Scientific Data Analysis for Post-Graduate Students Using R Programming Language.

Strengthening Research skills in Eastern and Southern Africa

RUFORUM
Capacity Building in Agriculture

MAKERERE UNIVERSITY

OVERVIEW OF SAMPLING SURVEYS

Facilitators:

Prof. Susan Balaba Tumwebaze

Dr. Thomas Odong

Dr. Hellen Namawejje

Overview of Sampling Surveys

Introduction to Sampling Method

Sample survey

- The main purpose of ***population survey*** is to draw an inference about a population using a subset of population, known as ***sample***.

- The inference is about a characteristics of the population: mortality, income, behavior, opinion, etc.

Sample surveys

- A good survey aims to select samples in such a way that they correctly represent the underlying population, i.e., *the sample characteristics are similar to population* characteristics. In other words, there is no bias in selection.

- This is usually done through a random process, which ensures no preferential treatment to any specific subject

Sample survey

- Nevertheless, some errors occur in sample surveys.

- One error – sampling error – is unavoidable. Because all elements of population are not selected, it is possible that the sample based estimated value may not exactly match with the population value.

- The discrepancy - difference between the estimated and true population values – is captured by ***sampling error/ margin of error***.

Sample surveys

- Given our resources (funding, time) , the survey sampling design is selected in such a way to minimize sampling errors.

- A population survey with efficient design has ***low sampling error***. With sample size calculation, we ensure that the survey's margin-of-error will be limited to an acceptable level (e.g., +/- 5%).

Point estimate

Study Population

An opinion poll on Uganda’s health concern was conducted by Gallup Organization between October 3-5, 2020, and the survey reported that 29% adults consider AIDS is the most urgent health problem of the US, with a *margin of error* of +/- 3%. The result was based on telephone interviews of 872 adults.

Sample Size (n)

Method of Survey Administration

Extent of Sampling Error

The outcome variable is **Bernoulli random variable.**

A binary variable have only yes/no category of responses.

"Do you consider AIDS is the most urgent health problem of the Uganda?"

29% responded "yes", and 71% responded "no".

Let p = 0.29, and q = 1 – p = 0.71.

So, 872 X 0.29 = 253 responded "Yes", and
872 X 0.71 = 619 responded "No".

In summary, from the raw data, Gallup's statistician *estimated* that

$$p = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum (1+0+1+1+.........+0)}{872} = \frac{253}{872} = 0.29$$

Here, 1 = "yes", and 0 = "no" responses.

How much confidence do we have on this "point estimate" (29%) ?

From our knowledge of basic statistics, we can construct a 95% confidence interval around p as:

$$\hat{p} \pm Z_{.05} * se(\hat{p})$$

That is,

$$\hat{p} \pm 1.96\sqrt{\frac{pq}{n}}$$

$$.29 \pm 1.96\sqrt{\frac{(.29)(.71)}{872}}$$

$$= .29 \pm 0.03$$

So, 95% CI of *p* ranges between (.26 to .32).

$$\hat{p} \pm Z_{.05} * se(\hat{p})$$

The above mathematical expression could be rephrased as:

$$estimate \pm margin\_of\_error$$

(29%) (3%)

$$margin\_of\_error = 1.96\sqrt{\frac{(.29)(.71)}{872}} = 0.03$$

$$margin\_of\_error^2 = 1.96^2 \frac{(.29)(.71)}{872}$$

$$872 = 1.96^2 \frac{(.29)(.71)}{margin\_of\_error^2}$$

$$n = 1.96^2 \frac{pq}{d^2}$$

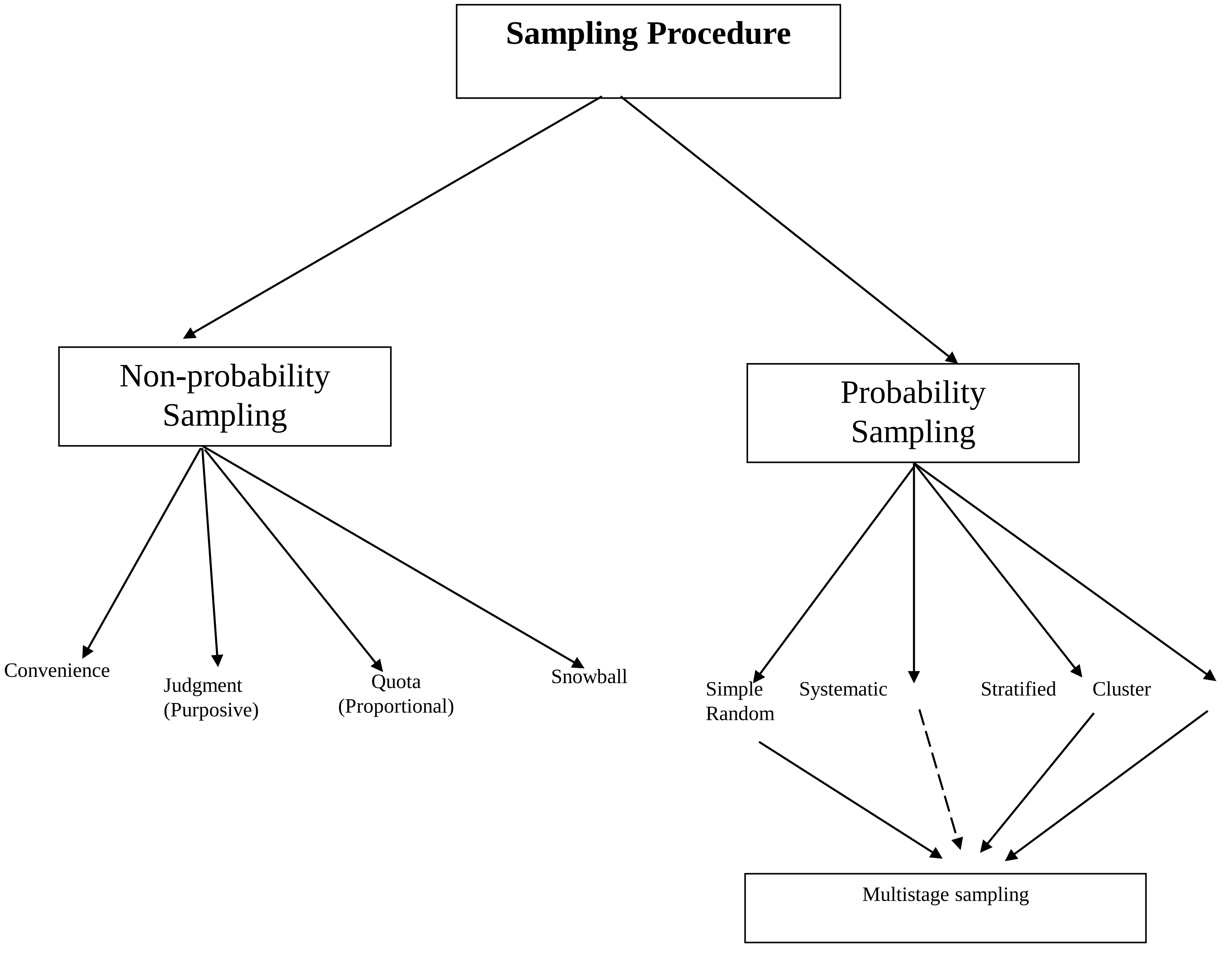
This example also shows that:

1. if we know/guess an estimated value (p), we can estimate the required sample size with a specified “margin of error”.

2. Help us to determine whether we need a large or small sample size.

Types of survey design

- Cross-sectional surveys: Data are collected at a point of time.
- Longitudinal surveys:
  - Trends: surveys of sample population at different points in time
  - Cohort: study of same population each time over a period (open cohort/closed cohort)
  - Panel: Study of sample of respondents at various time points.


Sampling Procedure
Non-probability Sampling
Probability Sampling
Convenience
Judgment (Purposive)
Quota (Proportional)
Snowball
Simple Random
Systematic
Stratified
Cluster
Multistage sampling

Characteristics of Good Survey Sampling

- Meets the requirements of the study objectives
- Provides reliable results
- The source is known – target population, probability of selection, missing responses
- Manageable/realistic: could be implemented
- Time consideration: reasonable and timely
- Cost consideration: economical
- Acceptable – not biased

Sampling Issues in Survey

- The sampling procedure produces best estimation
- Sample size: Not too low, not too large
- Minimum error/unbiased estimation
- Economic consideration
- Design consideration: best collection strategy

Methods in Sample Surveys

Simple Random Sampling

Simple Random Sampling (SRS)

- We employ some *randomization* process for sample selection so that there is no preferential treatment in selection that may introduce selectivity

- Simple random sampling (SRS) is simplest among sampling choices

- In SRS, each element has an equal probability of being selected from a list of all population units (sample of **n** from **N** population).

Systematic Sampling

"…systematic sampling, either by itself or in combination with some other method, may be the most widely used method of sampling."

Levy and Lemeshow, 1999

Systematic Sampling

- “Systematic sampling is perhaps the most widely known selection procedure.” - Leslie Kish, 1965
- An alternative method for random sampling

- In systematic sampling, only the first unit is selected at random,
- The rest being selected according to a predetermined pattern.
- to select a systematic sample of *n* units, the first unit is selected with a random start *r* from 1 to k sample, where $k=N/n$ sample intervals, k is the systematic sampling interval.

  and after the selection of first sample, every k^{th} unit is included where $1 \le r \le k$.

An example:

Let Population size(N)=100, sample size(n)=10, then k=100/10.

Then the random start *r* is selected between 1 and 10 (say, r=7).

So, the sample will be selected from the population with serial indexes of:

7, 17,27,.........,97

i.e., r, r+k, r+2k,......., r+(n-1)k

Systematic

N = 100

want n = 20

N/n = 5

select a random number from 1-5:
chose 4

start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Stratified Sampling

Stratified Sampling

In stratified sampling the population is *partitioned into groups*, called *strata*, and sampling is performed separately within each *stratum*.

In stratified sampling designs:

- stratum variables are *mutually exclusive* (non-over lapping), e.g., urban/rural areas, economic categories, geographic regions, race, sex, etc.

- the population (elements) should be *homogenous* within-stratum, and

- the population (elements) should be *heterogenous* between the strata.

The principal objective of stratification is to reduce sampling errors.

Two Basic Rules of Stratified Sampling

- A minimum of two-elements must be chosen from each stratum so that sampling errors can be estimated for all strata independently.
- The population (elements) should be *homogenous* within stratum, and the population (elements) should be *heterogenous* between the strata.

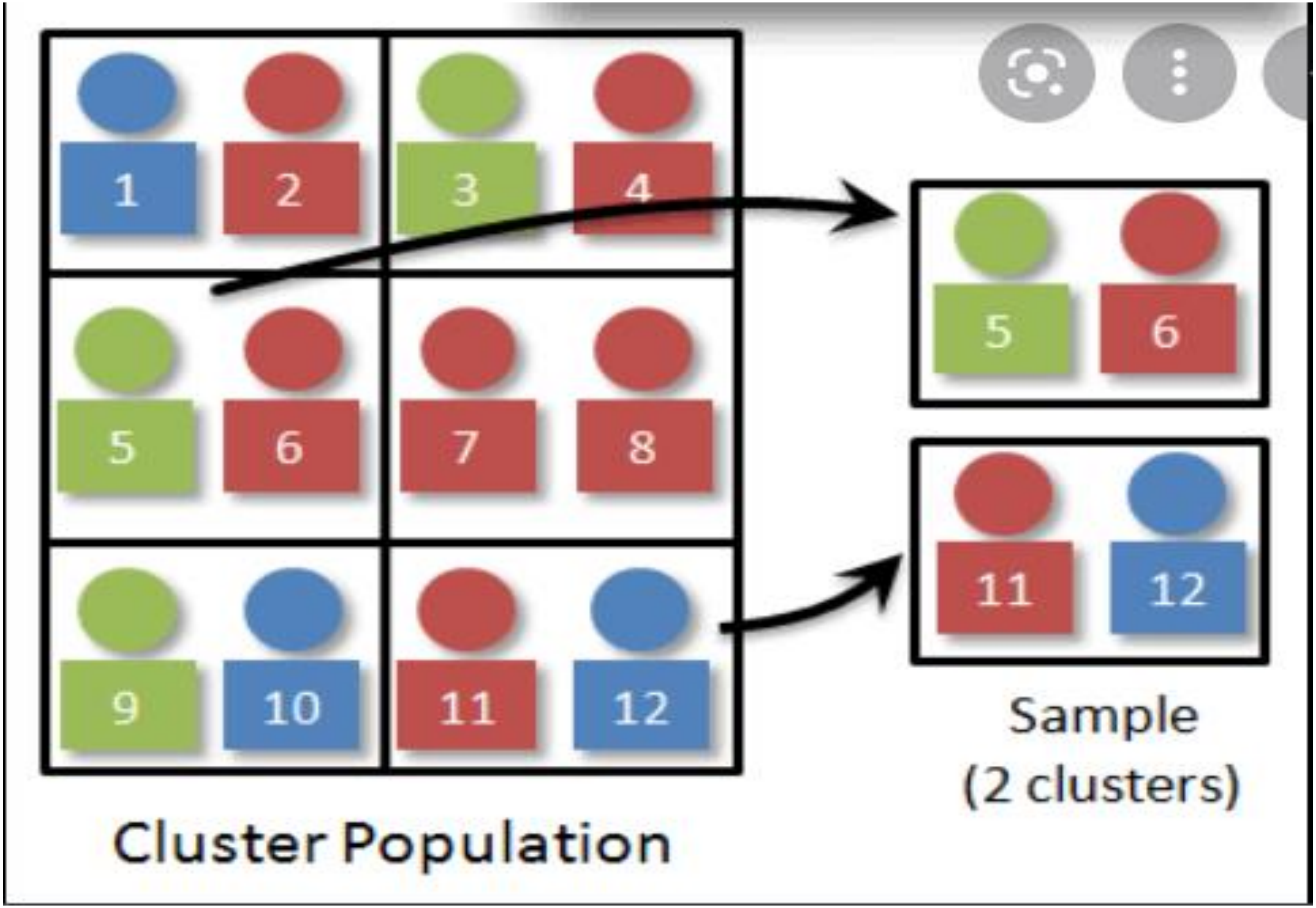
Cluster Sampling

Definition

In cluster sampling, <u>cluster,</u> i.e., a group of population elements<u>, constitutes the sampling unit</u>, instead of a single element of the population.

In cluster sampling, <u>clusters </u>are the first sampling units.

Cluster Sampling

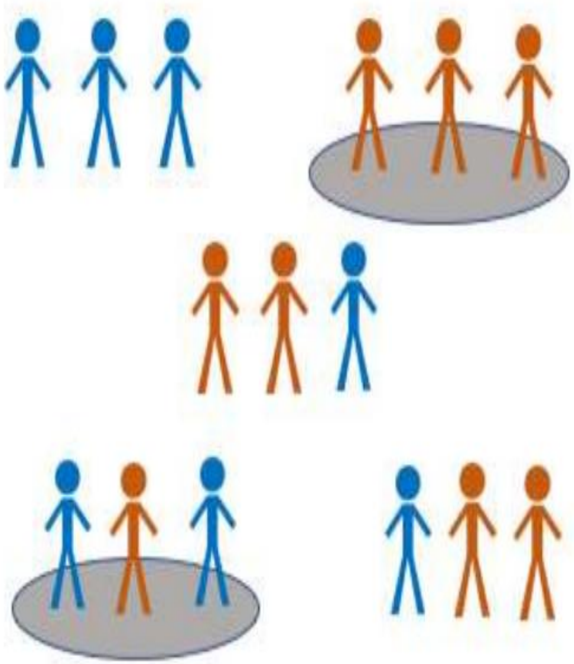


Cluster

Example: 10 schools have the same number of students across the county. We can randomly select 3 out of 10 schools as our clusters

Advantages: Readily doable with most budgets, does not require a sampling frame

Disadvantages: Results may not be reliable nor generalisable



Sampling selection procedure

- Primary sampling units (PSU): clusters
  - select the PSU's by using a specific *element* sampling techniques, such as simple random sampling, systematic sampling or by PPS sampling.
- Secondary sampling units (SSU): households/individual elements
  - select **all** SSU's for convenience, or
  - select **few** by using a specific element sampling techniques (such as simple random sampling, or systematic sampling).

Simple one-stage cluster sampling

- List all the clusters in the population, and from the list,
- select the clusters – usually with simple random sampling (SRS) strategy.
- **All units** (elements) in the sampled clusters are selected for the survey

Simple two-stage cluster sampling

- List all the clusters in the population.
- First, select the clusters, usually by simple random sampling (SRS).
- The elements in the selected clusters of the first-stage are then sampled in the second-stage, usually by simple random sampling (or often by systematic sampling).

Multi-stage sampling

- Sampling is done in more than one stage (step): select clusters → select households
- In practice, clusters are also stratified.
- **Stratified + Cluster + SRS/systematic = multistage sampling**
  - **[e.g., stratify the country into urban/rural, draw cluster sampling within each stratum independently, and finally select households from each cluster]**
- National level surveys are usually performed with multistage sampling design

Multi-stage sampling-Example

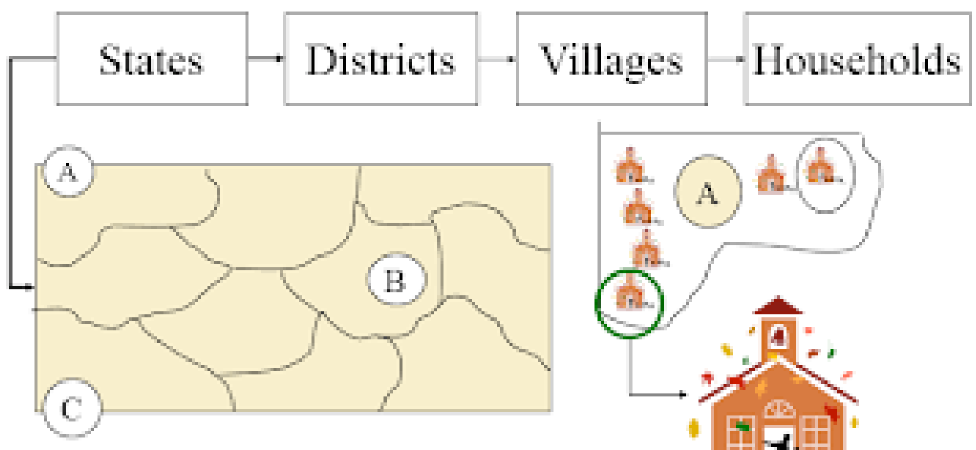
Multistage or Hierarchical Sampling

Several layers of the population

- Region – District – Village – Households

- It often combines several sampling methods (SRS, Stratified) depending on the situation at each level

-Practical reasons e.g. absence of good sampling frame for probability sampling



How to select clusters/PSUs? Two common procedures:

- Simple random sampling: all PSUs have same selection probability (*equal probability method*)
- Probability proportional to size (PPS): the probability of selection depends on the size of the PSU, e.g., larger the size of the PSUs, larger are the probabilities of selection (*unequal probability method*).

Probability proportional to size (PPS)

PPS is usually the preferred method of selecting clusters in cluster sampling design primarily for three reasons:

- Larger areas are more likely to be selected; larger areas means larger representation of the population: better representation, less costly

- Secondary Selection Units (SSUs) have same selection probability

Non-probability sampling techniques

Non-probability sampling techniques

Convenient Sampling

- Selection is often subjective (not probability sampling)

- It help researcher avoid the problem of sampling frame

Non-probability sampling techniques

Judgmental/purposive/authoritative sampling

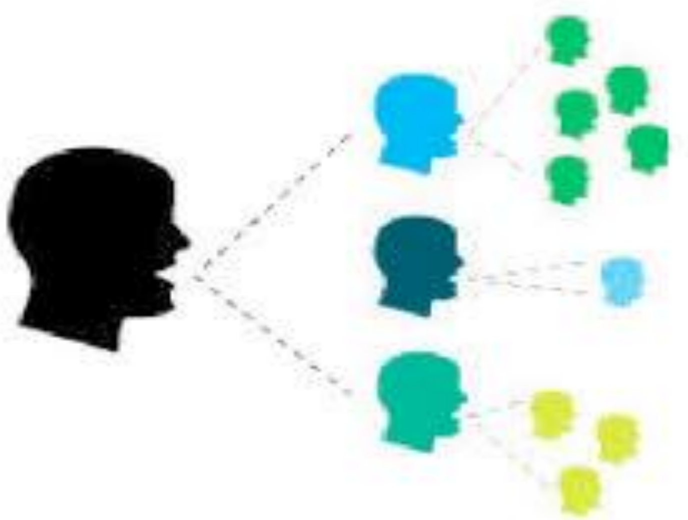
- Group people/elements are selected because they posses specific characteristics of interest to the researchers (villages known for growing bananas in Busoga) .

- Selection is based on the judgment of the researcher

Non-probability sampling techniques

Snowball sampling

- Existing subject recruit future subjects
- Used for hidden populations – e.g. drug users, prostitutes, etc



Ethical issues in sampling surveys

- Prior informed consent
- Sensitive topics
- Confidentiality
- Legal requirements
- Feedback to data providers

Sample Size and Power Estimation

In planning of a sample survey, a stage is always reached at which a decision must be made about the size of the sample. The decision is important. **Too large** a sample implies a waste of resources, and **too small** a sample diminishes the utility of the results.

Cochran, 1977

Sample size estimation

- What is the appropriate sample size for my study?

- What is the size of the sample which I can use to generalize the results on the population which I am going to study?

Possible Answers

- Purpose of your study
- The population size
- Risk of selecting a BAD sample
- Allowable sampling error

Sample Size Criteria

- According to Miaoulis & Michener, (1976), in addition to purpose of the study and population size
- The 3 criteria's usually needed to determine an appropriate sample size are

The 3 criteria's include

- Level of Precision
- Level of confidence or risk
- Degree of variability in attributes being measured

Level of precision

- Also known as the sampling error/ margin of error

- It is the range in which the true value of the population is estimated.

- This range is often expressed as percentage (±5%)

Example

- If a researcher finds that 70% of the students in the sample have adopted a recommended practice of submitting the assignment with a precision rate of ±5%, then he or she can conclude that between 65% and 75% of the students in the population have adopted the practice

Confidence interval

- Also known as risk level

- Based on the Central Limit Theorem, which means, when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value

- This is expressed in % points e.g. 95%

For example

If a 95% CI is selected, 95 out of 100 samples will have a true population value within the range of precision specified earlier.

Degree of Variability

- Refers to distribution of attributes in the population
- The more heterogeneous the population, the the larger the sample size required to obtain a given level of precision

- The less variable (more homogeneous) a population, the smaller the sample size

Strategies for Determining Sample Size

Strategies for determining sample size

- Using a census for small population
    - Use the entire population as sample
    - Have to consider cost and time
    - Usually used when population size is 200 or less
- Other sample size from similar studies (Literature view)
- Using published data
    - Published tables available e.g. (Yamane, 1967) formula

Strategies for determining sample size

- Using formulas
  - Cochran formula
  - Yamane formula
  - Formula for sample size for the mean or proportion

- Cochran and Yamane formulas are the commonly used formulas

Using formula for sample size for mean or proportion

Formula for sample size for mean

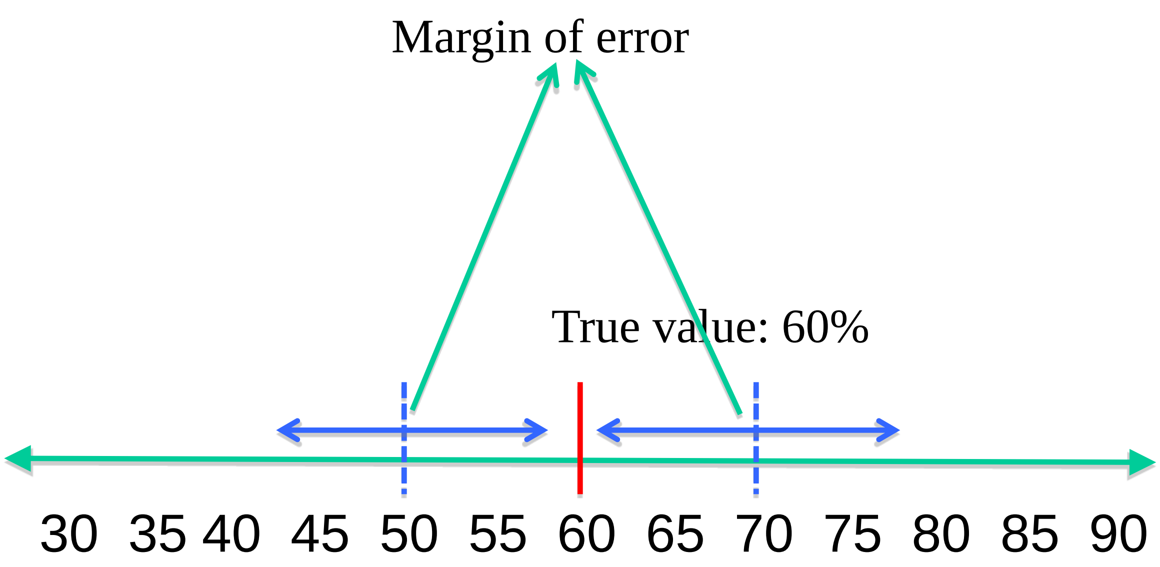
- According to Yamane (1967;86)

$$n = \frac{Z^2\sigma^2}{e^2}$$

- n is sample size, σ^2 is population variance
- e^2 is margin of error

Formula for sample size for proportion

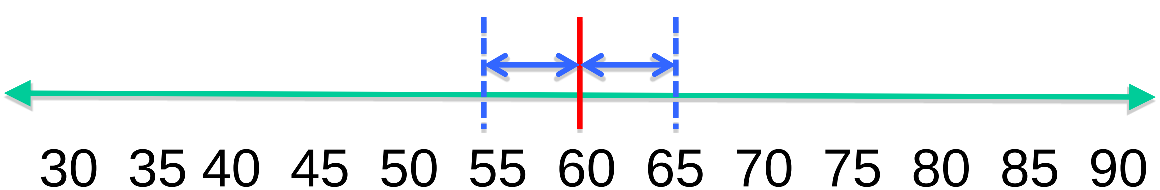
An example:
What proportion of women had skilled birth attendants (SBA) during their last delivery?



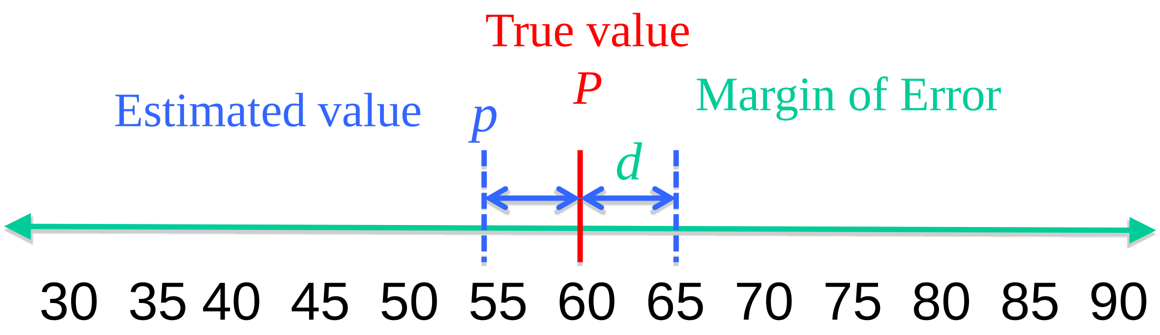
An example:
What proportion of women had skilled birth attendants (SBA) during their last delivery?

Smaller the Margin of Error- More we trust the value

True value: 60%



An example:
What proportion of women had skilled birth attendants (SBA) during their last delivery?



Stated mathematically:

- we want a sample size to ensure that we can estimate a value , say, ***p*** from a sample which corresponds to the population parameter, ***P***.

- Since we may not guarantee that p will be exact to ***P***, we allow some level of error

- Error level is limited to certain extent, that is this error should not exceed some specified limit, say ***d***.

- We may express this as:

  p - P = ± d,

  i.e., the difference between the estimated p and true ***P*** is not greater than ***d*** (allowable error: margin-of-error)

- But do we have any confidence that we can get a ***p***, that is not far away from the error of ***±d***?

- In other words, we want some confidence limits, say 95%, to our error estimate d.

  That is ***1-α*** = 95%

It is a common practice: α-error = 5%

From our basic statistical course, we know that we can construct a confidence interval for p by:

$$p \pm z_{1-\alpha/2} * se(p)$$

where z_{α} *denotes a value on the abscissa of a standard normal distribution (from an assumption that the sample elements are normally distributed) and* *se(p) = σ_p* is the standard error.

$$p \pm d = p \pm z_{1-\alpha/2}\sigma_p$$

Hence, we relate p $\pm$ d in probabilities such that:

$$\boxed{\begin{aligned} d &= Z_{1-\alpha/2}\sigma \\ &= Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \end{aligned}}$$

If we square both sides,

$$\boxed{\begin{aligned} d &= Z_{1-\alpha/2}\sigma \\ &= Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \end{aligned}}$$

$$\downarrow$$

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

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

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

Sample size calculation for point estimates (mean):

$$n = \frac{Z^2_{1-\alpha/2}\,p(1-p)}{d^2}$$

n = sample size

$Z_{0.05}$ = 1.96

p= proportions

d = margin of error

For the above example:

$$n = \frac{(1.96)^2 * 0.4 * 0.6}{(.10)^\wedge 2} = 92.2 \approx 93$$

Note that, the sample size requirement is highest when p=0.5. It is a common practice to take p=0.5 when no information is available about p for a conservative estimation of sample size.

As an example, p = 0.5, d = 0. 05
(5% margin-of-error), and α-error = 0.05:

$$n = \frac{(1.96)^2 * 0.5 * 0.5}{(.05)\text{^}2} = 384.16 = 385 \approx 400$$

Factors That Influence Sample Size Calculations

- Desired level of significance (*alpha*)
- Desired power (*1-beta*)
- The smallest difference
  - Smallest practically important difference
  - The difference that investigators think is worth detecting
  - The difference that investigators think is likely to be detected
- Justification of previous data
  - Published data
  - Previous work
  - Review of records
  - Expert opinion
- Software or formula being used

It is important to note that:

- Increasing the margin of error would reduce the sample size, it is always a trade off

- Reducing the margin of error would increase the sample size.

Formulae

Table 1: Formulae for Sample Size Calculations for Comparisons Between Means

Design	Hypothesis	H_0	H_a	Basic Rule
One-sample	Equality	$\mu-\mu_0=0$	$\mu-\mu_0\neq 0$	$n=\frac{\left(z_{\frac{\alpha}{2}}+z_\beta\right)^2\sigma^2}{(\mu-\mu_0)^2}$
	Superiority	$\mu-\mu_0\leq\delta$	$\mu-\mu_0>\delta$	$n=\frac{\left(z_\alpha+z_\beta\right)^2\sigma^2}{(\mu-\mu_0-\delta)^2}$
	Equivalence	$\lvert\mu-\mu_0\rvert\geq\delta$	$\lvert\mu-\mu_0\rvert<\delta$	$n=\frac{\left(z_\alpha+z_\beta\right)^2\sigma^2}{(\lvert\mu-\mu_0\rvert-\delta)^2}$
Two-sample Parallel	Equality	$\mu_1-\mu_2=0$	$\mu_1-\mu_2\neq 0$	$n_i=\frac{2\left(z_{\frac{\alpha}{2}}+z_\beta\right)^2\sigma^2}{(\mu_1-\mu_2)^2}$
	Non-inferiority	$\mu_1-\mu_2\geq\delta$	$\mu_1-\mu_2<\delta$	$n_i=\frac{2\left(z_\alpha+z_\beta\right)^2\sigma^2}{(\mu_1-\mu_2-\delta)^2}$
	Superiority	$\mu_1-\mu_2\leq\delta$	$\mu_1-\mu_2>\delta$	$n_i=\frac{2\left(z_\alpha+z_\beta\right)^2\sigma^2}{(\mu_1-\mu_2-\delta)^2}$
	Equivalence	$\lvert\mu_1-\mu_2\rvert\geq\delta$	$\lvert\mu_1-\mu_2\rvert<\delta$	$n_i=\frac{2\left(z_\alpha+z_\beta\right)^2\sigma^2}{(\lvert\mu_1-\mu_2\rvert-\delta)^2}$
Two-sample Crossover	Equality	$\mu_1-\mu_2=0$	$\mu_1-\mu_2\neq 0$	$n_i=\frac{\left(z_{\frac{\alpha}{2}}+z_\beta\right)^2\sigma^2}{2(\mu_1-\mu_2)^2}$
	Non-inferiority	$\mu_1-\mu_2\geq\delta$	$\mu_1-\mu_2<\delta$	$n_i=\frac{\left(z_\alpha+z_\beta\right)^2\sigma^2}{2(\mu_1-\mu_2-\delta)^2}$
	Superiority	$\mu_1-\mu_2\leq\delta$	$\mu_1-\mu_2>\delta$	$n_i=\frac{\left(z_\alpha+z_\beta\right)^2\sigma^2}{2(\mu_1-\mu_2-\delta)^2}$
	Equivalence	$\lvert\mu_1-\mu_2\rvert\geq\delta$	$\lvert\mu_1-\mu_2\rvert<\delta$	$n_i=\frac{\left(z_\alpha+z_\beta\right)^2\sigma^2}{2(\lvert\mu_1-\mu_2\rvert-\delta)^2}$

Table 2: Formulae for Sample Size Calculations for Comparisons Between Proportions

Design	Hypothesis	H_0	Basic Rule
One-sample	Equality	$\pi-\pi_0=0$	$n=\frac{\left(z_{\frac{\alpha}{2}}+z_\beta\right)^2\pi(1-\pi)}{(\pi-\pi_0)^2}$
	Superiority	$\pi-\pi_0\leq\delta$	$n=\frac{\left(z_\alpha+z_\beta\right)^2\pi(1-\pi)}{(\pi-\pi_0-\delta)^2}$
	Equivalence	$\lvert\pi-\pi_0\rvert\geq\delta$	$n=\frac{\left(z_\alpha+z_\beta\right)^2\pi(1-\pi)}{(\lvert\pi-\pi_0\rvert-\delta)^2}$
Two-sample Parallel	Equality	$\pi_1-\pi_2=0$	$n_i=\frac{\cancel{2}\left(z_{\frac{\alpha}{2}}+z_\beta\right)^2(\pi_1(1-\pi_2)+\pi_2(1-\pi_2))}{(\pi_1-\pi_2)^2}$
	Non-inferiority	$\pi_1-\pi_2\geq\delta$	$n_i=\frac{\cancel{2}\left(z_\alpha+z_\beta\right)^2(\pi_1(1-\pi_2)+\pi_2(1-\pi_2))}{(\pi_1-\pi_2-\delta)^2}$
	Superiority	$\pi_1-\pi_2\leq\delta$	$n_i=\frac{\cancel{2}\left(z_\alpha+z_\beta\right)^2(\pi_1(1-\pi_2)+\pi_2(1-\pi_2))}{(\pi_1-\pi_2-\delta)^2}$
	Equivalence	$\lvert\pi_1-\pi_2\rvert\geq\delta$	$n_i=\frac{\cancel{2}\left(z_\alpha+z_\beta\right)^2(\pi_1(1-\pi_2)+\pi_2(1-\pi_2))}{(\lvert\pi_1-\pi_2\rvert-\delta)^2}$
Two-sample Crossover	Equality	$\pi_1-\pi_2=0$	$n_i=\frac{\left(z_{\frac{\alpha}{2}}+z_\beta\right)^2\sigma_d^2}{2(\pi_1-\pi_2)^2}$
	Non-inferiority	$\pi_1-\pi_2\geq\delta$	$n_i=\frac{\left(z_\alpha+z_\beta\right)^2\sigma_d^2}{2(\pi_1-\pi_2-\delta)^2}$
	Superiority	$\pi_1-\pi_2\leq\delta$	$n_i=\frac{\left(z_\alpha+z_\beta\right)^2\sigma_d^2}{2(\pi_1-\pi_2-\delta)^2}$
	Equivalence	$\lvert\pi_1-\pi_2\rvert\geq\delta$	$n_i=\frac{\left(z_\alpha+z_{\beta/2}\right)^2\sigma_d^2}{2(\lvert\pi_1-\pi_2\rvert-\delta)^2}$

Statistical Terms

- *P-value:* Probability of obtaining an effect as extreme or more extreme than what is observed by chance
- *Significance level of a test (**alpha**):* cut-off point for the p-value (conventionally it is 5%)
- *Power of a test (**1-beta**):* correctly reject the null hypothesis when there is indeed a real difference or association (typically set at least 80%), i.e., prob of finding an effect that is there
- Effect size of clinical importance

Mathematical Formulae

$$\text{effect size}(ES) = \frac{(\mu_1 - \mu_2)}{\sigma}$$

- μ_1 is mean of group 1; μ_2 is mean of group 2

- σ^2 is the common error variance

- Cohen suggest that ES values of 0.2, 0.5 and 0.8 represent small, medium and large effect sizes respectively

Power Estimation

Power estimation

- It allows us to determine the sample size required to detect an effect of a given size
- The **power of a test** is the probability of rejecting the null hypothesis if it is false

The following **four quantities** have an intimate relationship:

1. sample size
2. effect size
3. significance level = P(Type I error) = probability of finding an effect that is not there
4. power = 1 - P(Type II error) = probability of finding an effect that is there

Given any three, we can determine the fourth.

Values of $Z_{1-\alpha/2}$ and Z_β corresponding to specified values of significance level and power

	Values	Two-sided	One-sided
Level (alpha)	1% 5% 10%	2.576 1.960 1.645	2.326 1.645 1.282
Power	80% 90% 95% 99%	0.84 1.282 1.645 2.326	

Power Analysis in R

pwr-package	Basic Functions for Power Analysis pwr
cohen.ES	Conventional effects size
ES.h	Effect size calculation for proportions
ES.w1	Effect size calculation in the chi-squared test for goodness of fit
ES.w2	Effect size calculation in the chi-squared test for association
plot.power.htest	Plot diagram of sample size vs. test power
pwr	Basic Functions for Power Analysis pwr
pwr.2p.test	Power calculation for two proportions (same sample sizes)
pwr.2p2n.test	Power calculation for two proportions (different sample sizes)
pwr.anova.test	Power calculations for balanced one-way analysis of variance tests
pwr.chisq.test	power calculations for chi-squared tests
pwr.f2.test	Power calculations for the general linear model
pwr.norm.test	Power calculations for the mean of a normal distribution (known variance)
pwr.p.test	Power calculations for proportion tests (one sample)
pwr.r.test	Power calculations for correlation test
pwr.t.test	Power calculations for t-tests of means (one sample, two samples and paired samples)
pwr.t2n.test	Power calculations for two samples (different sizes) t-tests of means

Power Analysis in R

The pwr package develped by Stéphane Champely, impliments power analysis as outline by Cohen (!988). Some of the more important functions are listed below.

function	power calculations for
pwr.2p.test	two proportions (equal n)
pwr.2p2n.test	two proportions (unequal n)
pwr.anova.test	balanced one way ANOVA
pwr.chisq.test	chi-square test
pwr.f2.test	general linear model
pwr.p.test	proportion (one sample)
pwr.r.test	correlation
pwr.t.test	t-tests (one sample, 2 sample, paired)
pwr.t2n.test	t-test (two samples with unequal n)

For each of these functions, you enter three of the four quantities (effect size, sample size, significance level, power) and the fourth is calculated.

Sample size and Power estimation in R

Sample size in R

Power calculations for one and two sample t tests

Description

Compute the power of the one- or two- sample t test, or determine parameters to obtain a target power.

Usage

```
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = 0.05,
             power = NULL,
             type = c("two.sample", "one.sample", "paired"),
             alternative = c("two.sided", "one.sided"),
             strict = FALSE, tol = .Machine$double.eps^0.25)
```

Calculate Power given sample size (n)

```
############Power for a given sample#############

#Aim: to compute the power of a study which aims to show
#a difference in means between group 1 (n=6) and group 2 (n=6)
#assuming that the magnitude of the difference is 0.3 units and
#the standard deviation is 0.28 units.

power.t.test(n=6,delta=0.3,sd=0.28,type="two.sample")
#Possible conclusion sentence:
#The power of the study is 39% to detect a difference in means of 0.3 units.
```

```r
power.t.test(delta=0.2, sd=0.5, power=0.8) # by default, it gives you a two.sample

#Example
power.t.test(power=0.9,delta=0.3,sd=0.28,type="one.sample") #one sample

#Aim: to compute the sample size needed to achieve a power of 90% in
#a study which aims to show a difference in means between two independent groups
#assuming that the magnitude of the difference is 0.3 units and
#the standard deviation is 0.28 units.

power.t.test(power=0.9,delta=0.3,sd=0.28,type="two.sample") # two sample
```

```
> power.t.test(power=0.9,delta=0.3,sd=0.28,type="one.sample")

     One-sample t test power calculation

              n = 11.24949
          delta = 0.3
             sd = 0.28
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

> power.t.test(power=0.9,delta=0.3,sd=0.28,type="two.sample")

     Two-sample t test power calculation

              n = 19.3192
          delta = 0.3
             sd = 0.28
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Using pwr package in R
Source: Cohen, J. (1988)

Power calculations for t-tests of means (one sample, two samples and paired samples)

Description

Compute power of tests or determine parameters to obtain target power (similar to power.t.test).

Usage

```
pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL,
    type = c("two.sample", "one.sample", "paired"),
    alternative = c("two.sided", "less", "greater"))
```

Arguments

`n`	Number of observations (per sample)
`d`	Effect size (Cohen's d) - difference between the means divided by the pooled standard deviation
`sig.level`	Significance level (Type I error probability)
`power`	Power of test (1 minus Type II error probability)
`type`	Type of t test : one- two- or paired-samples
`alternative`	a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less"

Using pwr package in R

```
> delta=10 #difference in mean (sample mean-hypothesized mean)# for one sample
> sigma=20 # standard error of the mean
> r=delta/sigma
> # d is the effect size, that is ratio of delta/sigma
> pwr.t.test(d=r, sig.level = 0.05, power = 0.9, type= "one.sample")

     One-sample t test power calculation

              n = 43.99548
              d = 0.5
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

Using pwr package in R

```
> delta=10 #difference in mean (sample mean of group 1- sample mean of group 2)# for two sample
> sigma=20 # standard error of the mean
> r=delta/sigma
> pwr.t.test(d=r, sig.level = 0.1, power = 0.8, type= "two.sample")

     Two-sample t test power calculation

              n = 50.14822
              d = 0.5
      sig.level = 0.1
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Power Analysis in R

- Discussed pwr.t.test() command for one sample, two sample t.tests

- Note: If you know the sample size(n) you can calculate the power and vice versa