Calculating sample size and power

Dr. R. Ramakrishnan
MAE, PhD

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Objectives

- Understand the relationship between sample size and power
- Determine sample size necessary to achieve a given level of power for estimating a simple proportion, and other measures of effect

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Steps in Estimating Sample Size

- Identify major study variable
- Determine type of estimate (%, mean, ratio,...)
- Indicate expected frequency of factor of interest
- Decide on desired precision of the estimate
- Decide on acceptable risk that estimate will fall outside its real population value
- Adjust for population size
- Adjust for estimated design effect
- Adjust for expected response rate

α and Confidence Level

- α: The significance level of a test: the probability of rejecting the null hypothesis when it is true (or the probability of making a Type I error).
- Confidence level: The probability that an estimate of a population parameter is within certain specified limits of the true value; commonly denoted by "1- α".

HEALTH RESEARCH FUNDAMENTALS
nie.gov.in

β and Power

- β: The probability of failing to reject the null hypothesis when it is false (or the probability of making a Type II error).
- Power: The probability of correctly rejecting the null hypothesis when it is false; commonly denoted by “1- β”

Precision

A measure of how close an estimate is to the true value of a population parameter. It may be expressed in absolute terms or relative to the estimate.

Sample Size Required for Estimating Population Mean

- Suppose we want an interval that extends *d* units on either side of the estimator

  d = (reliability coefficient) x (Standard error)

- If sampling is from a population sufficiently large size, the equation is:

$$d = z \frac{\sigma}{\sqrt{n}}$$

- When solved for *n* gives:

$$n = \frac{z^2 \sigma^2}{d^2}$$

Example 1 (1/2)
What Sample Size Do I Need If . . . ?

A health department nutritionist , wishing to conduct a survey among a population of teenage girls to determine the average daily protein intake

What information is needed to estimate the sample size?

- The nutritionist must provide three items of information: the desired width of the confidence interval, the level of confidence desired, and the magnitude of the population variance

Example 1 (2/2) What Sample Size Do I Need If . . . ?

- Solution: The nutritionist would like an interval about 10 units wide; that is, the estimate should be within about 5 units of the true value in either direction. A confidence coefficient of .95 is decided and on that, from past experience, the nutritionist feels that the population standard deviation is probably about 20 grams.
- Summarizing the information: $z = 1.96$, $\sigma = 20$, and $d = 5$
- Calculation:

$$n = \frac{(1.96)^2\,(20)^2}{(5)^2} = 61.47$$

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

A note on Population Standard Deviation σ

- The formulas for sample size require knowledge of σ^2. However, in general, the population variance is unknown and has to be estimated:
  - A pilot or preliminary sample. Observations used in the pilot can be counted as part of the final sample
  - Estimates may be available from previous studies
  - If thought that the population is approximately normally distributed, we may use the fact that the range (R) is approximately equal to 6 standard deviations.

$$\sigma \cong R/6$$

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Sample Size Required for Estimating Proportions

- The formula requires the knowledge of *p*, the proportion in the population possessing the characteristic of interest. However, this is what we are trying to estimate and is unknown
  - A pilot or preliminary sample. Observations used in the pilot study can be counted as part of the final sample
  - Estimates may be available from previous studies and the upper bound of *p* can be used in the formula
  - If impossible to come with a better estimate, set $p = 0.5$ in the formula to yield the maximum value of n

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Sample Size Required for Estimating Proportions

The method is essentially the same as for population mean. Assuming random sampling and approximate normality in the distribution of *p*, brings us to the formula for n if sampling is with replacement, from a population sufficiently large to warrant ignoring the finite population correction :

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

Where $q = 1 - p$

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Example 2 (1/3)
What Sample Size Do I Need If . . . ?

- I want to estimate the true immunization coverage in a community of school children
- Previous studies tell us that immunization coverage should be somewhere around 80%
- Precision (absolute): we'd like the result to be within 4% of the true value
- Confidence level: conventional = 95% = 1 - a; therefore, a = 0.05 and $z_{(1-a/2)}$ = 1.96 = value of the standard normal distribution corresponding to a significance level of 0.05 (1.96 for a 2-sided test at the 0.05 level)

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Example 2 (2/3)

- d = absolute precision = 0.04
- p = expected proportion in the population = 0.80
- $z_{(1-a/2)}$ = 1.96 = value of the standard normal distribution corresponding to a significance level of a (1.96 for a 2-sided test at the 0.05 level)

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Example 2 (3/3)
Sample Size

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

$$= \frac{(1.96)^2 (.80) (.20)}{(0.04)^2}$$

$$= 384$$

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Design Effect

- A bias in the variance introduced in the sampling design, by selecting subjects whose results are not independent from each other; relative change (increase) in the variance due to the use of clusters.
- The design effect can be calculated after study completion, but should be accounted for at the design stage.
  - The design effect is 1 (i.e., no design effect) when taking a simple random sample.
  - The design effect varies using cluster sampling; it is usually estimated that the design effect is 2 in immunization cluster surveys.

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

What You Need to Calculate Sample Size for Analytical Studies

- Desired values for the probabilities of a and b
- The proportion of the baseline (controls or non-exposed) population
  - EXPOSED (for case-control studies), or
  - DISEASED (for cohort/intervention studies)
  - Often based on previous studies or reports
- Magnitude of the expected effect (RR, OR)
  - Often based on previous studies or reports
  - Minimum effect that investigator considers worth detecting
- Formula: different formulae depending on study design, research question, and type of data

Example 3 (1/3) What Sample Size Do I Need If…?

- Cohort study of oral contraceptive (OC) use in relation to risk of MI among women of childbearing age
- Previous studies
  - Proportion of non-OC users who are at risk of disease = 0.15
  - Proportion of OC-users who are at risk of disease = 0.25
- Conventional a = 0.05 (two-sided)
- Conventional b = 0.20 (80% power to detect a difference if one truly exists)
- Assume equal sample sizes ($n_1 = n_2$)

HEALTH RESEARCH FUNDAMENTALS
nie.gov.in

Example 3 (2/3)

- p_0 = proportion of non-OC users who are diseased = 0.15
- p_1 = proportion of OC-users who are diseased = 0.25
- $q_0 = (1-p_0) = 1.0 - 0.15 = 0.85$
- $q_1 = (1-p_1) = 1.0 - 0.25 = 0.75$
- $z_{(1-a/2)} = 1.96$ = value of the standard normal distribution corresponding to a significance level of α (1.96 for a 2-sided test at the 0.05 level)
- $z_{(1-b)} = 0.84$ = value of the standard normal distribution corresponding to the desired level of power (0.84 for a power of 80%)

Example 3 (3/3)

HEALTH RESEARCH FUNDAMENTALS

$$\textbf{n (each group)} = \frac{(p_0q_0 + p_1q_1)(z_{1-a/2} + z_{1-b})^2}{(p_1-p_0)^2}$$

$$= \frac{[(.15)(.85) + (.25)(.75)][1.96 + 0.84]^2}{(0.25 - 0.15)^2}$$

$$= \frac{(0.315)(7.84)}{0.01} = 246.96$$

Therefore: 247 OC users (and 247 non-OC users)

nie.gov.in

Example 4 (1/3)
What Size Sample Do I Need If . . . ?

- Case-control study of oral contraceptive (OC) use in relation to risk of MI among women of childbearing age
- Previous studies: 10% of women use OCs
- OR of MI associated with current OC use = 1.8
- Conventional a = 0.05 (two-sided)
- Conventional b = 0.20 (80% power to detect difference if one truly exists)
- Assume equal sample sizes ($n_1=n_2$)

Example 4 (2/3)

- p_0 = proportion of controls who are current OC users = 0.10
- p_1 = proportion of cases who are current OC users = 0.18
- $q_0 = (1-p_0) = 1.0- 0.10 = 0.90$
- $q_1 = (1-p_1) = 1.0 - 0.18 = 0.82$
- $z_{(1-a/2)}$ = 1.96 = value of the standard normal distribution corresponding to a significance level of a (1.96 for a 2-sided test at the 0.05 level)
- $z_{(1-b)}$ = 0.84 = value of the standard normal distribution corresponding to the desired level of power (80%)

Example 4 (3/3)

$$n\text{ (each group)} = \frac{(p_0q_0 + p_1q_1)(z_{1-a/2} + z_{1-b})^2}{(p_1-p_0)^2}$$

$$\frac{[(.10)(.90) + (.18)(.82)][1.96 + 0.84]^2}{(0.18 - 0.10)^2}$$

$$\frac{(0.2376)(7.84)}{0.0064} = 291.06$$

Therefore: 291 cases and 291 controls

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Sample Sizes: Case-Control Study of OC Use and MI

OR	Required sample sizes
1.2	3834
1.3	1769
1.5	682
1.8	291
2.0	196
2.5	97
3.0	59

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

The 10% Rule

- Note that sample-size estimates should be interpreted as providing merely a MINIMUM estimate of the sample sizes necessary for the study
- The formula takes into account only the overall crude association between exposure & disease; i.e., no confounders are considered
- 10% rule: increase the sample size 10% for each confounder/variable added

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

SAMPLE SIZE : Free Soft wares for Sample Size

OpenEpi
Supported by Centers for Disease Control and Prevention, Atlanta
www.openepi.com

PS: Power and Sample Size Calculation by Department of Bio statistics Vanderbilt University
http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Thank you

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in