Sampling & Sample Size Estimation

Moazzam Ali MD, PhD, MPH

Department of Reproductive Health and Research

World Health Organization
Geneva, Switzerland

Presented at:
GFMER

September 16, 2014

Topics to be covered

- History of sampling
- Why sampling
- Sampling concepts and terminologies
- Types of sampling and factors affecting choice of sampling design
- Advantages of sampling

History of Sampling *(Contd)*

- Dates back to 1920 and started by Literary Digest, a news magazine published in the U.S. between 1890 and 1938.
- Digest successfully predicted the presidential elections in 1920, 1924,1928, 1932 but;
- Failed in 1936…

- The *Literary Digest* poll in 1936 used a sample of 10 million, drawn from government lists of automobile and telephone owners. Predicted Alf Landon would beat Franklin Roosevelt by a wide margin. But instead Roosevelt won by a landslide. The reason was that the sampling frame did not match the population. Only the rich owned automobiles and telephones, and they were the ones who favored Landon.

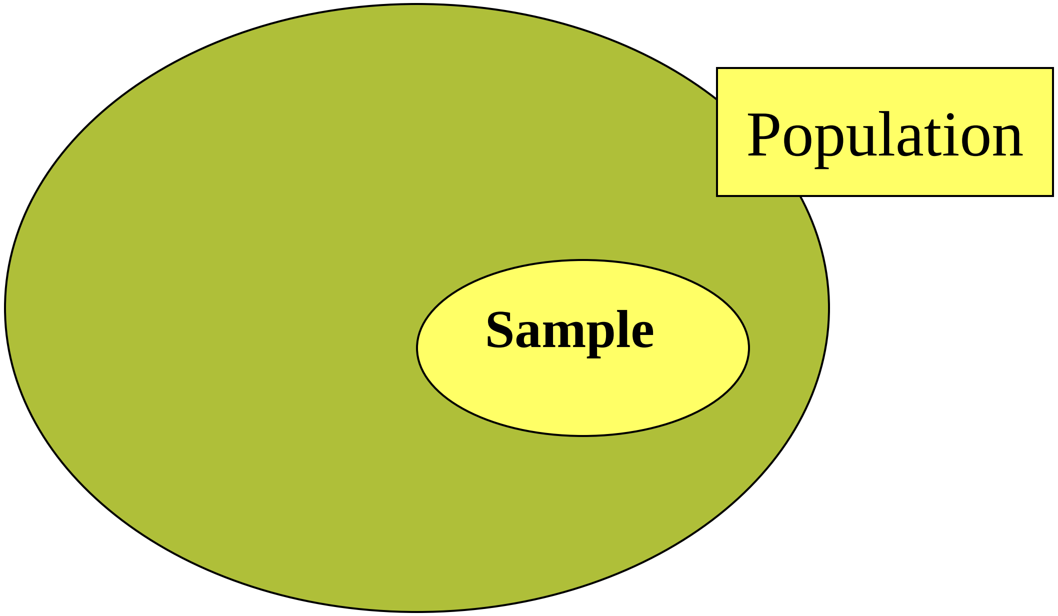
What is sampling

- A sample is some part of a larger body specially selected to represent the whole
- Sampling is then is taking any portion of a population or universe as representative of that population or universe
- Sampling is the process by which this part is chosen

Reasons for Drawing a Sample

- Less time consuming than a census
- Less costly to administer than a census
- Less cumbersome and more practical to administer than a census of the targeted population

Population and sample



Key Definitions

- A population (universe) is the collection of things under consideration
- A sample is a portion of the population selected for analysis
- A parameter is a summary measure computed to describe a characteristic of the population
- A statistic is a summary measure computed to describe a characteristic of the sample

A Census

- A survey in which information is gathered about all members of a population

- Gallup poll is able to develop representative samples of any adult population with interviews of approximately 1500 respondents

- That sample size allows them to be 95% confident that the results they obtain are accurate within + or – 3% points

Sampling concepts and terminologies

- Population/Target population
- Sampling unit
- Sampling frame

Population/Target Population

- *Target Population* is the collection of all individuals, families, groups organizations or events that we are interested in finding out about.
- Is the population to which the researcher would like to generalize the results. For example,all adults population of Myanmar aged 65 or older

Sampling unit/Element/ Unit of analysis

- The unit about which information is collected and that provides the basis of analysis

- Each member of a population is an element. (e.g. a child under 5)

- Sometimes it is household, e.g. any injury in the household in the last three months.

Sampling Frame

- The actual list of sampling units from which the sample, or some stage of the sample, is collected

- It is simply a list of the study population

Sample Design

- A set of rules or procedures that specify how a sample is to be selected
- This can either be probability or non-probability
- Sample size: The number of elements in the obtained sample

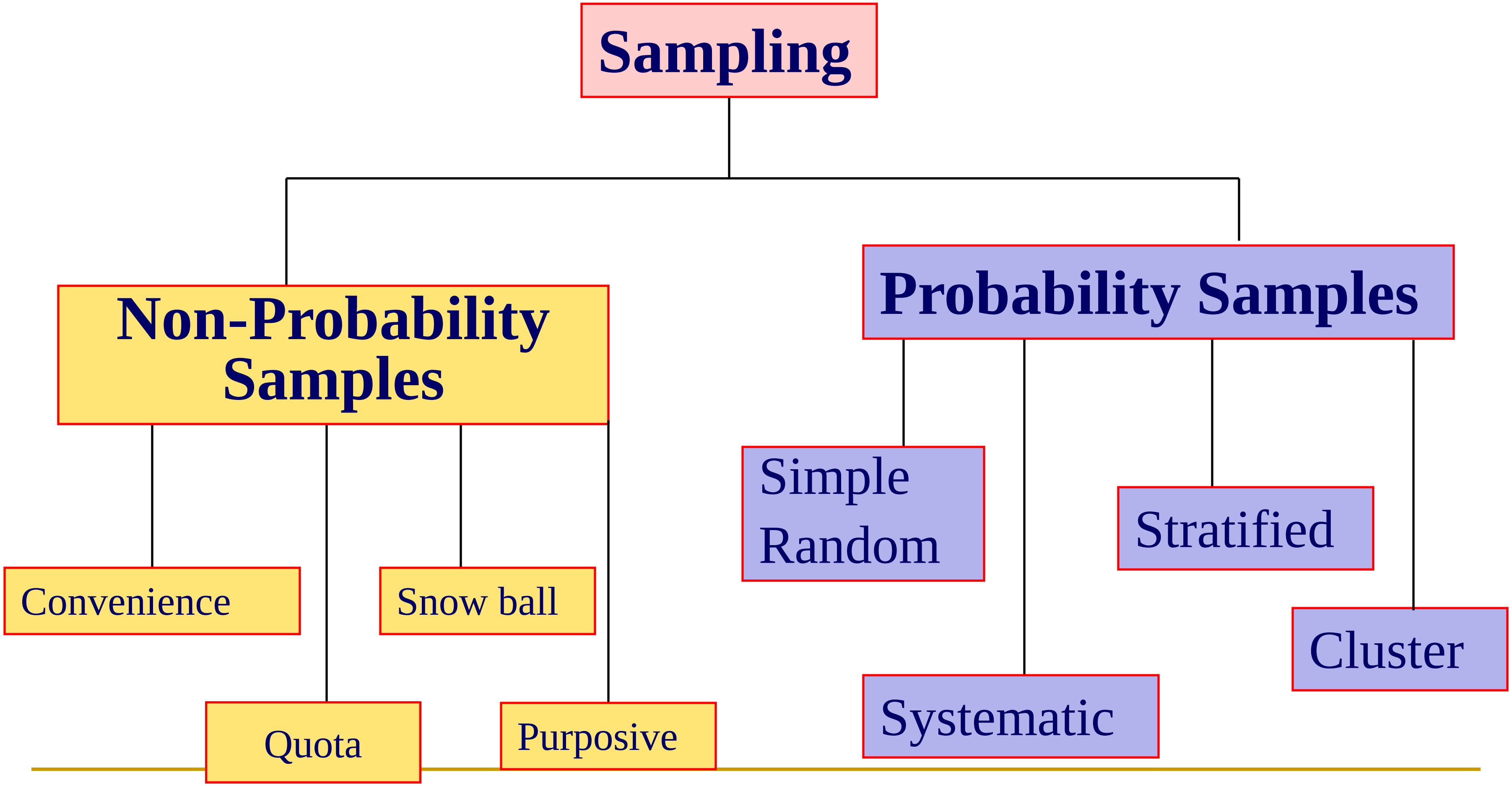
Types of sampling

Types of sampling

- There are two types of sampling techniques:

- Probability sampling (or Random)
- Non- Probability sampling (or non-Random)

Types of Sampling Methods



Sampling

Non-Probability Samples

Probability Samples

Convenience

Snow ball

Quota

Purposive

Simple Random

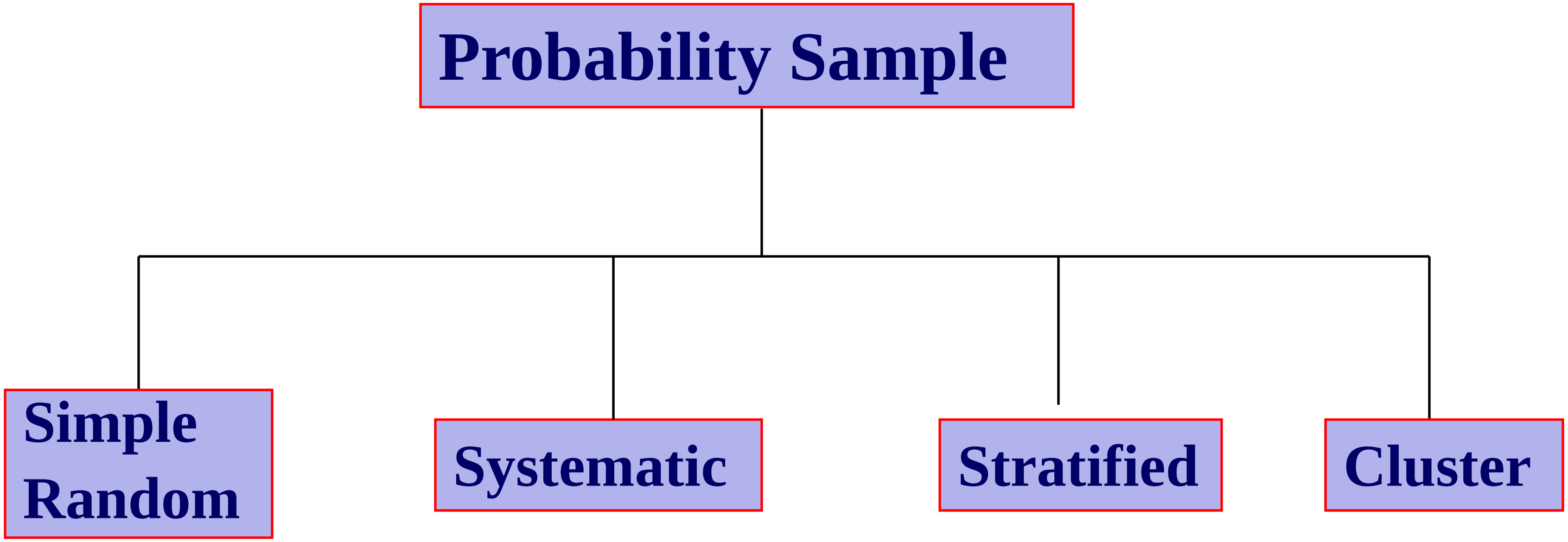
Stratified

Cluster

Systematic

Probability Sampling

- This is one in which each person in the population has a chance/probability of being selected

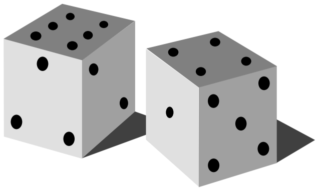


Types of Probability Sampling

- Simple random
- Systematic sampling
- Stratified random
- Cluster sampling
- Multi-stage sampling

Simple Random Samples

- Every individual or item from the frame has an equal chance of being selected
- Selection may be with replacement or, without replacement
- Samples obtained from table of random numbers or computer random number generators
- Random samples are unbiased and, on average, representative of the population

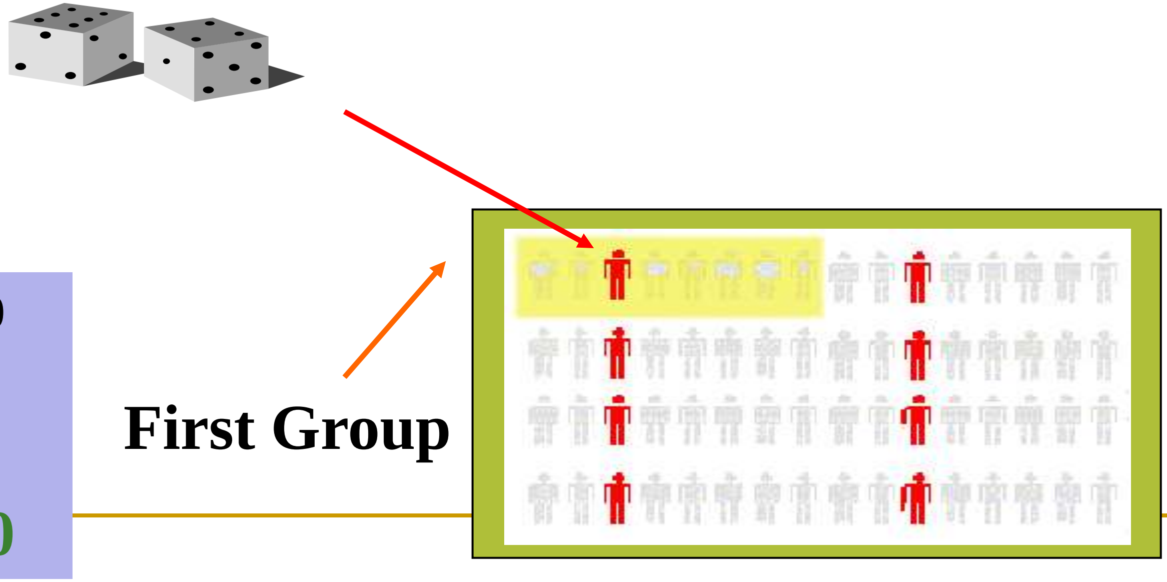


Systematic sample

- This method is referred to as a systematic sample with a random start.
- This is done by picking every 5th or 10th unit at regular intervals.
- For example to carry out a filarial survey in a town, we take 10% sample. If the total population of the town is about 5000. The sample comes to 500.

Systematic Samples

- Randomly select one individual from the 1st group
- Select every k-th individual thereafter
- We number the houses first. Then a number is taken at random; say 3.Than every 10th number is selected from that point onward like 3, 13, 23, 33 etc.



N = 500

n = 3

k = 10

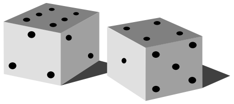
First Group

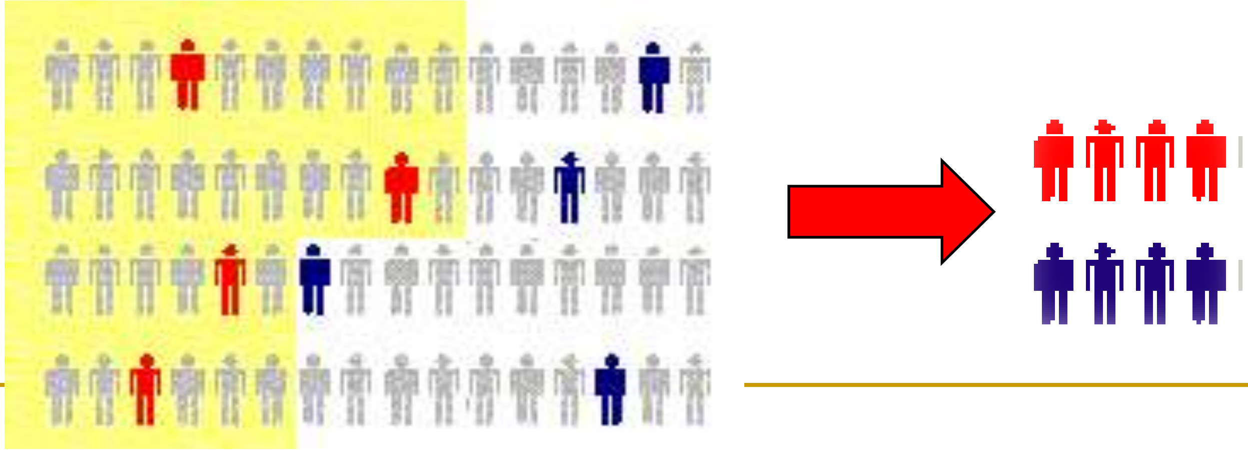
Stratified Random sample

- This involves dividing the population into distinct subgroups according to some important characteristics, such as age, or socioeconomic status, religion and selecting a random number from each subgroup. (e.g. African voodoo healers)
- Especially important when one group is so small (say, 3% of the population) that a random sample might miss them entirely.
- Population divided into two or more groups according to some common characteristic
- Simple random sample selected from each group
- The two or more samples are combined into one

Stratified Samples

- Procedure: Divide the population into strata (mutually exclusive classes), such as men and women. Then randomly sample within strata.
- Suppose a population is 30% male and 70% female. To get a sample of 100 people, we randomly choose males (from the population of all males) and, separately, choose females. Our sample is then guaranteed to have exactly the correct proportion of sexes.



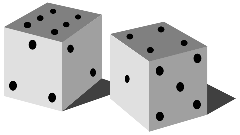


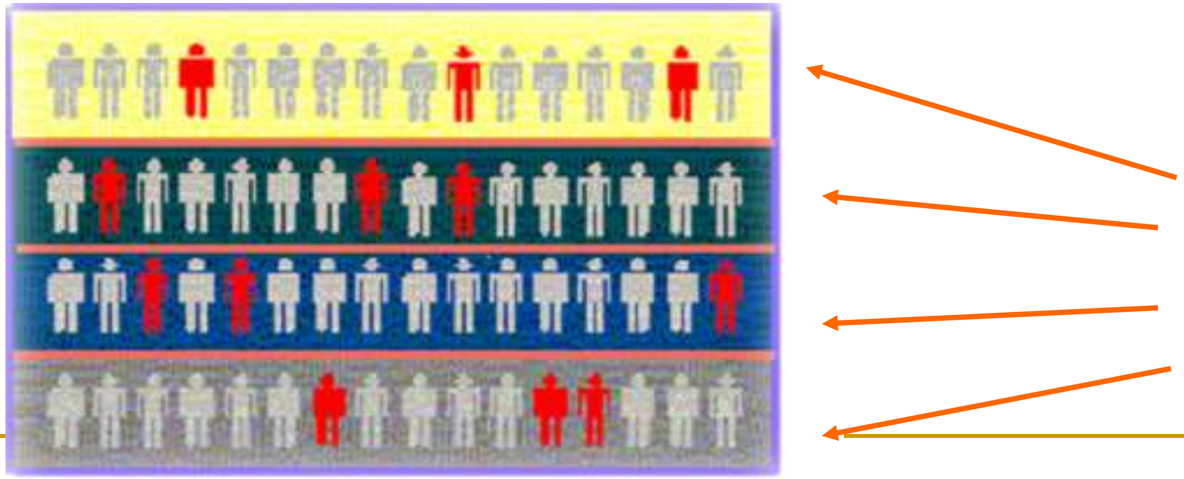
Cluster sample

- A sampling method in which each unit selected is a group of persons (all persons in a city block, a family, etc.) rather than an individual.
- Used when (a) sampling frame not available or too expensive, and (b) cost of reaching an individual element is too high
  - E.g., there is no list of automobile mechanics in the Myanmar. Even if you could construct it, it would cost too much money to reach randomly selected mechanics across the entire Myanmar : would have to have unbelievable travel budget
- In cluster sampling, first define large clusters of people. Fairly similar to other clusters. For example, cities make good clusters.
- Once you've chosen the cities, might be able to get a reasonably accurate list of all the mechanics in each of those cities. Is also much less expensive to fly to just 10 cities instead of 200 cities.
- Cluster sampling is less expensive than other methods, but less accurate.

Cluster Samples

- Population divided into several “clusters,” each representative of the population
- Simple random sample selected from each
- The samples are combined into one





Population divided into 4 clusters.

Non- Probability Sampling /(Non-Random)

- This is where the probability of inclusion in the sample is unknown.

- Convenience sampling
- Purposive sampling
- Quota sampling
- Snow ball sampling

Convenience Sample

- Man-in-the-street surveys and a survey of blood pressure among volunteers who drop in at an examination booth in public places are in the category.

- It is improper to generalize from the results of a survey based upon such a sample for there is no known way of knowing what sorts of biases may have been operating.

Convenience sample

- Whoever happens to walk by your office; who's on the street when the camera crews come out

- If you have a choice, don't use this method. Often produces really wrong answers, because certain attributes tend to cluster with certain geographic and temporal variables.
  - For example, at 8am in Tokyo, most of the people on the street are workers heading for their jobs.
  - At 10am, there are many more people who don't work, and the proportion of women is much higher.
  - At midnight, there are young people and muggers.

Quota

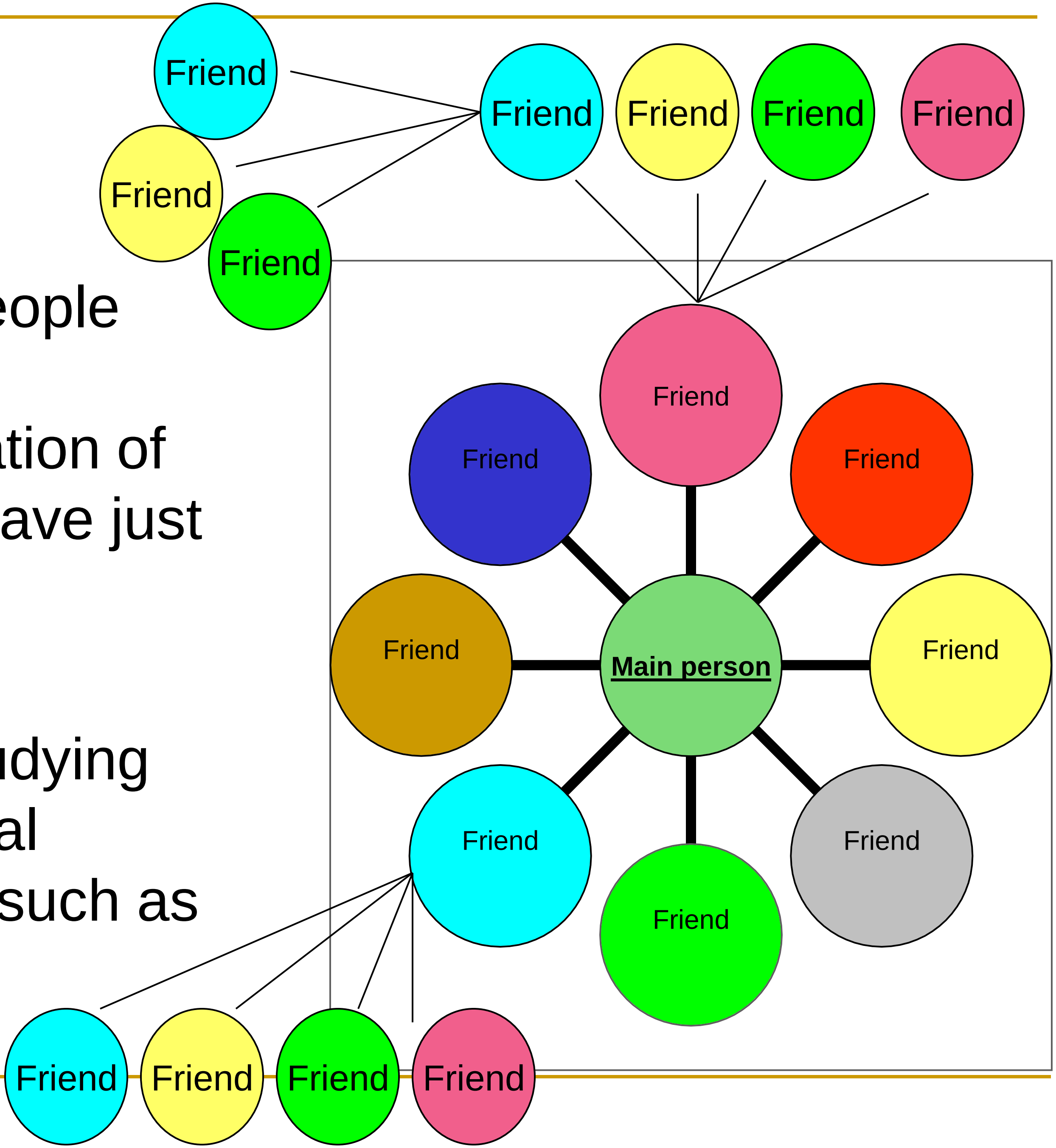
- Haphazard sampling within categories
- Is an improvement on convenience sampling, but still has problems.
- How do you know which categories are key?
- How many do you get of each category?

Purposive/Judgment

- Selecting sample on the basis of knowledge of the research problem to allow selection of appropriate persons for inclusion in the sample
- Expert judgment picks useful cases for study
- Good for exploratory, qualitative work, and for pre-testing a questionnaire.

Snowball

- Recruiting people based on recommendation of people you have just interviewed
- Useful for studying invisible/illegal populations, such as drug addicts

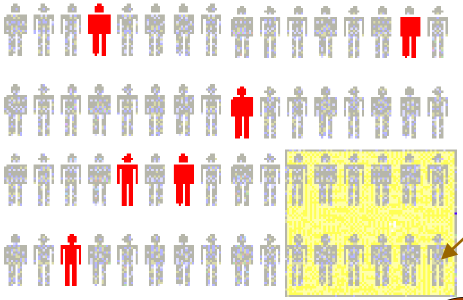


Non-sampling Errors

- An inadequate sampling frame (Non-coverage)
- Non-response from participants
- Response errors
- Coding and data entry errors

Types of Survey Errors

- Coverage error

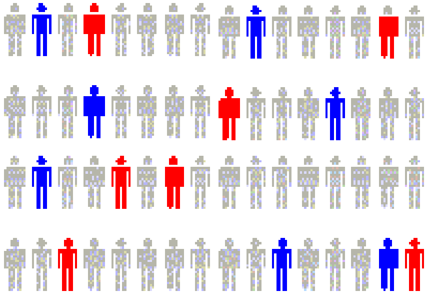


Excluded from frame.

- Non response error

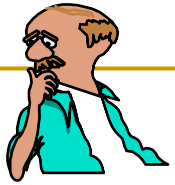
Follow up on non responses.

- Sampling error



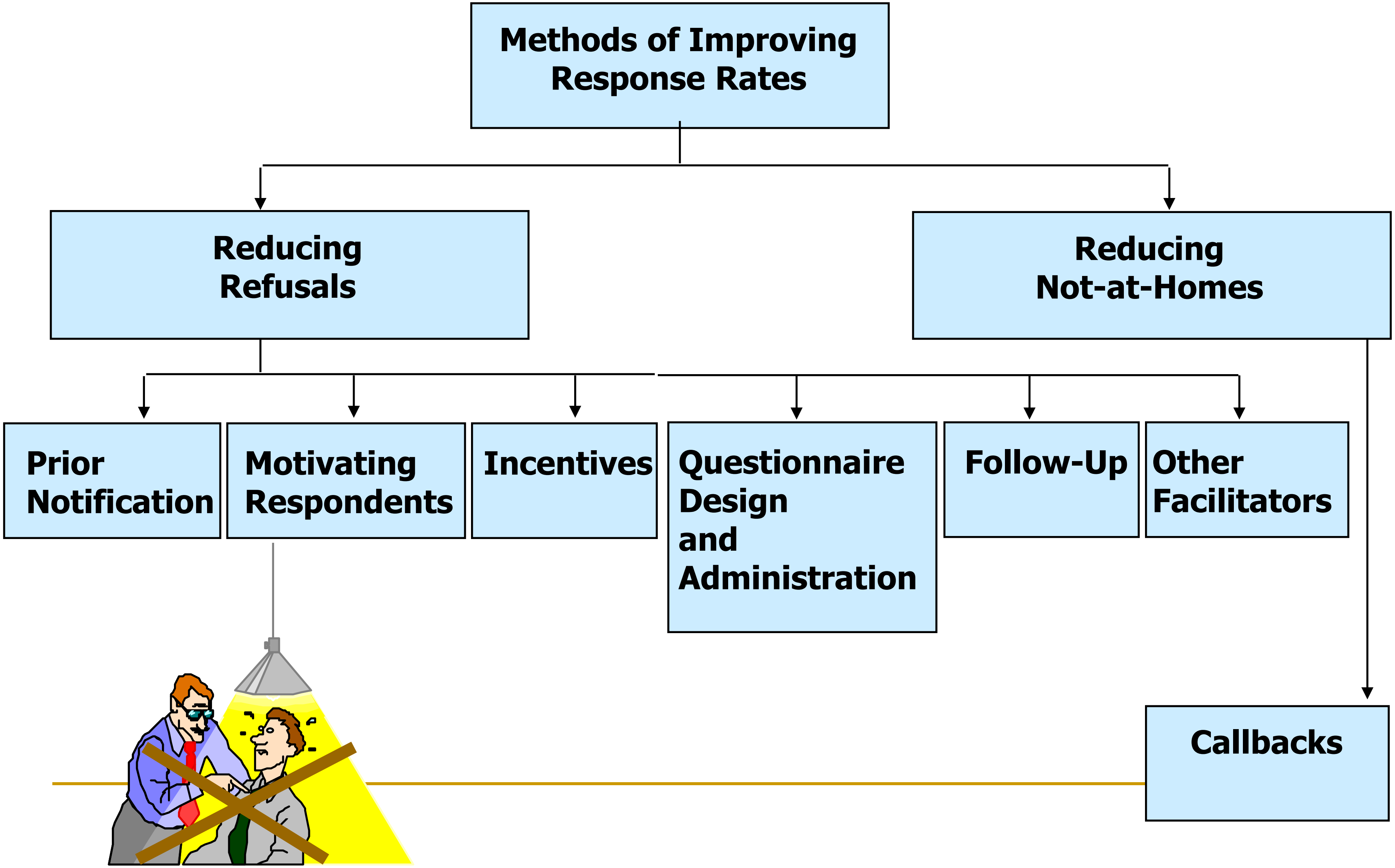
Chance differences from sample to sample.

- Measurement error



Bad Question!

Improving Response Rates



Methods of Improving Response Rates

Reducing Refusals

Reducing Not-at-Homes

Prior Notification

Motivating Respondents

Incentives

Questionnaire Design and Administration

Follow-Up

Other Facilitators

Callbacks

Evaluating Survey Worthiness

- What is the purpose of the survey?
- Is the survey based on a probability sample?
- Coverage error – appropriate frame
- Non-response error – follow up
- Measurement error – good questions elicit good responses
- Sampling error – always exists

Sample size estimation

Sample Size

- Sample size relates to how many people to pick up for the study

- The question often asked is: How big a sample is necessary for a good survey?

- The main objective is to obtain both a desirable accuracy and a desirable confidence level with minimum cost.

Determination of Sample Size

- Type of analysis to be employed
- The level of precision needed
- Population homogeneity /heterogeneity
- Available resources
- Sampling technique used

Sample Size Calculation

$$n = \frac{z^2 p q}{d^2}$$

- ***n:*** the desired sample size
- ***z:*** the standard normal deviate usually set at 1.96 (which corresponds to the 95% confidence level)
- ***p:*** the proportion in the target population to have a specific characteristic. If no estimate available set at 50% (or 0.50)
- ***q:****1-p*
- ***d:*** absolute precision or accuracy, normally set at 0.05.

Sample Size Calculation

$$n = \frac{(1.96)^2 (0.5) (0.5)}{(0.05)^2}$$

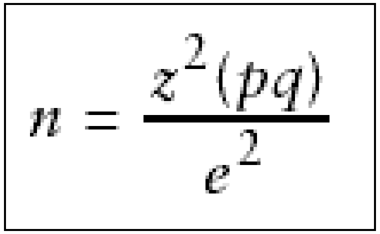
$$n = 384$$

Before you go to the field…

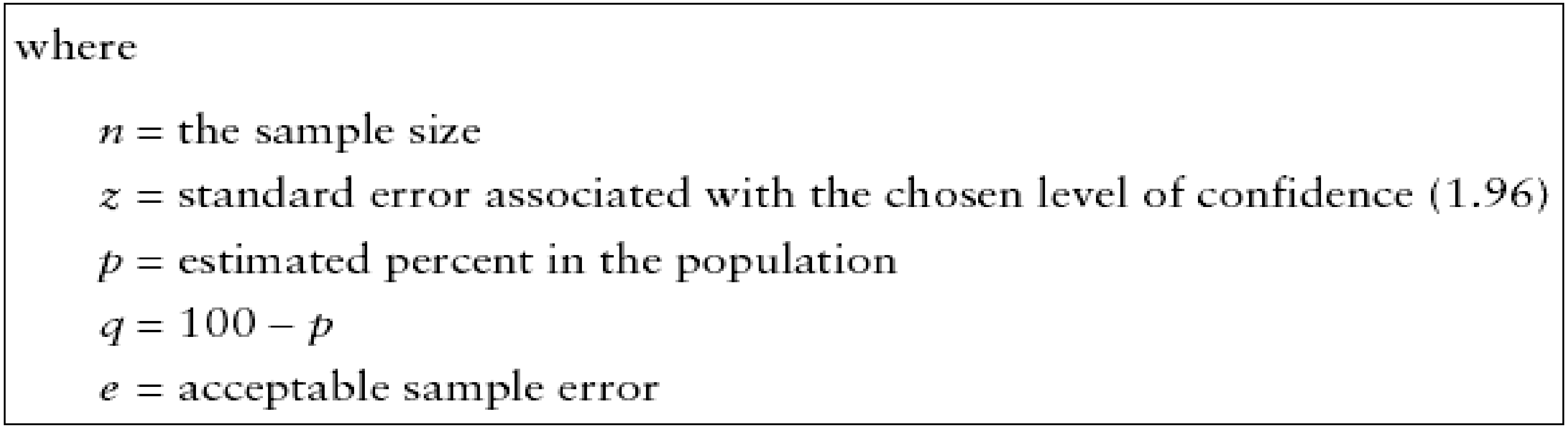
- Financing and budget
- Work plan
  - Time line
  - Field work logistics
- Develop instruments
- Drawing a sample of household
- Training manual
- Pilot test

Thank you

Sample Size Formula



$$n = \frac{z^2(pq)}{e^2}$$



where

n = the sample size

z = standard error associated with the chosen level of confidence (1.96)

p = estimated percent in the population

$q = 100 - p$

e = acceptable sample error