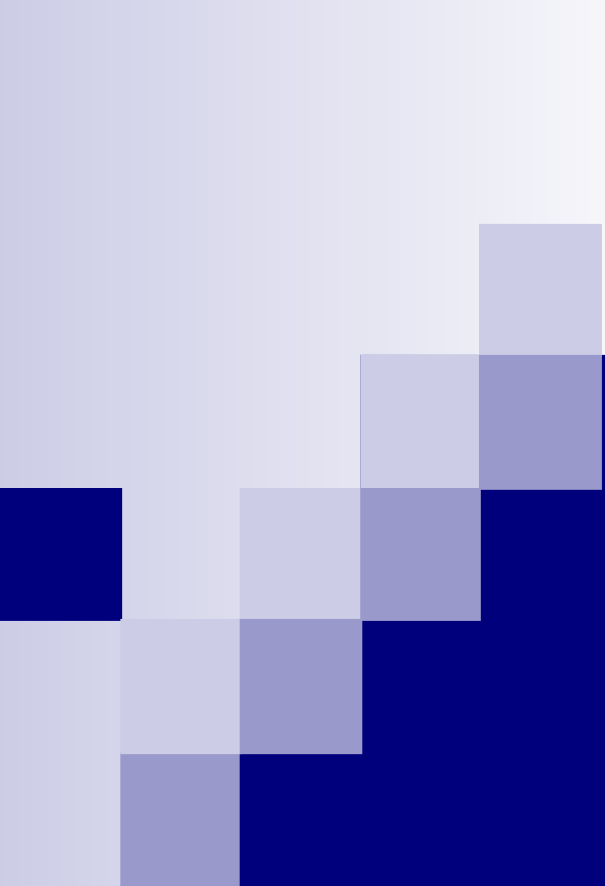




SAMPLE SIZE DETERMINATION

Ephrem Mannekulih (BSc,
MSc)

Biostatistics and Health
Informatics



Sample size determination

- ❖ ***Sample Size:*** The number of study subjects selected to represent a given study/source population.
- ❖ Sample size should be sufficient to represent the characteristics of the a given study population.

Cont....

- ❖ How many people need to be studied in order to answer the study objectives.
- ❖ If the size is too small;
 - We may fail to detect important effects
 - May estimate effects too imprecisely
- ❖ If the size is too large;
 - It may be infeasible in terms of resources.
- ❖ The eventual sample size is usually a compromise between what is **desirable** and what

Factors Affecting Sample size determination

- ❖ Determining the right number of subjects to be studied depends on the following factors:

1. Objective of the study

- Estimating single population mean or proportion
- Estimating difference between two population mean or proportion
- Testing hypothesis about single population mean or proportion
- Testing hypothesis about difference between two population mean or proportion
- Estimating the effect size of certain variable on outcome of interest



Cont....

2. Accuracy of the measurements to be made

- The allowed deviation from the true population parameter
- It can be within 1% or 5%, etc.

3. Degree of confidence within which the results to be conclude

- Commonly specified as 95%.

4. Degree of precision required for generalization



Cont....

5. Design of the study

- Cross sectional, case control, cohort etc.
- Sample size calculation depends on the type of the epidemiological study designs.
- Descriptive, observational and randomized controlled studies have different formulas to calculate sample size.

6. The size of the population that the sample is to represent.

- When population of size less than 10,000 or

Sample size based on objectives

- ❖ **The objective of the study could be;**
 - I. Estimating single population mean or proportion
 - II. Estimating difference between two population means or proportion
 - III. Testing hypothesis about single population mean or proportion
 - IV. Testing hypothesis about difference between two population means or proportion
 - V. Estimating the effect size of certain variable on outcome of interest

I. Sample size for estimating a single population mean

❖ **Objective:** To estimate population mean (μ) with narrow confidence interval and high precision

❖ **How :** Estimate $(\bar{x}) \pm d$ unit $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

where d = Margin of error =

= Measure of precision

= Half of the width (w) of CI

Steps:

1. Specify d (or $w = 2d$)

2. Use known population σ^2 or sample s^2

Cont....

❖ Example.

- To estimate the mean survival time of HIV infected patients taking ART

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

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

Cont...

❖ Where;

n = sample size.

$z_{\alpha/2}$ = Level of confidence.

σ^2 = variability of the variable of interest.

d = desired precision/accurate

Example:

- ❖ Find the minimum sample size needed to estimate the drop in mean heart rate (μ) for a new study using a higher dose of propranolol than the standard one. We require that the two-sided 95% CI for μ be no wider than 5 beats per minute and the sample sd for change in heart rate equals 10 beats per minute.
 - $n = (1.96)^2 10^2 / (2.5)^2 = 62$ patients

Cont....

- ❖ **What if the population σ^2 is unknown?**
 - Conducting pilot study
 - Use previous or similar studies finding

II. Sample size for estimating a single population proportion

- ❖ **Objective:** To estimate population proportion (P) with narrow confidence interval and high precision
- ❖ **How :** Estimate $() \pm d$ units
where d = Margin of error =
= Measure of precision
= Half of the width (w) of CI

Steps:

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

Cont....

❖ Example;

- To determine the proportion immunological treatment failure among HIV infected patients

$$\begin{aligned}d &= z_{\alpha/2} \sqrt{\frac{pq}{n}} \\d^2 &= (z_{\alpha/2})^2 \cdot \frac{pq}{n} \\n &= \frac{(z_{\alpha/2})^2 \cdot pq}{d^2}\end{aligned}$$

Cont...

❖ Where;

n = sample size.

$z_{\alpha/2}$ = Level of confidence.

p = Proportion of the variable of interest.

d = desired precision

Example

- 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

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

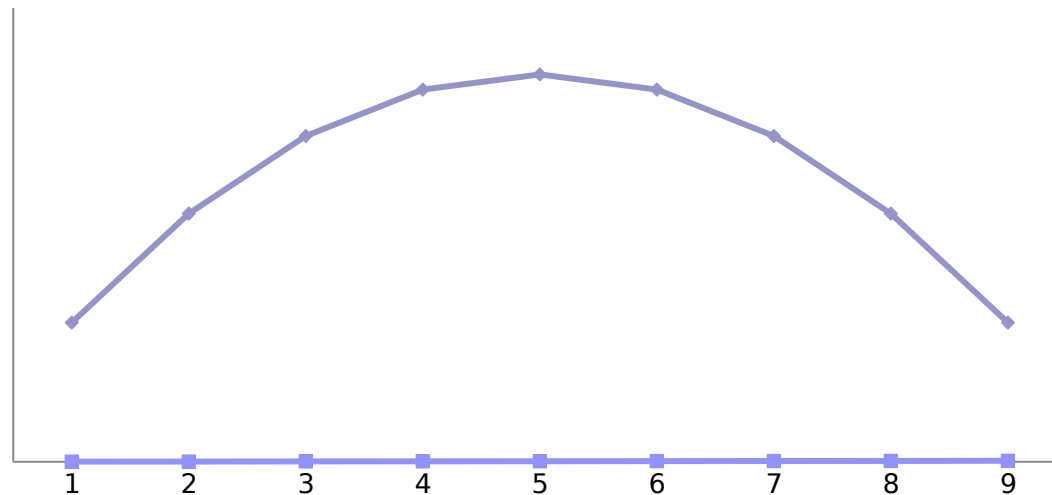
Example

- ❖ Suppose there is no prior information about the proportion (p) who breastfeed
- ❖ Assume $p=q=0.5$ (most conservative)
- ❖ Then the required sample size increases

$$n = \frac{(z_{\alpha/2})^2 \cdot pq}{d^2} = \frac{(1.96)^2 (0.5)(0.5)}{(0.03)^2} = 1,068$$

What if p is unknown or not available?

- ❖ Sample size should be calculated based on various assumptions for approximate values of p .



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

Example

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

III. Sample size for Estimating the difference between two population means

- ❖ **Objective:** To estimate the difference between two population means (-) with narrow confidence interval and high precision
- ❖ **How :** Estimate (-) $\pm d$ units
- ❖ Where $d =$
- ❖ Use σ_1^2 , σ_2^2 or estimate using s_1^2 and s_2^2

Cont....

❖ Example

- To estimate a difference of the mean survival time between HIV infected patients taking treatment X versus Y"
- If equal sample size in both groups is required, then

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

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

IV. Sample size for Estimating the difference between two population proportion

- ❖ **Objective**: To estimate the difference between two population proportion (-) with narrow confidence interval and high precision
- ❖ **How** : Estimate (-) $\pm d$ units
- ❖ Where $d =$
- ❖ Use estimates of p_1 , p_2 or (or $p_1 = p_2 = 0.5$ if unknown

Cont....

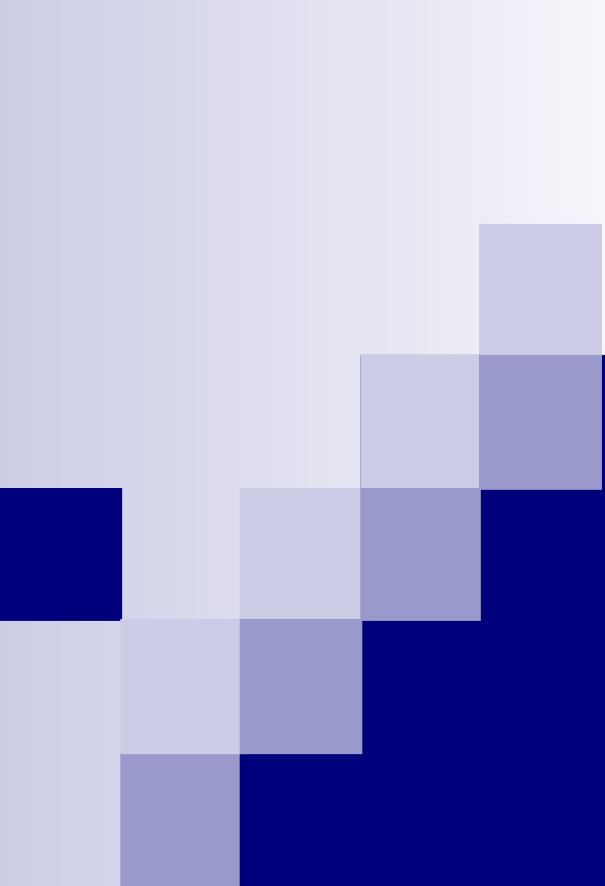
❖ Example

- To determine a difference in the proportion of immunological treatment failure among HIV infected patients taking treatment X versus Y

- If we want to detect a difference of d between the two groups, then

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

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



Sample Size for Testing Hypothesis

Sample Size for Testing Hypothesis

- ❖ The method of determining sample size based on hypothesis testing considers the probability of both **type I** and **type II** errors
- ❖ The aim is to maintain low probability of a **Type I error (α)** and low probability of a **Type II error (β)** and
- ❖ To have enough samples to detect a difference in population means or proportions

Cont....

- ❖ **Type I error (α)** = The probability of rejecting H_0 when it is true
 - / true)
 - Significance level of a test = α = Type I error
- ❖ **Type II error (β)** = The probability of fail to reject or accept H_0 when it is false
 - / false)
 - $1 - \beta$ = Power

Cont....

- ❖ **Power ($1-\beta$)** = the probability is rejected given that it is false
 - Power / false)
- ❖ If the power of a test is low, then there is little chance of detecting a difference even if one really exists
- ❖ Most of the studies recommend power of 80%
 - ❖ (Power ($1 - \beta$) = 80%, $Z_{\beta} = 0.84$)

Factors affecting the power

- ❖ If α decreases, the power decreases
- ❖ When the difference between μ_1 and μ_2 increases, then the power increases
- ❖ When σ increases, then the power decreases
- ❖ If the sample size (n) increases, the power increases

Factors affecting the sample size

- ❖ The sample size increases as σ increases
- ❖ The sample size increases as the significance level (α) is made smaller (α decreases)
- ❖ The sample size increases as the required power increases
- ❖ The sample size decreases as the absolute value of the difference between the μ and μ_0 increases



Sample Size for Testing Hypothesis

- ❖ Sample size to test hypothesis about single population mean
- ❖ Sample size to test hypothesis about single population proportion
- ❖ Sample size to test hypothesis about the difference between two population mean (paired or independent)
- ❖ Sample size to test hypothesis about the difference between two population proportion (paired or independent)

.....Sample Size for Testing Hypothesis

I. Sample size for testing hypothesis about single population mean

❖ Notation used

- = test value of the population mean under
- = hypothesized value of the population mean
- = variability in the variable of interest in the population
- $100(\)\%$ = level of significance
- $100(\)\%$ = power of test

Cont....

❖ = there is no difference between the two mean

0 : =

❖ = there is difference between the two mean

0 : $\neq$ for two sided test

0 : $>$ or $<$ for one sided test

Cont....

Testing hypothesis about single population mean

❖ **For one sided test**

$$0 \text{ } n =$$

❖ **For two sided test**

$$0 \text{ } n =$$

.....Sample Size for Testing Hypothesis

II. Sample size for testing hypothesis about two population mean

❖ Notation used

- $t = 0$ test value of the difference between two population means under
- μ_1 and μ_2 = hypothesized value of the two population means
- σ_1 and σ_2 = variability in the variable of interest in the two populations
- $100(\alpha)\%$ = level of significance

Cont....

- ❖ = there is no difference between the two population means
 - $\mu_1 - \mu_2 = 0$
- ❖ = there is difference between the two population means
 - $\mu_1 - \mu_2 \neq 0$ for two sided test
 - $\mu_1 - \mu_2 > 0$ or $\mu_1 - \mu_2 < 0$ for one sided test

Cont....

Comparison between two means (Equal sample sizes)

❖ **For one sided test**

$$0 = =$$

❖ **For two sided test**

$$0 = =$$

Cont....

Comparison between two means (Unequal sample sizes)

❖ **For one sided test**

$$0 =$$

❖ **For two sided test**

$$0 =$$

❖ Where,

$$0 = \lambda,$$

.....Sample Size for Testing Hypothesis

III. Sample size for testing hypothesis about single population proportion

❖ Notation used

- = test value of the population proportion under
- = hypothesized value of the population proportion
- 100()% = level of significance
- 100()% = power of test

Cont....

❖ = there is no difference between the two proportion

$$0 : =$$

❖ = there is difference between the two proportion

$$0 : \neq \text{ for two sided test}$$

$$0 : > \text{ or } < \text{ for one sided test}$$

Cont....

❖ **For one sided test**

$$n =$$

❖ **For two sided test**

$$n =$$

.....Sample Size for Testing Hypothesis

IV. Sample size for testing hypothesis about two population proportion

❖ Notation used

- z_0 = test value of the difference between two population proportion under
- p_0 and q_0 = hypothesized value of the population proportions
- α = level of significance
- $1 - \alpha$ = power of test

Cont....

- ❖ = there is no difference between the two proportion

- : - = 0

- ❖ = there is difference between the two proportion

- : $\neq$ for two sided test

- : - > 0 or - < 0 for one sided test

Cont....

Comparison between two proportions (Equal sample sizes)

❖ **For one sided test**

$$0 \leq n =$$

❖ **For two sided test**

$$0 \leq n =$$

❖ Where =

Cont....

Comparison between two proportions (Unequal sample sizes)

❖ **For one sided test**

$$0 =$$

❖ **For two sided test**

$$0 =$$

❖ Where $=$

$$= \lambda$$

.....Sample Size for Testing Hypothesis

V. Sample size for paired data difference in mean

❖ Notations used

- n = sample size
- σ_d = standard deviation of the within pair difference
- μ_1 and μ_2 = hypothesized value of the two population means
- $100(\alpha)\%$ = level of significance

Cont....

❖ = there is no difference between the two proportion

$$0 : = - = 0$$

❖ = there is no difference between the two proportion

$$0 : \neq 0 \text{ for two sided test}$$

$$0 : > 0 \text{ or } < 0 \text{ for one sided test}$$

Cont....

❖ **For one sided test**

$$0 \leq n =$$

❖ **For two sided test**

$$0 \leq n =$$

.....Sample Size for Testing Hypothesis

VI. Sample size for paired data difference in proportion

❖ Notations used

- n = sample size
- p_1 and p_2 = hypothesized value of the two population proportion
- $100(\alpha)\%$ = level of significance
- $100(1-\alpha)\%$ = power of test

Cont....

❖ = there is no difference between the two proportion

$$H_0 : p_1 - p_2 = 0$$

❖ = there is no difference between the two proportion

$$H_0 : p_1 - p_2 \neq 0 \text{ for two sided test}$$

$$H_0 : p_1 - p_2 > 0 \text{ or } p_1 - p_2 < 0 \text{ for one sided test}$$

Cont....

❖ **For one sided test**

$$0 \leq n =$$

❖ **For two sided test**

$$0 \leq n =$$

Cont....

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

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

$$P_1 = P_2 * RR$$



Sample size for Different study Designs

Sample size calculation for case control study

- ❖ When the exposure variable is qualitative/categorical
- ❖ Comparing odds of exposure between case and control

$$\text{Sample size} = \frac{r+1}{r} \frac{(p^*)(1-p^*)(Z_\beta + Z_{\alpha/2})^2}{(p_1 - p_2)^2} \quad \text{d abuse}$$

Cont....

Where;

r = ratio of control to cases

$P1$ = proportion of exposure in cases

$P2$ = proportion of exposure in controls

p^* = Average proportion of exposure

$P1 - P2$ = Expected difference in proportion
between cases and controls

= Level of significance

= Power

Cont...

- ❖ When the exposure variable is quantitative
- ❖ Comparing odds of exposure between case and control
 - Eg. To see the link between birth weight with diabetes in adulthood

$$\text{Sample size} = \frac{r+1}{r} \frac{SD^2 (Z_{\beta} + Z_{\alpha/2})^2}{d^2}$$

- NB. This formula is for independent case control study

Cont...

❖ Where; r = ratio of control to cases

SD = Standard deviation

d = Expected mean difference

between cases and

$Z_{\alpha/2}$ controls

Z_{β} = Level of significance

= Power

Sample size calculation for cohort study

- ❖ Comparing the rate of events between people with and without exposure
 - Eg. To see the impact of physical exercise on cardiovascular mortality

$$n = \frac{\left[Z_{\alpha} \sqrt{(1 + 1/m) \bar{p}(1 - \bar{p})} + Z_{\beta} \sqrt{p_0(1 - p_0)/m + p_1(1 - p_1)} \right]^2}{(p_0 - p_1)^2}$$

$$\bar{p} = \frac{p_1 + m p_0}{m + 1}$$

- When
- NB. This formula is for independent cohort study

Cont...

❖ Where r = ratio of control to cases

P_0 = proportion of event in unexposed group

P_1 = proportion of event in exposed group

m = the number of unexposed per exposed group

$P_1 - P_2$ = Expected difference in proportion between

Sample size for interventional studies

- ❖ When the event of interest is quantitative
- ❖ To see the effect of intervention on particular outcome of interest
 - Eg. To see the effect of antihypertensive drug on the mean blood pressure of an individual

$$\text{Sample size} = \frac{2SD^2(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$$

Cont...

❖ Where

SD = Standard deviation

d = Expected mean difference between
interventional

$Z_{\alpha/2}$ and control group

Z_{β} = Level of significance

= Power

Sample size calculation interventional study

- ❖ When Event of interest is qualitative
- ❖ To see the effect of intervention on particular outcome of interest
 - Eg. To see the protective effect of drug on mortality in patients with myocardial

ir.

$$\text{Sample size} = \frac{2(Z_{\alpha/2} + Z_{\beta})^2 P(1-P)}{(p_1 - p_2)^2}$$

Cont...

❖ Where

SD = Standard deviation

$P_1 - P_2$ = Expected difference in the proportion of events between interventional and control group

$Z_{\alpha/2}$ = average prevalence of events in two groups

= Level of significance

= Power



Sample size for qualitative studies

- ❖ There are no fixed rules for sample size in qualitative research.
- ❖ The size of the sample depends on what you try to find out, and from what different informants or perspectives you try to find that out
- ❖ Saturation of idea – ending

Points for Consideration

- ❖ 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.
- ❖ If sampling is from a finite population of size N , then
$$\left(1 + \frac{n_0}{N}\right)$$

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.

- ❖ Design effect for complex cluster sampling.



Reading Assignment

❖ Design Effect



Google for more!!!