12 SAMPLE SIZE CALCULATION

The smaller the sample size, the more likely that it is unrepresentative of the wider population.

— Michael Lewis

The majority of research investigations involve the observation of a sample from a target population that has been predetermined. For instance, a sample of people is monitored in epidemiological research for exposure to various risk factors, health outcomes, and other associated variables. The study's conclusions are frequently based on extrapolating the sample's findings to the complete population that served as the sample. The quality of the sample collection and, in particular, the sample's representation of the population, will determine how accurate the results are.

The size of the researcher's sample is one of the most challenging considerations he must make. Research investigations frequently use analytical and empirical methodologies. Using sample sizes that have been employed in comparable research is a component of the empirical approach. This has no scientific foundation and can only be accepted if the generalisation mistakes in the earlier studies were kept below reasonable bounds and the present study's scope is substantially similar (objectives, design, study population, etc.). It's not advised to use this approach. Determining the size of the sample is one of the most challenging decisions that a researcher has to make. The two most common approaches in this regard are: *empirical* and *analytical*. Using sample sizes that have been employed in comparable research is a component of the empirical approach. However, this has no scientific foundation and can only be accepted if the generalisation errors in the earlier studies were kept below reasonable bounds and the present study's scope is substantially similar (objectives, design, study population, etc.). The analytical (scientific) method of choosing the right sample size for the study hinges on the evaluation of inference mistakes and a desire to reduce "Sampling error".

The question, "How large a sample must I take"—is one that statisticians are regularly asked. Statistics and common sense need to be combined to determine the appropriate sample size for any investigation. In actuality, the sample size is typically determined by the number of accessible subjects, the expense, or the allotted amount of time. The calculations for sample size are nevertheless helpful, even when the sample size is selected based on practical factors.

By allowing for the analysis of what size differences are expected to be found in the planned experiment or survey, the sample size might indicate if the experiment or survey is likely to be worthwhile.

ADEQUACY OF SAMPLE SIZE

The sample size for research must be calculated at the time the study is proposed; a sample that is too large is unethical and unneeded, while a sample that is too small is unscientific and also unethical. A statistical formula can be used to compute the required sample size, depending on certain assumptions. If no assumptions can be made, the sample size for a pilot study is selected arbitrarily. To determine an adequate sample size, we need to consider:

- *Degree of difference:* The difference between two groups, such as the difference between two means or proportions (the control value and the expected test result), is referred to as the degree of difference.
- *Confidence level (CL):* The likelihood that a population parameter estimates falls within a set of predetermined limits of the true value. It is represented by $1–\alpha$ or $(1–\alpha) \times 100$ (in terms of

%). The confidence Level is automatically fixed when α is selected. If α is 5% then CL 95%. If α is 10% then CL is 90%.

- *Type I error*: Alpha error is defined as "the likelihood of identifying a difference that does not exist" (False Positive). Typically, it is set at 0.05 or 5%. It is also called the "Level of significance".
- *Type II error:* The likelihood of *not* detecting a difference that actually exists is known as beta(β) error (False Negative). Usually, it is set as 0.2 or 20%. Power is 1- β (1-0.2 = 0.8).
- *Power of the test*: This refers to the likelihood that a false null hypothesis will be correctly rejected. It is calculated as 1- β. *The power of the test* is automatically fixed when option b is chosen.
- *Level of precision (d* or *e):* It is known by a variety of names like *sampling error* or *margin of error* or *precision*. It is the range within which the true value of the population is estimated to be. *For example*: if the estimated mean age is 20 years and the level of precision is ±5, then the true population mean would be between 15 and 25 years.
- *Effect size:* The effect size is the smallest difference between the groups under study that the researcher hopes to find, or the difference between an estimated and an unmeasured parameter that the researcher hopes to measure.
- *Design Effect:* When a sampling method (such as cluster sampling, respondent-driven sampling, or stratified sampling) produces bigger sample sizes (or wider confidence intervals) than one might anticipate with simple random sampling, the adjustment is known as a design effect (expressed as DEFF). The ratio of the actual variance to the variance anticipated with SRS is known as the design effect. A DEFF of 2 indicates that the variance is twice as great as what SRS would predict. Additionally, it implies that you would require twice as many samples if you had employed cluster sampling. The formula to find the design effect is: $DEFF = 1 + \delta(n - 1)$. Here, δ stands for *Interclass correlation,* which measures the reliability of measurements for clusters—data that has been collected as a group or sorted into groups. It is also customary to consider a design effect of 1.5 to 3 in cluster sampling.
- *Z value:* Also known as the *critical value.* At different levels of significance, the Z-value can be determined by consulting the Z-distribution table. Following are some commonly used 'z' values at a 95% confidence level:
  - *Two-sided test*: $Z_{\alpha/2} = 1.96$ and one-sided test: $Z_{\alpha} = 1.65$.
  - *At 90% Confidence level:* Two-sided test: $Z_{\alpha/2} = 1.65$ and one-sided test: $Z_{\alpha} = 1.28$
  - *At 90% Power*: $Z_{\beta} = 1.28$ and at 80% Power: $Z_{\beta} = 0.84$.

IMPORTANCE OF SAMPLE SIZE CALCULATION IN RESEARCH

Sample size calculation is a foundational element of rigorous research design. It ensures that studies are methodologically sound, ethically responsible, and statistically valid. By determining the appropriate number of participants, researchers can enhance the reliability, credibility, and applicability of their findings while using resources efficiently and respecting the rights of participants.

1. Ensuring Statistical Power: An accurate sample size ensures that the study has sufficient statistical power to detect a true effect if it exists. Underpowered studies may produce false negatives, failing to reveal meaningful associations or differences. This weakens the study's contribution to scientific knowledge and may lead to missed opportunities for innovation or policy impact.

2. Controlling Error Rates: Sample size calculation helps control the risks of Type I error (incorrectly concluding an effect exists) and Type II error (failing to detect a real effect). This balance is achieved by setting thresholds for statistical significance (e.g., α = 0.05) and desired power (e.g., 80% or 90%). Without this step, statistical conclusions may be unreliable.

3. Improving Precision and Reliability: Larger, well-calculated samples produce narrower confidence intervals, resulting in more precise effect estimates. This is especially critical in public health, where minor variations in risk or prevalence estimates can

substantially affect resource allocation and intervention design decisions.

Importance of sample size calculation

- Ensuring Statistical Power
- Controlling Error Rates
- Improving Precision and Reliability
- Enhancing Generalizability
- Supporting Ethical Research
- Optimising Resources
- Meeting Regulatory and Funding Requirements
- Ensuring Consistency Across Multi-Site or Multi-Phase Studies

4. Enhancing Generalisability: A sample that is large enough and appropriately selected improves external validity, making it more likely that the study findings will apply to the broader population. Generalisability is crucial for informing national policies, scaling interventions, and applying findings across different demographic or geographic contexts.

5. Supporting Ethical Research: Ethical research practice demands that participant involvement is meaningful and scientifically justified. Underpowered studies may waste participants' time and expose them to interventions without producing valid knowledge. Over-sampling, by contrast, may expose more individuals to potential risks than necessary. Sample size calculation ensures a proper ethical balance.

6. Optimising Resources: Accurate sample size planning helps avoid unnecessary use of time, money, staff, and materials. A study that overestimates or underestimates the required sample may experience delays, higher costs, or produce inconclusive results. Resource optimization is particularly important in low-resource settings or time-sensitive research contexts.

7. Meeting Regulatory and Funding Requirements: Research ethics committees, institutional review boards (IRBs), and funding agencies often require a justified sample size calculation as part of the proposal approval process. This demonstrates scientific rigour and accountability, increasing the likelihood of funding approval and ethical clearance. Failure to justify the sample size can delay or disqualify a study from proceeding.

8. Ensuring Consistency Across Multi-Site or Multi-Phase Studies: In multi-centre trials, program evaluations, or longitudinal studies, consistent sample size planning ensures that comparisons across sites, groups, or time points are valid and statistically robust. This consistency is essential when aggregating results or performing subgroup analyses. Without uniform sample size considerations, inter-site or inter-phase variability may confound the findings.

METHODS OF SAMPLE SIZE CALCULATION

Sample size calculation is an important aspect of designing a study or an experiment, as it determines the number of participants needed to achieve a desired level of statistical power and precision. The sample size is dependent on several factors such as the study design, research question, level of significance, etc. There are various methods to determine the sample size like:

- Random numbers (not recommended).
- Earlier studies (may or may not be accurate).
- Nomograms and tables can be utilised; however, they might not be precise or adaptable.
- Formulas, however, change depending on the study's design and analysis.
- User-friendly computer applications .

CALCULATION OF SAMPLE SIZE

In quantitative research, it is tough for researchers to access a large population size; therefore, researchers need to reduce the population size into a correct, adequate, and appropriate sample size for collecting data. The population can be classified by the scope of population into two types. Such as a) finite population is one where all the numbers can be completely counted, and b) infinite population, where all the numbers can't be completely counted.

SAMPLE SIZE CALCULATION: BASED ON POPULATION CHARACTERISTICS

In biomedical research, sample size is estimated using two formulas. They are:

a) Cochran Formula:

1. If the population size is unknown but a lot, the population proportion is known, the formula is:

$$Sample\ size\ \text{'n'} = \frac{z^2pq}{d^2}$$

Here, d = level of precision.

z = value of the standard normal distribution corresponding to a significance level of α.

p = expected proportion in the population

q = 1-p

For example, if a researcher wants to estimate the proportion of individuals in a population who have a certain disease (hypertension) with 95% confidence and 5% precision, and the expected proportion is 0.50 (or 50%), the Cochran formula can be used to calculate the required sample size as follows:

$$n = \frac{(1.96)^2 \times 0.5 \times (1\text{-}0.5)}{(.05)^2}$$

n = 384.16 ≈ 384

Therefore, a sample size of at least 384 individuals would be required to estimate the proportion of individuals with the disease with 95% confidence and 5% precision.

2. If the population size is unknown but a lot, the population proportion is unknown, the formula is:

$$Sample\ size\ \text{'n'} = \frac{z^2}{4d^2}$$

Here, d = level of precision.

z = value of the standard normal distribution corresponding to a significance level of α.

This formula is assuming that the sample size is less than 5% of the population size, which is usually the case for very large populations. The formula provides a conservative estimate of the required sample size, which can ensure a certain level of precision and confidence in the study results.

For example, if a researcher wants to estimate the proportion of individuals in a population who have diabetes with 95% confidence and 5% precision, the formula can be used to calculate the required sample size as follows:

$$n = \frac{(1.96)^2}{4 \times (.05)^2}$$

n = 384.16 ≈ 384

Therefore, a sample size of at least 384 individuals would be required to estimate the population proportion with 95% confidence and 5% precision, assuming the population size is very large.

3. If the population size is unknown and the estimated population mean is known, the formula is:

$$Sample\ size\ \text{'n'} = \frac{z^2\sigma^2}{d^2}$$

Here, d = level of precision.

z = value of the standard normal distribution corresponding to a significance level of α.

σ = Standard Deviation of sample.

For example, if a researcher wants to estimate the mean weight of adult women in a city with 95% confidence and a margin of error of 1 kg, and the estimated population standard deviation is 5 kg, the formula can be used to calculate the required sample size as follows:

$$n = \frac{(1.96)^2 \times (5)^2}{(1)^2}$$

n = 96.04 ≈ 96

Therefore, a sample size of at least 96 adult women would be required to estimate the mean weight of adult women in the city with 95% confidence and a margin of error of 1 kg, assuming the estimated population standard deviation is 5 kg.

4. If the population size is known with an estimated population proportion, the formula is:

$$Sample\ size\ \text{'n'} = \frac{p\ (1-p)}{\frac{d^2}{z^2} + \frac{p\ (1-p)}{N}}$$

Here, d = level of precision.

z = value of the standard normal distribution corresponding to a significance level of α.

N = Population size

p = The population proportions

For example, if a researcher wants to estimate the proportion of children (<5 years) in a population who

have malnutrition in a city with 95% confidence and 5% precision, the population size is 10,000 and previous studies showed a proportion of 12% malnutrition among children, the formula can be used to calculate the required sample size as follows:

$$n = \frac{0.12 \times (1-0.12)}{\frac{(0.05)^2}{(1.96)^2} + \frac{0.12\,(1-0.12)}{10000}}$$

n =159.67 ≈ 160

Therefore, a sample size of at least 160 children (<5 years), would be required to estimate the proportion of malnutrition in the city with 95% confidence and 5% precision.

5. If the population size is known with an estimated population mean, the formula is:

$$Sample\ size\ \text{'n'} = \frac{Nz^2\sigma^2}{(N-1)d^2 + z^2\sigma^2}$$

Here, d = level of precision.

z = value of the standard normal distribution corresponding to a significance level of α.

N = population size

σ = standard deviation of sample.

b) Taro Yamane Formula:

A method for calculating or determining sample size in relation to the population under study was developed by mathematician Taro Yamane, allowing inferences and conclusions from the survey to be applied to the entire population from which the sample was drawn.

The following is Taro Yamane's statistical formula:

$$Sample\ size\ \text{'n'} = \frac{N}{1 + Nd^2}$$

Here, d = level of precision.

N = Population size.

For example, a health organisation wants to conduct a survey to estimate the proportion of adults in a certain urban region who have been vaccinated against COVID-19. The region has a population of 500,000 people. The organisation intends to be 99% confident that their estimate is within 2% of the true population proportion.

Using the Taro Yamane formula, we can calculate the required sample size as:

$$n = \frac{500000}{1 + \{500000 \times (0.02)\}^2}$$

n = 2487.6 ≈ 2488

Therefore, we would need a sample size of at least 2488 adults to estimate the proportion of adults in the region who have been vaccinated against COVID-19 with 99% confidence and a margin of error of 2%, assuming that the population size is known.

SAMPLE SIZE CALCULATION FOR ONE SAMPLE

a) Estimating sample size with absolute precision:

Example:

To estimate the true immunization coverage in a community of under five children, previous studies indicated that immunization coverage is around 85%. With a level of precision at 5% and confidence level at 95% the formula for sample size calculation would be—

$$Sample\ size,\ n = \frac{z^2 p\,(1-p)}{d^2}$$

Here, d = level of precision = 5% = 0.05

z = 1.96 = value of the standard normal distribution corresponding to a significance level of α.

p = expected proportion in the population = 85% = 0.85

1-p = 1-0.85 = 0.15

Hence,

$$n = \frac{(1.96)^2 \times 0.85 \times 0.15}{0.05^2}$$

n = 195. 92 ≈ 196

So, the sample size is 196.

Therefore, the researcher would need a sample size of at least 196 under five children to estimate the true immunization coverage in a community.

b) Estimating sample size with relative precision:

Example:

To estimate the true immunization coverage among

the under five children in a community, where previous studies indicated that immunization coverage is around 80%, relative precision, i.e., within 10% of the true value at confidence level = 95%.

The sample size calculation would be—

Here, *Relative Precision (d)* = 10%= 0.10 = Relative difference between sample coverage and true population coverage (we have decided that this would be ±10% of the anticipated population proportion of 80%).

p = Expected proportion in the population = 80% or 0.8.

z = 1.96 (at 95% confidence level.)

So, for a relative precision of 10%, the formula will be—

Sample size, n = $\frac{z^2\, p\,(1-p)}{(d.p)^2}$

So, $n = \frac{(1.96)^2 \times (0.8) \times (1-0.8)}{(0.1 \times 0.8)^2}$

n = 96.

Therefore, the researcher would need a sample size of at least 96 under five children to estimate the true immunization coverage among the under five children in a community with a relative precision of 10%.

Another widely used formula is:

Sample size, n = $\frac{z^2 q}{r^2 p}$

Following the above example where,

Here,

p = Expected proportion in the population = 80% or 0.8.

q = 1-p = 0.2

Relative Precision ® = 10%

z = 1.96 (at 95% confidence level.)

$n = \frac{(1.96)^2 \times (0.2)}{(0.1)^2 \times 0.8}$

n = 96.04 ≈ 96

If the researcher wishes to estimate parameters for each sub-category with the same level of precision, the calculated sample size should be applied to each category.

c) Estimating sample size using simple random sampling

The formula for estimating sample size using simple random sampling is:

$$n = \frac{z^2 \times s^2}{d^2}$$

Here,

z = the standard normal deviate associated with the desired level of confidence

s = the estimated standard deviation of the population parameter being estimated (if unknown, it can be estimated from a pilot study or previous research)

d = the desired level of precision (expressed as a decimal)

For example, a researcher wants to estimate the average blood pressure of a population of individuals. Previous research suggests that the standard deviation of the population is around 10 mmHg. The researcher would like to estimate the population mean with a margin of error of 2 mmHg and be 95% confident in the estimate.

Using the formula, we can calculate the required sample size as:

$$n = \frac{(1.96)^2 \times (10)^2}{(2)^2}$$

n = 96.04 ≈ 96.

Therefore, the researcher would need a sample size of at least 96 individuals to estimate the average blood pressure of the population with a margin of error of 2 mmHg and a confidence level of 95%, assuming that the standard deviation of the population is known.

d) Estimating sample size using cluster sampling

The formula for estimating sample size using cluster sampling is:

Sample size, n = $\frac{gz^2\, PQ}{d^2}$

Here, P = Expected Prevalence

Q = 1-P

d = Absolute precision

g = Design effect (usually estimated as 2 in immunization cluster surveys).

For example, to estimate the true immunization coverage among under five children of a community, previous studies showed that immunization coverage around 70% with a level of precision at 5% and a

confidence level at 95%.

$$n = \frac{2 \times (1.96)^2 \times 0.7 \times 0.3}{0.05^2}$$

$n = 645.39 \approx 645$

So, a total of 645 samples is needed.

e) Estimating sample size for a one-sided test Example:

A survey among school children in the rural areas of Rajshahi division of Bangladesh found that the prevalence rate of dental caries was 25%. How many children should be included in a new survey designated to test for a decrease in the prevalence of dental caries, if it is desired to be 90% sure of detecting a rate of 20% at a 5% level of significance?

Here,

- Test caries rate, P_o = 25%
- Anticipated caries rate, P_a = 20%
- Level of significance, α = 5%
- The alternate hypothesis is (one-sided) = (less than 25%)

Power of the test (1–β) = 90% The formula is *Sample size,* $n = \frac{[\{z\,(1-\alpha)\,.\sqrt{P_o\,(1-P_o)}\}+\{z\,(1-\beta)\,\sqrt{P_a\,(1-P_a)}\}\,]^2}{(P_o-P_a)^2}$

$$n = \frac{[1.645\,.\,\sqrt{25 \times 75} + 1.283\,\sqrt{20 \times 80}\,]^2}{(25-20)^2}$$

$n = 601$

So, the sample size would be 601 school children.

f) Estimating sample size for a two-sided test

Example: In a hospital, the success rate of surgical treatment was 70%. How many patients are to be studied to test an alternative hypothesis that is not 70% at a 5% level of significance? An investigator wants to have 90% power of detecting a difference between the success rate of 10 percentage points more in either direction.

Here,

- Test success rate, P_o = 70%
- Anticipated success rate, P_a = 60 or 80%
- Level of significance = 5%
- Power of the test = 90%
- Alternate hypothesis (two-sided) = 70%
- $Z_{1-\alpha/2}$ = The standard normal deviate associated with the desired level of confidence, (1.96 at 95% CI)
- $Z_{1-\beta}$ = 1.283 at 90% power.

The formula is, $n = \frac{[\{Z\,(1-\alpha/2)\,.\sqrt{P_o\,(1-P_o)}\}+\{Z\,(1-\beta)\,\sqrt{P_a\,(1-P_a)}\}\,]^2}{(P_o-P_a)^2}$

$$n = \frac{[1.96\,.\,\sqrt{30 \times 70} + 1.283\,\sqrt{40 \times 60}\,]^2}{(70-60)^2}$$

$n = 233$

So, 233 patients are to be studied to test an alternative hypothesis.

SAMPLE SIZE CALCULATION FOR TWO SAMPLE

a) Estimating sample size for the difference between two proportions with absolute precision

Example: It is found that 40% of 50 coal mine workers in a coal mining factory at the digging point had anthracosis. On the other hand, 32% of 50 coal mine workers in a coal mining factory at the office point had anthracosis. Estimate the true anthracosis risk difference within 5 percentage points with 95% confidence.

Here,

a. Anticipated population proportions:

P_1 = 40% = 0.4

P_2 = 32% = 0.32

b. Confidence level = 100 (l–α) % = 95%

c. Absolute precision required on either side of the true value of the difference between the proportions (percentage points), d = 5% = 0.05

The formula is—

Sample size, $n = \frac{Z^2\,[\{P_1(1-P_1)\}+\{P_2(1-P_2)\}]}{d^2}$

$$n = \frac{(1.96)^2 \times \{(0.40 \times 0.60)+(0.32 \times 0.68)\}}{0.05^2}$$

$n = 703.16 \approx 703$

Therefore, we need to sample at least 703 workers from each group to compare the proportions.

b) Estimating sample size for the difference between two proportions with standard error

Example: In a national survey, it was found that

doctors leaving service and joining in another country for better prospects were as follows: From District "A" 30% of doctors, while in District "B" 15% were leaving the job. The investigator wished to find out the difference between two districts at 95% confidence limits with standard error.

Here, s.e. = $\sqrt{\frac{30\times70}{n}+\frac{15\times85}{n}}$

We know, s.e. = s.e. of difference between proportions= 5 (given)

So, s.e. = $\sqrt{\frac{2100}{n}+\frac{1275}{n}}$

$(5)^2 = (\sqrt{\frac{3375}{n}})^2$

$25 = \frac{3375}{n}$

$n = \frac{3375}{25}$

n = 135.

Therefore, we need to sample at least 135 doctors from each group to compare the proportions.

c) Estimating sample size for two population proportions with a one-tail test

Example:

It is believed that the proportion of patients who develop complications after undergoing one type of surgery is 5% while the proportion of patients who develop complications after a second type of surgery is 15%. How large should the sample size be in each of the two groups of patients if an investigator wishes to detect, with a power of 90%, whether the second procedure has a complication rate significantly higher than the first at the 5% level of significance?

The formula is:

$$n = \frac{\{Z_{1-\alpha} \times \sqrt{[2P\ (1-\ P)} + Z_{1-\beta}\sqrt{[\ P_1\ (1-P_1) + P_2\ (1-P_2)]}\}^2}{(\ P_1-P_2)^2}$$

Here,

Proportion of patients developing complications after one type of surgery (P_1) = 0.05.

Proportion of patients developing complications after another type of surgery (P_2) = 0.15

Determine that the second procedure has a significantly higher complication rate than the first procedure at a level of significance (α) = 5% (one-tail),

Power (1–β) = 90%

$Z_{1-\alpha}$ = At a confidence level of 95% = 1.645 (as it is a one-sided test).

Z_β = 1.283 (at 90% Power)

$P = \frac{(P_1 + P_2)}{2}$

So, $n = \frac{\{[1.645\times\sqrt{2\times0.10\times0.90}]+[1.283\times\sqrt{(0.05\times0.95)+(0.15\times0.85]}\}^2}{(0.05-0.15)^2}$

n = 153

Therefore, a sample size of 153 would be needed in each group.

d) Estimated sample Size for two population proportions with two-tail test

It is believed that the proportion of patients who develop complications after undergoing one type of surgery is 5% while the proportion of patients who develop complications after a second type of surgery is 15%. How large should the sample size be in each of the two groups of patients if an investigator wishes to detect, with a power of 80%, whether there is a significantly different rate of complication rate among the types of surgery at the 5% level of significance?

It can be calculated by the following formula:

$$n = \frac{\{Z_{1-\alpha} \times \sqrt{[2\bar{P}\ (1-\ \bar{P})} + Z_{1-\beta}\sqrt{[\ P_1\ (1-P1) + P_2\ (1-P_2)]}\}^2}{(\ P_1-P_2)^2}$$

Here,

Proportion of patients developing complications after one type of surgery (P_1) = 0.05.

Proportion of patients developing complications after another type of surgery (P_2) = 0.15.

Determine that the second procedure has a significantly higher complication rate than the first procedure at level of significance (α) = 5% (one-tail),

Power (1–β) = 80%

$Z_{1-\alpha}$ = At a confidence level of 95% = 1.96 (as it is two-sided).

Z_β = 0.84 (at 80% Power)

$P = \frac{(P_1 + P_2)}{2}$

So, $n = \frac{\{[1.96\times\sqrt{2\times0.10\times0.90}]+[0.84\sqrt{(0.05\times0.95)+(0.15\times0.85]}\}^2}{(0.05-0.15)^2}$

n ≈ 136

Therefore, a sample size of 136 would be needed in each group.

SAMPLE SIZE CALCULATION: BASED ON THE STUDY DESIGN

DESCRIPTIVE STUDIES

Studies that aim to estimate population parameters, often proportions or percentages and means, are referred to as descriptive studies. A distinction between studies with finite populations and those with infinite populations is made among these studies.

a) Finite populations

i. Estimation of a sample size using proportion:

The formula is—

$$\textit{Sample size, } n = \frac{Z^2pqN}{(N-1)d^2+Z^2pq}$$

For example, a researcher is interested in estimating the proportion of patients in a hospital who are satisfied with the quality of care they receive. The hospital has a total of 10,000 patients. The researcher wants to estimate this proportion with a 95% confidence level and a margin of error of 3%.

Here,

Z=1.96 for a 95% confidence level

P is the estimated proportion of patients who are satisfied with the quality of care. Suppose this proportion is 0.8 (80%), based on prior studies or expert opinion.

Q = 1-P = 1-0.8 = 0.2

N = the population size = 10,000

d = desired margin of error 3% = 0.03

So, sample size would be

$$n = \frac{(1.96)^2 \times 0.8 \times 0.2 \times 10000}{\{(10000-1)\times 0.03^2\} + \{(1.96)^2 \times 0.8 \times 0.2\}}$$

n = 640

Therefore, we need to sample at least 640 patients from the hospital to estimate the proportion of patients who are satisfied with the quality of care with a 95% confidence level and a margin of error of 3%.

ii. Estimation of a sample size using a mean:

The formula is—

$$\textit{Sample size, } n = \frac{Z^2s^2N}{(N-1)d^2+Z^2s^2}$$

Here, n = Sample size to be calculated.

N = Size of the population from which the sample is drawn.

s^2 = Variance of the variable for which we want to estimate the mean

d = Accepted margin of error (usually between 5 and 10%).

z = At a confidence of level 95% = 1.96.

b) Infinite populations

The formulas pertaining to sample size can be simplified in the situation of infinite populations because the size of the population has no bearing on the results.

i. Estimation of a sample size using a proportion:

The formula is—

$$\textit{Sample size, } n = \frac{Z^2pq}{d^2}$$

Here, p = expected proportion.

q = 1-p

d = Accepted margin of error (typically between 5 and 10%).

z = At a confidence level of 95% = 1.96.

ii. Estimation of a sample size using a mean:

The formula is—

$$\textit{Sample size, } n = \frac{Z^2s^2}{d^2}$$

Here, n = Sample size to be calculated.

s^2 = Variance of the variable for which we want to estimate the mean

d = Accepted margin of error (often between 5 and 10%).

z = At a confidence level of 95% = 1.96.

ANALYTICAL STUDIES

a) Case-control study:

i. Sample size for independent case-control studies:

The formula for estimating sample size *n* for an independent case-control study is as follows:

$$\textit{Sample size, } n = \frac{(r+1)}{r}\ \frac{(\overline{p})(1-\overline{p})(Z_\beta+Z\alpha_2)^2}{(p_1-p_2)^2}$$

Here,
n = sample size
r = Ratio of control to case is 1
p_2 = Proportion of control exposed is 48.5% = 0.485
p_1 = Proportion of case exposed
Where,

$$p_1 = \frac{OR \times p_2}{p_2(OR-1)+1}$$

$$= \frac{2.25 \times 0.485}{0.485(2.25-1)+1}$$

= 0.68
OR= Usually taken from a previous study = 2.25
$\overline{p}$ = The average proportion of exposed $= \frac{p_1+p_2}{2}$ = 0.58
z_β = 80% power =0.84
$z_{\alpha/2}$ = for 0.05 significant level, $z\alpha/_2$ = 1.96
So, finally, our sample, size would be,

$$n = \frac{2 \times 0.58 \times 0.42 \times 7.84}{0.038} = 100.5 \approx 101$$

Considering 10% nonresponse rate,
n = 101/(1-0.10)
n = 101/0.9
n = 112.2 ≈ 112
Therefore, we need to 112 cases and 112 controls.

ii. Sample size for matched case-control studies:

In a matched case-control study, cases are paired with controls based on specific characteristics such as age, gender, or hospital admission date to minimise the influence of confounding variables. Because the data in these studies are paired (correlated), standard sample size formulas for independent groups are inappropriate. Instead, the calculation focuses on the number of discordant pairs needed to detect a statistically significant association. A discordant pair occurs when the case and the control in a matched pair have different exposure histories (e.g., the case was exposed but the control was not, or vice versa). Concordant pairs, where both the case and control have the same exposure status (both exposed or both unexposed), do not contribute to the calculation of the odds ratio in a matched analysis (such as McNemar's test) and therefore do not contribute to the statistical power of the study regarding the hypothesis test.

The sample size calculation involves two stages. First, the researcher must calculate the number of discordant pairs (m) required to achieve the desired statistical power. Second, the researcher must estimate the total number of pairs (N) needed to find that many discordant pairs, based on the expected prevalence of the exposure in the population.

The sample size for a matched case-control study can be estimated using the following steps:

$$m = \frac{(Z_\beta + Z\alpha_2)^2 \times (OR+1)^2}{(OR-1)^2}$$

Here,
m = Number of discordant pairs required.
$Z_{\alpha 2}$=1.96 (for 95% confidence).
Z_β=0.84 s(for 80% power).
OR = Expected Odds Ratio.
The total sample size (N) is then estimated by dividing *m* by the expected probability of a pair being discordant (p_e).

$$N = \frac{m}{p_e}$$

For example, if a researcher intends to conduct a 1:1 matched case-control study to investigate the association between a history of smoking (exposure) and the development of a specific rare respiratory condition (outcome). The cases (patients with the condition) are matched with controls (patients without the condition) by age and sex.

The researcher makes the following assumptions for the calculation:
Significance level (α): 5%
$Z_{\alpha/2}$: 1.96
Power (1−β): 80%
Z_β=0.84
Expected Odds Ratio (OR): 3.0 (the researcher hypothesizes that smokers are 3 times more likely to develop the condition).
Using the formula

$$m = \frac{(0.84+1.96)^2 \times (3+1)^2}{(3-1)^2}$$

$$=\frac{(2.8)^2\times(4)^2}{(2)^2}$$

$$=31.36 \approx 32$$

So, the researcher requires 32 discordant pairs to detect an odds ratio of 3.0 with 80% power. It is important to note that this is not the total sample size. To find 32 pairs that disagree on exposure status, the researcher will need to recruit a larger number of total pairs (N). If the prevalence of smoking in the control population is estimated to be 30% (P_0=0.3), statistical tables or further formulas would be used to estimate that the probability of a pair being discordant is approximately 0.46.

Calculation of p_e:

We know the exposure rate in controls (p_0) is 30% (0.3) and the OR is 3.0. We use these to find the expected exposure rate in cases:

$$p1 = \frac{OR \times p_0}{p_0(OR-1)+1} = 0.5625$$

In other words, roughly 56% of the cases are expected to be exposed.

A pair is discordant if the case is exposed and control is not *or* if the case is not exposed and control is.

If we assume the groups are independent (unmatched), the calculation is:

$$p_e = [P_1 \times (1 - P_0)] + [P_0 \times (1 - P_1)]$$

=[0.5625×0.7]+[0.3×0.4375]

= 0.525

However, because this is a matched study, the case and control are similar (e.g., same age, same sex). This creates a correlation ® between them. Matching makes pairs more likely to be the same (concordant) and less likely to be discordant. Epidemiologists typically adjust for this correlation (often estimating r≈0.1 to 0.2).

To adjust for this correlation, we must calculate the correlation correction factor and substract it from the unadjusted probability. For that, first we calculate a measure of variance (σ) based on the exposure probabilities of the two groups.

$\sigma = \sqrt{\{p_1 (1-p_1) \times p_0 (1-p=)\}}$

$= \sqrt{\{0.5625 (0.4375) \times 0.3 (0.7)\}}$

=0.227

The formula for correlation factor is

CF = 2 × r × σ

= 2 × 0.15 × 0.227

= 0.068

When we adjust for this correlation factor, the probability of finding a discordant pair drops from 0.525 to approximately 0.46.

Therefore, the total number of pairs required would be 32/0.46≈70 pairs (70 cases and 70 controls).

b) Cohort study:

The formula for estimating sample size '*n*' for a cohort study is as follows:

$$n = \frac{[Z_\alpha\sqrt{(1+\frac{1}{m})p^*(1-p^*)} + Z_\beta\sqrt{p_1(1-p_1)/m + p_2(1-p_2)}]^2}{(p_1-p_2)^2}$$

Here,

Z_α = Z value at a given confidence level.

Z_β = Z value at a given power.

m = Number of controls per experimental subjects

p_1 = Probability of events in the control group

p_2 = Probability of events in experimental group

$$p^* = \frac{p_2 + mp_1}{m+1}$$

For example, if a researcher wants to find out the association between walking exposure and diabetes mortality. According to previous studies, the proportion of death due to diabetes in the exposed group (p_1) was 20% and in the unexposed group (p_2) it was 40%. Hence sample size calculation for 5% significance level and 80% power with equal number of exposed and unexposed groups will be:

$$n = \frac{[1.96\sqrt{2\times 0.3(0.7)} + 0.84\sqrt{0.2(0.8)+0.4(0.6)}]^2}{(0.2-0.4)^2}$$

=81

Therefore, 81 participants will be in each group.

SAMPLE SIZE CALCULATION FOR SIGNIFICANT DIFFERENCE BETWEEN TWO GROUPS

a) Randomised controlled trial:

i. When outcome is expressed as proportion:

The formula for sample size calculation is

$$n = \frac{P_1(1-P_1)+P_2(1-P_2)}{(P_1-P_2)^2} \times (Z_\alpha + Z_\beta)^2$$

For example, a researcher wants to conduct an RCT to compare the proportion of individuals in the intervention group who experience an adverse event to the proportion of individuals in the control group who experience the same adverse event. He or she hypothesizes that the proportion of individuals in the intervention group who experience the adverse event will be lower than the proportion in the control group. The researcher expects that 20% of individuals in the control group will experience the adverse event. A statistical power of 80% and a significance level of 5% is required.

Here:

Z_α = the Z-score for the desired significance level (e.g., for a 5% significance level, 1.96)

Z_β = the Z-score for the desired power (for 80% power 0.84)

P_1 = the expected proportion of individuals in the control group who experience the adverse event = 0.20.

P_2 = the expected proportion of individuals in the intervention group who experience the adverse event. The expected difference between P_1 and P_2 is the effect size, which can be determined based on previous research or expert opinion. In this case, suppose the researcher expects the proportion in the intervention group to be 10%. Therefore, $P_2 = 0.10$.

$P_1 - P_2 = 0.20 - 0.10 = 0.10$.

So, $n = \frac{0.20\,(1-0.20)+0.10\,(1-0.10)}{(0.10)^2} \times (1.96 + 0.84)^2$

n = 196

Considering a 10% potential loss to follow up, the sample size would be: n = (196 / 1-0.1)

n =196/0.s9

n =217.7 ≈ 218

Therefore, we need a sample size of at least 218 individuals in each group (i.e., the control and intervention groups) to have a statistical power of 80% and a significance level of 5% to detect a 10% difference in proportions between the two groups.

ii. When outcome is expressed as mean:

$$n = \frac{\{\sigma_1^2 + \sigma_2^2)}{(\mu_1-\mu_2)^2} \times (Z_\alpha + Z_\beta)^2$$

Here,

For example, a researcher wants to conduct an RCT to compare the mean score on a depression scale in the intervention group to the mean score in the control group. He or she hypothesizes that the mean score in the intervention group will be lower than the mean score in the control group. The researcher expects that the mean score in the control group will be 15, the mean score in the intervention group will be 12, and the SD of the control group will be 3, the SD in the intervention group will be 4, a statistical power of 80% and a significance level of 5% is required.

Here:

Z_α = the Z-score for the desired significance level (e.g., for a 5% significance level, 1.96)

Z_β = the Z-score for the desired power (for 80% power 0.84)

σ_1 = Standard deviation from intervention group = 4

σ_2 = Standard deviation from control group = 3

μ_1 = Mean of intervention group = 12

μ_2 = Mean of control group = 15

So, the sample size would be:

$$n = \frac{\{4^2 + 3^2)}{(12-15)^2} \times (1.96 + 0.84)^2$$

n =21.77

Considering a 10% potential loss to follow up, the sample size would be: n = (21.77 / 0.9)

n =24.2 ≈ 24

Therefore, we need a sample size of at least 24 individuals in each group (i.e., the control and intervention groups

b) Comparative cross-sectional studies

i) The formula for sample size calculation is:

$$n = \frac{P_1(1-P_1)+P_2(1-P_2)}{(P_1-P_2)^2} \times (Z_\alpha + Z_\beta)^2$$

For example, a researcher wants to compare the prevalence of cervical cancer between urban and rural women. From the previous study, the proportion of the two groups is obtained, i.e., 40% and 20%, respectively. The researcher wants 95% CI and 80% power for the study.

Here,

Z_α = the Z-score for the desired significance level (e.g., for a 5% significance level, 1.96)

Z_β = the Z-score for the desired power (for 80% power 0.84)

P_1 = the proportion in urban group = 40% = 0.40.

P_2 = the proportion in rural group = 20% = 0.20.

So, $n = \frac{0.40\,(1-0.40)+0.20\,(1-0.20)}{(0.40-0.20)^2} \times (1.96 + 0.84)^2$

n =78.5

Considering a 10% potential loss, the sample size would be: n = (78.5 /0.9)

n =87.2 ≈ 87

Therefore, the researcher requires a minimum of 87 participants for the study in each group.

ii) There is another formula that is designed to make the most of the relative precision ® when calculating sample size. This formula is typically used for estimating the sample size needed to estimate a population proportion with a specified level of relative precision. However, it can also be used in comparative cross-sectional studies. You would have to calculate the sample size using to sets of proportion and relative precision (usually the same for both proportions), and consider the greater sample size. However, use this formula for comparative cross-sectional studies with caution as this might cause a loss of precision or confidence, or both.

The formula is:

$$n = \frac{z^2 \times q}{r^2 \times p}$$

Here,

n = sample size.

z = the z-score for the desired significance level (e.g., for a 5% significance level, 1.96)

p = population proportion

q = 1-p

For example, a comparative cross-sectional study is being conducted to estimate the proportions of urban and rural pregnant women in a population who seek antenatal care with a margin of error no greater than 10% and a 95% confidence level. Here, we would need a large sample size. It is estimated that the proportion of women seeking such care in the urban area is 45%, and in rural area it is 25% using previous studies.

So, for urban pregnant women:

p = the proportion in urban group = 0.45

r = relative precision = 0.1 (10%)

q = 1-p = 0.55

Z = the Z-score for the desired significance level (e.g., for a 5% significance level, 1.96)

So,

$$n_1 = \frac{(1.96)^2 \times 0.55}{(0.1)^2 \times 0.45}$$

n_1= 470

And for rural pregnant women:

p = the proportion in rural group = 0.25

r = relative precision = 0.1 (10%)

q = 1-p = 0.75

Z = the Z-score for the desired significance level (e.g., for a 5% significance level, 1.96)

So,

$$n_2 = \frac{(1.96)^2 \times 0.75}{(1)^2 \times 0.25}$$

n_2 = s1152

As $n_2 > n_1$, the second sample size (n_2) should be taken for each group.

c) Quasi-experimental study:

The formula for sample size calculation of quasi-experimental study depends on on several factors, such as the research question, the specific quasi-experimental design being used, the desired level of statistical power, and the expected effect size. There is no single formula that applies to all quasi-

experimental studies. However, the Pretest-Posttest design is most commonly used in quasi-experimental study.

The formula is:

$$n = \frac{(Z_\alpha + Z_\beta)^2 \times 2\sigma^2}{\delta^2}$$

For example, a researcher wants to know: does a new medication reduce blood pressure in patients with hypertension? So, he or she used a Pretest-Posttest Design with Two Groups.

Group 1: Patients with hypertension receiving the new medication (intervention group).

Group 2: Patients with hypertension receiving standard care without the new medication (control group).

Here,

Z_α = 1.96 (for alpha level of 0.05, two-tailed)

Z_β = 0.84 (for power of 0.80)

σ^2 = Estimated population variance (σ^2) of blood pressure = 100 (from previous studies)

δ = Expected effect size of reduction in systolic blood pressure = 5 mmHg

$$\text{So, } n = \frac{(1.96+0.084)^2 \times (2\times 100)}{(5)^2}$$

$$n = 62.70 \approx 63.$$

The estimated sample size for each group would be 63. Therefore, a total sample size of 126 (63 in each group) would be needed for this quasi-experimental study to detect a statistically significant difference in blood pressure between the intervention and control groups.

d) Superiority trials

A superiority trial is the most common form of randomised controlled trial. The primary objective is to demonstrate that a new intervention is statistically significantly better than a comparator, which may be a placebo or a standard-of-care treatment. The null hypothesis in this context assumes that there is no difference between the two groups, while the alternative hypothesis assumes that a difference exists. Determining the appropriate sample size requires the researcher to pre-specify the:

- acceptable Type I error rate (alpha, usually set at 0.05),
- the statistical power (1-beta, usually set at 0.80 or 0.90),
- and the minimum clinically important difference (effect size) they wish to detect.

Continuous variables

When the primary outcome of the trial is a continuous variable, such as blood pressure, cholesterol levels, or weight, the sample size calculation relies on comparing the means of the two groups. The formula requires an estimate of the standard deviation of the outcome variable within the population. It is assumed for the purpose of calculation that the variance is the same in both groups.

The formula for the sample size per group (n) is:

$$n = \frac{2(Z_{\alpha/2} + Z_\beta)^2 \sigma^2}{(\mu_1 - \mu_2)^2}$$

In this formula,

$Z_{\alpha/2}$: 1.96 (for a 5% significance level)

Z_β: 0.84 (for 80% power)

σ: The standard deviation

$(\mu_1 - \mu_2)$: Expected difference between the two groups.

Example

Consider a researcher designing a trial to test a new antihypertensive drug. The goal is to show that the new drug lowers systolic blood pressure more than the standard drug. Based on previous literature, the standard deviation of systolic blood pressure in this population is estimated to be 15 mmHg. The researcher determines that a difference of 5 mmHg is clinically significant. The significance level is set at 0.05 and the power at 80%.

Using the values:

$Z_{\alpha/2}$: 1.96; Z_β: 0.84; σ: 15; $(\mu_1 - \mu_2)$: 5

Placing these into the formula: $n = \frac{2\,(1.96+0.84)^2\,15^2}{(5)^2}$

= 141.12

Rounding up, the researcher requires 142 participants per group, or 284 participants in total, to detect a 5 mmHg difference with 80% power.

Categorical variables

When the primary outcome is categorical, typically binary (e.g., cured vs. not cured, dead vs. alive), the calculation compares two proportions. This requires an estimate of the expected proportion of events in the control group and the

intervention group.

The formula for the sample size per group (n) is:

$$n = \frac{P_1(1-P_1)+P_2(1-P_2)}{(P_1-P_2)^2} \times (Z_\alpha + Z_\beta)^2$$

Here, P_1 is the expected proportion of the event in the control group, and P_2 is the expected proportion in the intervention group. The denominator represents the square of the difference between these two proportions.

Example

A trial is designed to test whether a new surgical dressing reduces the rate of post-operative infection compared to the standard dressing. Historical data suggests the infection rate with the standard dressing is 20% (0.20). The researchers hope the new dressing will reduce this rate to 10% (0.10). The parameters are set with a significance level of 0.05 and a power of 80%.

Placing these into the formula: $n = \frac{(0.2 \times 0.8)+(0.1 \times 0.9)}{(0.2-0.1)^2} \times (1.96+0.84)^2$

$n = 1.96 / 0.01$

$n = 196$

Therefore, the study requires 196 participants per group, for a total sample size of 392, to detect a 10% reduction in infection rates with 80% power.

e) Non-inferiority trials

A non-inferiority trial aims to demonstrate that a new treatment is not worse than an active control by more than a pre-specified amount, known as the non-inferiority margin (Δ or δ). Unlike superiority trials, these studies typically employ a one-sided hypothesis test. Consequently, the calculation uses Zα (one-sided) rather than $Z_{(\alpha/2)}$ (two-sided). The statistical power (1-β) remains standard, typically 80% or 90%. The critical driver of sample size in these trials is the margin; a smaller, stricter margin requires a significantly larger sample size.

Continuous variables

For continuous outcomes, the sample size is calculated to ensure that the lower limit of the confidence interval for the difference between means does not cross the non-inferiority margin. The calculation generally assumes that there is truly no difference between the groups (i.e., the mean difference is 0), and the study is powered to prove that the difference is not worse than -Δ.

The formula for the sample size per group (n):

$$n = \frac{2(Z_\alpha + Z_\beta)^2\sigma^2}{\Delta^2}$$

Where:

Zα: is the critical value for a one-sided significance level (usually 0.025 for a 97.5% confidence interval, which equates to 1.96)

Z_β: is the critical value for power (0.84 for 80% power).

σ: is the standard deviation (assumed to be the same for both groups).

Δ: is the non-inferiority margin (the maximum acceptable difference).

Example

A researcher wishes to test a new, cheaper analgesic (painkiller) against a standard drug. The outcome is pain relief measured on a continuous scale of 0–100. The standard deviation (σ) in this population is known to be 10 points. The clinician decides that the new drug is non-inferior if it is no more than 3 points worse than the standard drug (Δ=3).

So, the sample size $n = \frac{2(1.96+0.84)^2(10)^2}{3^2}$

= 174.22

Rounding up, the study requires 175 participants per group (350 total) to prove non-inferiority within a 3-point margin.

Categorical variables

For binary outcomes, the researcher looks at the difference in proportions. The goal is to show that the event rate in the intervention group is not higher (or lower, depending on whether the event is good or bad) than the control group by more than the margin (δ).

The formula for the sample size per group (n) is:

$$n = \frac{(Z_\alpha + Z_\beta)^2[p_1(1-p_1) + p_2(1-p_2)]}{\delta^2}$$

Where,

p_1 and p_2 are the expected proportions in the standard and new treatment groups (often assumed to be equal, i.e., $p_1=p_2$, if the new drug is expected to perform similarly to the standard). δ (delta) is the non-inferiority margin (absolute difference in proportions).

Example

A trial compares a new short-course antibiotic to a standard long-course antibiotic for treating pneumonia. The cure rate for the standard treatment is 90% (0.90). The researchers assume the new treatment has the same cure rate (0.90). They define the non-inferiority margin as 5% (0.05); meaning, they will accept the new drug if its cure

rate is not lower than 85%.

So, the sample size

$$n = \frac{(1.96 + 0.84)^2[0.90(0.10) + 0.90(0.10)]}{0.05^2}$$

=564.48

Rounding up, the study requires 565 participants per group (1,130 total).

f) Equivalence trials

An equivalence trial is designed to confirm that two treatments are therapeutically indistinguishable. The objective is to show that the difference between the new treatment and the standard treatment lies entirely within a specific, pre-defined range known as the equivalence margins (-δ to +δ). Because the study must rule out both the possibility that the new drug is inferior and the possibility that it is superior, equivalence trials essentially involve two simultaneous hypothesis tests (two one-sided tests, or TOST). Therefore, the sample size calculation is generally based on the sum of the critical values for the significance level and power, similar to superiority trials, but it requires a larger sample size because the equivalence margins are typically narrower than the effect sizes used in superiority studies.

Continuous variables

For continuous outcomes, the calculation ensures that the confidence interval for the mean difference falls strictly within the equivalence margins. The formula assumes there is no true difference between the means ($\mu_1 - \mu_2 = 0$).

The formula for the sample size per group (n) is:

$$n = \frac{2(Z_\alpha + Z_{\beta/2})^2\sigma^2}{\delta^2}$$

Where,

Z_α is the critical value for the significance level (typically 1.96 for a 5% significance level, applied as two one-sided tests at 2.5%).

$Z_{\beta/2}$ is the critical value for power. Note that for equivalence trials, specific adjustments to power calculations sometimes require careful consideration, but frequently the standard Z_β (0.84 for 80% power) is used in simplified formulas. However, strictly speaking, because two tests are being performed, the power calculation is more complex. A commonly used approximation for the total power uses $(Z_\alpha + Z_{\beta/2})^2$. Here,

σ = is the standard deviation.

δ = (delta) is the equivalence margin.

Example

A pharmaceutical company wants to show that a generic blood pressure medication is bioequivalent to the brand-name drug. The outcome is the reduction in systolic blood pressure. The standard deviation (σ) is 12 mmHg. The equivalence margin (δ) is set at 5 mmHg. The significance level is 5% (Z_α=1.96) and power is 80% (Z_β≈0.84), or more precisely calculated depending on the specific TOST power curve, but 0.84 is often used for approximation).

$$n = \frac{2(1.96+0.84)^2(12)^2}{5^2} = 90.31$$

Rounding up, the study requires 91 participants per group (182 total).

Categorical variables

For binary outcomes, the study aims to show that the difference in proportions between the two groups is within the equivalence limits. The formula for the sample size per group (n) is:

$$n = \frac{(Z_\alpha + Z_\beta)^2[p_1(1 - p_1) + p_2(1 - p_2)]}{\delta^2}$$

Where,

p_1 and p_2 are the expected proportions (usually assumed to be equal, so $p_1=p_2=p$)

δ is the equivalence margin (the acceptable difference in proportions).

Example

A study compares two types of sutures to ensure they have equivalent wound closure rates. The standard suture has a success rate of 95% (p=0.95). The researcher defines equivalence as the rates being within 3% of each other (δ=0.03).

So, sample size

$$n = \frac{(1.96 + 0.84)^2[0.95(0.05) + 0.95(0.05)]}{0.03^2}$$

= 827.55

Rounding up, the study requires 828 participants per group (1,656 total).

OTHER FORMULAS OF SAMPLE SIZE CALCULATION

In these instances, small letters represent the following:

n = Sample size

σ = Standard deviation

d = Required size of standard error

r = Rate

p = Proportion/percentage

a) *Formula for estimation of sample size from a single mean:* $n = \frac{z^2\sigma^2}{d^2}$

b) Formula for estimation of sample size from a single proportion: $n = \frac{z^2\, p\,(1-p)}{d^2}$

c) *Formula for estimation of sample size for difference between two means:* $n = \frac{z^2\,(\sigma_1{}^2+\sigma_2{}^2)}{d^2}$, in each group

d) *Formula for estimation of sample size for difference between two rates:* $n = \frac{z^2\,(r_1+r_2)}{d^2}$, in each group

e) *Formula for estimation of sample size for difference between two proportions:* $n = \frac{z^2\,[P_1\,(1-P_1)+P_2\,(1-P_2)]}{d^2}$, in each group.

CORRECTION FOR FINITE POPULATION

When the sample size is ≥10% of the total population, then we can use a finite population correction factor (FPC). The extra precision we obtain when the sample size approaches the population size is measured by the finite population correction factor. The Fpc formula is:

Fpc = $\sqrt{\frac{N-n}{N-1}}$. Where, *N* is the size of the population and *n* is the size of the sample.

For example, a researcher might want to estimate the proportion of patients with a particular medical condition in a hospital population. Suppose that the hospital has a population of 10,000 patients and the researcher wants to estimate the proportion of patients with the condition to within a margin of error of 3% with 95% confidence. The formula for the sample size needed for an infinite population is:

$n = \frac{Z^2pq}{d^2}$

Here:

Z = the z-score corresponding to the desired level of confidence (1.96 for 95% confidence)

p = the estimated proportion of patients with the condition (0.5 for a conservative estimate)

q=1-p

d = the desired margin of error (0.05 in this case).

So, the sample size needed for an infinite population is, $n_0 = \frac{(1.96)^2\times 0.5\times 0.5}{(0.05)^2} = 384$

However, since the hospital population is finite, the sample size needs to be adjusted. The correction factor for a finite population is:

Fpc = $\sqrt{\frac{10000-384}{10000-1}} = 0.98$

So, the adjusted sample size is, $n = \frac{384}{0.98} = 391.83 \approx 392$

We can also use the following formula to correct the sample size in the case of a finite population:

New sample, $n = \frac{n_o}{1+\frac{n_o-1}{N}}$

Here, N = population size,

n = calculated sample size,

n_0 = calculated sample size of the finite population before correction.

USE OF STATISTICAL SOFTWARE AND ONLINE CALCULATORS FOR SAMPLE SIZE CALCULATION

The landscape of research methodology has shifted dramatically from the manual application of algebraic formulas to the utilisation of sophisticated digital tools. While understanding the mathematical derivation of sample size formulas is essential for grasping the theoretical underpinnings of research design, the practical execution of these calculations is rarely performed by hand in modern scientific inquiry. Manual calculation is not only time-consuming but also prone to human error, particularly when dealing with complex study designs such as cluster-randomised trials, time-to-event analyses, or non-inferiority studies. Consequently, the use of statistical software and online calculators has become the standard practice for ensuring accuracy, efficiency, and reproducibility in research planning.

DEDICATED STATISTICAL SOFTWARE FOR SAMPLE SIZE CALCULATION

In contemporary research, the manual calculation of

sample size is increasingly being replaced by the use of dedicated statistical software. These specialized programmes are designed to navigate the complex mathematical underpinnings of power analysis, providing researchers with accurate and reproducible estimations that are essential for ethical and scientific validity. Unlike general spreadsheet applications, dedicated sample size software contains validated algorithms for hundreds of different study designs, ranging from simple descriptive surveys to complex, multi-stage clinical trials. This ensures that the researcher selects the correct parameters for their specific design, reducing the risk of underpowered studies.

*G*Power*

GPower is a free, standalone software package developed by experimental psychologists at the University of Düsseldorf. It is perhaps the most widely used tool for sample size calculation in the social, behavioural, and biomedical sciences because it is free and covers a broad range of common statistical tests, including t-tests, F-tests, and chi-square tests. A defining feature of GPower is its ability to perform five different types of power analysis. The most common is 'a priori' analysis, where the researcher inputs the desired power and effect size to calculate the required sample size. However, it also performs 'post-hoc' analysis, calculating the achieved power of a study after it is completed, and 'compromise' analysis, which is useful when resources are strictly limited. Its graphical interface allows researchers to plot power curves easily, visually demonstrating how increasing the sample size yields diminishing returns in power. Because it is standalone, it does not perform data analysis, but its ease of use makes it a favourite for students and early-career researchers.

IBM SPSS Statistics

For decades, SPSS has been the dominant statistical package in the social sciences due to its user-friendly, point-and-click interface. Historically, SPSS required a separate add-on module called 'SamplePower' for these calculations. However, recent versions (version 25 and onwards) have integrated a dedicated power analysis module directly into the main software. This integration allows researchers to stay within a familiar environment. SPSS is particularly strong in calculating power for means (t-tests, ANOVA) and proportions. While it may lack the exhaustive library of study designs found in PASS, its menu-driven approach removes the need for programming skills, making power analysis accessible to clinicians and social scientists who may be intimidated by code-based software.

STATA is widely favoured in epidemiology, economics, and public health for its precision and reproducibility. Stata handles sample size calculations through its dedicated power command suite. Unlike the rigid menus of SPSS, STATA allows users to write scripts (do-files) for their calculations. This is a significant advantage for scientific rigour because the exact code used to calculate the sample size can be saved, shared, and included in the research protocol, providing a transparent audit trail. STATA is highly flexible; researchers can easily adjust parameters for cluster-randomised trials, stratified designs, and survival analysis. Furthermore, it excels at producing publication-quality graphs that display how power varies across different alpha levels or effect sizes, which is often required for grant applications.

R is a free, open-source programming language and software environment for statistical computing. It represents the pinnacle of flexibility but comes with the steepest learning curve. Unlike the other tools, R does not have a single 'button' for sample size; instead, it relies on community-developed packages such as pwr or TrialSize. Because R is a programming

language, it allows researchers to go beyond standard formulas. For novel or complex study designs where no standard formula exists, researchers can write custom code to perform 'Monte Carlo simulations'. This involves simulating a clinical trial thousands of times virtually to estimate the required sample size empirically. While this requires advanced coding skills, it provides a solution for almost any experimental design imaginable, making R the tool of choice for statisticians developing new methodologies.

Epi Info

Epi Info is a suite of software tools designed by the Centers for Disease Control and Prevention (CDC) specifically for the global public health community. Within this suite, the 'StatCalc' module serves as a highly accessible tool for calculating sample sizes for common epidemiological designs. It is particularly favored in field epidemiology and public health surveillance because it is free to use and does not require advanced statistical training to operate. StatCalc allows researchers to quickly estimate the number of participants needed for population surveys, cohort studies, and case-control studies.

For example, in a descriptive study, the user simply inputs the estimated population size, the expected frequency of the factor under study, and the desired margin of error. The software then instantly provides the sample size required for various confidence levels. For analytical studies, such as unmatched case-control studies, it allows the user to input the desired power, the ratio of controls to cases, and the expected exposure rates. A distinct advantage of Epi Info is its integration with data entry and analysis modules, allowing a researcher to use the same platform from the planning phase through to the analysis of the final data. Its simplicity and accessibility make it the standard choice for students, government health workers, and researchers in resource-limited settings.

PASS (Power Analysis and Sample Size)

On the other end of the spectrum lies PASS, a commercial software package developed by NCSS. PASS is widely considered the industry standard for planning clinical trials and complex pharmaceutical research. Unlike Epi Info, which focuses on core epidemiological designs, PASS is exhaustive, offering over 1,100 different scenarios for sample size calculation. It covers highly specialized designs including group-sequential trials, equivalence and non-inferiority trials, and designs using generalised linear mixed models.

The primary strength of PASS is its rigour and documentation. For every calculation routine, the software provides the exact mathematical formulas used and the primary literature references that validate those formulas. This feature is indispensable when writing research protocols for regulatory bodies, such as the FDA or local ethics committees, which often require a detailed justification of the sample size methodology. Furthermore, PASS allows researchers to generate power curves and reports that visually demonstrate how sensitive the sample size is to changes in the input parameters, such as the standard deviation or the effect size. While it requires a paid license, its depth makes it the preferred tool for biostatisticians and academic researchers conducting high-stakes experimental studies.

ONLINE CALCULATORS

For researchers who require a quick estimate for standard study designs, particularly descriptive cross-sectional surveys or simple comparisons of proportions, web-based online calculators offer a convenient alternative. Tools such as OpenEpi, Raosoft, and ClinCalc are accessible via any web browser and do not require software installation. These calculators typically feature user-friendly interfaces where the researcher inputs key parameters—such as the population size, expected prevalence, margin of error, and confidence level—

and the calculator instantly generates the required sample size.

OpenEpi is particularly popular in the public health and epidemiology sectors. It provides calculators for a wide range of epidemiological statistics, including sample sizes for unmatched case-control studies, cohort studies, and cross-sectional surveys. These web tools are excellent for teaching purposes and for quick checks during the initial concept phase of a proposal. However, researchers must exercise caution when using online calculators. Unlike validated software packages, the underlying algorithms of some web tools may not be visible or fully documented. It is crucial to verify that the calculator uses the correct formula for the specific study design being proposed. For example, a calculator designed for a simple random sample will yield an incorrect result if applied to a cluster-sampling design without applying a design effect correction.

THE BLACK BOX PHENOMENON AND PROPER REPORTING

A significant risk associated with the reliance on digital tools is the 'black box' phenomenon. This occurs when a researcher simply plugs numbers into a software interface without understanding the meaning of the inputs or the logic of the output. Software is incapable of judging whether the input parameters are clinically realistic; it merely performs the mathematical operation requested. If a researcher inputs an inflated effect size to reduce the required sample size, the software will output a number that is mathematically correct but scientifically flawed. Therefore, the use of software does not remove the need for statistical literacy. The researcher must still possess a firm grasp of concepts such as alpha, beta, standard deviation, and effect size to use these tools effectively.

Furthermore, the use of software entails a responsibility for transparency and reproducibility. When writing a research protocol or a published manuscript, it is insufficient to simply state that 'sample size was calculated statistically'. The standard of academic reporting requires the author to specify exactly which software or calculator was used, including the version number. Additionally, all the input parameters used in the calculation—such as the anticipated effect size, the source of the variance estimate, the alpha level, and the target power—must be listed. This transparency allows peer reviewers and other scientists to replicate the calculation to verify the validity of the study design.

ADJUSTING FOR DROPOUTS AND NON-RESPONSE

Adjustment for dropouts and non-response is an essential component of practical sample size planning. Theoretical sample size formulas assume complete data from all selected study units. In real-world research, this assumption rarely holds. Some individuals decline participation, withdraw after enrolment, are lost to follow-up, or provide incomplete or unusable data. Failure to account for such losses can result in an achieved sample size that is smaller than required, leading to reduced precision, inadequate statistical power, and potentially biased estimates.

Dropouts and non-response represent distinct but related phenomena. Non-response typically refers to failure to participate at the outset, while dropouts occur after initial inclusion, particularly in longitudinal or interventional studies. Both reduce the effective sample size and must be anticipated during the design phase rather than addressed retrospectively.

Implications for precision and power

The primary statistical consequence of dropouts and non-response is loss of information. In descriptive studies, this results in wider confidence intervals and reduced precision of estimates. In analytical and experimental studies, it reduces power to detect true effects, increasing the risk of false-negative findings. When losses are differential across exposure or outcome groups, they may also introduce systematic bias, undermining internal validity.

From a methodological perspective, it is insufficient

to rely solely on analytical adjustments at the analysis stage. While techniques such as weighting, imputation, or sensitivity analysis can partially address missing data, they do not fully compensate for an inadequately sized sample. Therefore, proactive inflation of sample size at the design stage remains the most robust approach.

ANTICIPATED NON-RESPONSE

Anticipated non-response or dropout rate refers to the proportion of selected or enrolled participants expected not to contribute complete data. Estimation of this rate should be evidence-based rather than arbitrary. Common sources include prior studies conducted in similar settings, pilot studies, registry data, or operational experience from comparable surveys or trials.

Rates of non-response vary by study design, population, data collection method, and study duration. Cross-sectional household surveys may experience moderate non-response at recruitment, while cohort studies and clinical trials often face cumulative loss to follow-up over time. Studies involving sensitive topics or invasive procedures may encounter higher refusal rates.

It is important to distinguish between total non-response and analytically relevant loss. Some participants may partially respond, but fail to provide data on key variables. Only losses that affect the primary outcome or exposure should be considered when inflating the sample size.

METHODS OF SAMPLE SIZE INFLATION

The most common approach to adjusting for non-response is simple sample size inflation. After calculating the required analytical sample size, the researcher increases this number based on the anticipated non-response proportion. The adjusted sample size is obtained by dividing the required sample size by the expected response proportion.

For example, if a study requires 400 completed observations and a non-response rate of 20 percent is anticipated, the adjusted sample size is calculated as 400 divided by 0.80, resulting in 500 individuals to be approached or enrolled. This method assumes that non-response is random and independent of the outcome of interest.

In longitudinal studies, adjustment may need to account for cumulative attrition across multiple follow-up points. In such cases, anticipated dropout rates at each stage should be combined to estimate overall retention. Failure to account for cumulative loss can substantially reduce the final sample size.

DIFFERENTIAL NON-RESPONSE AND STRATIFIED ADJUSTMENT

In many studies, non-response is not uniform across subgroups. Certain strata defined by age, sex, socioeconomic status, or geographic location may experience higher refusal or dropout rates. If the study design involves stratified sampling or subgroup analysis, uniform inflation may be inadequate.

In such situations, differential inflation by stratum is recommended. Sample size requirements are calculated separately for key strata, and each is inflated according to its expected non-response rate. This approach ensures adequate representation and precision within subgroups, particularly when comparisons across strata are central to the research objectives.

Differential adjustment is especially important when the primary outcome is rare or concentrated in specific subpopulations. Failure to account for subgroup-specific non-response may compromise both internal and external validity.

DROPOUTS IN INTERVENTIONAL AND LONGITUDINAL STUDIES

In interventional and cohort studies, dropouts pose additional challenges because loss may be related to treatment allocation, outcome status, or adverse events. Such losses are often non-random and can bias effect estimates.

Sample size adjustment in these studies should consider not only the expected proportion of dropouts, but also their potential impact on effect size

estimation. Conservative assumptions regarding attrition are generally preferred, particularly when follow-up periods are long or participant burden is high.

Retention strategies, such as frequent contact, incentives, and flexible follow-up procedures, should be considered alongside sample size inflation. Ethical and logistical feasibility of enrolling additional participants must be balanced against the risk of excessive attrition.

RELATIONSHIP WITH MISSING DATA MECHANISMS

Adjustment for non-response at the design stage does not eliminate the need to consider missing data mechanisms during analysis. Missing data may be missing completely at random, missing at random, or missing not at random. Sample size inflation primarily addresses loss of precision due to reduced sample size, not bias arising from systematic missingness.

Researchers should therefore align design-stage adjustments with planned analytical strategies. Studies expecting substantial non-response should predefine approaches such as weighting, multiple imputation, or sensitivity analyses. Transparent reporting of both anticipated and observed non-response strengthens methodological credibility.

Practical limits and over-inflation

While adjustment for non-response is essential, excessive inflation may be impractical or unethical. Recruiting substantially more participants than necessary increases cost, field burden, and participant exposure without proportional scientific benefit. Over-inflation may also complicate logistics and data management.

Researchers should justify their chosen non-response rate and demonstrate that inflation is proportionate and evidence-based. Where uncertainty exists, conservative but reasonable assumptions are preferable to arbitrary large adjustments.

In some cases, it may be more efficient to invest resources in improving response and retention rather than substantially increasing the initial sample size. Design decisions should therefore integrate statistical reasoning with operational planning.

REPORTING AND TRANSPARENCY

All assumptions regarding anticipated non-response and dropout should be explicitly reported in study protocols and final publications. This includes the source of the assumed rate, the method of adjustment, and the final target sample size. Clear documentation allows readers and reviewers to assess whether the achieved sample size was adequate and whether deviations from planned recruitment may have affected results.

CONCLUSION

Sampling is a fundamental component of research that enables researchers to obtain accurate and reliable estimates of population parameters. The choice of sampling technique depends on various factors, including the research question, the population of interest, and the available resources. Good sampling requires careful consideration of the sampling frame, appropriate sampling methods, and the avoidance of sampling biases and errors. Additionally, the use of reliable and valid measurement instruments, proper training of data collectors, and consideration of ethical concerns are also essential for the success of a sampling study. Hence, sampling is an important tool for researchers to obtain representative and Generalisable findings, and it plays a critical role in the advancement of scientific knowledge across a range of fields.