Detailed Summary of Lecture 17: Sampling

1 Introduction to Sampling

- Sampling is the process of selecting a subset of individuals, items, or data points from a larger population for the purpose of estimating characteristics of the whole population.
- Studying an entire population is often not feasible due to constraints in time, cost, and effort. Sampling provides a practical alternative.
- A **sample** is a smaller, manageable representation of the larger population.
- A **population** refers to the complete set of individuals or items that possess some common characteristic under investigation.
- Types of populations:
  - **Finite Population**: Has a countable number of elements (e.g., students in a college).
  - **Infinite Population**: The total number of elements cannot be counted (e.g., coin toss outcomes).
- A well-selected sample ensures that generalizations made about the population are reliable and valid.

2 Sample Design and Terminologies

- **Sampling Frame**: A complete list or map that includes all members of the population from which the sample will be drawn (e.g., voter list, class roster).
- **Sampling Unit**: The basic element or group of elements considered for selection in a sampling process (e.g., person, household, farm).
- **Sample Size**: The number of units to be included in the sample. A larger sample size improves accuracy but increases cost and complexity.
- **Sample Statistics**: Values computed from the sample data, such as sample mean, variance, and proportion. These are used to estimate population parameters.
- **Population Parameters**: True values (e.g., population mean, variance) that describe the entire population. Often unknown and estimated using sample statistics.

3 Principal Steps in a Sample Survey

1. **Define Objectives**:
    - Clearly define the goals and purposes of the survey.
    - Objectives should be practical, specific, and aligned with available resources.
2. **Define the Population**:
    - Clearly specify what constitutes inclusion in the population.
    - Ambiguity in defining the population can lead to biased results.
3. **Frame and Sampling Units**:
    - Ensure that all elements in the population are covered.
    - Units must be distinct and non-overlapping.
    - A good sampling frame is essential for effective sample selection.
4. **Data to Collect**:
    - Decide in advance what data are necessary.
    - Avoid collecting irrelevant or excessive information.
    - Draft an outline of the tables you want to produce post-survey.
5. **Design Questionnaire or Schedule**:
    - Prepare clear, concise, and unbiased questions.
    - The questionnaire may be self-administered or interviewer-administered.
    - Include clear instructions and ensure logical flow of questions.
6. **Data Collection Methods**:
    - *Interview Method*: Investigator meets respondents directly and fills out the questionnaire.
    - *Mailed Questionnaire*: Sent to respondents to be filled out and returned.
    - Choose the method based on literacy level, cost, accuracy, and coverage.
7. **Non-response Handling**:
    - Some selected respondents may not respond due to absence or refusal.
    - Investigators must record reasons and take steps to minimize non-response bias.
8. **Select Sampling Design**:
    - Choose an appropriate sampling method (e.g., simple random, stratified, cluster).
    - Estimate an appropriate sample size for desired precision and confidence level.

- Consider cost and logistics before finalizing the design.

9. **Organize Field Work**:
   - Train enumerators to locate and interact with sampling units.
   - Pretesting helps to identify potential issues in questionnaire and methodology.
   - Proper supervision ensures data quality.
10. **Data Analysis**:
    - **Editing and Scrutiny**: Check for inconsistencies or incomplete responses.
    - **Tabulation**: Summarize data into tables for easier analysis.
    - **Statistical Analysis**: Apply estimation formulas and compute confidence intervals or significance levels.
    - **Reporting**: Document methodology, findings, and limitations clearly.
11. **Learning for Future Surveys**:
    - Record lessons learned regarding design, execution, and analysis.
    - Use these insights to improve accuracy and efficiency in future surveys.

Lecture 18: Parameter, Statistic, Sampling and Non-Sampling Errors

1 Parameter and Statistic

- **Parameter**: Statistical constants of the population, such as mean, variance, etc., are called parameters.
- **Statistic**: Measures computed from sample observations alone (e.g., sample mean, sample variance) are called statistics.
- In practice, parameter values are usually unknown and are estimated based on sample values. Thus, a statistic is an estimate of the parameter, derived solely from sample values.
- Since there are multiple possible samples from a population, a statistic varies from sample to sample.
- Characterizing the variation in the values of a statistic obtained from different samples (attributed to chance or sampling fluctuations) is a fundamental problem in sampling theory.

2 Sampling Distribution

- The number of possible samples of size n that can be drawn from a finite population of size N is ${}^{N}C_{n}$.
- For each sample, a statistic (e.g., mean, variance) can be computed, and these values will vary from sample to sample.
- The aggregate of all such values (one from each sample) forms a frequency distribution known as the **sampling distribution** of the statistic.
- Thus, we can have the sampling distribution of the sample mean $\bar{x}$, the sample variance, etc.

3 Standard Error

- The standard deviation of the sampling distribution of a statistic is called its **Standard Error (S.E.)**.
- The standard errors of some well-known statistics are commonly tabulated, where n is the sample size, σ^2 is the population variance, P is the population proportion, and $Q = 1 - P$.

4 Sampling and Non-Sampling Errors

The errors in data collection, processing, and analysis can be broadly classified as:

1. **Sampling Errors**
2. **Non-sampling Errors**

4.1 Sampling Errors

- Originate from the use of only a part of the population (sample) to estimate population parameters.
- Absent in a complete enumeration (census) survey.
- Sampling biases can arise due to:
  (1) **Faulty selection of the sample:** Use of defective sampling techniques (e.g., purposive or judgment sampling) may introduce bias. This can be minimized by using simple random sampling or random sampling with restrictions that do not introduce bias.
  (2) **Substitution:** Replacing a selected unit with a convenient member if difficulties arise in enumeration introduces bias, as the substitute may have different characteristics.
  (3) **Faulty demarcation of sampling units:** Especially significant in area surveys (e.g., agricultural experiments), where discretion in including borderline cases can introduce bias.
  (4) **Constant error due to improper choice of statistic:** For example, using the sample variance formula

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

as an estimate of the population variance σ^2 is biased, whereas

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is an unbiased estimate.
- **Sample Size:** Increasing the sample size generally decreases sampling error.

4.2 Non-sampling Errors

- Arise at the stages of observation, ascertainment, and processing of data.
- Present in both sample surveys and complete enumeration surveys.
- Major sources include:

  (1) **Planning Errors:**

  - Inadequate or inconsistent data specification.
  - Errors in locating units, measurement, recording, or due to poorly designed questionnaires.
  - Lack of trained investigators or supervisory staff.

  (2) **Response Errors:**

  - *Accidental errors:* Misunderstanding questions.
  - *Prestige bias:* Upgrading or downgrading responses due to pride.
  - *Self-interest:* Providing incorrect information to protect one's interests.
  - *Interviewer bias:* Influence of interviewer's beliefs or recording style.
  - *Recall failure:* Inability to accurately remember past events.

  (3) **Non-response Biases:** Occur when full information is not obtained for all units (e.g., respondent not at home, refusal to answer).

  (4) **Coverage Errors:**

  - Inclusion of units that should not be included or exclusion of units that should be included due to unclear objectives or definitions.

  (5) **Compiling Errors:** Errors during data processing (editing, coding, tabulation, summarizing). Can be controlled by verification and consistency checks.

  (6) **Publication Errors:**

  - Mechanical errors in publication (proofing, printing).
  - Failure to point out limitations of the statistics.

Lecture 19: Sampling Frame and Sampling Techniques

Sampling Frame

The **sampling frame** (also known as the “sample frame” or “survey frame”) is the actual collection of units from which a sample is drawn. A basic random sample gives all units in the frame an equal probability of being selected.

A sampling frame is a complete list or collection from which sample participants will be drawn in a predetermined manner. This list is organized so that each member of the population has an individual identity and a contact mechanism, allowing for categorization and coding of known segmentation features.

Having a sampling frame means we have a supply or list of all individuals in the target population, as well as a process for selecting the sample. Any resource that enables access to every individual in the targeted group qualifies as a sampling frame.

Characteristics of a Good Sampling Frame

When selecting lists, ensure the sample frame is large enough for the requirements. A good sampling frame should:

- Include everyone in the target demographic.
- Exclude anyone not in the target group.
- Contain factual information to reach specific people.
- Assign a unique identification to each member (e.g., a number code from 1 to 3000).
- Be free of duplicates.
- Be well organized (e.g., sorted alphabetically).
- Be up to date, with regular checks for address or contact changes.

Examples of the Sampling Frame

Studying every individual in a population is often impractical. For example, to learn about the opinions of Nepalese bankers about vehicle ownership, surveying every bank would be too time-consuming and expensive. In such cases, a sample is investigated.

Before choosing a sample, construct a sampling frame-a list of all units in the population of interest. Study findings can only be generalized to the population identified by the sampling frame.

Conclusion

A sampling frame is a researcher's list or device to specify the population of interest. A basic random sample gives all units an equal probability of selection. Units can be people, organizations, or records. It is critical to be as detailed as possible when describing the population.

Issues of Choosing Appropriate Sampling Technique(s)

Choosing a sampling strategy is essential to ensure data reliability and representativeness. Consider a survey on characteristics (tax, education, etc.) of residents in five towns, totaling 3,200 households. These households form the target population.

Step One: Define Sample and Target Population

Sometimes a survey covers the entire target population (a census survey), but this is often impractical. Instead, a smaller, representative sample is chosen to reflect the population's characteristics. A survey on a smaller number is called a *sample survey*; findings from this can be generalized to the whole population.

Step Two: Define Sample Size

There are no strict rules for sample size; it depends on objectives, time, budget, and desired precision. To select an appropriate sample size, determine the degree of accuracy (confidence interval and confidence level):

- **Confidence interval** (margin of error): The range within which the true value is expected to fall (e.g., ±5).
- **Confidence level**: The probability that the sample reflects the population (e.g., 95%).

Example: If 65% of sampled households say "yes" to a question, with a ±5 confidence interval and 95% confidence level, we can say that between 60% and 70% of all 3,200 households would also answer "yes."

Sample size can be determined using a standard calculator. For 3,200 households:

- **Option A:** 5% confidence interval, 95% confidence level → sample size = 345 households.

- **Option B:** 5% confidence interval, 99% confidence level → sample size = 551 households.

The improvement in accuracy diminishes as sample size increases, so the choice should balance objectives and resources.

Step Three: Define Sampling Technique

After choosing the sample size, select a sampling technique based on the project's nature and objectives. Sampling techniques are broadly divided into:

- **Random Sampling**
- **Non-random Sampling**

Random Sampling

Random sampling means selecting the sample randomly from a population, often from a list or at the survey location.

- **Simple random sampling without replacement:** Each unit can be selected only once.
- **Simple random sampling with replacement:** Units may be selected multiple times.
- **Systematic sampling:** Divide the population by the sample size to get the sampling interval. E.g., for 3,200 households and a sample of 345, the interval is 9 ($3200/345 \approx 9$); select every ninth household.
- **Stratified random sampling:** For heterogeneous populations, divide into strata and sample each in proportion to its size.

Example of stratified random sampling:

Location	Population size	Proportion (%)	Stratified sample size
Town 1	1200	38%	129
Town 2	900	28%	97
Town 3	800	25%	86
Town 4	180	6%	19
Town 5	120	4%	13
Total	3200	100%	345

Non-random Sampling

Non-random sampling selects samples based on specific characteristics. Used when the sample must meet certain criteria, such as owning a car or having young children. Methods include:

- Convenience sampling
- Judgment sampling
- Quota sampling

- Snowball sampling

Step Four: Minimize Sampling Error

Efforts should focus on reducing sampling error and making the sample as representative as possible. The robustness of the sample depends on minimizing error, which varies by technique.

For random samples, results are reported with the ± sampling error. In non-random samples, the sampling error is unknown.

Summary: Use random sampling when inferring proportions of the population. Use non-random sampling when targeting specific perceptions or characteristics, especially when speed or specificity is needed.

Note: Without a sampling strategy, data may be biased or unrepresentative, rendering it invalid.

Lecture 20: Sample Size Determination

Sample Size

The size of the sample is an important factor as it directly affects the accuracy, estimation, cost, and administration of a survey. Large samples have lower sampling error, whereas small samples have higher sampling error. However, unnecessarily large samples increase costs, so an optimum sample size should be chosen to ensure efficiency, representativeness, reliability, and flexibility.

Factors Affecting Sample Size

Sample size depends on several factors:

1. **Nature of population:** If the population is homogeneous, a small sample size suffices; if heterogeneous, a larger sample is needed for representativeness.
2. **Number of classes:** More classes in the classification require a larger sample.
3. **Nature of study:** Studies that take longer may benefit from smaller samples due to financial and analytical constraints.
4. **Types of sampling used:** Simple random sampling typically needs a larger sample, while stratified sampling can achieve representativeness with a smaller sample.
5. **Degree of accuracy:** Higher accuracy requires a larger sample.

Testing Reliability of the Sample

A sample is considered reliable if it is representative of the population. Reliability can be tested by:

1. **Drawing parallel samples:** Draw another sample parallel to the original and compare measures such as average, dispersion, skewness, and kurtosis. If measures are similar, the sample is reliable.
2. **Comparing sample with population:** Compare sample statistics with population parameters. If they match closely, the sample is reliable.

3. **Drawing sub-samples from the main sample:** Compare measures from sub-samples with the main sample to detect errors due to faulty selection.

Methods of Estimating Sample Size

Estimation Using the Mean

Let $\bar{x}$ be the sample mean from a random sample of size n drawn from a population with mean μ and standard deviation σ. For a confidence level $(1-\alpha)$ and margin of error d:

$$P\left(|\bar{x}-\mu| \leq Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Setting $d = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, we get:

$$n = \frac{Z_{\alpha/2}^2\sigma^2}{d^2}$$

If σ is unknown, use the sample standard deviation s.

For a finite population of size N:

$$n = \frac{Z_{\alpha/2}^2\sigma^2}{d^2 + \frac{Z_{\alpha/2}^2\sigma^2}{N}}$$

Estimation Using Proportion

Let p be the sample proportion from a random sample of size n drawn from a population with proportion P and $Q = 1-P$:

$$P\left(|p-P| \leq Z_{\alpha/2}\sqrt{\frac{PQ}{n}}\right) = 1-\alpha$$

Setting $d = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}$, we get:

$$n = \frac{Z_{\alpha/2}^2PQ}{d^2}$$

If P is unknown, use $P = p$.

For a finite population of size N:

$$n = \frac{Z_{\alpha/2}^2PQ}{d^2 + \frac{Z_{\alpha/2}^2PQ}{N}}$$

Examples

Example 1: Determine the minimum sample size required so that the sample estimate lies within 10% of the true value with 95% confidence, given a coefficient of variation (CV) of 60%.

Solution: CV $= 0.6$, margin of error $d = 0.1\mu$, $Z_{0.025} = 1.96$.

$$0.1\mu = 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{1.96 \times \sigma}{0.1\mu}\right)^2 = 384.16 \times (0.6)^2 = 138.29 \approx 138$$

Example 2: A psychologist wants to be 99% confident that the error in estimating reaction time (standard deviation 0.05 seconds) does not exceed 0.01 seconds.

Solution: $Z_{0.005} = 2.58$, $d = 0.01$, $\sigma = 0.05$.

$$n = \frac{(2.58)^2 \times (0.05)^2}{(0.01)^2} = 166.4 \approx 167$$

Example 3: A survey in Kathmandu Valley wants to estimate the proportion of disabled persons (population proportion $p = 0.1$, desired error $d = 0.02$, 95% confidence).

Solution: $Z_{0.025} = 1.96$, $Q = 0.9$.

$$n = \frac{(1.96)^2 \times 0.1 \times 0.9}{(0.02)^2} = 864.36 \approx 865$$

Example 4: For $p = 0.2$, $d = 0.05$, $Z = 2$, find n. If $N = 1000$, find corrected n.

Solution:

$$n = \frac{4 \times 0.2 \times 0.8}{(0.05)^2} = 256$$

For $N = 1000$:

$$n_{corr} = \frac{256}{1 + \frac{256}{1000}} = 204$$

Example 5: Mean systolic blood pressure is 125 mm Hg, standard deviation 15 mm Hg. Find sample size for 5% significance and error not exceeding 2.

Solution: $Z_{0.025} = 1.96$, $d = 2$, $\sigma = 15$.

$$n = \frac{(1.96)^2 \times 15^2}{2^2} = 216.09 \approx 216$$

For $N = 500$:

$$n_{corr} = \frac{216}{1 + \frac{216}{500}} = 151$$

Yamane Formula

For a finite population, Yamane's formula is:

$$n = \frac{N}{1 + Ne^2}$$

where n = sample size, N = population size, e = acceptable sampling error.

Example: For $N = 10,000$, $e = 0.05$:

$$n = \frac{10,000}{1 + 10,000 \times (0.05)^2} = \frac{10,000}{26} = 385$$

Standard Error of the Mean

- For an infinite population:

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

- For a finite population of size N:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Note

When the original sample is more than 5% of the population, use the finite population correction. The sample size should be manageable, cost-effective, and representative.

Lecture 21: Statistical Analysis and Data Preparation

Statistical Analysis

Analysis refers to the categorizing, ordering, manipulating, and summarizing of data to answer research questions. The purpose is to reduce data to an intelligible and interpretable form so that relationships between research problems can be studied and tested. Statistics are used to manipulate and summarize numerical data, and to compare results with chance expectations. The researcher should plan analysis paradigms or models when formulating problems and hypotheses.

Data Editing

Editing is the process of examining collected data for errors and omissions, and making necessary corrections. Editing occurs in two stages:

1. **Field editing:** Review of reporting forms by the enumerator or investigator for completing what was written in abbreviated form during data collection.
2. **Central editing:** Editing of obvious errors (e.g., entries in the wrong place, missing replies) by an editor after all forms have been returned to the office.

Data Coding

Coding is the process of assigning numerals or symbols to responses so they can be grouped into a limited number of classes or categories. Quantitative data collected via questionnaires or schedules is already numeric and may not require coding. For qualitative data, numeric codes are assigned before analysis. This allows qualitative responses to be converted into numerical figures suitable for statistical treatment.

Classification of Data

Classification arranges related facts or data into groups or classes according to similarities, making data easily understandable. Classification should be:

- According to the research problem

- Exhaustive
- Mutually exclusive
- Independent

Objectives of classification:

- To condense large amounts of data
- To facilitate comparison
- To highlight features of the data at a glance
- To enable statistical analysis

Types of classification:

1. **Chronological classification:** Data is arranged by time (years, months, weeks, etc.), usually in ascending order.
   Example:

Year	1970	1971	1972	1973	1974	1975	1976
Birth rate	36.8	36.9	36.6	34.6	34.5	35.2	34.2

2. **Geographical classification:** Data is classified by region or place.
   Example:

Country	America	China	Denmark	France	Nepal
Yield of wheat (kg/acre)	1925	893	225	439	862

3. **Qualitative classification:** Data is classified by attributes or qualities such as sex, literacy, religion, employment, etc. Attributes cannot be measured on a scale. When two or more attributes are considered, a manifold classification is formed.
   Example: Classifying population by sex and employment:
   - Male employed
   - Male unemployed
   - Female employed
   - Female unemployed

   Further attributes (e.g., marital status) can extend the classification.

4. **Quantitative classification:** Data is classified by measurable characteristics such as height, weight, etc.

Data Entry into Spreadsheet

A spreadsheet is an interactive computer application for organizing and analyzing data in tabular form. Spreadsheets simulate paper accounting worksheets, with data in cells organized as rows and columns. Each cell can contain numeric or text data, or formulas that automatically calculate values. Users can change values and instantly see the effects on calculations, making spreadsheets useful for ”what-if” analysis. Modern spreadsheets can have multiple sheets and display data as text, numbers, or graphs.

Management of Missing and Inconsistent Information

Researchers often face:

- Missing data
- Impossible values
- Inconsistencies
- Transcription errors

Missing and inconsistent data are common in research. Strategies for handling them include:

- Careful questionnaire design and data entry
- Training data entry personnel
- Cross-checking responses
- Checking for impossible values
- Recording and creating composite variables
- Documenting any logical changes to raw data
- Using coding systems to reduce errors
- Data cleaning with standard software (e.g., SAS, SPSS)
- Labeling values and merging cells as needed
- Formatting data for analysis
- Maintaining a master dataset for all analysts

Listwise Deletion

Cases with any missing values are deleted from analysis. Only cases with complete data are retained. This is the default in many statistical programs but is generally not recommended.

Pairwise Deletion

The maximum amount of available data is retained. Cases are excluded only from analyses where required data is missing. For example, a case missing data on one variable is excluded from analyses involving that variable, but included elsewhere.

Descriptive Statistical Measures

Types of average:

- Arithmetic Mean
- Geometric Mean
- Harmonic Mean
- Median
- Mode

Measures of Dispersion:

- Absolute and relative measures
- Range
- Quartile deviation
- Mean deviation
- Standard deviation
- Coefficient of Variation

Skewness and Kurtosis:

$$\text{Kurtosis } K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} \qquad K = 0.263$$

Correlation and Regression

Lecture 22: Inferential Statistics and Hypothesis Testing

Inferential Statistics

Inferential statistics involve drawing conclusions about a population based on sample data. A key component is hypothesis testing, which allows us to make decisions or inferences about population parameters.

Testing of Hypothesis

Hypothesis testing is a statistical method that uses sample data to evaluate a hypothesis about a population parameter. The general steps are:

1. Formulate the null hypothesis (H_0) and alternative hypothesis (H_1).
2. Select an appropriate test statistic.
3. Determine the level of significance (α).
4. Find the critical value(s) from statistical tables.
5. Make a decision: reject or fail to reject H_0.

Z Test

The Z test is an important parametric test based on the assumption of normality. It is used when the sample size is large ($n > 30$), or when the population variance is known. For large samples, the sampling distribution of the statistic is approximately normal, even if the population is not strictly normal.

The Z test statistic is defined as:

$$Z = \frac{t - E(t)}{SE(t)} \sim N(0, 1)$$

where t is a statistic, $E(t)$ its expected value, and $SE(t)$ its standard error.

Z tests are used for:

- Significance of a single mean
- Significance of difference between two means
- Significance of a single proportion
- Significance of difference between two proportions
- Significance of difference between sample and population correlations
- Significance of difference between independent sample correlations

Test of Significance of a Single Mean

Suppose a sample of size $n > 30$ is drawn from a normal population $N(\mu, \sigma^2)$. The sample mean $\bar{x}$ is approximately normally distributed.

Steps:

1. **State the hypotheses:**
   - $H_0 : \mu = \mu_0$ (sample is from population with mean μ_0)
   - $H_1 : \mu \neq \mu_0$ (two-tailed), or $H_1 : \mu > \mu_0$ (right-tailed), or $H_1 : \mu < \mu_0$ (left-tailed)
2. **Test statistic:**
$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \text{(for known variance)}$$
$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \text{(for unknown variance, large } n\text{)}$$
   For finite population size N:
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$
3. **Level of significance:** Usually $\alpha = 0.05$ unless specified.
4. **Critical value:** Obtain from standard normal tables according to α and the alternative hypothesis.
5. **Decision:** Reject H_0 if $|Z| > Z_{\text{tabulated}}$; otherwise, accept H_0.

Example:

A sample of 400 students has a mean height of 170 cm. Can it be regarded as a sample from a large population with mean 169.5 cm and standard deviation 3.5 cm?

Solution:
$$Z = \frac{170 - 169.5}{3.5/\sqrt{400}} = \frac{0.5}{0.175} = 2.857$$

At $\alpha = 0.05$, $Z_{\text{tab}} = 1.96$. Since $2.857 > 1.96$, reject H_0. The sample cannot be regarded as from the population with mean 169.5 cm.

Test of Significance of Difference Between Two Means

Suppose two independent samples of sizes n_1 and n_2 are drawn from populations with means μ_1, μ_2 and variances σ_1^2, σ_2^2. Let $\bar{x}_1$, $\bar{x}_2$ be the sample means.

Steps:

1. **State the hypotheses:**
   - $H_0 : \mu_1 = \mu_2$
   - $H_1 : \mu_1 \neq \mu_2$ (two-tailed), or one-tailed as appropriate
2. **Test statistic:**
$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$
   If population variances are unknown (large samples), use sample variances s_1^2, s_2^2.
3. **Level of significance:** Usually $\alpha = 0.05$.
4. **Critical value:** Obtain from standard normal tables.
5. **Decision:** Reject H_0 if $|Z| > Z_{\text{tabulated}}$.

Example:

Sample 1: $n_1 = 500$, $\bar{x}_1 = 20$, $\sigma_1 = 4$
Sample 2: $n_2 = 400$, $\bar{x}_2 = 15$, $\sigma_2 = 4$

$$Z = \frac{20 - 15}{\sqrt{\frac{16}{500} + \frac{16}{400}}} = \frac{5}{0.27} = 18.51$$

At $\alpha = 0.05$, $Z_{\text{tab}} = 1.96$. Since $18.51 > 1.96$, reject H_0. The samples are not from the same population.

Test of Significance of Difference Between Two Proportions

Let P_1 and P_2 be the population proportions, p_1 and p_2 the sample proportions from independent samples of sizes n_1 and n_2.

Steps:

1. **State the hypotheses:**
   - $H_0 : P_1 = P_2$
   - $H_1 : P_1 \neq P_2$ (two-tailed), or one-tailed as appropriate
2. **Test statistic:**
$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$, $Q = 1 - P$.

3. **Level of significance:** Usually $\alpha = 0.05$.
4. **Critical value:** Obtain from standard normal tables.
5. **Decision:** Reject H_0 if $|Z| > Z_{\text{tabulated}}$.

Example:

Machine 1: $x_1 = 21$ defectives out of $n_1 = 500$
Machine 2: $x_2 = 3$ defectives out of $n_2 = 100$

$$p_1 = \frac{21}{500} = 0.042, \quad p_2 = \frac{3}{100} = 0.03$$

$$P = \frac{500 \times 0.042 + 100 \times 0.03}{600} = 0.04, \quad Q = 0.96$$

$$Z = \frac{0.042 - 0.03}{\sqrt{0.04 \times 0.96 \left(\frac{1}{500} + \frac{1}{100}\right)}} = 0.571$$

At $\alpha = 0.01$, $Z_{\text{tab}} = 2.58$. Since $0.571 < 2.58$, accept H_0. There is no significant difference in performance.

t Test

The t test is used when the sample size is small ($n \leq 30$), the sample is drawn from a normal population, and the population standard deviation is unknown. The t distribution is similar to the normal distribution but has heavier tails. As the sample size increases, the t distribution approaches the normal.

t tests are used for:

- Significance of a single mean
- Significance of difference between means
- Significance of correlation coefficient
- Significance of regression coefficient

Other Tests

- Chi-square test
- ANOVA (Analysis of Variance)
- Run test
- Sign test
- Mann-Whitney U test

- Kruskal-Wallis test

Lecture 23: Preparation of Research Report

Research Report

A research report is a concise, clear communication of the important findings of research work. It is a statement or description of things that have already occurred, compiling information as a result of research and data analysis. Reports focus on transmitting information with a clear purpose to a specific audience. Good reports are accurate, objective, complete, well-written, clearly structured, and expressed in a way that holds the reader's attention and meets their expectations.

Purpose and Importance of Research Reports

- To organize data, analysis, and conclusions in a form that can be used for academic or application purposes.
- The research report is the main tangible product of your work and is the primary basis for evaluation by examiners or committees.
- The quality of your research will largely be judged by the report, not by your fieldwork.
- A good report demonstrates your performance, skills, and thoughts, which are vital for assessment and grading.
- Reports are used by organizations, professors, researchers, and students for various purposes.
- Writing the report helps you:
  - Monitor progress and spot problems in time.
  - Reflect on progress, consolidate arguments, and identify gaps in knowledge, data, or methodology.
  - Develop an appreciation of standards and learn to monitor your own progress.
  - Practice academic report writing and discourse.
  - Form a basis for future project work or journal articles.

Research Report Process

The process of preparing a research report involves several steps, where raw data collected from different sources is gradually compressed and organized. The main stages are:

1. **Collection of Raw Data:** Gather data from various sources.
2. **Processing and Analysis:** Process and analyze the collected data to extract meaningful information.
3. **Interpretation:** Interpret the results in the context of the research objectives.
4. **Presentation:** Organize and present the findings in a structured report.
5. **Communication:** Communicate the findings effectively to the intended audience.

Qualities of a Good Research Report

A good research report should be:

- **Accurate:** Free from errors and faithfully represents the research findings.
- **Objective:** Unbiased and based on facts, not personal opinions.
- **Complete:** Covers all relevant aspects of the research.
- **Clear and Concise:** Written in simple, straightforward language.
- **Well-structured:** Logically organized with a clear flow from introduction to conclusion.
- **Relevant:** Focused on the research objectives and the needs of the audience.

General Structure of a Research Report

Although formats may vary, a typical research report includes the following sections:

1. **Title Page:** Title of the research, author's name, affiliation, and date.
2. **Abstract:** A brief summary of the research objectives, methods, findings, and conclusions.
3. **Table of Contents:** List of chapters and sections with page numbers.
4. **Introduction:** Background, statement of the problem, objectives, and significance.
5. **Literature Review:** Summary of previous research and theoretical background.
6. **Methodology:** Description of research design, sampling, data collection, and analysis methods.
7. **Results:** Presentation of findings using tables, charts, and descriptive text.
8. **Discussion:** Interpretation of results, implications, and comparison with previous studies.

9. **Conclusion and Recommendations:** Summary of findings and suggestions for future work or policy.
10. **References:** List of sources cited in the report.
11. **Appendices:** Supplementary material such as questionnaires, raw data, or additional tables.

Tips for Writing a Good Research Report

- Use clear, simple, and precise language.
- Be objective and avoid personal bias.
- Present data and analysis logically.
- Use tables, figures, and charts for clarity.
- Revise and proofread to eliminate errors.
- Ensure proper citation and referencing.
- Tailor the report to the needs and expectations of the audience.

Lecture 24: Conventions and Layout of Academic Research Writing

Conventions of Academic Writing

- Write direct, positive sentences using familiar words and short, simple constructions.
- Avoid unessential, overly technical, or unusual words and phrases.
- Well-constructed, natural, and direct sentences are a mark of skill in writing.

Presentation

- Label charts, sections, sub-sections, and tables adequately.
- Keep the system of headings and subheadings simple.
- Ensure the sequence of sections and subsections is logical and clear.
- Use past tense for introduction, data analysis, and findings; present tense for conclusions; and future tense for recommendations.

Use of the First Person

- Academic reports should be written in the third person.
- Avoid pronouns such as I, my, mine, our, ours, we, us, and me.
- If necessary, refer to yourself as ”the writer” or ”the investigator.”

Use Gender-neutral Language

- Select terminology that treats all genders equally.
- Do not make assumptions about one gender over another.
- Use ”he or she” or similar constructions when necessary.

Avoid Emotional Terms

- Prefer factual statements over emotional or subjective descriptions.
- For example, state the percentage increase in sales rather than using terms like ”tremendous” or ”fantastic.”

Label Opinions

- Facts are preferred, but specialist opinions may be included when facts are unavailable.
- When presenting opinions, reveal the background and identity of the person if relevant.
- Opinions can substantiate explanations and conclusions, especially when data is inconclusive.

Use of Notes and Footnotes

- Supplementary material inappropriate for the main text may be included in footnotes (bottom of page) or notes (end of chapter).

Non-English Terms and Expressions

- Commonly used non-English terms in English do not need italics.
- Less familiar expressions (e.g., *chakka jam*) should be italicized.
- Consider whether the term will be understood by most readers.

English and American Spellings

- Use a consistent spelling format (either -ise or -ize endings) throughout the report.
- Both British and American spellings are acceptable, but consistency is essential.

Abbreviations

- Use abbreviations sparingly, as they can disrupt the flow and readability.
- On first use, provide the full term followed by the abbreviation in parentheses.
- Use the abbreviation alone thereafter.

Confidentiality and Anonymity

- Maintain confidentiality and anonymity as much as possible.
- Use fictional names for case studies if needed to protect identities.

Consistency

- Maintain consistency in spelling, abbreviations, style, and formatting throughout the report.
- Choose a format and adhere to it strictly.

Typing the Research Report

Paper

- Use white Xerox paper, size 8.6 by 11 inches.
- Type on only one side of the paper.

Chapter Page

- Center the chapter number about two inches from the top of the page.
- Place the chapter title in capital letters two spaces below the chapter number.
- Begin the first line of text four spaces below the title.

Margins

- Use 1-inch margins on all sides (top, bottom, left, right) as per APA guidelines.

Spacing

- Double-space the main text.
- Single-space indented quotations and footnotes.
- Use the same style and size of font throughout.

Page Number

- Place the page number at the top right corner, one inch from the top and right edges.
- The first line of text should be two spaces below the page number.

Pagination

- Number pages consecutively in Arabic numerals from the first page of text to the end, including appendices.
- Introductory sections (preface, table of contents, etc.) use small Roman numerals (i, ii, iii, etc.), placed one inch from the bottom.
- Page numbers should stand alone, without periods, hyphens, or dashes.

Proofreading

- Review the manuscript critically for inaccuracies, omissions, inconsistencies, and errors in quotations, footnotes, tables, figures, paragraphing, sentence structure, headings, spelling, style, and bibliography.
- Mark corrections clearly for the typist.

Layout of the Research Report

A comprehensive layout should include:

(a) Preliminary Pages

- Title and date
- Acknowledgements (preface or foreword)
- Table of contents
- List of tables and illustrations

(b) Main Text

- **Introduction:** State objectives, background, hypotheses, definitions, methodology, sample details, statistical analysis, scope, and limitations.
- **Statement of findings and recommendations:** Present findings and recommendations in non-technical language; summarize if extensive.
- **Results:** Detailed presentation of findings with supporting data, tables, charts, and validation. Present all relevant results in logical sequence.
- **Implications of the results:** Discuss inferences, limitations, unanswered questions, and suggestions for further research.
- **Summary:** Briefly restate the research problem, methodology, major findings, and conclusions.

(c) End Matter

- Appendices (technical data, questionnaires, sample info, mathematical derivations, etc.)
- Bibliography of sources consulted
- Index (alphabetical listing of names, places, topics with page references)

Lecture 25: Research Proposal

Research Proposal

A research proposal presents the author's plan for the research they intend to conduct. Its purpose is to demonstrate the relevance, necessity, and feasibility of the proposed research, often for approval by a supervisor, department, or funding body. In some cases, submitting a research proposal is a requirement for graduate school applications or for securing funding.

A well-written research proposal shows:

- How and why the research is relevant to the field.
- That the research fills a gap, supports, or adds new knowledge to existing literature.
- The author's capability and academic background to carry out the research.
- The academic merit and feasibility of the proposed ideas.
- The methodology, tools, and procedures for data collection, analysis, and interpretation.
- Consideration of budget and institutional or programmatic constraints.

The proposal must also include a literature review, which details the sources to be used, how they will be used, and their relevance to the research.

Length of Proposal

Research proposals for bachelor's and master's theses are typically a few pages long, while those for Ph.D. dissertations or funding requests are longer and more detailed. The focus should be on including all necessary elements, not on meeting a specific word or page count.

Research Proposal Structure

A standard research proposal generally includes the following sections:

Introduction

- Introduces the research topic.
- States the problem statement and research questions.

- Provides context for the research.

In some cases, an abstract and/or table of contents may precede the introduction.

Background Significance

- Explains why the research is necessary.
- Relates the research to existing studies.
- Clearly defines the problems addressed.
- Outlines the scope and any related questions not covered.

Literature Review

- Introduces and discusses all sources relevant to the research.
- Explains how each source will be used and its significance.
- Goes beyond listing sources by analyzing and contextualizing them.

Research Design, Methods, and Schedule

- Specifies the type of research (qualitative, quantitative, experimental, correlational, descriptive).
- Describes the data, population, and subject selection.
- Details data collection tools and methods (sampling frame, sampling method, statistics, experiments, surveys, observations).
- Justifies the chosen methods.
- Includes a research timeline, budget, and anticipated obstacles with plans to address them.

Suppositions and Implications

- Discusses how the research may challenge existing theories or assumptions.
- Explains the potential for future research based on the findings.
- Describes the practical value and possible impact of findings (for practitioners, educators, policymakers, etc.).
- States how findings could be implemented and their transformative potential.

Conclusion

- Summarizes the proposal and reinforces the research's purpose.

Bibliography

- Lists all sources referenced in the proposal, formatted according to the relevant academic style (APA, IEEE, Chicago, etc.).
- Sometimes a references list is sufficient for shorter proposals.

How to Write a Research Proposal

- Write in a formal, objective, and concise academic tone.
- Follow the standard structure so the reader can easily follow and evaluate your proposal.
- Ensure every section answers potential questions from the reader.

Common Mistakes to Avoid

- **Being too wordy:** Keep writing brief and to the point.
- **Failing to cite relevant sources:** Reference landmark studies and connect your work to the field.
- **Focusing too much on minor issues:** Address only the key questions and issues central to your research.
- **Failing to make a strong argument:** Clearly persuade the reader of your research's importance and feasibility.

Polishing Your Proposal

- Ensure the proposal is free from spelling and grammatical errors.
- Maintain an appropriate, consistent academic tone.
- Revise awkward phrasing to strengthen clarity and credibility.