Research Methodology

Unit II: Data Collection and Analysis

Dr. Ratnesh Prasad Srivastava
Department of CSIT, GGV, Bilaspur (C.G)

Academic Year 2026-2027

Preface

These lecture notes build upon Unit I, where we explored research formulation. Now we move to the heart of empirical research: data collection and analysis. We'll continue using the **Electrical Chakki** (electric grinder) as our running example to demonstrate how theoretical concepts apply to real-world research problems.

Contents

1 Foundations of Data Collection

1.1 Opening Discussion

Key Question: If research questions are the "destination," what is data?

Data

Data is the vehicle, fuel, and map that powers our research journey. It transforms abstract questions into empirical answers.

Electrical Chakki Example: What Data Might We Collect?

If researching Electrical Chakki performance and durability, potential data includes:

- **Technical Data:** Motor RPM, power consumption (watts), temperature (°C), vibration frequency
- **Usage Data:** Hours of operation per day, types of grains ground, frequency of use
- **Failure Data:** Time to failure, type of failure (motor, stone, electrical), repair costs
- **Contextual Data:** Voltage fluctuations in the area, humidity levels, user demographics
- **User Experience Data:** Satisfaction ratings, preferences, reported problems

Key Insight: Different research questions require different types of data. The quality of your conclusions depends directly on the quality of your data.

2 Aspects of Method Validation

Before collecting any data, we must ensure our methods are valid. This is like checking your measuring instruments before starting an experiment.

2.1 The Four Pillars of Method Validation

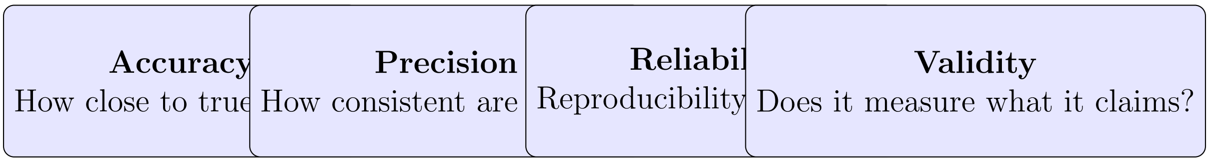

Figure 1: The Four Pillars of Method Validation

2.2 Interactive Demonstration: Understanding Accuracy vs. Precision

Thermometer Demonstration

Consider three thermometers measuring actual temperature of 23°C:

Thermometer	Readings	Characteristic	Description
A	Consistently 25°C	Reliable but not accurate	Systematic error (bias)
B	Varies $20 - 26$°C	Neither reliable nor accurate	Random error
C	Exactly 23°C	Both reliable and accurate	Ideal

Electrical Chakki Application:

- **Accuracy:** Does our power consumption meter show the true watts used?
- **Precision:** Do repeated measurements under same conditions give similar readings?
- **Reliability:** If different researchers measure the same chakki, do they get similar results?
- **Validity:** Does measuring RPM actually tell us about grinding efficiency?

2.3 Field-Specific Examples of Validation

Table 1: Validation Approaches Across Disciplines

Field	Validation Method	Chakki Example
Social Sciences	Survey validation through pilot testing	Pilot test user satisfaction questionnaire with 20 households
Engineering	Equipment calibration against standards	Calibrate power meter against certified reference device
Medicine	Diagnostic test validation against gold standard	Compare new failure prediction method against actual failures
Computer Science	Algorithm validation on benchmark datasets	Test predictive model on historical chakki failure data
Business Research	Instrument validation through factor analysis	Validate survey items measuring "user satisfaction" construct
Education	Test validation through reliability analysis	Ensure test questions consistently measure understanding

3 Observation and Collection of Data

Data collection begins with observation. But observation itself is a skill that needs training.

3.1 Types of Observation

1. **Participant Observation:** Researcher becomes part of the group being studied

   Chakki Example: Participant Observation

   A researcher lives in a village for 3 months, using the chakki daily with families, participating in their cooking routines to deeply understand usage patterns, maintenance practices, and cultural attitudes toward the appliance.
   Strength: Deep contextual understanding, trust with participants
   Challenge: Potential bias, time-consuming, researcher may ”go native”

2. **Non-participant Observation:** Researcher observes without involvement

   Chakki Example: Non-participant Observation

   A researcher sits in a kitchen corner, quietly noting how family members use the chakki—who operates it, when, for how long, what safety practices they follow—without interacting or interfering.
   Strength: More objective, less influence on behavior
   Challenge: May miss contextual understanding, participants may act differently when watched

3. **Structured Observation:** Using predefined categories/checklists

   Chakki Example: Structured Observation

   Observer uses a checklist every 5 minutes:

   - Is chakki running? Yes/No
   - If running, what grain? Wheat/Rice/Spices/Other
   - Duration since last pause: ___ minutes
   - Any unusual noise? Yes/No
   - Temperature: Cool/Warm/Hot (touch test)

4. **Unstructured Observation:** Open-ended, noting everything

   Chakki Example: Unstructured Observation

   ”10:15 AM: Mother starts chakki for wheat grinding. Child comes to watch, mother warns to stay away. Chakki makes grinding sound, occasional louder noise when larger grains enter. After 15 minutes, motor seems to slow down slightly...”

3.2 Ethical Considerations in Observation

Ethical Imperative: Observation without consent is unethical unless in public spaces where privacy isn't expected. Even then, ethical considerations apply.

Chakki Research: Ethical Guidelines

- Always obtain informed consent before observing in private homes
- Explain the purpose of observation clearly
- Allow participants to withdraw at any time
- Don't observe sensitive activities (bathroom, personal spaces)
- Anonymize data in reporting
- Consider if observation might change behavior (Hawthorne effect)
- In public spaces (chakki repair shops), observation may be acceptable but still requires discretion

4 Methods of Data Collection

The choice of data collection method depends on your research question, resources, and context.

4.1 Traditional Data Collection Methods

Chakki Example: Choosing Methods

Research Question: "Impact of voltage fluctuations on Electrical Chakki lifespan in rural Karnataka"
Recommended Mixed-Methods Approach:

1. **Surveys (Quantitative):** Distribute to 500 households
   - When did you buy your chakki?
   - How many times has it broken?
   - What were the repair costs?
   - Do you experience power cuts? How often?
2. **Technical Measurements (Quantitative):** Install voltage loggers in 50 homes
   - Record voltage every minute for 6 months
   - Document when chakkis fail
   - Correlate failures with voltage events

3. **Interviews (Qualitative):** Conduct 20 in-depth interviews
   - Understand user experiences with failures
   - Learn about maintenance practices
   - Explore coping strategies during power issues
4. **Focus Groups (Qualitative):** 5 groups of 8-10 women
   - Discuss what features they value
   - Share stories about chakki problems
   - Generate ideas for improvements
5. **Document Analysis:** Review repair shop records
   - Analyze 2 years of chakki repair invoices
   - Identify most common failure types
   - Track seasonal patterns

4.2 Modern Data Collection Methods

- **Web Scraping:** Automated collection from websites
- **Sensor Data:** IoT devices, wearables
- **Social Media Analytics:** Mining platforms like Twitter
- **Mobile Data Collection:** Using smartphones for surveys

Chakki Example: Modern Methods

- **Web Scraping:** Collect online reviews of chakkis from e-commerce sites
- **Sensor Data:** Install IoT sensors in chakkis to monitor usage remotely
- **Social Media Analytics:** Analyze Twitter discussions about chakki problems
- **Mobile Data Collection:** Use smartphone app for daily usage diaries

5 Sampling Methods

We rarely study entire populations. Sampling is the art of selecting a representative subset.

5.1 The Sampling Decision Tree

5.2 Probability Sampling Methods

1. **Simple Random Sampling:** Every member has equal chance

Table 2: Comparison of Data Collection Methods

Method	**Best For**	**Advantages**	**Limitations**
Surveys	Large samples, quantitative	Cost-effective, anonymous	Low response rates
Interviews	In-depth understanding	Rich data, flexible	Time-consuming, bias
Focus Groups	Group dynamics	Interaction sparks ideas	Groupthink issues
Experiments	Establishing causality	Control over variables	Artificial setting
Document Analysis	Historical research	Non-reactive data	Authenticity issues
Observation	Natural behavior	Contextual data	Observer bias

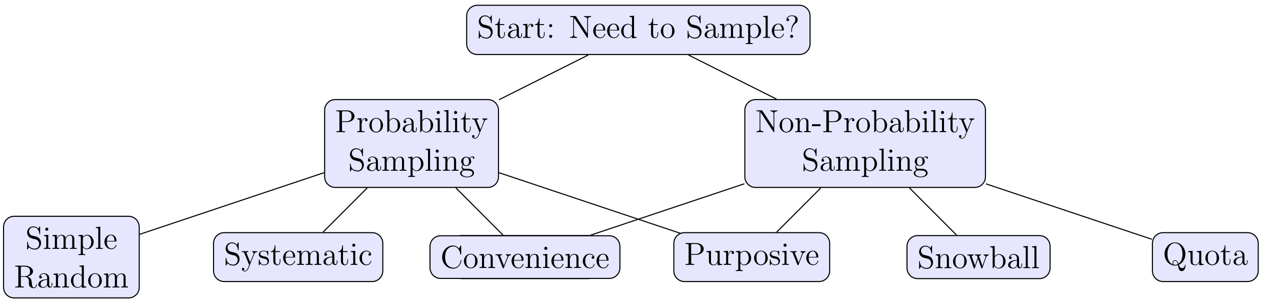

Figure 2: The Sampling Decision Tree

Chakki Example: Simple Random

From a list of 10,000 households in a district, use random number generator to select 400 for a survey.
Advantage: Unbiased, statistically sound
Disadvantage: May miss subgroups; need complete list

2. **Systematic Sampling:** Selecting every kth element

Chakki Example: Systematic

From 1000 repair records sorted by date, select every 10th record (k=10) for detailed analysis.
Advantage: Easy to implement
Risk: Periodicity bias (if records have pattern every 10th entry)

3. **Stratified Sampling:** Divide into strata, sample from each

   Divide population by:

   - Urban households (60%) → sample 240
   - Rural households (40%) → sample 160

   Ensure proportional representation from both groups.

   Advantage: Ensures subgroup representation

   Randomly select 20 villages from 200 total. Study ALL households in those 20 villages.

   Advantage: Cost-effective geographically

   Disadvantage: Higher sampling error

S

5.3 Non-Probability Sampling Methods

When to Use Non-Probability Sampling:

- Exploratory research
- When probability sampling is impossible
- Qualitative research
- Limited resources

1. **Convenience Sampling:** Selecting readily available

Purposive Sampling: Based on specific characteristics

Chakki Example: Purposive

Interview only:

- Chakki repair technicians (10 experts)
- Manufacturers (5 company representatives)
- Long-term users (20 women who've used chakkis for 10+ years)

Snowball Sampling: Participants refer others

Studying traditional stone chakki makers (a rare profession):

- Find one artisan, interview them
- Ask them to refer other artisans they know
- Continue until saturation

Useful for: Hidden or hard-to-reach populations

Target: 200 respondents with:

- 100 urban, 100 rural (matches population)
- Within each, 50% male, 50% female

But selection within quotas is non-random.

F

5.4 Sample Size Determination

bP 95% confidence level with margin of error e:

$$n = \frac{z^2 \times p \times (1-p)}{e^2}$$

Where:

- $z = 1.96$ for 95% confidence
- $p =$ estimated proportion (use 0.5 for maximum variability)
- $e =$ margin of error (e.g., $0.05 = \pm 5\%$)

For large populations, $n = 384$ is sufficient for 95% confidence with ±5% margin.

Chakki Example: Sample Size

For a survey estimating proportion of households experiencing chakki failures:

$$n = \frac{1.96^2 \times 0.5 \times 0.5}{0.05^2} = \frac{3.84 \times 0.25}{0.0025} = \frac{0.96}{0.0025} = 384$$

Interpretation: Survey 384 households to estimate failure rates with ±5% accuracy at 95% confidence.

6 Data Processing and Analysis Strategies

Raw data is useless. Processing transforms it into information.

6.1 Data Processing Pipeline

6.2 Data Cleaning: The Critical First Step

Chakki Example: Data Cleaning

Raw survey data might have:

- Missing: Some respondents didn't answer "repair cost"
- Outliers: One person reported "500 hours daily usage" (impossible)
- Inconsistencies: "5 years" vs "60 months" vs "2019 purchase"
- Duplicates: Same household surveyed twice

Cleaning Steps:

1. Remove impossible values (500 hours/day → delete)
2. Impute missing repair costs with median value
3. Standardize all dates to "YYYY-MM-DD"
4. Remove duplicate entries based on household ID

6.3 Data Transformation

- **Normalization:** Scaling to [0,1] or [-1,1]
- **Standardization:** z-scores (mean = 0, SD = 1)
- **Encoding:** Categorical to numerical (one-hot encoding)
- **Feature Engineering:** Creating new variables from existing ones

Chakki Example: Data Transformation

- **Normalization:** Scale power consumption (300-1500W) to [0,1]
- **Standardization:** Convert usage hours to z-scores
- **Encoding:** Grain type (Wheat=001, Rice=010, Spices=100)
- **Feature Engineering:** Create "stress index" = (voltage drop × duration × frequency)

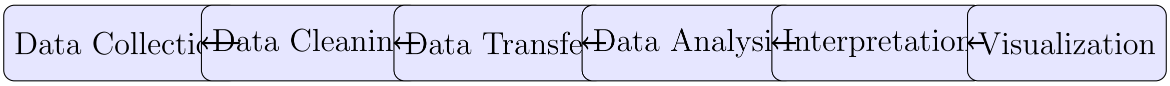

Figure 3: Data Processing Pipeline

Table 3: Common Data Quality Issues and Solutions

Issue	**Detection Method**	**Solution**
Missing Values	Incomplete records	Imputation (mean, median, mode) or deletion
Outliers	Statistical tests (z-score ¿ 3), box plots	Investigate, transform, or remove if error
Inconsistencies	Cross-field validation	Standardization rules
Duplicates	Record matching algorithms	Remove duplicates
Format issues	Visual inspection	Standardize formats

Table 4: Statistical Software Comparison

Package	**Best For**	**Strengths**	**Learning Curve**
SPSS	Social sciences	User-friendly, good docs	Gentle
SigmaStat	Engineering	Good graphs	Moderate
R	Advanced stats	Free, powerful, community	Steep
Python	Data science	Versatile, integrates	Moderate
Excel	Basic analysis	Ubiquitous, familiar	Easy

7 Data Analysis with Statistical Packages

7.1 Statistical Software Comparison

7.2 SPSS Interface Overview

8 Statistical Tests: t-test and ANOVA

8.1 Student's t-test

When to use: Comparing means of TWO groups

[t-test Formula] Independent t-test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

8.2 Assumptions of t-test

1. Continuous dependent variable
2. Independent observations
3. Normal distribution (approximately)
4. Equal variances (for independent t-test)

Chakki Example: Independent t-test

Research Question: Do chakkis with stone grinding plates last longer than those with steel plates?

- Group 1 (Stone): n=30, mean lifespan = 8.2 years, SD = 1.5
- Group 2 (Steel): n=30, mean lifespan = 6.8 years, SD = 1.8
- t-statistic = 3.24, p = 0.002

Interpretation: Since p ¡ 0.05, we reject null hypothesis. Stone plates significantly outlast steel plates.

8.3 ANOVA (Analysis of Variance)

When to use: Comparing means of THREE OR MORE groups

[F-ratio]

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}}$$

Interpretation:

- Large F (p ¡ 0.05): Means differ significantly
- Small F (p ¿ 0.05): No significant difference

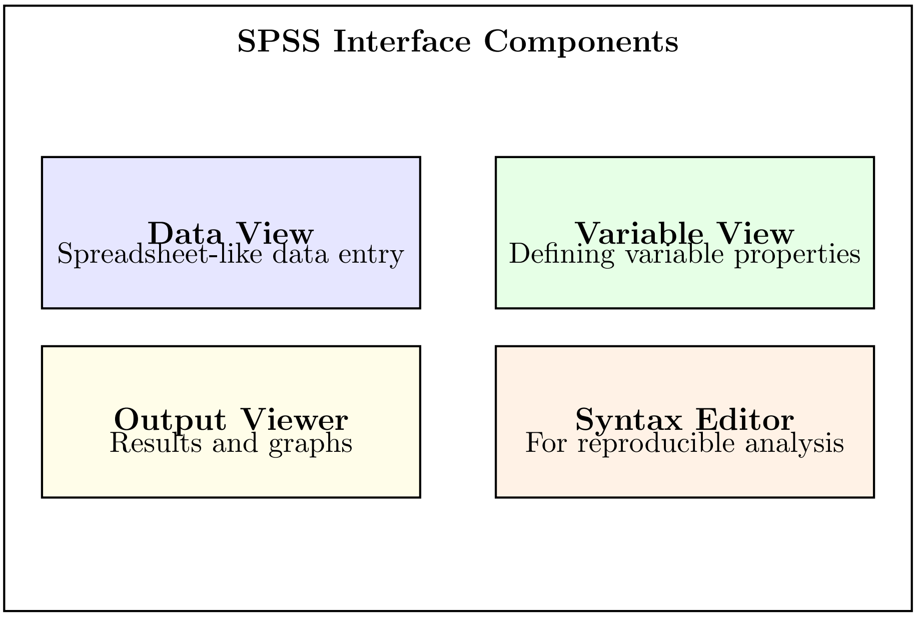
SPSS Interface Components
Data View
Spreadsheet-like data entry
Variable View
Defining variable properties
Output Viewer
Results and graphs
Syntax Editor
For reproducible analysis

Figure 4: SPSS Interface Components

Table 5: Types of t-tests

Type	**Use Case**	**Chakki Example**
Independent t-test	Two separate groups	Compare motor RPM: Brand A vs Brand B
Paired t-test	Same group, two times	Grinding time: Before vs after maintenance
One-sample t-test	Compare to known value	Average lifespan vs manufacturer claim of 5 years

Table 6: Types of ANOVA

Type	**Use Case**	**Chakki Example**
One-way ANOVA	One IV with 3+ levels	Compare 4 different motor brands
Two-way ANOVA	Two independent variables	Motor brand × Grain type on grinding time
Repeated Measures	Same subjects, multiple times	Performance at 1, 6, 12 months
MANOVA	Multiple DVs	Effect on both grinding time AND power consumption

Chakki Example: One-way ANOVA

Research Question: Do four different chakki brands have different lifespans?

- Brand A: n=25, mean = 5.2 years
- Brand B: n=25, mean = 6.8 years
- Brand C: n=25, mean = 4.9 years
- Brand D: n=25, mean = 7.1 years

ANOVA Results:

- F(3,96) = 8.42, p ¡ 0.001

Interpretation: Significant differences exist among brands. Post-hoc tests needed to identify which pairs differ.

8.4 Common Pitfalls in ANOVA

- Violating assumptions (normality, homogeneity of variance)
- Not checking post-hoc tests after significant F
- Misinterpreting interactions in factorial designs
- Ignoring effect sizes (focusing only on p-values)

9 Hypothesis Testing

This connects back to our first lecture on research formulation.

9.1 The Hypothesis Testing Process

9.2 Key Concepts in Hypothesis Testing

Key Terms

- **Null Hypothesis (H):** No effect, no difference, no relationship
- **Alternative Hypothesis (H):** Effect exists, difference present
- **p-value:** Probability of observing data if H is true
- **(alpha):** Rejection threshold (typically 0.05)
- **Type I Error ():** False positive — rejecting true H
- **Type II Error ():** False negative — failing to reject false H
- **Power (1-):** Probability of detecting true effect

Chakki Example: Hypothesis Testing

Research Question: Does adding a voltage stabilizer extend chakki lifespan?

- **H:** No difference in lifespan (with stabilizer - without = 0)
- **H:** Lifespan differs (with stabilizer - without 0)
- **:** 0.05
- **Test:** Independent t-test
- **Data:** With stabilizer: mean = 7.2 years (n=30); Without: mean = 5.8 years (n=30)
- **Result:** t = 3.56, p = 0.001

Decision: Since p = 0.001 ¡ 0.05, reject H.
Interpretation: Voltage stabilizers significantly increase chakki lifespan.

Misconception Alert: ”p ¡ 0.05” does NOT mean: ”95% chance the effect is real.”
It means: ”If there were truly no effect, we'd see results this extreme only 5% of the time.”

9.3 Effect Sizes: Beyond p-values

Statistical significance Practical significance

Chakki Example: Effect Size

From our t-test example:

- Mean difference = 1.4 years
- Pooled SD = 1.65
- Cohen's d = 1.4/1.65 = 0.85

Interpretation: d = 0.85 is a LARGE effect size. The voltage stabilizer doesn't just make a statistically significant difference—it makes a practically important difference of nearly a year of additional life.

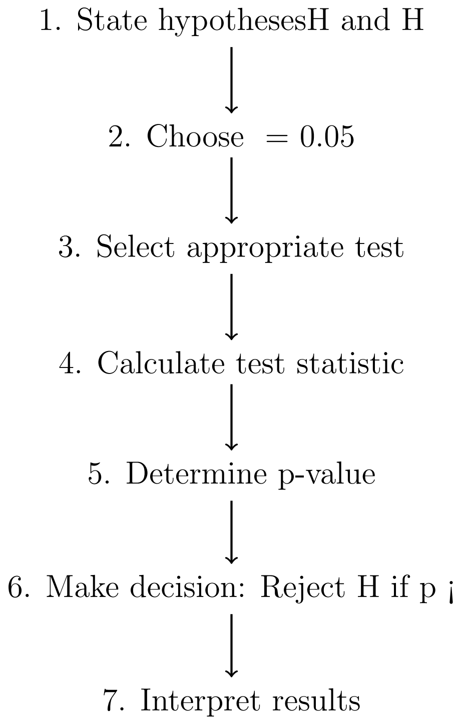
1. State hypothesesH and H
2. Choose = 0.05
3. Select appropriate test
4. Calculate test statistic
5. Determine p-value
6. Make decision: Reject H if p ¡
7. Interpret results

Figure 5: The Hypothesis Testing Process

Table 7: Common Effect Size Measures

Metric	**Interpretation**	**Thresholds**
Cohen's d	Standardized mean difference	Small: 0.2, Medium: 0.5, Large: 0.8
r	Correlation coefficient	Small: 0.1, Medium: 0.3, Large: 0.5
^{2} (eta-squared)	Variance explained	Small: 0.01, Medium: 0.06, Large: 0.14
Odds Ratio	Association in logistic	1 = no effect, ¿1 = increased odds

10 Practice Activities

10.1 Comprehensive Case Study Exercise

Electrical Chakki Research: Complete Design

You are designing a study to investigate the relationship between usage patterns and chakki failure rates.

Answer the following:

1. **Sampling Method:** What sampling method would you use and why?
2. **Data Collection:** What methods would you employ?
3. **Statistical Test:** If comparing failure rates between urban and rural areas, what test?
4. **Hypothesis:** State H and H for the urban-rural comparison
5. **Additional Analyses:** What other factors might you analyze?

Guidance:

- Population: All households in a district (50,000 households)
- Resources: Limited budget, 6 months, team of 4 researchers
- Research question: ”Do rural households experience higher chakki failure rates than urban households?”

10.2 Sample Solution

Suggested Approach

1. **Sampling Method:** Stratified random sampling
   - Why: Ensures representation from both urban and rural areas
   - Strata: Urban (60%) → 240 households; Rural (40%) → 160 households
   - Total: 400 households (meets sample size requirement)
2. **Data Collection Methods:**
   - Survey: Collect demographic data, chakki age, failure history
   - Technical inspection: Examine chakkis in 100 subsample
   - Voltage monitoring: Install loggers in 50 homes across both areas
   - Repair shop records: Analyze local repair data
3. **Statistical Test:** Independent t-test

- Group 1: Urban households (n=240) — mean failure rate
- Group 2: Rural households (n=160) — mean failure rate

4. **Hypotheses:**

- H: $_urban =_r ural(Nodifferenceinfailurerates)$
- H: $_urban_rural(Failureratesdiffer)$

5. **Additional Analyses:**

- Correlation: Voltage fluctuation severity vs failure rate
- ANOVA: Compare across multiple grain types
- Regression: Predict failure from usage hours, voltage, maintenance
- Effect size: Cohen's d for urban-rural comparison

11 Summary and Key Takeaways

- **Method Validation:** Ensure accuracy, precision, reliability, and validity before collecting data
- **Observation Types:** Choose participant/non-participant, structured/unstructured based on research needs
- **Data Collection Methods:** Select methods that best answer your research questions
- **Sampling:** Probability sampling for generalizability; non-probability for exploratory/qualitative research
- **Sample Size:** n = 384 provides 95% confidence with ±5% margin for large populations
- **Data Processing:** Clean, transform, and prepare data before analysis
- **t-test:** For comparing means of TWO groups
- **ANOVA:** For comparing means of THREE OR MORE groups
- **Hypothesis Testing:** Follow the 7-step process; understand Type I/II errors
- **Effect Sizes:** Always report effect sizes alongside p-values

Thought for the Day

”Without data, you're just another person with an opinion.”
— W. Edwards Deming

Chakki Connection:
Behind every broken chakki lies data waiting to be collected. Good research transforms everyday observations into actionable knowledge.

Assignment for Next Class

Homework Assignment

Part A: Design Your Data Collection Plan

1. Choose a research problem related to Electrical Chakki (or your field)
2. Design a complete data collection plan including:
   - Sampling method with justification
   - Sample size calculation
   - Data collection instruments (survey questions, observation checklist, etc.)
   - Validation approach for your methods
   - Planned statistical analyses

Part B: Software Preparation

1. Install SPSS (trial version available) or R/RStudio
2. Explore the interface and basic functions
3. Bring a dataset (even 20-30 rows) to next class for hands-on practice

Reading: "Advanced Statistical Methods" chapter for next week

Preview of Next Lecture

Advanced Analysis Techniques and Research Reporting

We'll explore:

- Regression analysis
- Non-parametric tests
- Research report writing
- Presenting findings effectively