

Sample size estimation

Presented by :Dr. Reshma.S
Moderator : Dr. Subodh S Gupta

Framework

- Why do we need sample size calculation?
- When should we calculate sample size?
- Basic principles for sample size calculation
- Derivation of sample size formula
- Description of some commonly used terms
- Practical issues in calculating sample size
- Procedure for calculating sample size
- Formulae for calculation of sample size in different study design



Why do we need sample size calculation?

- Sample Size (n) is the number of individuals in a group under study
- If the sample size is too small, even the most rigorously executed study may fail to answer its research question, may **fail to detect important effects or associations**, or may estimate those effects or associations too **imprecisely**
- If the sample size is too large, the study will be more difficult and **costly**, and may even **lead to a loss in accuracy**, as it is often difficult to maintain high data quality.
- Hence, it is necessary to estimate the optimum sample size for each individual study

When should we calculate sample size?

- Sample size can be addressed at two stages of the actual conduct of the study
- Firstly, calculate the optimum sample size required during the **planning stage**, while **designing the study**, using appropriate approaches and some information on parameters.
- Secondly, sample size can also be calculated in between the study. For example in rare case study sample size can be re evaluated after mid-term evaluation

Basic principles for sample size calculation

- Sample size primarily depend on:
 - Availability of resource
 - Proposed plan of analysis
- Sample size estimation requires:
 - Estimate of the variable of interest (e.g. mean, proportion, OR, RR)
 - Desired precision
- Select the confidence level for the interval (e.g. 95% or 99%) and power of the study
- State the null hypothesis and alternative hypothesis.
- Loss to follow up

Derivation of sample size formula:

▶ Standard error, $SE = \frac{SD}{\sqrt{n}}$

▶ $n = \frac{SD^2}{SE^2}$ -----(1)

if 'd' is the unit on either side of point estimate, then,

$$d = Z (1-\alpha) * SE \text{ of mean}$$

$$\text{Then , } SE = \frac{d}{Z(1-\alpha)}$$

Now, putting the value of SE in (1), we have

$$n = \frac{Z^2(1-\alpha)*SD^2}{d^2}$$



Description of some commonly used terms:

- Random error
- Systematic error (bias)
- Precision (Reliability)
- Null hypothesis
- Alternative hypothesis
- Type I error (α)
- Type II error (β)
- Hypothesis Testing
- Power of the study ($1 - \beta$)
- Design effect

Random error

- It describes the role of chance
- Sources of random error include:
 - sampling variability
 - subject to subject differences
 - measurement errors
- It can be controlled and reduced to acceptably low levels by:
 - Increasing the sample size
 - Repeating the experiment


Systematic error (bias)

- It describes deviations that are not a consequence of chance alone
- Several factors including **patient selection criteria** might contribute to it.
- These factors may not be amenable to measurement
- Removed or reduced by good design and conduct of the experiment
- A strong bias can yield an estimate very far from the true value

Precision (Reliability)

- Degree to which a variable has the same value when measured several times
- It is a measure of consistency
- It is a function of :
 - random error
(the greater the error, the less precise the measurement)
 - sample size
 - confidence interval required &
- A larger sample size would give precise estimates

Accuracy (Validity)

- It indicates the degree to which the variable actually represents what it is supposed to represent
 - It is a function of systematic error
 - The greater the error the less accurate the variable
- 

Null hypothesis

- Null hypothesis is a hypothesis which states that there is no difference among groups or that there is no association between the predictor and the outcome variables
- This hypothesis needs to be tested

Alternative hypothesis

- It assumes that there is a difference among the groups or there exists an association between the predictor and outcome variable
- There are two types of alternative hypothesis:
 - one-tailed (one-sided) hypothesis &
 - two-tailed (two-sided) hypothesis
- One-tailed hypothesis specifies the difference (or effect or association) in one direction only.
- Two-tailed hypothesis specifies the difference (or effect or association) in either direction.

Hypothesis testing

		Truth In the population	
		H_0 (False)	H_0 (True)
Results in the study	Reject hypothesis	Correct	Type I error
	Accept hypothesis	Type II error	correct

Type I error

Type 1 error

- ▶ Rejecting a null hypothesis actually true in the population
- ▶ probability of erroneously finding a disease exposure association, when none exists in reality.

Type II error

Type 2 error

- ▶ Fails to reject a null hypothesis that is actually false in the population
- ▶ probability of not erroneously finding disease exposure association, when it exists in reality.

Power ($1 - \beta$)

- This is the probability that the test will correctly identify a significant difference or effect or association in the sample that exist in the population.
- The larger the sample size, the study will have greater power to detect significance of difference or effect or association


Design effect

- It is the ratio of the variance when other sampling method is used other than SRS to the variance when simple random sampling is used
- The sample sizes for simple random samples are multiplied by the design effect to obtain the sample size for the clustered sample

Practical issues in calculating sample size

- **Multiple outcomes:** The usual approach is sample size calculation on primary outcome. Alternate approach is to make calculation for each outcome and then to use largest size for planning study
- **Dropout :**subject who are enrolled but in whom outcome status cannot be ascertained do not count in the sample size.
Anticipating the dropout rate sample size can be calculated.
- **Number of sub-groups to analyze:** If multiple sub-groups in a population are going to be analyzed, the sample size should be increased to ensure that adequate numbers are obtained for each sub-group

Procedure for calculating sample size:

- Use of formula
 - Readymade tables
 - Nomograms
 - Computer software
- 

Ready made table:

Table 3: Estimating an incidence rate with specified relative precision [Formula: $n = (Z_{1-\alpha/2} / \epsilon)^2$]


Relative precision (ϵ)	Confidence level		
	99%	95%	90%
0.01	66358	38417	27061
0.02	16590	9605	6766
0.03	7374	4269	3007
0.04	4148	2402	1692
0.05	2655	1537	1083
0.06	1844	1068	752
0.07	1355	785	553
0.08	1037	601	423
0.09	820	475	335
0.10	664	385	271
0.12	461	267	188
0.14	339	197	139
0.16	260	151	106
0.18	205	119	84
0.20	166	97	68
0.22	138	80	56
0.24	116	67	47
0.26	99	57	41
0.28	85	50	35
0.30	74	43	31
0.32	65	38	27
0.34	58	34	24
0.36	52	30	21
0.38	46	27	19
0.40	42	25	17
0.42	38	22	16
0.44	35	20	14
0.46	32	19	13
0.48	29	17	12
0.50	27	16	11

Computer software for estimating sample size

Software

- **PASS**
(www.ncss.com/pass.html)
- **nQUERY**
(www.statsolusa.com/nquery/nquery.htm)
- **EPI-INFO**
(www.cdc.gov/epiinfo)
- **EPIDAT**
(www.paho.org/English/SHA/epidat.htm)

Formulae for calculation of sample size in different study design :

- Cross sectional study
 - Case control study
 - Cohort study
 - Clinical trial
- 

Cross sectional study

Estimating the population proportion with specified absolute precision

➤ Required information:

- Population proportion – p
- Confidence level – $100(1-\alpha)\%$
- Absolute precision required on either side of proportion – d

If it is not possible to estimate p , the figure of 0.5 should be used since the sample size required largest when $p = 0.5$

$$\text{Formula: } n = Z^2_{1-\alpha/2} p(1-p)/d^2$$

Example :

- A local health department wishes to estimate the prevalence of tuberculosis among children under five years of age in its locality.
- How many children should be included in the sample, so that the prevalence may be estimated to within 5 percentage points of the true value with 95% confidence, if it is known that the true rate is unlikely to exceed 20% ?

Solution :

- Anticipated population proportion = 20 % ($p=0.20$)
- Confidence level = 95%
- Absolute precision (15 %-25 %) = 5 percentage points ($d=0.05$)

By using the above formula, we have

$$\begin{aligned}n &= 1.96^2 \times 0.2 \times (1 - 0.2) / (0.05)^2 \\ &= 245.86\end{aligned}$$

i.e. 246



Estimating a population proportion with specified relative precisions

Required information:

- Anticipated population proportion = P
- Confidence level = $100(1-\alpha)\%$
- Relative precision = ε

Formula:

$$n = Z_{1-\alpha/2}^2 (1-p) / \varepsilon^2 P$$

Example:

- An investigator working for the national program of immunization seeks to estimate the proportion of children in country who are receiving appropriate childhood vaccinations.
- How many children must be studied if the resulting estimate is to fall within 10% of true proportion with 95% confidence?
The vaccination coverage is not expected to be below 50%.

Solution :

- Anticipated population proportion = 50%
- Confidence level = 95%
- Relative precision (45%-55%) = 10% of 50% ($\epsilon = 0.10$)
- $$\mathbf{n = Z^2_{1-\alpha/2} (1-P) / \epsilon^2 P}$$
$$\mathbf{= 1.96^2 \times (1 - 0.5) / 0.10^2 \times 0.5}$$
$$\mathbf{= 384.16}$$

Example :

- Previous surveys have demonstrated that the usual prevalence of dental caries among school children in a particular community is about 25%.
- How many children should be included in a survey designed 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 the 5% level of significance?

Solution:

- Test caries rate = 25% ($P_o = 0.25$)
- Anticipated caries rate = 20% ($P_a = 0.20$)
- Level of significance = 5%
- Power of test = 90%
- Alternative hypothesis (one-sided test): caries rate < 25%

Substituting the value in the formula

$$\begin{aligned}n &= \{Z_{1-\alpha} \sqrt{[P_o(1-P_o)]} + Z_{1-\beta} \sqrt{[P_a(1-P_a)]}\}^2 / (P_o - P_a)^2 \\ &= 597\end{aligned}$$

Estimating the difference between two population proportions with specified absolute precision

Required information:

- Anticipated population = p_1 and p_2
- Confidence level = $100(1-\alpha)\%$
- Absolute precision required on either side of the true value of the difference between the proportions (in percentage points) = d
- Intermediate value = $v = [p_1(1-p_1) + p_2(1-p_2)]$
- For any value of d , the sample size required will be largest when both p_1 and p_2 are equal to 50% therefore if it is not possible to estimate either population proportion, the safest choice of 0.5 should be used in both cases.
- **Formula:** $n = Z_{1-\alpha/2}^2 [p_1(1-p_1) + p_2(1-p_2)] / d^2$

Example:

- What sample size should be selected from each of two groups of people to estimate a risk difference to within 5 percentage points of the true difference with 95% confidence, when no reasonable estimate of p_1 and p_2 can be made?

Solution:

- Anticipated population proportion, $p_1=50\%$, $p_2=50\%$
- Confidence level = 95 %
- Absolute precision = 5 %
- Intermediate value = 0.50 { $v=[p_1(1-p_1) + p_2(1-p_2)]$ }

By using the formula $n = Z^2_{1-\alpha}[p_1(1-p_1)+p_2(1-p_2)]^2/d^2$

we get $n = 768$

Sample size calculation in case control studies

Requirements:

- Anticipated prevalence of exposure in the control group, P_0
- A hypothesized odds ratio associated with exposure that would have sufficient biologic or public health importance to warrant its detection, R
- The desired level of significance, α
- The desired study power, $1-\beta$

Unmatched case control study

Formula for case control study with equal number of cases and control, the required sample size for each group (n per group) is calculated as

$$\mathbf{n = [Z\alpha\sqrt{2pq+Z\beta\sqrt{(p_1q_1+p_0q_0)}}]^2/[(p_1-p_0)^2]}$$

where,

$$p_1 = [p_0R]/[1+p_0(R-1)] \quad p = (1/2)(p_1+p_0)$$

$$q = 1-p \quad q_1 = 1-p_1 \quad q_0 = 1-p_0$$

- ▶ $Z\alpha$ is the value from the standard normal distribution corresponding to α
- ▶ $Z\beta$ is the value from the standard normal distribution corresponding to β
- ▶ A simpler formula for practical purpose is given by

$$\mathbf{n = [(2pq) (Z\alpha + Z\beta)^2] / [(p_1 - p_0)^2]}$$

Example :

▶ Case control study:

- Congenital heart defects
- Women using oral contraceptives occurring around the time of conception.

30% of women of child bearing age will have an exposure to within 3 months of conception

EXAMPLE:

Congenital heart defects

- Women using oral contraceptives occurring around the time of conception
- 30% of women of child bearing age will have an exposure to within 3 months of conception

Here,

- $p_0 = 0.30$
- $\alpha = 0.05$ (two sided) $Z\alpha=1.96$
- $\beta = 0.10$ $Z\beta=1.28$
- $R = 3$

Now

- $p_1 = [0.3 \times 3] / [1 + 0.3(3-1)] = 0.5625$
- $P = (1/2) (0.3 + 0.5625) = 0.43125$
- $n = [1.96\sqrt{(0.4905)} + 1.28\sqrt{(0.2461+0.21)}] / [(0.2625)^2]$
 $= 73$

Sample size formula in a case control study with 'c' control per case is given by:

$$\mathbf{n = [Z\alpha\sqrt{(1+1/c)pq} + Z\beta\sqrt{(p_1q_1+p_0q_0/c)}]^2 / [(p_1-p_0)^2]}$$

where,

- ▶ $p = (p_1 + cp_0) / (1 + c)$
- ▶ $p_1 = [p_0R] / [1 + p_0(R - 1)]$

Equivalent simpler formula is

$$\mathbf{n = [(1+1/c)pq] (Z\alpha + Z\beta)^2 / [(p_1-p_0)^2]}$$

Cohort study–Hypothesis test for a relative risk

Required information

- For a two sided test:
 - Test value of the relative risk under the null hypothesis, $H_0 : RR=1$
 - Vs the alternative hypothesis, $H_a : RR \neq 1$

Two of the following should be known

- Anticipated probability of disease in people exposed to factor of interest = P_e
- Anticipated probability of disease in people not exposed to factor of interest P_c
- Anticipated relative risk RR
- Level of significance = α
- Power of the test = $1-\beta$
- For determining sample size for a cohort study when $RR > 1$, the values of both P_c and RR are needed. If P_e is known this can be calculated
- $RR = P_e/P_c$ and $P_c = P_e/RR$
- $P_e = RR \times P_c$
- If $RR < 1$ the values P_e and $1/RR$ should be used
 - Sample size formula :
$$\frac{\{ Z_{1-\alpha/2} \sqrt{2P(1-2P)} + Z_{1-\beta} \sqrt{[(1-P_e)P_e + (1-P_c)P_c]} \}^2}{(P_e - P_c)^2}$$
 -
- Where $p = (P_e + P_c)/2$

Example:

- Two competing therapies for a particular cancer are to be evaluated by a cohort study.
- Treatment A is a new therapy that will be widely used if it can be demonstrated that it halves the risk of recurrence in the first five years after treatment. 35% recurrence is being reported in patients with treatment B.
- How many patients should be studied in each of the two treatment groups if the investigator wishes to be 90% confident of correctly rejecting the null hypothesis if it is false, at a 5% level of significance?



Solution:

- Test value of the relative risk under the null hypothesis, $H_0: RR=1$
Vs the alternative hypothesis, $H_a : RR \neq 1$ (2 sided)
- Number of exposure groups = 2
- Outcome measure - recurrence of cancer
- Follow up period
- Anticipated of probability of disease given B, $P_c=0.35$
- Anticipated $RR = 0.5$
- Power of the study = 90%
- Level of significance, $\alpha =0.05$
- $RR<1$, $1/RR=2$ and Anticipated of probability of disease given A,
 $P_e =0.35/2=0.175$
- $P = (0.175+0.35)/2= 0.2625$

Hence the required sample size in each group

$$n = \frac{\{1.96 \times \sqrt{2 \times 0.2625(1-0.2625)} + 1.282 \sqrt{[(1-0.175) \times 0.175 + (1-0.35) \times 0.35]}\}^2}{(0.175-0.35)^2}$$

$$=130$$

Estimating an incidence rate with a specified relative precision

Required information

- The following should be known
 - relative precision, ϵ
 - confidence level, $(1-\alpha)$
 - Sample size formula

$$n = [Z / \epsilon]^2$$

where Z value corresponds to appropriate level of significance

Example:

- ▶ How large a sample of patients should be followed up if an investigator wishes to estimate the incidence rate of a disease to within 10% of its true value with 95% confidence?

Solution

- ▶ Relative precision, $\varepsilon = 0.10$
- ▶ Confidence level, $(1-\alpha) = 0.95$
- ▶ Required sample size is

$$n = [1.96/.10]^2$$
$$= 384$$

Hypothesis test for an incidence rate

Required information

- Test value of the incidence rate under the null hypothesis, $H_0: \lambda = \lambda_0$
 - Vs the alternative hypothesis, $H_a: \lambda \neq \lambda_0$ (or $\lambda = \lambda_a$)
 - Or $H_0: \lambda_0 = \lambda_a$ Vs $H_a: \lambda_0 \neq \lambda_a$
- Anticipated value of the population incidence rate = λ_a
- Power of test $1 - \beta$
- Level of significance α
- Sample size formula

$$n = \frac{(Z_{1-\alpha/2} \lambda_0 + Z_{1-\beta} \lambda_a)^2}{(\lambda_0 - \lambda_a)^2}$$

Example:

- On the basis of a five year follow up study of a small number of people, the annual incidence rate of a particular disease is reported to be 40%.
- What minimum sample size would be needed to test the hypothesis that the population incidence rate is different from 40% at the 5% level of significance?
- It is desired that the test should have a power of 90% of detecting a true annual incidence rate of 50%.

Solution:

- Test value of the population incidence rate under the null hypothesis,

$$H_0: \lambda_0 = .40$$

- Anticipated value of the population incidence rate (under H_a), $\lambda_a = 0.50$

- Power of test $1 - \beta = 0.90$

- Level of significance, $\alpha = 0.05$


- Required sample size is

$$n = \frac{(1.96 \times .40 + 1.282 \times .50)^2}{(0.40 - 0.50)^2}$$

$$(0.40 - 0.50)^2$$

$$= 203$$

Sample size estimation for Clinical trial

- **Design specifications affecting sample considerations in clinical trials**
 - **Number of treatment groups**
 - **Outcome measures**
 - **Length of follow-up**
 - **Alternative hypothesis**
 - **Treatment difference**
 - **Type I and Type II error protection**
 - **Allocation ratio**
- 

Design specifications affecting sample considerations in clinical trials Cont.

- **Rate of loss to follow up**
- **Noncompliance rate**
- **Treatment lag time**
- **Degree of stratification for baseline risk factors**
- **α and β levels adjustment for multiple comparisons**
- **α and β levels adjustment for multiple looks**
- **α and β levels adjustment for multiple outcomes**

Sample size for binary outcome measures:

- ▶ **Situation1: uniform allocation ($\lambda=1$)**

- ▶ Sample size formula

- ▶
$$n_c = \frac{(Z_{\alpha} \sqrt{2pq} + Z_{\beta} \sqrt{p_c q_c + p_t q_t})^2}{\Delta A^2}$$

- ▶ $n_t = n_c$ and $n = (r+1) n_c$

where,

- ▶ r = number of test groups

- ▶ $\lambda = n_t / n_c$ (allocation ratio)

- ▶ n_c = sample size required for the control treatment group

- ▶ n_t = sample size required for each of the treatment group

- ▶ P_c = event rate in the control group

- ▶ P_t = event rate in the treatment group

- ▶ $q_c = 1 - p_c$

- ▶ $q_t = 1 - p_t$

- ▶ p = weighted average of 2 events rates = $(p_c + \lambda p_t) / (1 + \lambda)$

- ▶ $q = 1 - p$ and

- ▶ ΔA = absolute difference in 2 events rates = $p_c - p_t$

Example : (cardiovascular mortality study)

➤ Design specification:

- number of treatment group= 2
- outcome measure: 5 year mortality
- alternative hypothesis: one sided
- detectable treatment difference : 10% difference in 5 year mortality of two group
- $p_c=0.40$
- $p_t=0.30$
- error protection: $\alpha(\text{one sided})=0.05$, $\beta=0.05$
- allocation ratio: 1:1, $\lambda=1$
- loss due to drop out and non compliance:d=20%

Solution:

$$\begin{aligned} \text{➤ } n_c &= \frac{1.645\sqrt{2(0.35)(0.65)} + 1.645\sqrt{(0.4 \times 0.6 + 0.3 \times 0.7)^2}}{(0.10)^2} \\ &= 490 \end{aligned}$$

Adjusting for 20% losses,

$$\text{➤ } n_c = 490 \times (1/.8)$$

$$\text{➤ } = 613$$

$$\text{➤ } n_t = 613$$

$$\begin{aligned} \text{Total sample size } n &= 613 + 613 \\ &= 1226 \end{aligned}$$

non uniform allocation ($\lambda \neq 1$)

Sample size formula

➤ $n_c = \frac{(Z\alpha\sqrt{(\lambda+1)/\lambda} + Z\beta\sqrt{(p_cq_c + p_tq_t/\lambda)})^2}{\Delta A^2}$

$$\Delta A^2$$

➤ $n_t = \lambda n_c$

➤ and $n = r n_t + n_c$

Example: coronary drug project

▶ Design specification:

- ❖ number of treatment group= 6 (1 control and 5 test treatments)
- ❖ outcome measure: 5 year mortality
- ❖ alternative hypothesis: one sided
- ❖ detectable treatment difference : 25% difference in 5 year mortality of test group in relation to control group
- ❖ $p_c=0.30$, $\Delta A == p_c - p_t / p_c = 0.25$
- ❖ i.e. $p_t == 0.225$
- ❖ error protection: α (one sided)=0.01, $\beta=0.05$
- ❖ allocation ratio: 1:1:1:1:2.5, $\lambda=1/2.5$
- ❖ loss due to drop out and non-compliance: $d=30\%$ after 5 year follow up

▶ **Sample size calculation**

▶ $n_c =$

$$\frac{(2.326\sqrt{0.28 \times 0.72(0.4+1+1.645\sqrt{(0.3 \times 0.7+0.225 \times 0.775/0.4)^2}})}{(0.075)^2}$$

$$=1906$$

▶ $n_t = 1906 \times (1/2.50) = 762$

▶ Adjusting for 30% losses,

▶ $n_c = 1906 \times (1/1-0.3) = 2723$

▶ $n_t = 763 \times (1/.7) = 1089$

▶ Total sample size = $5(1089) + 2723 = 8168$

Calculation of sample size in diagnostic test:

- The sample size in a diagnostic test study is done in two stages
- *First, specify the expected “sensitivity” of the test and specify the “acceptable deviation” from this sensitivity on either side of the expected sensitivity. Then,*

$$a = \frac{\text{sensitivity} (1-\text{sensitivity})}{(\text{deviation})^2}$$

- Let us say, we are validating ELISA test for HIV infection. Our rough estimate is that the sensitivity would be 95% (i.e. 0.95) and we accept a deviation of 3% on either side (i.e. acceptable range of sensitivity to be detected by the present study sample = 92% to 98%); thus $d = 3\%$ (i.e. 0.03).

$$a = \frac{0.95(1-0.95)}{(0.03)^2} = 53$$

➤ *Now, the actual sample size 'N' is calculated by the*

Formula, $N = a / \text{prevalence}$

Let us say the expected prevalence of HIV infection in the population we are doing our study (say, professional blood_donors) is 5% (i.e. 0.05)

Thus,

$$N = 53 / 0.05$$

$$= 1060$$

References :

1. Lwanga SK, Lemeshow S. Sample size determination in health studies - A practical manual. 1st ed. Geneva: World Health Organization; 1991.
2. Zodpey SP, Ughade SN. Workshop manual: Workshop on Sample Size Considerations in Medical Research. Nagpur: MCIAPSM; 1999
3. Zodpey SP. Sample size and power analysis in medical research. Indian J Dermatol Venerol Leprol 2004;70(2):123-28
4. Rao Vishweswara K. Biostatistics A manual of statistical methods for use in health , nutrition and anthropology. 2nd edition. New Delhi: Jaypee brothers;2007
5. Bhalwar R et al. Text book of Public Health and Community Medicine 1st ed. Pune :Department of Community Medicine Armed Forces Medical College; 2009