the class is an example of a simple random sample: any group of three names is as equally likely as picking any other group of three names.
- VERY EASY TO DEFINE!
- VERY, VERY DIFFICULT TO DO!
- Random sample of 100 cokes bottles today at the coke plant.
- Random sample of 50 pine trees in a 1000 acre forest.
- Random sample of 5 deer in a national forest.

Simple Random Sampling…

- A government income tax auditor must choose a sample of 5 of 11 returns to audit…[Can do many different ways]

Person	Generate Random #
baker	0.87487
george	0.89068
ralph	0.11597
mary	0.58635
sally	0.34346
joe	0.24662
andrea	0.47609
mark	0.08350
greg	0.53542
aaron	0.37239
kim	0.73809

	Person	Sorted Random #
1	mark	0.08350
2	ralph	0.11597
3	joe	0.24662
4	sally	0.34346
5	aaron	0.37239
	andrea	0.47609
	greg	0.53542
	mary	0.58635
	kim	0.73809
	baker	0.87487
	george	0.89068

Stratified Random Sampling…

- A ***stratified random sample*** is obtained by separating the population into mutually exclusive sets, or strata, and then drawing simple random samples from each stratum.

Strata 1 : Gender
Male
Female

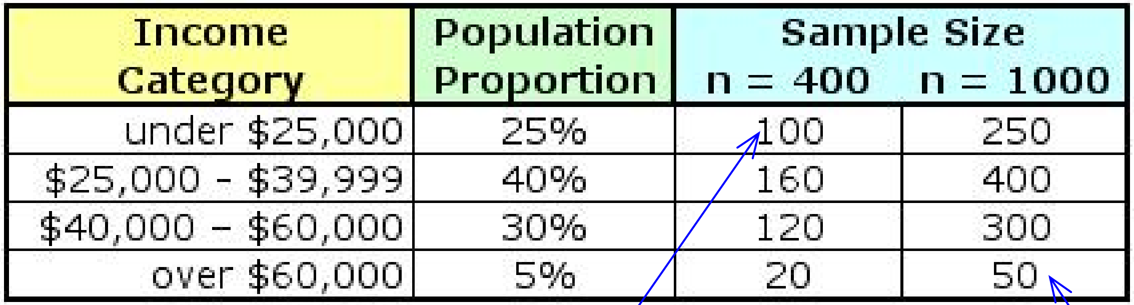
Strata 2 : Age
< 20
20-30
31-40
41-50
51-60
> 60

Strata 3 : Occupation
professional
clerical
blue collar
other

We can acquire about the total population,
make inferences within a stratum
or make comparisons across strata

Stratified Random Sampling…

- After the population has been stratified, we can use ***simple random sampling*** to generate the complete sample:



Income Category	Population Proportion	Sample Size	
		n = 400	n = 1000
under $25,000	25%	100	250
$25,000 - $39,999	40%	160	400
$40,000 – $60,000	30%	120	300
over $60,000	5%	20	50

If we only have sufficient resources to sample 400 people total, we would draw 100 of them from the low income group…

…if we are sampling 1000 people, we'd draw 50 of them from the high income group.

Cluster Sampling…

- A ***cluster sample*** is a simple random sample of groups or clusters of elements (vs. a simple random sample of individual objects).

- This method is useful when it is difficult or costly to develop a complete list of the population members or when the population elements are widely dispersed geographically. Used more in the "old days".

- Cluster sampling may increase sampling error due to similarities among cluster members.

Sample Size…

- Numerical techniques for determining sample sizes will be described later, but suffice it to say that the larger the sample size is, the more accurate we can expect the sample estimates to be.

Sampling and Non-Sampling Errors…

- Two major types of error can arise when a sample of observations is taken from a population:
- ***sampling error*** and ***nonsampling error***.
- ***Sampling error*** refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample. Random and we have no control over.
- ***Nonsampling errors*** are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly. Most likely caused be poor planning, sloppy work, act of the Goddess of Statistics, etc.

Sampling Error…

- ***Sampling error*** refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample.
- Increasing the sample size **will** reduce this type of error.

Nonsampling Error…

- ***Nonsampling errors*** are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly. Three types of nonsampling errors:

- Errors in data acquisition,
- Nonresponse errors, and
- Selection bias.

- Note: increasing the sample size **will not** reduce this type of error.

Errors in data acquisition…

- …arises from the recording of incorrect responses, due to:

- — incorrect measurements being taken because of faulty equipment,
- — mistakes made during transcription from primary sources,
- — inaccurate recording of data due to misinterpretation of terms, or
- — inaccurate responses to questions concerning sensitive issues.

Nonresponse Error…

- …refers to error (or ***bias***) introduced when responses are not obtained from some members of the sample, i.e. the sample observations that are collected may not be representative of the target population.

- As mentioned earlier, the ***Response Rate*** (i.e. the proportion of all people selected who complete the survey) is a key survey parameter and helps in the understanding in the validity of the survey and sources of nonresponse error.

Selection Bias…

- …occurs when the sampling plan is such that some members of the target population cannot possibly be selected for inclusion in the sample.

Sampling Numericals

Q1. How would you split a dataset of 10,000 customer reviews for sentiment analysis using random sampling into training (70%), validation (15%), and testing (15%) sets?

- A1. Training: 7000
- Validation: 1500
- Testing: 1500

Q2. In a classification dataset with 1200 cat images and 800 dog images, how many of each should be selected for a stratified sample of 400 images?

- A2. Cats: 240
- Dogs: 160

Q3. If system logs are recorded every second and total 90,000 entries, how many entries would you get by sampling every 30th record (systematic sampling)?

- A3. 90,000 / 30 = 3000 records

Q4. From 50 edge devices generating 1000 readings each, if you randomly select 10 devices and use all their data (cluster sampling), how many readings do you get?

- A4. 10 × 1000 = 10,000 readings

Q5. If training data from 2021 had 50% gaming and 50% social media users, but test data from 2025 has 80% gaming and 20% social media, what issue arises?

- A5. Sampling bias due to data drift – poor model generalization

Q6. To estimate a 30% spam message rate with 95% confidence and ±2% margin of error, how many samples are needed?

- A6. Use formula: $n = (Z^2 \times p \times (1 - p)) / E^2$
- $Z = 1.96$, $p = 0.3$, $E = 0.02 \rightarrow n \approx 2018$

Q7. In bootstrapping, if you generate 1000 samples (size 300 each) from a dataset of 300, how many unique samples are expected in each bootstrap sample?

- A7. Approx. $300 \times (1 - 1/e) \approx 190$ unique samples

Q8. A city survey includes 1200 households across 30 localities. You randomly choose 5 localities and survey all households there. What sampling method is this?

- A8. Cluster Sampling – entire clusters (localities) are selected and fully surveyed.

Q9. From a batch of 500 concrete cubes tested, you want to select a representative sample of 50 using systematic sampling. What should be the interval?

- A9. Interval = 500 / 50 = every 10th cube

Q10. A production line manufactures 10,000 bolts daily. You inspect every 100th bolt for quality. What sampling technique is this?

- A10. Systematic Sampling – every 100th item is checked.

Q11. You want to assess machine wear from 500 machines, where 100 are randomly selected and grouped by age for analysis. What is this approach?

- A11. Stratified Random Sampling – grouped (stratified) by machine age.

Q12. From a power plant with 200 sensors, you want to select a representative sample for maintenance analysis. You use a random number generator to pick 40 sensors. Identify the sampling type.

- A12. Simple Random Sampling

Q13. In a signal processing experiment, you record 100,000 samples. You take every 500th sample to analyze signal stability. What type of sampling is used?

- A13. Systematic Sampling – selecting at regular intervals.