Choosing the Right Sample Size in Social Science Studies: A Methodological Review

Zairemmawia Renthlei*
C. Lallawmkima**

Abstract

*Sample size calculation is an integral part of social science research especially in quantitative studies. The problems of sample size determination has befuddled many researchers and has led to confusions and heated discussions as a proper theoretical understanding of the matter has often eluded researchers. The present paper is an attempt to catalogue various methods and techniques that are utilised by academicians for calculating sample sizes in various situations and conditions. The paper has included brief discussions on Cochran's Formula, Yamane's Formula, Krejcie amd Morgan's Table Calculation Procedure, Samuel B Green's Formula as well as a brief introduction to G*Power software for sample calculations. Assumptions and necessary conditions as well as applications have been highlighted wherever possible.*

Key words: *Sampling size, calculation, Cochran, Yamane, Krejcie and Morgan, Green, G*Power*

Introduction

Sample size determination is a critical aspect of social science research, influencing the validity, reliability, and generalizability of study findings. An appropriately chosen sample ensures that research conclusions accurately reflect the target population while minimizing errors and biases. Conversely, an inadequate sample size can lead to misleading interpretations, reduced statistical power, and difficulties in hypothesis testing (Cohen, 1992).

In social science research, determining the right sample size depends on multiple factors, including research design, population characteristics, statistical techniques,

*Dr. Zairemmawia Renthlei, Assistant Professor, Institute of Advanced Studies in Education;
Ph – 9774089972, email – jimzrenthlei76@gmail.com
**C. Lallawmkima, Research Scholar, Institute of Advanced Studies in Education;
Ph – 9774089972, email - lomkim08@gmail.com

and study objectives (Babbie, 2020). Researchers often rely on various methodological approaches to calculate sample size, ranging from heuristic rules to sophisticated statistical formulas. Commonly used methods include Cochran's formula (Cochran, 1977), Yamane's formula (Yamane, 1967), and Power analysis (Faul et al., 2009), each suited to different research scenarios.

Despite the availability of established sample size determination methods, many researchers face challenges in selecting the most appropriate approach for their specific study. Constraints related to time, resources, ethical considerations, and population accessibility further complicate the process (Fowler, 2013).

This methodological review aims to provide a comprehensive overview of sample size calculation techniques in social science research. It explores key theoretical concepts, statistical methods, and practical considerations, offering guidance on choosing the most suitable approach based on research objectives. By examining the strengths and limitations of various sample size determination methods, this study seeks to support researchers in making informed methodological decisions and enhancing the credibility of their research findings.

The subsequent sections will discuss the fundamental principles of sampling, factors influencing sample size selection, and an in-depth analysis of different estimation techniques applicable to social science research.

Theoretical Framework & Key Concepts

Sampling is a foundational element of research methodology in social sciences, allowing researchers to draw conclusions about a population without studying every individual (Babbie, 2020). Since studying entire populations is often impractical, sampling methods enable researchers to generalize findings while maintaining feasibility and accuracy (Cochran, 1977). The selection of an appropriate sample size is crucial for ensuring the reliability and validity of research outcomes.

Several key factors influence the determination of an adequate sample size in social science research:

1. Confidence Level and Margin of Error: Higher confidence levels (e.g., 95% or 99%) require larger sample sizes to ensure precision, whereas a larger margin of error allows for smaller sample sizes (Fowler, 2013).
2. Effect Size: The magnitude of the expected effect influences the required sample size; smaller effects demand larger samples to achieve statistical significance (Cohen, 1992).
3. Population Variability: Greater heterogeneity within a population requires a larger sample to capture its diversity adequately (Yamane, 1967).

4. Statistical Power: A study should have at least 80% power to detect true effects while minimizing the risk of Type II errors (Faul et al., 2009).
5. Sampling Method: Probability sampling techniques (e.g., stratified, cluster sampling) often require different sample size considerations than non-probability methods (Krejcie & Morgan, 1970).

Sampling Techniques and Their Impact on Sample Size

Sampling methods significantly affect the required sample size and generalizability of research findings:

- Probability Sampling: Ensures random selection and generalizability; methods include simple random sampling, stratified sampling, and cluster sampling (Cochran, 1977).
- Non-Probability Sampling: Includes convenience, quota, and purposive sampling, often requiring careful justification for sample size selection due to potential bias (Babbie, 2020).

Understanding these theoretical concepts provides a framework for selecting an appropriate sample size methodology, balancing statistical rigor with practical research constraints.

Methods for Sample Size Calculation:

A. Rule of Thumb Methods

1. 10% Rule

The 10% rule proposes that selecting a sample equivalent to 10% of the total population can serve as a practical approach to estimate population parameters, especially when conducting complex statistical calculations is not practical.

Example: Suppose a researcher is examining higher secondary commerce students in Mizoram, with a total population of 5,000 students. Applying the 10% rule would suggest a sample size of 500 students.

Although this method offers simplicity and ease of application, it may not always be efficient for large populations, as it can yield excessively large samples. In such instances, more refined sampling techniques or established guidelines—such as those developed by Cochran (1977) or Krejcie and Morgan (1970)—may provide a better balance between statistical accuracy and practical constraints.

2. Minimum Sample Size Recommendations

- Surveys: To ensure statistical reliability and generalizability of findings, it is generally advised to include a minimum of 100 to 200 respondents in survey-based research (Krejcie & Morgan, 1970).
- Experimental Research: For meaningful statistical comparisons, especially when employing inferential tests, it is recommended to have at least 30 participants in each group (Roscoe, 1975; Cohen, 1992).
- Qualitative Studies: The ideal sample size in qualitative research typically ranges from 5 to 30 participants, depending on the nature and depth of the study. Smaller samples allow for in-depth exploration and thematic saturation (Guest, Bunce, & Johnson, 2006; Creswell, 2013).
- Structural Equation Modeling (SEM): A minimum sample size of 200 is considered appropriate to ensure stability and accuracy of parameter estimates in SEM analyses (Jackson, 2003; Kline, 2015).

B. Statistical Approaches

1. Cochran's Formula (for Large/Unknown Populations):

Determining an appropriate sample size is essential in research disciplines such as social sciences, education, health, and market studies. Cochran's Formula, introduced by William G. Cochran in 1977, is a widely accepted method for calculating the minimum sample size needed to ensure data accurately represents a population, particularly when the population is large or its characteristics are not precisely known.

This formula is especially useful when estimating proportions and allows researchers to achieve a desired level of confidence and precision. Key advantages of using Cochran's formula include:

1. Statistical Validity: It ensures that the sample is representative, making the findings generalizable.
2. Reduced Sampling Error: It minimizes the likelihood of inaccuracies caused by random data fluctuations.
3. Efficient Resource Use: It prevents unnecessary data collection, saving time and cost.
4. Proportion-Based Analysis: It is ideal for studies aiming to estimate population proportions.

For Large or infinite population

Cochran's formula for determining sample size is:

$$n_0 = \frac{Z^2 pq}{e^2}$$

Where:

n_0 = Required sample size (for a large or infinite population)

Z = Z-score (standard normal deviation) corresponding to the desired confidence level

p = Estimated proportion of the population that has the characteristic of interest

q = 1 - p (proportion of the population without the characteristic)

e = Margin of error (or precision level), which represents the acceptable level of sampling error

Each component in Cochran's formula has a specific role in determining the sample size:

i) Z-score (Z): Confidence Level Representation

The Z-score represents how many standard deviations a data point is from the mean in a standard normal distribution. It corresponds to the confidence level, which reflects how certain we are that our sample represents the entire population.

Common confidence levels and their corresponding Z-scores:

- 90% Confidence Level → Z=1.645
- 95% Confidence Level → Z=1.96 (most commonly used)
- 99% Confidence Level → Z==2.576

A higher confidence level results in a larger sample size.

ii) Proportion of Population (p):

Represents the expected proportion of people in the population who have a specific characteristic. If no prior information about p is available, p =0.5 (i.e. 50%), is often used because it provides the largest possible sample size, ensuring the most conservative estimate. If historical data suggests a different proportion, that value should be used.

iii) Complement of q:

Since q= 1 - p, it represents the proportion of the population that does not have the characteristic being studied.

iv) Margin of Error (e):

Also called the precision level, it represents the amount of error the researcher is willing to accept. Common choices for margin of error:

e =0.05 (5%) → Used in most social science studies.

e =0.01 (1%) → Used when high precision is needed.

(A smaller margin of error increases the required sample size.)

For Finite Population (Adjusting Cochran's Formula)

Cochran's formula assumes an infinitely large population. However, if the actual population size (N) is small or finite, an adjustment is needed using the finite population correction (FPC) formula:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

Where:

n = Adjusted sample size for a finite population

N = Total population size

n_0 = Initial sample size (calculated using Cochran's formula)

This correction reduces the required sample size when the population is small, preventing unnecessary data collection.

Example 1: Large Population

A researcher wants to determine the sample size needed to estimate a proportion with 95% confidence and a 5% margin of error. Since no prior data is available, the researcher assumes p=0.5

Given: Z==1.96 (95% confidence level), p=0.5, q=1−0.5= 0.5, e=0.05

Using the formula:

$$n_0 = \frac{(1.96)^2(0.5)(0.5)}{(0.05)^2}$$

$$n_0 = \frac{3.8416 \times 0.25}{0.0025}$$

$$n_0 = \frac{0.9604}{0.0025} = 384.16$$

Thus, the required sample size is 385 respondents.

Example 2: Finite Population (e.g., 1000 people)

If the total population is only 1000, we apply the finite population correction:

$$n = \frac{385}{1 + \frac{385 - 1}{1000}}$$

$$n = \frac{385}{1 + \frac{384}{1000}}$$

$$n = \frac{385}{1.384} = 278.3$$

Thus, the required sample size for a population of 1000 is 278 respondents.

Key Takeaways

1. *If the population is large (approaching infinity), use Cochran's formula directly.*
2. *For a small population, apply the finite population correction.*
3. *A higher confidence level and lower margin of error increase the required sample size.*
4. *Using p=0.5, provides the most conservative estimate but can be adjusted if prior knowledge is available.*
5. *If non-response is expected, the sample size should be increased accordingly.*

2. Yamane's Formula (for when Population Size is Known)

Yamane's formula provides a simplified approach for calculating the required sample size when the total population size (N) is known. This formula is widely used in social science research to determine an appropriate sample size while considering a margin of error.

The formula for determining sample size is:

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

where:

n = Required sample size
N = Total population size
e = Margin of error (expressed as a decimal, e.g., 5% = 0.05)

Understanding the Formula Components:

1. Population Size (N): The total number of people in the group you are studying.
2. Margin of Error (e): The level of precision required in the estimate, typically ranging from 1% to 10%. A lower margin of error requires a larger sample size.
3. Denominator Interpretation: The term 1+Ne^2increases as N grows, ensuring that larger populations require proportionally smaller samples.

Example:

Let's assume we want to conduct a study on a population of 5000 individuals and we choose a margin of error of 5% (e =0.05).

$$n = \frac{5000}{1 + 5000(0.05)^2}$$

$$n = \frac{5000}{1 + 5000(0.0025)}$$

$$n = \frac{5000}{1 + 12.5}$$

$$n = \frac{5000}{13.5} = 370.37$$

Since sample size should be a whole number, we round up to 371 respondents.

When to Use Yamane's Formula

- *When total population size (N) is known.*
- *When a simplified method is preferred over complex statistical formulas.*
- *When a basic margin of error-based approach is sufficient.*
- *Useful for social sciences, business research, and field studies.*

Yamane's formula is a practical and straightforward approach to determining sample size when the population size is known. It enables researchers to achieve a balance between precision and practicality, ensuring a statistically valid sample for their research.

3. Krejcie and Morgan's Table

Krejcie and Morgan's Table is a widely recognized statistical tool used for determining the appropriate sample size for surveys or research studies based on a specified population size. Developed by Robert V. Krejcie and Daryle W. Morgan in their

1970 publication, "Determining Sample Size for Research Activities", in the Educational and Psychological Measurement journal, the table offers a convenient reference for researchers to ensure their sample size is sufficient for obtaining statistically significant results.

Key Concepts of Krejcie and Morgan's Table:

i) Population Size (N): This refers to the total number of individuals or units within the population under study.

ii) Sample Size (S): This represents the number of individuals or units selected from the population to participate in the study.

iii) Confidence Level: A typical confidence level of 95% is applied, meaning that, if the population were repeatedly sampled, the sample mean would fall within the confidence interval 95% of the time.

iv) Margin of Error (e): This indicates the range within which the true population parameter is expected to lie, often set at 5%, which corresponds to a 95% confidence level.

How the Table Functions:

i) Population Size (N): The table lists different population sizes in one column.

ii) Sample Size (S): In the adjacent column, the corresponding recommended sample sizes are provided.

The sample sizes in the table are calculated using a formula that takes into account the desired level of precision, confidence, and variability in the population. The formula used is:

$$n = \frac{\chi^2 \cdot N \cdot P \cdot (1 - P)}{d^2 \cdot (N - 1) + \chi^2 \cdot P \cdot (1 - P)}$$

Where:

n = required sample size

χ = chi-square value for the desired confidence level (e.g., 3.8416 for 95% confidence, corresponding to Z = 1.96)

N = population size

P = estimated population proportion (assumed to be 0.5 for maximum sample size)

d = margin of error or degree of accuracy (e.g., 0.05 for ±5%)

Practical Application

- Ease of Use: Researchers can simply look up the population size in the table and find the corresponding sample size without performing complex calculations.
- Accuracy: The table ensures that the sample size is sufficient to make reliable inferences about the population.
- Standardization: It provides a standardized method to determine sample size, promoting consistency across different studies.

Example: Imagine you are conducting a survey to understand the job satisfaction levels of employees at Higher Education Institutions (HEIs). The total number of employees at the HEIs (population size, N) is 1,200.

Using Krejcie and Morgan's Table:

- Population Size (N): Find the row in the table that corresponds to a population size of 1,200.
- Sample Size (n): Look at the recommended sample size for a population of 1,200.

According to Krejcie and Morgan's Table, for a population size of 1,200, the recommended sample size (S) is approximately 291. This is the number of employees you need to survey to obtain statistically significant results with a 95% confidence level and a margin of error of ±5%.

$$n = \frac{\chi^2 \cdot N \cdot P \cdot (1 - P)}{d^2 \cdot (N - 1) + \chi^2 \cdot P \cdot (1 - P)}$$

Parameters: N = 1,200; χ^2 $\approx$ 3.841 (for 95% confidence level); P = 0.5 (population proportion, for maximum variability); d =0.05 (degree of accuracy, or margin of error)

$$n = \frac{3.841 \cdot 1200 \cdot 0.5 \cdot (1 - 0.5)}{(0.05)^2 \cdot (1200 - 1) + 3.841^2 \cdot 0.5 \cdot (1 - 0.5)}$$

$$n = 291$$

4. Samuel B. Green's Formula (1991) (Sample Size in Multiple Regression)

Samuel B. Green (1991) proposed a heuristic approach to estimate the minimum required sample size for multiple regression analysis. His approach includes two different rules of thumb depending on whether the focus is on testing individual predictors or the overall model.

A. Formula for Testing Individual Predictors: When the goal is to test the significance of individual predictors, the required sample size (n) is given by:

$$n \geq 104 + k$$

B. Formula for Testing the Overall Model: If the research focuses on testing the overall model rather than individual predictors, the required sample size is:

$$n \geq 50 + 8k$$

where:

n = Required sample size

k = Number of predictors (independent variables)

Example:

Suppose a researcher is using 5 predictors in a multiple regression model (k =5).

Step 1: Calculate Sample Size for Individual Predictors

$$n \geq 104 + 5 = 109$$

Step 2: Calculate Sample Size for the Overall Model

$$n \geq 50 + 8 \times 5$$

$$n \geq 50 + 40 = 90$$

Step 3: Select the Larger Sample Size

To ensure adequate power for both tests, the researcher chooses the larger of the two values:

n = 109

Thus, the researcher should collect data from at least 109 participants.

5. G*Power Software

G*Power is a free statistical software used to determine the appropriate sample size for various statistical tests and to conduct power analysis.

Steps for Sample Size Calculation in G*Power:

1) Define Research Goals & Hypothesis:

- Identify the research question and whether the test is one-tailed or two-tailed.
- State the null (H_0) and alternative (H_1) hypotheses.

2) Select the Statistical Test:

- Choose from tests like: t-tests, ANOVA, Chi-square tests, Correlation analysis, Regression analysis, etc.
- G*Power categorizes tests based on:
  - Distribution-based approach (e.g., FFF-tests, ttt-tests)
  - Design-based approach (e.g., comparing means, correlations)

3) Choose the Type of Power Analysis: G*Power provides five types of power analysis:

- A priori: Determines the required sample size (N) before conducting the study.
- Post hoc: Calculates the statistical power given a fixed sample size.
- Compromise: Finds an optimal balance between Type I (α) and Type II (β) errors.
- Criterion: Determines the significance level (α) for a given power and sample size.
- Sensitivity: Identifies the effect size detectable with a given sample size and power.

4) Input Key Parameters:

- Effect Size (d,f,r): Determines the magnitude of the effect to detect.
- Significance Level (α): Commonly set at 0.05
- Power (1−β): Typically 0.8, meaning an 80% chance of detecting a true effect.
- Sample Allocation Ratio (N_2/N_1): Defines group sizes if comparing two groups.

5) Calculate the Sample Size: Click the "Calculate" button, and G*Power provides the required sample size.

6) Interpreting Results: The output shows:

- Computed sample size
- Achieved power
- Critical test values
- The "X-Y plot for a range of values" allows visualization of how changes in effect size, power, or α affect the required sample size.

Example: A Priori Sample Size Calculation for an Independent t-Test.

Suppose you are comparing two groups with:

- Effect size: d=0.5 (medium)
- Significance level: α=0.05
- Power: 1−β=0.80
- Equal group sizes: $N_1 = N_2$

*Steps in G*Power:*

1. Choose t-tests → Means: Two independent groups (t-test).
2. Select A Priori.
3. Input:

   Effect size d=0.5

   Significance level α=0.05

   Power $1-\beta$=0.80

   Allocation ratio N_2/N_1 =1 (equal group sizes)

   Click Calculate.

Result:

The required sample size per group: $N_1 = N_2 = 64$

Total sample size: $N_{total} = N_1 + N_2 = 128$

G*Power is an efficient tool for ensuring research studies have sufficient participants to detect meaningful effects while avoiding unnecessary large samples that waste resources.

Conclusion

The authors have attempted to compile a list of methods and techniques with a few illustrative examples for practice. This list is neither comprehensive nor complete as there are many other methods for sample size calculation. However, it is the belief of the authors that such a theoretical treatment is urgently necessary to enhance the research capabilities of educationists and other social science researchers in Mizoram and beyond. This paper may serve as a reference for future as well as current researchers and it is the hope of the authors that the scientific validity of our research will be improved in a small extent by this paper.

References

Babbie, E. (2020). The practice of social research (15th ed.). Cengage Learning.

Cochran, W. G. (1977). Sampling techniques (3rd ed.). John Wiley & Sons.

Cohen, J. (1992a). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1992b). A power primer. Psychological Bulletin, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Creswell, J. W. (2013). Qualitative inquiry and research design: Choosing among five approaches (3rd ed.). SAGE Publications.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fowler, F. J. (2013). Survey research methods (5th ed.). SAGE Publications.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. Field Methods, 18(1), 59–82. https://doi.org/10.1177/1525822X05279903

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. Structural Equation Modeling, 10(1), 128–141. https://doi.org/10.1207/S15328007SEM1001_6

Kline, R. B. (2015). Principles and practice of structural equation modeling (4th ed.). Guilford Press.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. Educational and Psychological Measurement, 30(3), 607–610. https://doi.org/10.1177/001316447003000308

Roscoe, J. T. (1975). Fundamental research statistics for the behavioral sciences (2nd ed.). Holt, Rinehart & Winston.

Yamane, T. (1967). Statistics: An introductory analysis (1st ed.). Harper & Row.