

# **Sample Size Determination**

**An essential part of planning any study is to decide how many people need to be studied**

# Sample Size

- ***Sample Size***: The number of study subjects selected to represent a given study population.
- Important to make inferences based on the findings from the sample.
- Should be sufficient to represent the characteristics of interest of the study population.

- In estimating a certain characteristic of a population, sample size calculations are important to ensure that **estimates** are obtained with required precision or confidence
- The accuracy of the envisaged results determine the size of the sample

# Example

- A prevalence of 10% from a sample size of 20
  - would have a 95% CI of 1% to 31%,
  - which is not very precise or informative.
- But, a prevalence of 10% from a sample of size 400
  - would have a 95% CI of 7% to 13%,
  - which may be considered sufficiently accurate.

- In studies concerned with **detecting an effect** (e.g. a difference between two groups), sample size calculations are important to ensure the detection of whether association exists or not.
- Large sample size can result in statistical significance when  $\Delta$  is very small (not of practical, clinical or public health importance).

- Small sample size can result in a non statistically significant finding even when  $\Delta$  is large (of practical, clinical or public health importance).
- If the sample is too small, then even if large differences are observed, it will be impossible to show that these are due to anything more than sampling variation.

# Sample size determination depends on the:

- Objective of the study
- Design of the study
  - Descriptive/Analytic
- Accuracy of the measurements to be made
- Degree of precision required for generalization
- Plan for statistical analysis
- Degree of confidence with which to conclude

- **Common questions:**

- ***“How many subjects should I study?”***

- **Too small sample = Waste of time and resources**

- = Results have no practical use**

- **Too large sample = Waste of resources**

- = Data quality compromised**

# When deciding on sample size:

PRECISION COST



Sample size = Precision = Cost

- **The feasible sample size is also determined by the availability of resources:**

- time**
- manpower**
- transport**
- available facility, and**
- money**

# 1. Sample Size: Single Sample

- The aim is to have a large enough sample with which to estimate a population mean or proportion within a narrow interval with high reliability.
- Concerned with the precision of the estimate (“narrowness of the CI”).



**estimate  $\pm$  d units**

# **Sample size for single sample includes:**

- A. Sample size for estimating a single population mean
- B. Sample size to estimate a single population proportion

# A. Sample size for estimating a single population mean

- **AIM:** Estimate  $\mu$
- **WANT:** Estimate  $(\bar{x}) \pm d$  units  
where  $d = \text{Margin of error}$   
 $= \text{Absolute precision}$   
 $= \text{Half of the width (w) of CI}$

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

## Steps:

1. Specify  $d$  (or  $w = 2d$ )
2. Use known  $\sigma^2$  or estimate using  $s^2$

Standard error of the estimator of the parameter of interest

$$d = z_{\alpha/2} \sqrt{\sigma^2/n}$$
$$d^2 = (z_{\alpha/2})^2 \cdot \sigma^2/n$$

$$n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{d^2}$$

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{e^2} = \left( \frac{z_{\alpha/2} \sigma}{e} \right)^2$$

Where  $d = e$  in some text books

# Example:

- 1. Find the minimum sample size needed to estimate the drop in heart rate ( $\mu$ ) for a new study using a higher dose of propranolol than the standard one. We require that the two-sided 95% CI for  $\mu$  be no wider than 5 beats per minute and the sample sd for change in heart rate equals 10 beats per minute.**

$$n = (1.96)^2 10^2 / (2.5)^2 = 62 \text{ patients}$$

**2. Suppose that for a certain group of cancer patients, we are interested in estimating the mean age at diagnosis. We would like a 95% CI of 5 years wide. If the population SD is 12 years, how large should our sample be?**

$$n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{d^2} = \frac{(1.96)^2(144)}{(2.5)^2} = 88.5 \approx 89$$

- Suppose  $d=1$
- Then the sample size increases

$$n = \frac{z^2 \cdot \sigma^2}{d^2} = \frac{(1.96)^2 (144)}{(1)^2} = 553.2 \approx 554$$

3. A hospital director wishes to estimate the mean weight of babies born in the hospital. How large a sample of birth records should be taken if she/he wants a 95% CI of 0.5 wide? Assume that a reasonable estimate of  $\sigma$  is 2. *Ans: 246 birth records.*

But the population  $\sigma^2$  is most of the time unknown

As a result, it has to be estimated from:

- *Pilot or preliminary sample:*
  - Select a pilot sample and estimate  $\sigma^2$  with the sample variance,  $s^2$
- *Previous or similar studies*

## B. Sample size to estimate a single population proportion

- **Aim:** Estimate  $p$
- **Want:** Estimate  $\pm d$  units where  $d = Z \cdot SE$

(95% CI of width= $2d$ )

### Steps:

1. Specify  $d$  (or  $w = 2d$ )
2. Use estimated  $p$  (use  $p=0.5$  if no information)

### 3. Solve for n

$$d = z_{\alpha/2} \sqrt{\frac{pq}{n}}$$

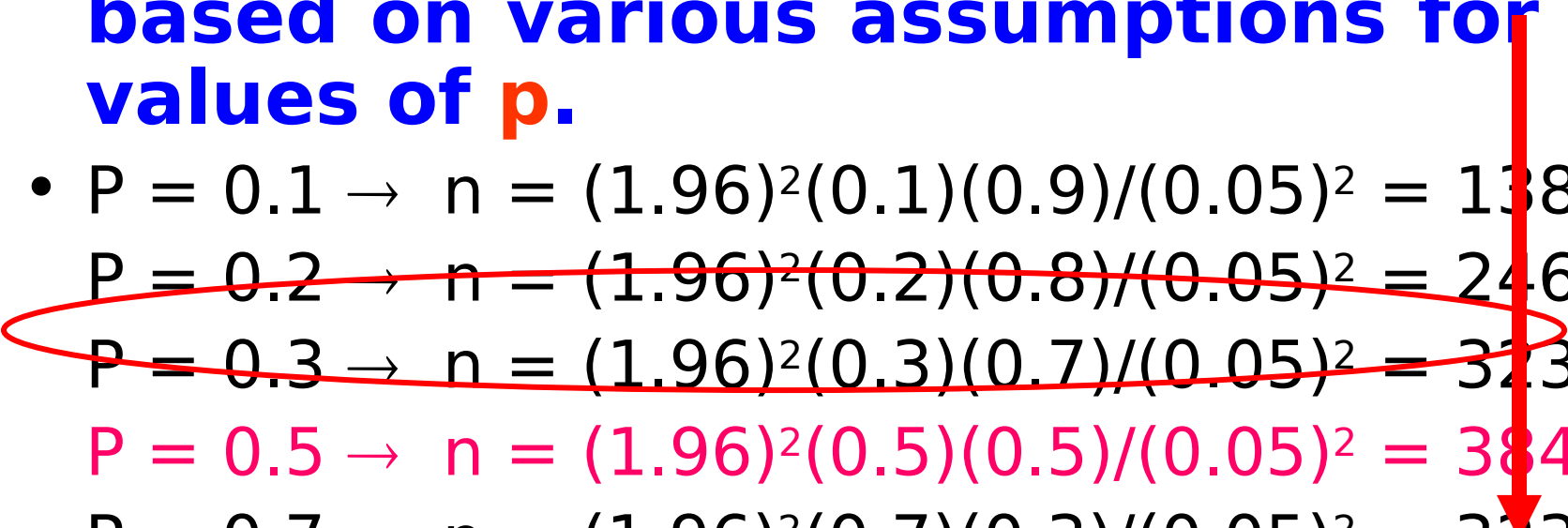
$$d^2 = (z_{\alpha/2})^2 \cdot \frac{pq}{n}$$

$$n = \frac{(z_{\alpha/2})^2 \cdot pq}{d^2}$$

- 1. Suppose that you are interested to know the proportion of infants who breastfed >18 months of age in a rural area. Suppose that in a similar area, the proportion (p) of breastfed infants was found to be 0.20. What sample size is required to estimate the true proportion within  $\pm 3\%$  points with 95% confidence? Let  $p=0.20$ ,  $d=0.03$ ,  $\alpha=5\%$**

$$n = \frac{(z_{\alpha/2})^2 \cdot pq}{d^2} = \frac{(1.96)^2 (0.2)(0.8)}{(0.03)^2} = 683$$

- Suppose there is no prior information about the proportion (**p**) who breastfeed
- Assume **p=q=0.5** (most conservative)
- **T**  
**i**  $n = \frac{(z_{\alpha/2})^2 \cdot pq}{d^2} = \frac{(1.96)^2 (0.5)(0.5)}{(0.03)^2} = 1,068$  **e**

- An estimate of **p** is not always available.
  - However, the formula may also be used for sample size calculation based on various assumptions for the values of **p**.
  - $P = 0.1 \rightarrow n = (1.96)^2(0.1)(0.9)/(0.05)^2 = 138$
  - $P = 0.2 \rightarrow n = (1.96)^2(0.2)(0.8)/(0.05)^2 = 246$
  - $P = 0.3 \rightarrow n = (1.96)^2(0.3)(0.7)/(0.05)^2 = 323$
  - $P = 0.5 \rightarrow n = (1.96)^2(0.5)(0.5)/(0.05)^2 = 384$
  - $P = 0.7 \rightarrow n = (1.96)^2(0.7)(0.3)/(0.05)^2 = 323$
  - $P = 0.8 \rightarrow n = (1.96)^2(0.8)(0.2)/(0.05)^2 = 246$
- 

- For a fixed absolute precision ( $d$ ), the required sample size increases as  $P$  increases from 0 to 0.5, and then decreases in the same way as the prevalence approaches 1.

**2. A survey is planned to determine what proportion of the medical students have regularly chewed *khat*. If no estimate of  $p$  is available and a pilot sample cannot be drawn, what sample size would be required if a 95% confidence is desired, and  $d=0.04$  is to be used.**

**Ans: 600 students**

## 2. Sample Size: Two Samples

- A. Estimation of the difference between two population means
- B. Estimation of the difference between two population proportions

# A. Sample size for estimating a difference in two means

- **Aim:** Estimate  $\mu_1 - \mu_2$
- **Want**  $\bar{X}_1 - \bar{X}_2$  within  $\pm d$  units,  
where  $d = Z_{\alpha/2} \cdot SE$   
(95% CI of width =  $w = 2d$ )

- If equal sample size in both groups is required, then:

$$d = z \cdot \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}$$

$$n = \frac{z^2 \cdot (\sigma_1^2 + \sigma_2^2)}{d^2}$$

- Use  $\sigma_1^2$ ,  $\sigma_2^2$  or estimate using  $s_1^2$  and  $s_2^2$

## B. Sample size for estimating a difference in two proportions

- **Aim:** Estimate  $p_1 - p_2$
- **Want**  $\hat{p}_1 - \hat{p}_2$  within  $\pm d$  units  
where  $d = Z_{\alpha/2} \cdot SE$   
(95% CI of width =  $w = 2d$ )

- If equal sample sizes in both groups, then:

$$d = z \cdot \sqrt{\frac{p_1 q_1}{n} + \frac{p_2 q_2}{n}}$$
$$n = \frac{z^2 \cdot (p_1 q_1 + p_2 q_2)}{d^2}$$

- Use estimates of  $p_1$ ,  $p_2$  or (or  $p_1=p_2=0.5$  if unknown)

# Points for Consideration

1. Sample size estimates might need to be adjusted to compensate for non-response rate, patient dropout or loss to follow-up, lack of compliance, etc.
2. If sampling is from a finite population of size  $N$ , then:

$$n = \frac{n_0}{\left(1 + \frac{n_0}{N}\right)}$$

where  $n_0$  is the sample from an infinite population. When  $N$  is large in comparison to  $n$ , (i.e.,  $n/N \leq 0.05$ ), the finite population correction may be ignored.

3. Design effect for complex cluster sampling. Common values: multiply  $n$  by 2, 3, ...5.

### 3. Sample Size Based on Hypothesis Testing

- The method of determining sample size in the preceding sections takes into account the probability of a **type I** error, but not a **type II** error since the level of confidence is determined by the confidence level  $(1-\alpha)$ .
- However, in many statistical inference procedures, **type II** and **type I** errors are considered when determining the sample size.

# Significance Difference Between Two Groups

- Using power of a study to determine sample size = significant difference = Hypothesis testing
- Aim: Have large enough samples to detect a difference in population means (or in population proportions)

- We would like to maintain low probability of a **Type I error ( $\alpha$ )** and low probability of a **Type II error ( $\beta$ )** [**high power =  $1 - \beta$** ].

**Significance level of a test =  $\alpha$  = Type I error**

**$1 - \alpha$  = Confidence**

**$1 - \beta$  = Power**

- **Type I error ( $\alpha$ )** = The probability of rejecting  $H_0$  when it is true

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

- **Type II error ( $\beta$ )** = The probability of not rejecting  $H_0$  when it is false

$$\beta = P(\text{do not reject } H_0 \mid H_0 \text{ false})$$

- **Power ( $1-\beta$ )** = the probability  $H_0$  is rejected given that it is false  
=  $P(\text{rejecting } H_0/H_1 \text{ is true})$
- If the power of a test is low, then there is little chance of detecting a difference even if one really exists
- **Power** is an important part of the design of a study

- Power  $(1 - \beta) = 50\%$ ,  $Z_\beta = 0.00$
- Power  $(1 - \beta) = 75\%$ ,  $Z_\beta = 0.67$
- Power  $(1 - \beta) = 80\%$ ,  $Z_\beta = 0.84$
- Power  $(1 - \beta) = 90\%$ ,  $Z_\beta = 1.28$

- Power is one-sided and  $Z_\beta$  is always one-sided
- Most of the studies recommend power of 80%.

# Factors affecting the power

- If  $\alpha$  decreases, the power decreases
- When the difference between  $H_0$  and  $H_A$  increases, then the power increases
- When  $\sigma$  increases, then the power decreases
- If the sample size ( $n$ ) increases, the power increases

# Factors affecting the sample size

- The sample size increases as  $\sigma^2$  increases
- The sample size increases as the significance level ( $\alpha$ ) is made smaller ( $\alpha$  decreases)
- The sample size increases as the required power increases
- The sample size decreases as the absolute value of the difference between the  $H_0$  and  $H_A$ ) increases

**Ho** = There is no difference between  
the

two groups

$$\text{Ho: } \mu_1 - \mu_2 = 0$$

$$P_1 - P_2 = 0$$

**HA** = There is a difference between  
the

two groups

$$\text{HA: } \mu_1 - \mu_2 \neq 0$$

$$P_1 - P_2 \neq 0$$

## A. Comparison between two means (Equal sample sizes)

$$n_1 = n_2 = \frac{(z_{\alpha/2} + z_{\beta})^2 (\sigma_1^2 + \sigma_2^2)}{\Delta^2}$$

$$\Delta = |\mu_1 - \mu_2|$$

The means and variances of the two respective groups are  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$ .

# Example

1. Determine the sample sizes required to detect a difference of 5 mm in mean blood pressure between individuals receiving placebo and those receiving drug with  $\alpha = 5\%$  and power of 0.80

- Assume  $\sigma_1 = \sigma_2 = 15$  mm in each group.

- We are interested in testing:

$$H_0: \mu_1 - \mu_2 = 5, H_A: \mu_1 - \mu_2 \neq 5$$

$$n = \frac{(z_\alpha + z_\beta)^2 (\sigma_1^2 + \sigma_2^2)}{\Delta^2} = \frac{(1.96 + 0.84)^2 (15^2 + 15^2)}{5^2} = 141.1$$

- We would need **142** individuals in each group

2. Suppose that the true blood pressure distribution among OC users is normal with  $\mu_1$  and  $\sigma_1^2$ . Similarly, for non-users the distribution is normal with  $\mu_2$  and  $\sigma_2^2$ . We wish to test the hypothesis that  $H_0: \mu_1 = \mu_2$  **versus**  $\mu_1 \neq \mu_2$ . Determine the appropriate sample size for the study using  $\alpha = 0.05$  and a power of 80%. It was revealed by the small study that: sample mean1=132.86,  $s_1=15.34$ , sample mean2=127.44, and  $s_2=18.23$ . Use the sample data to estimate population parameters.

$$n = (15.34^2 + 18.23^2)(1.96 + 0.84)^2 / (132.86 - 127.44)^2 = \mathbf{152 \text{ in each group}}$$

## B. Comparison between two means (Unequal sample sizes)

$$n_1 = \frac{(z_{\alpha/2} + z_{\beta})^2 (\sigma_1^2 + \sigma_2^2 / \lambda)}{\Delta^2}$$

$$n_2 = \lambda n_1$$

$$\lambda = n_2 / n_1$$

In some text books,  $\lambda = k = r$

3. Suppose we anticipate twice as many non OC users as OC users entering the study using the previous example. Determine the sample size to achieve an 80% power in the study using  $\alpha=0.05$ .  $\lambda = 2$ .

$$\begin{aligned} n_1 &= \\ (15.34^2 + 18.23^2 / 2)(1.96 + 0.84)^2 / (5.42)^2 \\ &= 108 \text{ OC users} \\ \text{and } n_2 &= 2(108) = 216 \text{ non-OC users.} \end{aligned}$$

## C. Comparison between two proportions (Equal sample sizes)

- To test the hypothesis,  
Ho:  $p_1 - p_2 = 0$  vs  $H_A: p_1 - p_2 \neq 0$ ,  
 $|p_1 - p_2| = \Delta$   
with  $\alpha$  and power  $(1 - \beta)$

$$n_1 = n_2 = \frac{\left( z_{\alpha/2} \sqrt{2\bar{p}\bar{q}} + z_{\beta} \sqrt{p_1 q_1 + p_2 q_2} \right)^2}{\Delta^2}$$

Where

$$\bar{p} = \frac{p_1 + p_2}{2} \quad \bar{q} = 1 - \bar{p}.$$

$$\Delta = p_1 - p_2$$

- **Let  $p_1=0.35$ ,  $p_2=0.25$ , and  $\Delta=p_1-p_2=0.35-0.25=0.10$**

$$n_1 = n_2 = \frac{\left(1.96\sqrt{2(0.30)(.70)} + 0.84\sqrt{(0.35)(0.65) + (0.25)(0.75)}\right)^2}{(0.10)^2}$$

$$n_1 = n_2 = 328.1$$

- **We would need approximately 329 subjects in each group**

## D. Comparison between two proportions (Unequal sample sizes)

$$n_1 = \frac{\left[ Z_{\alpha/2} \sqrt{\bar{p}\bar{q} (1 + 1/\lambda)} + Z_{\beta} \sqrt{p_1 q_1 + p_2 q_2 / \lambda} \right]^2}{\Delta^2}$$

Where  $n_2 = n_1 \lambda$ ,  $\bar{p} = (p_1 + \lambda p_2) / (1 + \lambda)$   $\bar{q} = 1 - \bar{p}$ .

Note: This formula is quite general, and applies to cross-sectional, case-control and cohort studies.

# Example

- A study is proposed to study the effect of a new anticoagulant therapy. Patients are to be randomly divided into two groups: one receives the anticoagulant, and the other placebo. The groups are then followed for the incidence of major bleeding events over 3 years. Suppose that 5% of treated patients and 22% of controls are anticipated to experience a major event over 3 years. How large sample should such a study be to have an 80% chance of finding a significance difference at a ratio of 1:2 for treated and control at  $\alpha = 5\%$ .

# Solution

$$p_1=0.05, p_2=0.2 \quad \bar{p} = (0.05+2*0.22)/(1+2) = 0.16$$

$$q_1=0.95, q_2=0.78, \Delta = 0.22-0.05 = 0.17$$

$$n_1 = \frac{\left[ 1.96\sqrt{0.16*0.84(1+1/2)} + 0.84\sqrt{0.05*0.95+0.22*0.78/2} \right]^2}{(0.17)^2}$$

$$= 49$$

$$n_2 = 2*n_1 = 2*49 = 98$$

We would approximately need 147 subjects

- If the OR or RR and one of the proportions are known, we can compute the unknown proportion by:

$$P_1 = \frac{P_2}{P_2 + \frac{1 - P_2}{OR}}$$

$$P_1 = P_2 * RR$$

# Example

- A case-control study to compare the efficacy of a vaccine for the prevention of child-hood tuberculosis with a placebo. Let the proportion of unvaccinated children is 30%, with an estimated OR of at least 2.

$$P_2 = 0.3, q_2 = 0.7, OR = 2.0$$

$$P_1 = 0.3/(0.3+0.7/2) = 0.462$$

- With equal cases and controls, what sample size is required to detect, with 80% power and at  $\alpha$  5%?  
= 140 in each group

# Summary

- Sample size calculations depend on a number of assumptions:
  - the hypothesized difference of interest,  $\Delta$
  - the probability of Type I error ( $\alpha$ )
  - the probability of Type II error ( $\beta$ )
  - the variance
- Choice of sample size depends on a balance of reasonable assumptions, time, effort, and expense
- Sample sizes provide a minimum estimate of the desired sample sizes for the study