Sampling Methods & Sample Size Calculation

Presented by:

Mohammad Moqaddasi Amiri

Assistant Professor of Biostatistics

Sirjan Faculty of Medical Sciences

April 2021

Objectives

- Understanding the concept of sampling and various types of sampling methods

- Learning the calculation of sample size in medical studies

Outline

- Introduction
- Sampling Methods
  - Non-Probability
  - Probability
- Sample Size Calculation
  - Proportion
  - Mean

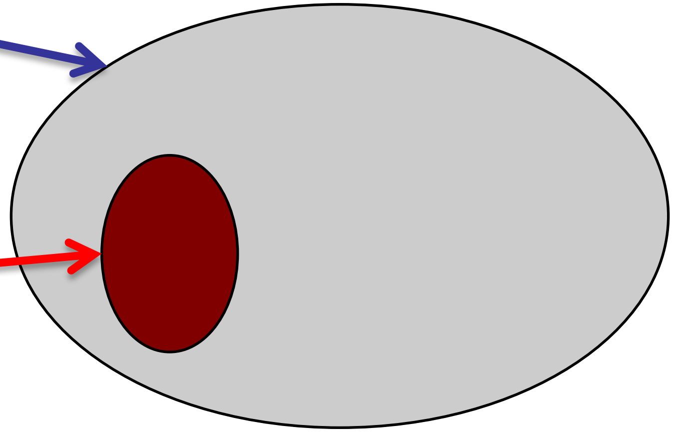
Introduction

Statistics: the science of

- Information/data collection
- Description of data
- Analysis and inference of data

Introduction

- Data collection
  - Population
    - a set which includes all measurements of interest to the researcher
  - Sample
    - A subset of the population



  - Census
    - complete enumeration of a population

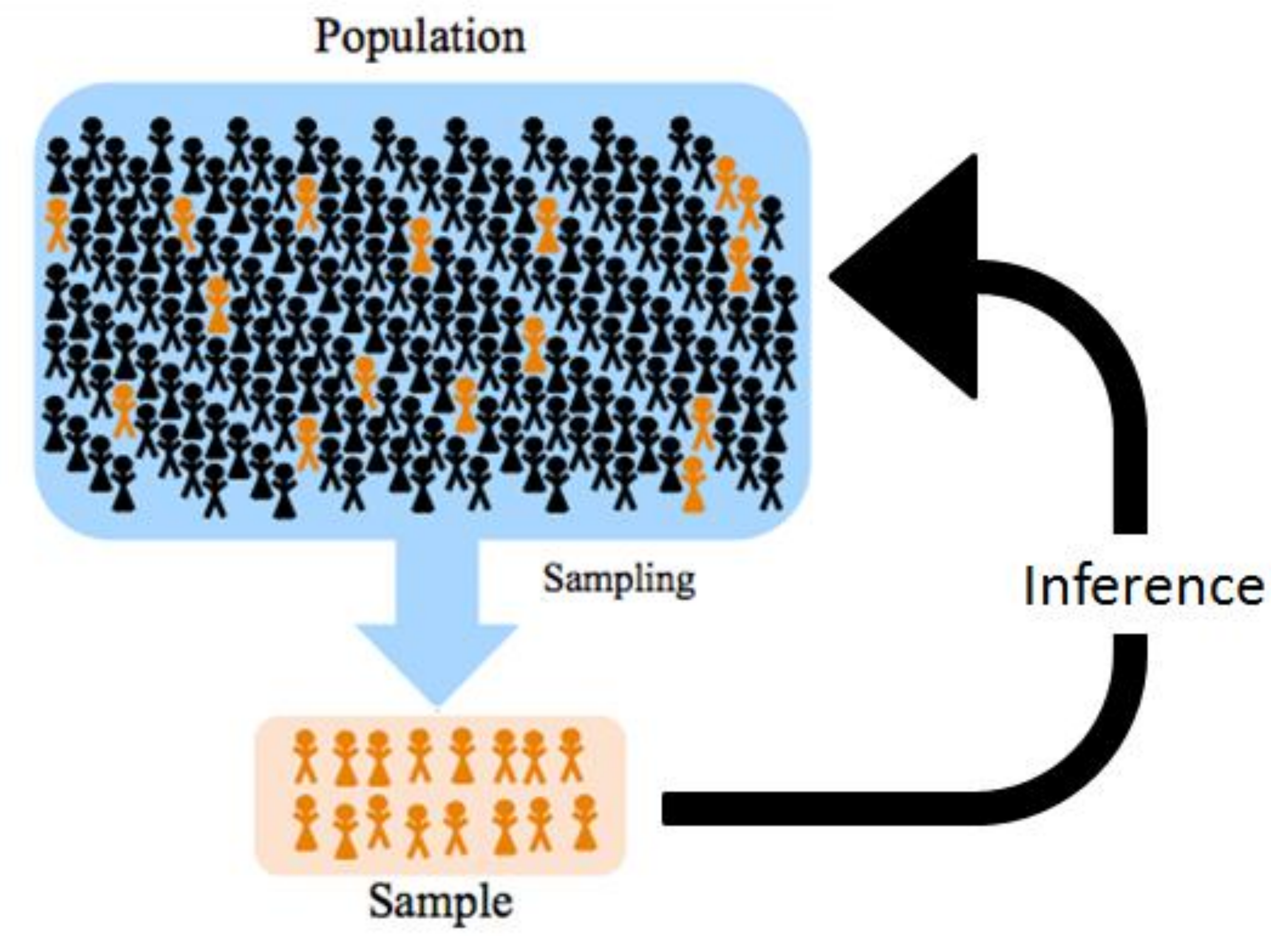
Introduction

- Sampling Advantages
  - Less costs
  - Less field time
  - More accuracy and precision

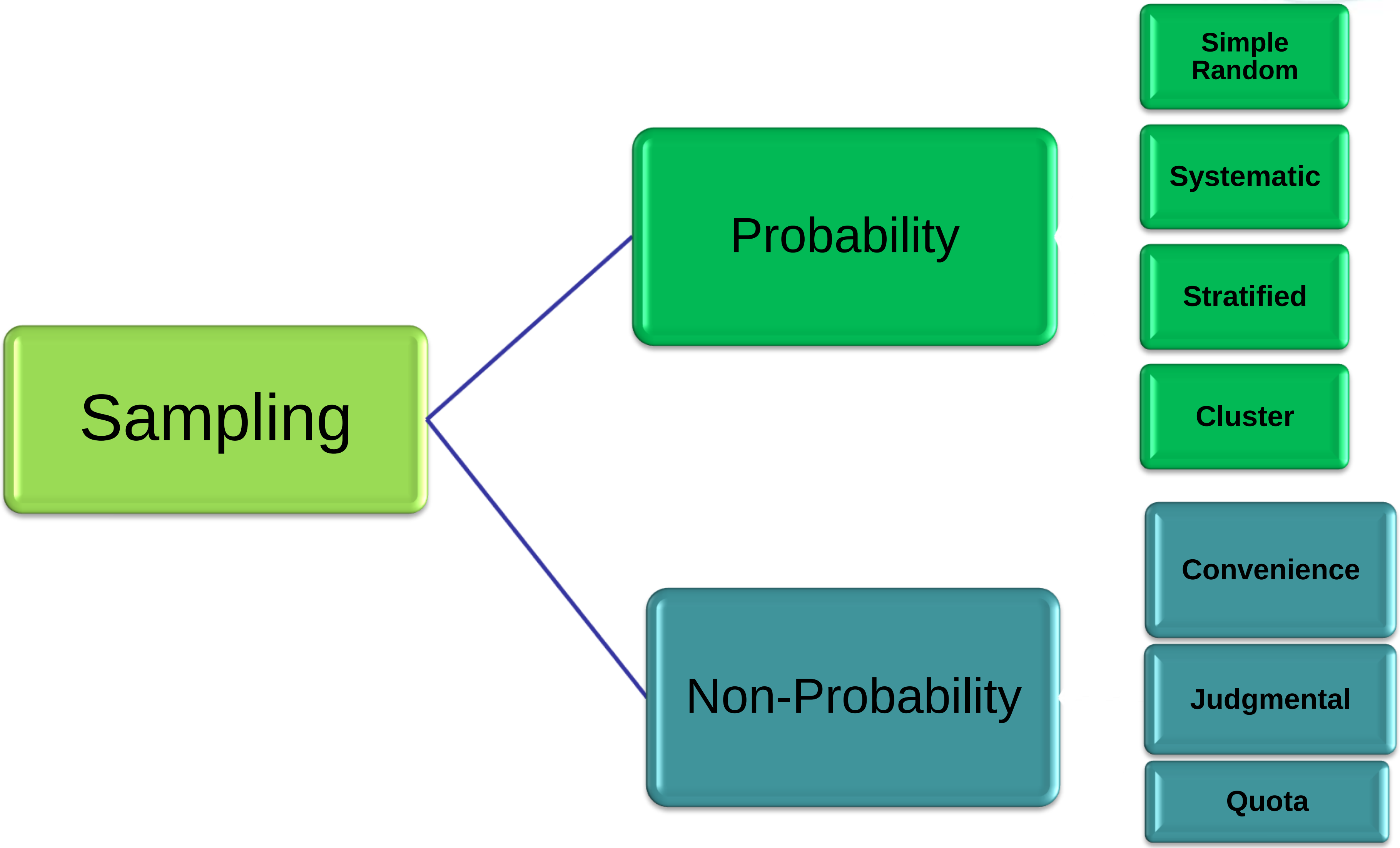
Introduction

- Target Population
  - The population to be studied
- Sampling Unit
  - smallest unit from which sample can be selected
- Sampling Frame
  - List of all the sampling units from which sample is drawn
- Sampling Scheme
  - Method of selecting sampling units from sampling frame

Introduction



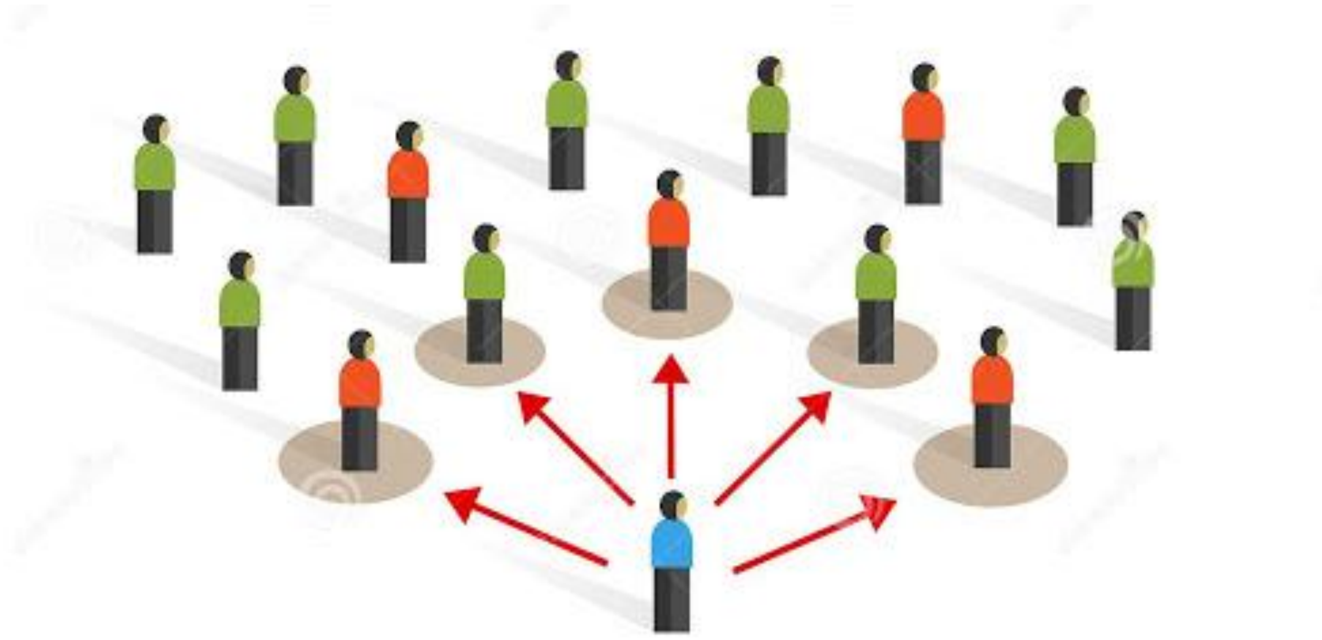
Sampling Methods



Non-probability Sampling

Convenience (ease of access)

elements of a population that are easily accessible



Non-probability Sampling

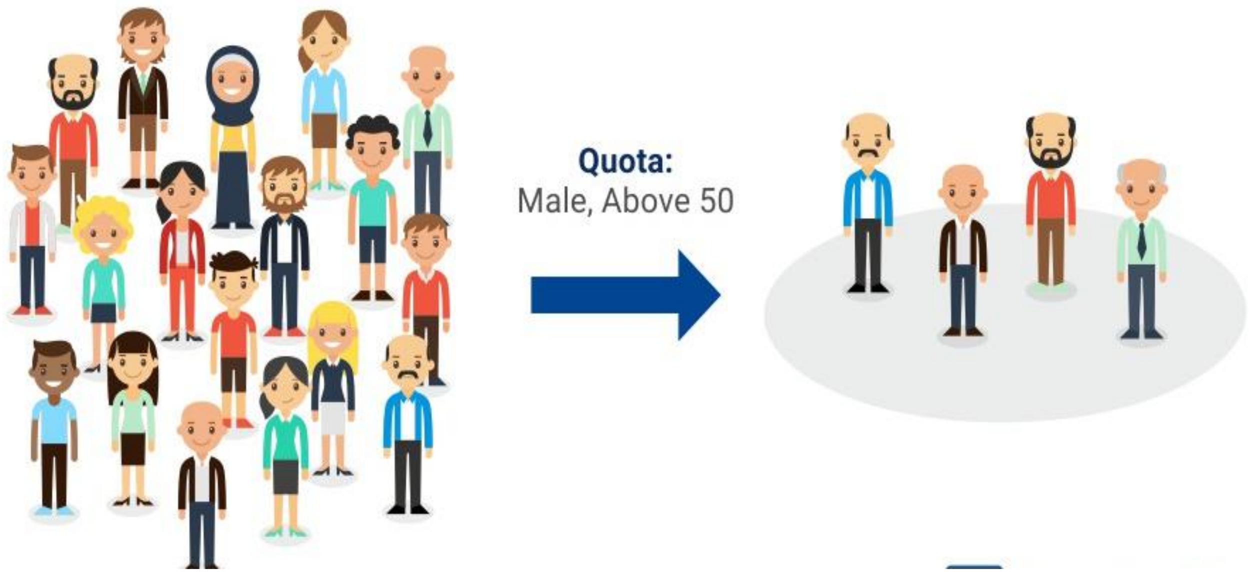
Judgmental (purposive)

You chose who you think should be in the study

Non-probability Sampling

Quota

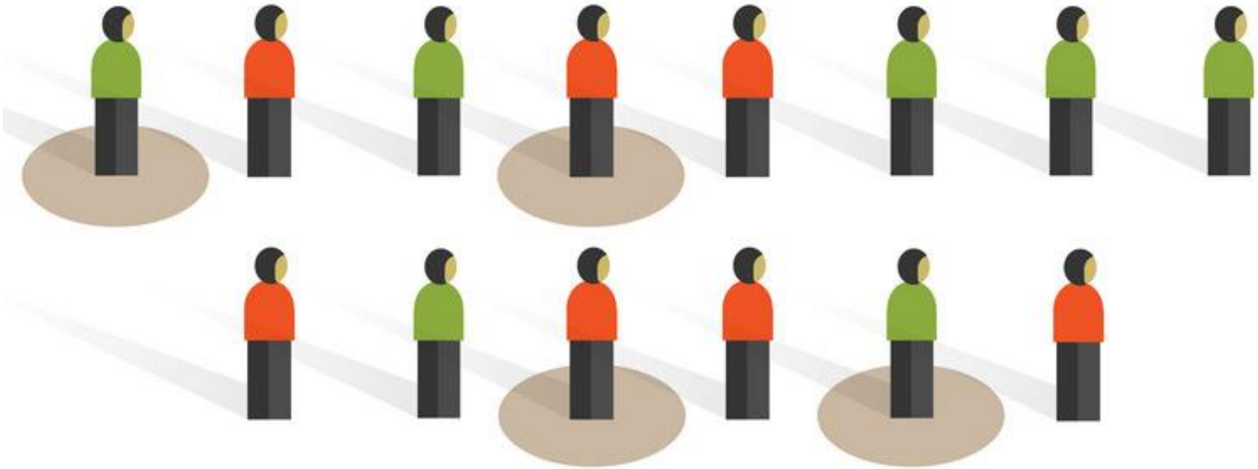
create a **sample** involving individuals that represent a population



Probability Sampling

Simple Random Sampling

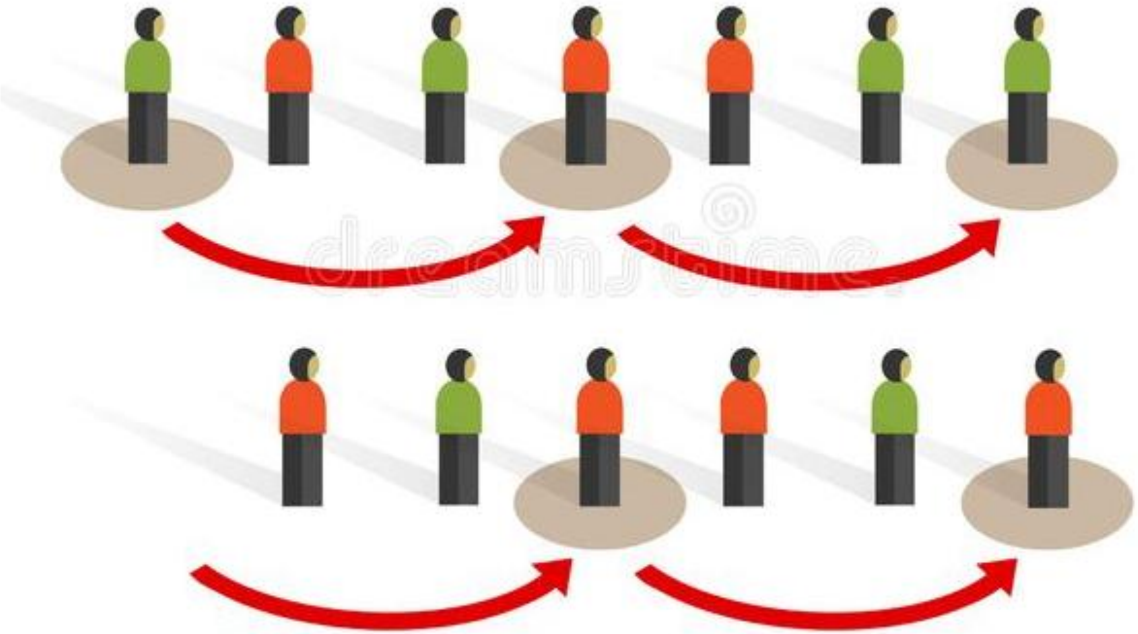
Each subject has a known probability of being selected



Probability Sampling

Systematic Sampling

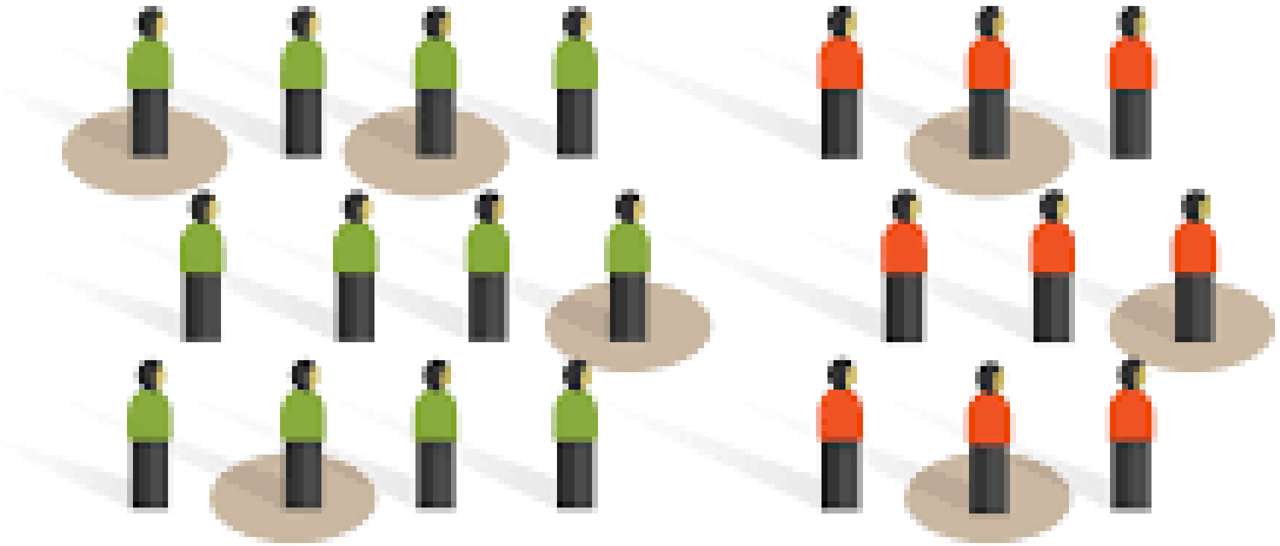
sampling fraction: Ratio between sample size and population size



Probability Sampling

Stratified Sampling

Draw a sample from each stratum



Probability Sampling

Cluster Sampling

Cluster: a group of sampling units close to each other

Probability Sampling

Cluster Sampling

- the population is very large
- impossible to construct an accurate frame
- the population is highly dispersed

Design effect: the ratio of variance with cluster sampling and variance with simple random sampling

$1 < deff < 2$

Sample Size Calculation

- Dispersion (variance) of variable
- Precision or error of estimation
- Confidence level (1-α)
- Power (1-β)
- Population size (N)
- Sampling method (Design Effect)

Sample size calculation

Precision:

$$\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

precision

$$\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + \underbrace{z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}}_{d}$$

$$\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + \underbrace{z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{d}$$

precision

Example:

Prevalence of diabetes estimated 0.2 in a research with 100 samples.

$$0.2-1.96\sqrt{\frac{0.2(1-0.2)}{100}} \le p \le 0.2+1.96\sqrt{\frac{0.2(1-0.2)}{100}}$$

$$0.1216 \le p \le 0.2784$$

$$precision = 0.2-0.1216 = 0.2784-0.2 = \boxed{0.0784}$$

Sample size calculation

- Mean (quantitative)
  - Mean estimation
  - Mean comparison

- Proportion (qualitative)
  - Proportion estimation
  - Proportion comparison

Sample size calculation

- Mean Estimation

$$d = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \frac{Z^2_{1-\frac{\alpha}{2}} \sigma^2}{d^2}$$

d= estimation error

σ^2= variance

α= error level

Mean Estimation

σ^2= ???

– Previous similar studies

– Researcher opinion

– Pilot study

– considering $\frac{1}{4}$ or $\frac{1}{6}$ of variation rang

as the standard deviation(sd)

$$SE = \frac{sd}{\sqrt{n}}$$

Mean Estimation

d= ???

- Researcher opinion
- Pilot study

- d< 30% of sd

Z table

α (significance level)	0.1	0.05	0.01
1-α (confidence level)	0.9	0.95	0.99
Z(1-α/2)	1.64	1.96	2.58

Mean Estimation

Example:

How many samples are needed for the estimation of the serum Cholesterol with 5% significance level and d=5? the standard deviation of the serum Cholesterol is equal to 40 in a similar study.

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}\sigma^2}{d^2}$$

$$n = \frac{1.96^2 \times 40^2}{5^2} = 245.9$$

Sample size calculation

- Mean comparison

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)}{d^2}$$

d= mean difference of two groups

σ^2= variance

α= error level

1-β= power of study

Z table

α (significance level)	0.1	0.05	0.01
1-α (confidence level)	0.9	0.95	0.99
Z(1-α/2)	1.64	1.96	2.58

1-β (power)	0.8	0.9	0.95
Z(1-β)	0.84	1.28	1.64

جدول شماره ۱۵

$$Z=(z_{1-\frac{\alpha}{2}}+z_{1-\beta})^{2}$$

$1-\beta$	α	۰/۲۰	۰/۱۰	۰/۰۵	۰/۰۲	۰/۰۱	۰/۰۰۲	۰/۰۰۱	۰/۰۰۰۱
	Z	۱/۲۸	۱/۶۴	۱/۹۶	۲/۳۳	۲/۵۸	۳/۰۹	۳/۲۹	۳/۸۹
۰/۵۰	۰/۰۰	۱/۶۴	۲/۷۱	۳/۸۴	۵/۴۳	۶/۶۶	۹/۵۵	۱۰/۸۲	۱۵/۱۳
۰/۶۰	۰/۲۵۳	۲/۳۵	۳/۵۸	۴/۹۰	۶/۶۷	۸/۰۳	۱۱/۱۸	۱۲/۵۵	۱۷/۱۶
۰/۷۰	۰/۵۲۵	۳/۲۶	۴/۶۹	۶/۱۸	۸/۱۵	۹/۶۴	۱۳/۰۷	۱۴/۵۵	۱۹/۴۹
۰/۷۵	۰/۶۷۵	۳/۸۴	۵/۳۶	۶/۹۴	۹/۰۳	۱۰/۶۰	۱۴/۱۸	۱۵/۷۲	۲۰/۸۴
۰/۸۰	۰/۸۴	۴/۵۰	۶/۱۷	۷/۸۴	۱۰/۰۲	۱۱/۶۷	۱۵/۴۴	۱۷/۰۶	۲۲/۳۸
۰/۸۵	۱/۰۴	۵/۳۸	۷/۱۸	۹/۰۰	۱۱/۳۶	۱۳/۱۰	۱۷/۰۶	۱۸/۷۵	۲۴/۳۰
۰/۹۰	۱/۲۸	۶/۵۷	۸/۵۷	۱۰/۵۱	۱۳/۰۲	۱۴/۸۸	۱۹/۱۱	۲۰/۹۱	۲۶/۷۶
۰/۹۵	۱/۶۴	۸/۵۷	۱۰/۸۲	۱۲/۹۹	۱۵/۷۷	۱۷/۸۲	۲۲/۴۲	۲۴/۳۶	۳۰/۶۵
۰/۹۸	۲/۰۵	۱۱/۰۹	۱۳/۶۲	۱۶/۰۸	۱۹/۱۸	۲۱/۴۴	۲۶/۴۲	۲۸/۵۲	۳۵/۲۸
۰/۹۹	۲/۳۳	۱۳/۰۲	۱۵/۷۷	۱۸/۳۷	۲۱/۶۴	۲۴/۰۳	۲۹/۳۳	۳۱/۵۵	۳۸/۶۵
۰/۹۹۸	۲/۸۸	۱۷/۳۱	۲۰/۴۳	۲۳/۴۳	۲۷/۱۴	۲۹/۸۱	۳۵/۶۴	۳۸/۰۷	۴۵/۸۳
۰/۹۹۹	۳/۰۹	۱۹/۱۱	۲۲/۴۲	۲۵/۵۰	۲۹/۳۳	۳۲/۱۰	۳۸/۱۹	۴۰/۷۲	۴۸/۷۳
۰/۹۹۹۹	۳/۷۱	۲۶/۷۶	۳۰/۶۵	۳۴/۲۳	۳۸/۶۵	۴۱/۸۲	۴۸/۷۳	۵۱/۵۸	۶۰/۵۶

Mean Comparison

Example

A researcher want to compare the mean of the blood pressure in two groups. The mean and sd of the blood pressure of two groups was obtained from a pilot study as follows. calculate the sample size of this study with $\alpha=0.05$ and $1-\beta=0.9$.

$$\mu_1 = 11, \mu_2 = 14$$

$$\sigma_1 = 2, \sigma_2 = 6$$

Mean Comparison

Example

A researcher want to compare the mean of the blood pressure in two groups. The mean and sd of the blood pressure of two groups was obtained from a pilot study as follows. calculate the sample size of this study with α=0.05 and 1-β=0.9.

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 (\sigma_1^2 + \sigma_2^2)}{d^2} = \frac{10.51 \times 40}{9} = 46.7$$

Sample size calculation

- Proportion Estimation

$$d = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \quad \Rightarrow \quad n = \frac{Z^2_{1-\frac{\alpha}{2}} p(1-p)}{d^2}$$

d= estimation error

p= proportion or prevalence of the variable

α= error level

Proportion Estimation

p= ???

- Previous similar studies
- Researcher opinion
- Pilot study
- p= 0.5

Proportion Estimation

d= ???

- Researcher opinion
- Pilot study

- d< 20% of *p*

Proportion Estimation

Example

A researcher want to estimate the prevalence of the Tuberculosis in Kermanshah province. The global prevalence was estimated 0.003. how many samples are needed for this goal with 1-α=0.95?

d= 0.15p

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}\, p(1-p)}{d^2} = \frac{1.96^2 \times 0.003 \times (1-0.003)}{(0.15 \times 0.003)^2} = 56742$$

Sample size calculation

- Proportion Comparison

$$n = \frac{\{Z_{1-\frac{\alpha}{2}}\sqrt{2\overline{p}(1-\overline{p})} + Z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\}^2}{d^2}$$

d= p_1-p_2

α= error level

1-β= power of study

$$\overline{p} = \frac{p_1 + p_2}{2}$$

Proportion Comparison

Example:

There are two surgical procedures for the treatment of a special disease. Based on a pilot study, the proportion of patients with side effects are 0.05 and 0.15 for each procedure, respectively. Calculate the sample size with α=0.05 and 1-β=90%.

$$n = \frac{\{Z_{1-\frac{\alpha}{2}}\sqrt{2\overline{p}(1-\overline{p})} + Z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\}^2}{d^2}$$

Proportion Comparison

Example:

$$\bar{p} = \frac{0.15 + 0.05}{2} = 0.1$$

$$n = \frac{\{1.96\sqrt{2 \times 0.1 \times 0.9} + 1.28\sqrt{(0.15 \times 0.85) + (0.05 \times 0.95)}\}^2}{(0.15 - 0.05)^2} = 186.9$$

Sample size calculation

- Unbalanced groups

$$n_2 = K \times n_1$$

- Mean

$$n_1 = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2 (\sigma_1^2 + \sigma_2^2 / K)}{d^2}$$

Sample size calculation

- Unbalanced groups

$$n_2 = K \times n_1$$

- Proportion

$$n_1 = \frac{\{Z_{1-\frac{\alpha}{2}}\sqrt{\bar{p}(1-\bar{p})(1+\frac{1}{K})} + Z_{1-\beta}\sqrt{p_1(1-p_1)+\frac{p_2(1-p_2)}{K}}\}^2}{d^2}$$

$$\bar{p} = \frac{p_1 + Kp_2}{1+K}$$

Unbalanced groups

Example

A researcher want to compare the mean of the blood pressure in two groups. The mean and sd of the blood pressure of two groups was obtained from a pilot study as follows. calculate the sample size of this study with $\alpha=0.05$ and $1-\beta=0.9$. the size of group2 be 2 times of group1

$$\mu_1 = 11, \mu_2 = 14$$

$$\sigma_1 = 2, \sigma_2 = 6$$

$$n_1 = \frac{10.51(4 + {}^{36}\!/_{2})}{9} = 25.7$$

$$n_2 = 2 \times 25.7 = 51.4$$

Unbalanced groups

Example:

There are two surgical procedures for the treatment of a special disease. Based on a pilot study, the proportion of patients with side effects are 0.05 and 0.15 for each procedure, respectively. Calculate the sample size with α=0.05 and 1-β=90%. the size of group2 be 4 times of group1

$$n=\frac{\left\{Z_{1-\frac{\alpha}{2}}\sqrt{\bar{p}(1-\bar{p})(1+\frac{1}{K})}+Z_{1-\beta}\sqrt{p_1(1-p_1)+\frac{p_2(1-p_2)}{K}}\right\}^2}{d^2}$$

$$\bar{p}=\frac{p_1+Kp_2}{1+K}$$

Unbalanced groups

Example

$$\bar{p} = \frac{p_1 + Kp_2}{1+K} = \frac{0.15 + (4\times 0.05)}{1+4} = 0.13$$

$$n_1 = \frac{\{1.96\sqrt{0.13(1-0.13)(1+\frac{1}{4})} + 1.28\sqrt{0.15(1-0.15) + \frac{0.05(1-0.05)}{4}}\}^2}{0.1^2} = 110$$

$$n_2 = 4\times 110 = 440$$

Sample size calculation

Limited population size (N)

$$\frac{n}{N} > 0.05$$

$$f.p.c = \frac{N - n}{N}$$

$$n' = f.p.c \times n = \frac{n}{1 + \frac{n}{N}}$$

Limited population size

Example:

How many samples are needed for the estimation of the serum Cholesterol with 5% significance level and d=5. the standard deviation of the serum Cholesterol is equal to 40 in a similar study? (the population size is 3000)

$$n = 246$$

$$\frac{246}{3000} = 0.082 > 0.05$$

$$n' = \frac{246}{1 + 0.082} = 228$$

Sample size calculation

Comparison of mean in multiple groups

$$n = \frac{\lambda_{g,\alpha,1-\beta}}{\Delta}$$

$$\Delta = \frac{1}{\sigma^2}\sum_{i=1}^{k}\left(\mu_i - \overline{\mu}\right)^2$$

$$\overline{\mu} = \frac{1}{k}\sum_{j=1}^{k}\mu_j$$

Comparison of mean in multiple groups

- جدول ۴.۱ ـ مقادیر $\lambda_{\alpha,1-\beta}$ بر حسب خطای آلفای ۱ و ۵ درصد و خطای بتای ۱۰ و ۲۰ درصد

g	$1-\beta=0.80$		$1-\beta=0.90$	
	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$
2	11.68	7.85	14.88	10.51
3	13.89	9.64	17.43	12.66
4	15.46	10.91	19.25	14.18
5	16.75	11.94	20.74	15.41
6	17.87	12.83	22.03	16.47
7	18.88	13.63	23.19	17.42
8	19.79	14.36	24.24	18.29
9	20.64	15.03	25.22	19.09
10	21.43	15.65	26.13	19.83
11	22.18	16.25	26.99	20.54
12	22.89	16.81	27.80	21.20
13	23.57	17.34	28.58	21.84
14	24.22	17.85	29.32	22.44
15	24.84	18.34	30.34	23.03
16	25.44	18.82	30.73	23.59
17	26.02	19.27	31.39	24.13
18	26.58	19.71	32.04	24.65
19	27.12	20.14	32.66	25.16
20	27.65	20.65	33.27	25.66

Comparison of mean in multiple groups

Example: $$\lambda_{4,\alpha=0.01,1-\beta=0.9} = 19.25$$

Comparison of mean in multiple groups

Example: $\lambda_{4,\alpha=0.01,1-\beta=0.9}=19.25$

$\mu_1 = 70\,mmHg$ $\mu_2 = 77\,mmHg$ $\mu_3 = 85\,mmHg$ $\mu_4 = 68\,mmHg$

$$\sigma^2 = 14^2\,mmHg$$

Comparison of mean in multiple groups

Example: $\lambda_{4,\alpha=0.01,1-\beta=0.9} = 19.25$

$\mu_1 = 70\,mmHg$ $\mu_2 = 77\,mmHg$ $\mu_3 = 85\,mmHg$ $\mu_4 = 68\,mmHg$

$$\sigma^2 = 14^2\,mmHg$$

$$\overline{\mu} = \frac{1}{k}\sum_{j=1}^{k}\mu_j \qquad \overline{\mu} = \frac{1}{4}(70+77+85+68) = 75$$

$$\Delta = \frac{1}{\sigma^2}\sum_{i=1}^{k}\left(\mu_i - \overline{\mu}\right)^2 = \frac{1}{14^2}\left[(70-75)^2 + (77-75)^2 + (85-75)^2 + (68-75)^2\right] = 0.908$$

Comparison of mean in multiple groups

Example: $$\lambda_{4,\alpha=0.01,1-\beta=0.9} = 19.25$$

$$\mu_1 = 70\,mmHg \qquad \mu_2 = 77\,mmHg \qquad \mu_3 = 85\,mmHg \qquad \mu_4 = 68\,mmHg$$

$$\sigma^2 = 14^2\,mmHg$$

$$\overline{\mu} = \frac{1}{k}\sum_{j=1}^{k}\mu_j \qquad \overline{\mu} = \frac{1}{4}(70+77+85+68) = 75$$

$$\Delta = \frac{1}{\sigma^2}\sum_{i=1}^{k}\left(\mu_i - \overline{\mu}\right)^2 = \frac{1}{14^2}\left[(70-75)^2 + (77-75)^2 + (85-75)^2 + (68-75)^2\right] = 0.908$$

$$n = \frac{\lambda_{g,\alpha,1-\beta}}{\Delta} = \frac{19.25}{0.908} = 21.2 = 22$$

Sample size calculation

- Stratified sampling
  - Proportional allocation

$$n_h = \frac{N_h}{N} \times n$$

Sample size calculation

- Stratified sampling
  - Proportional allocation

$$n_h = \frac{N_h}{N} \times n$$

$$N_1 + N_2 + N_3 = N$$

Sample size calculation

- Stratified sampling
  - – Proportional allocation

$$n_h = \frac{N_h}{N} \times n$$

$$N_1 + N_2 + N_3 = N$$

$W_1 = N_1/N$ → $W_2 = N_2/N$ → $W_3 = N_3/N$

Sample size calculation

- Stratified sampling
  - Proportional allocation

$$n_h = \frac{N_h}{N} \times n$$

$N_1 + N_2 + N_3 = N$

$W_1 = N_1/N$ → $W_2 = N_2/N$ → $W_3 = N_3/N$

n ⇒ $n_1 = W_1 \times n$ + $n_2 = W_2 \times n$ + $n_3 = W_3 \times n$ = n

Stratified sampling

Example

A researcher believes that the prevalence of obesity in COVID-19 patients is about 0.2. He want to investigate 3 cities A, B, and C for estimating the prevalence of obesity in COVID-19 patients. The population size of each city is 2000, 3000, and 5000, respectively. Calculate the sample size of each stratum with $d=0.03$ and $\alpha=0.05$ by proportional allocation.

Stratified sampling

Example

$$n = \frac{1.96^2 \times 0.2 \times 0.8}{0.03^2} = 682.95$$

City	Population size (N_h)	$W_h=N_h/N$	$n_h=W_h \times n$
A	2000	0.2	137
B	3000	0.3	205
C	5000	0.5	342
Total	10000	1	684

Sample size calculation

- Cluster sampling

$1<deff<2$

$$n' = deff \times n$$

Sample size calculation

- Cluster sampling

  $1 < \mathit{deff} < 2$

$$n' = \mathit{deff} \times n$$

Example:

n=1000

deff=1.5

Sample size calculation

- Cluster sampling

$1<deff<2$

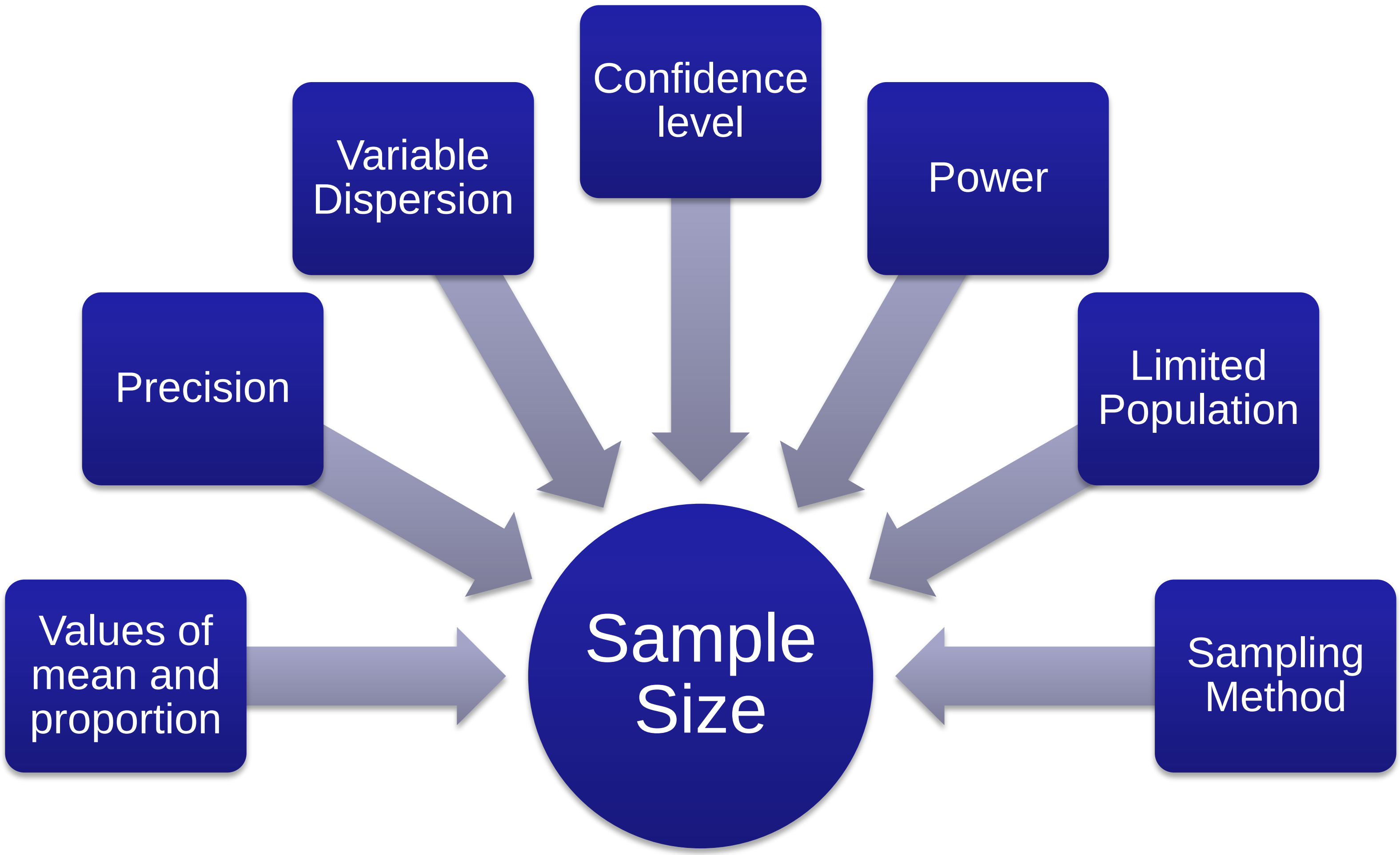
$$n' = deff \times n$$

Example:

$$n' = 1.5 \times 1000 = 1500$$

n=1000

deff=1.5


Confidence level
Variable Dispersion
Power
Precision
Limited Population
Values of mean and proportion
Sample Size
Sampling Method

Thank You