UNIT-III

Basic concepts of probability, random variables, probability distributions, sampling and estimation, statistical inference

Random Variables and Probability Distributions

- **Random Variables** - Random outcomes corresponding to subjects randomly selected from a population.

- **Probability Distributions -** A listing of the possible outcomes and their probabilities (discrete r.v.^{s}) or their densities (continuous r.v.^{s})

- **Normal Distribution** - Bell-shaped continuous distribution widely used in statistical inference.

- **Sampling Distributions -** Distributions corresponding to sample statistics (such as mean and proportion) computed from random samples.

Random Variables

- A variable was defined as a characteristic that can assume different values.
- **A random variable** was a numerical measure of the outcome from a probability experiment, so its value is determined by chance. Random variables are denoted using letters such as ***X***.

Example: Consider an experiment of tossing a fair of coin three times.
Let the random variable X be the number of heads in three tosses, then find X?

$$\Rightarrow S = \{(HHH),(HHT),(HTH),(HTT),(THH),(THT),(TTH),(TTT)\}$$

$$\Rightarrow X(HHH) = 3,\quad X(HHT) = X(HTH) = X(THH) = 2,$$

$$X(HTT) = X(THT) = X(TTH) = 1$$

$$X(TTT) = 0$$

⇒ X = {0, 1, 2, 3}

Random Variables

Random variables are of two types:

Discrete random variable: are variables which can assume only a specific number of values. They have values that can be counted .

- Examples:
  - Toss a coin n time and count the number of heads.
  - Number of children in a family.
  - Number of car accidents per week.
  - Number of defective items in a given company.
  - Number of bacteria per two cubic centimeter of water.

Continuous random variable: are variables that can assume all values between any two give values.

- Examples:
  - Mark of a student.
  - Life time of light bulbs.
  - Length of time required to complete a given training.

EXAMPLE *Distinguishing Between Discrete and Continuous Random Variables*

Determine whether the following random variables are discrete or continuous. State possible values for the random variable.

(a) The number of light bulbs that burn out in a room of 10 light bulbs in the next year.

Discrete; $x = 0, 1, 2, \ldots, 10$

(b) The number of leaves on a randomly selected oak tree.

Discrete; $x = 0, 1, 2, \ldots$

(c) The length of time between calls to 911.

Continuous; $t > 0$

Probability distribution

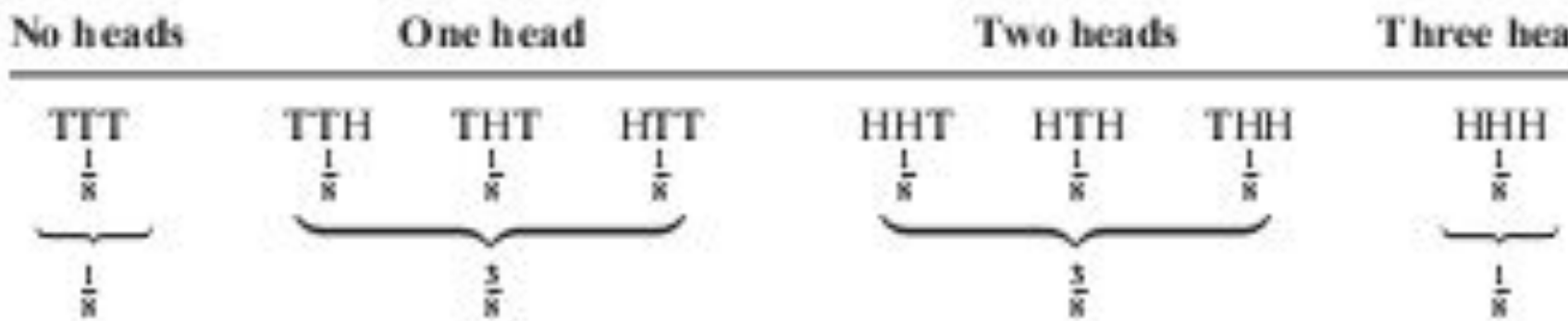
- **Probability distribution:-**consists of a value a random variable can assume and the corresponding probabilities of the values .

A **probability distribution** provides the possible values of the random variable *X* and their corresponding probabilities. A probability distribution can be in the form of a table, graph or mathematical formula.

- Probability distribution can be discrete or continues.
- **Discrete probability distribution:-** A table used to specify all possible values of the discrete random variable **X** along with their respective probabilities.

- Example: Consider the experiment of tossing a coin three times.
- Let X be the number of heads. Construct the probability distribution of X.

Probabilities for the values of X can be determined as follows:



Hence, the probability of getting no heads is $\frac{1}{8}$, one head is $\frac{3}{8}$, two heads is $\frac{3}{8}$, and three heads is $\frac{1}{8}$. From these values, a probability distribution can be constructed by listing the outcomes and assigning the probability of each outcome, as shown here.

Number of heads X	0	1	2	3
Probability $P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

EXAMPLE A Discrete Probability Distribution

The table to the right shows the probability distribution for the random variable X, where X represents the number of movies streamed on Netflix each month.

x	$P(x)$
0	0.06
1	0.58
2	0.22
3	0.10
4	0.03
5	0.01

Rules for a Discrete Probability Distribution

Let $P(x)$ denote the probability that the random variable X equals x; then

1. $\Sigma P(x) = 1$
2. $0 \leq P(x) \leq 1$

EXAMPLE *Identifying Probability Distributions*

Is the following a probability distribution?

x	$P(x)$
0	0.16
1	0.18
2	0.22
3	0.10
4	0.30
5	0.01

No. $\Sigma P(x) = 0.97$

EXAMPLE *Identifying Probability Distributions*

Is the following a probability distribution?

x	$P(x)$
0	0.16
1	0.18
2	0.22
3	0.10
4	0.30
5	– 0.01

No. $P(x = 5) = -0.01$

EXAMPLE *Identifying Probability Distributions*

Is the following a probability distribution?

x	$P(x)$
0	0.16
1	0.18
2	0.22
3	0.10
4	0.30
5	0.04

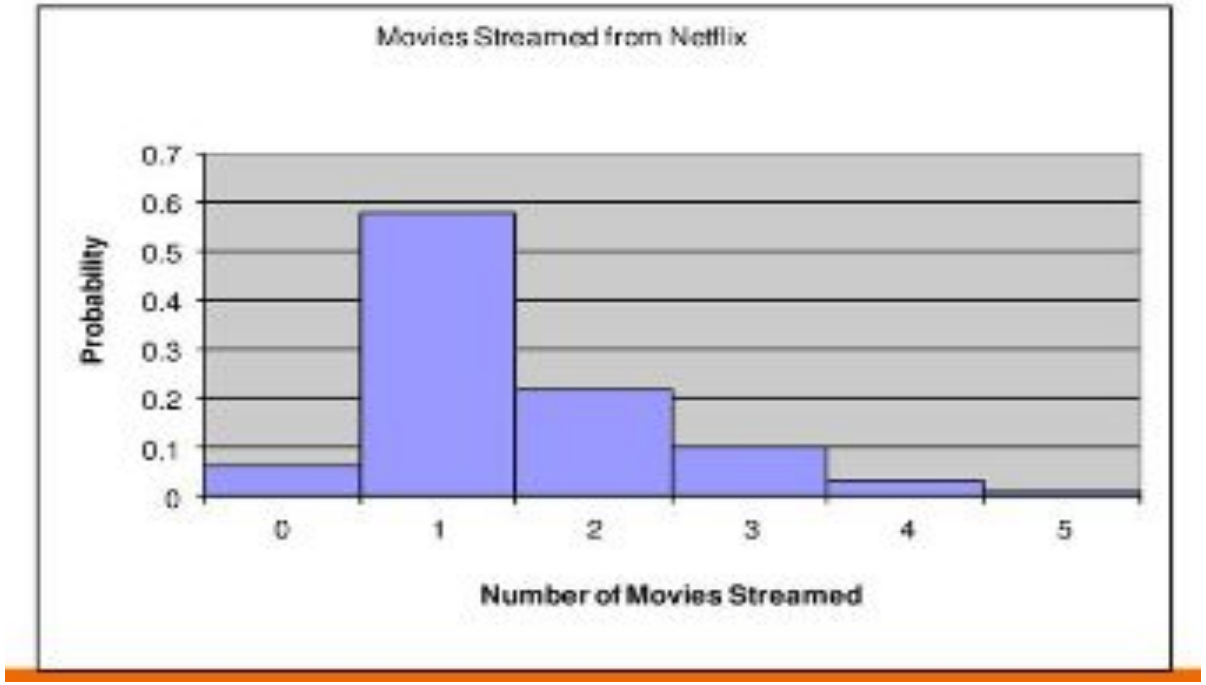
Yes. $\Sigma P(x) = 1$

A **probability histogram** is a histogram in which the horizontal axis corresponds to the value of the random variable and the vertical axis represents the probability of that value of the random variable.

EXAMPLE Drawing a Probability Histogram

Draw a probability histogram of the probability distribution to the right, which represents the number of movies streamed on Netflix each month.

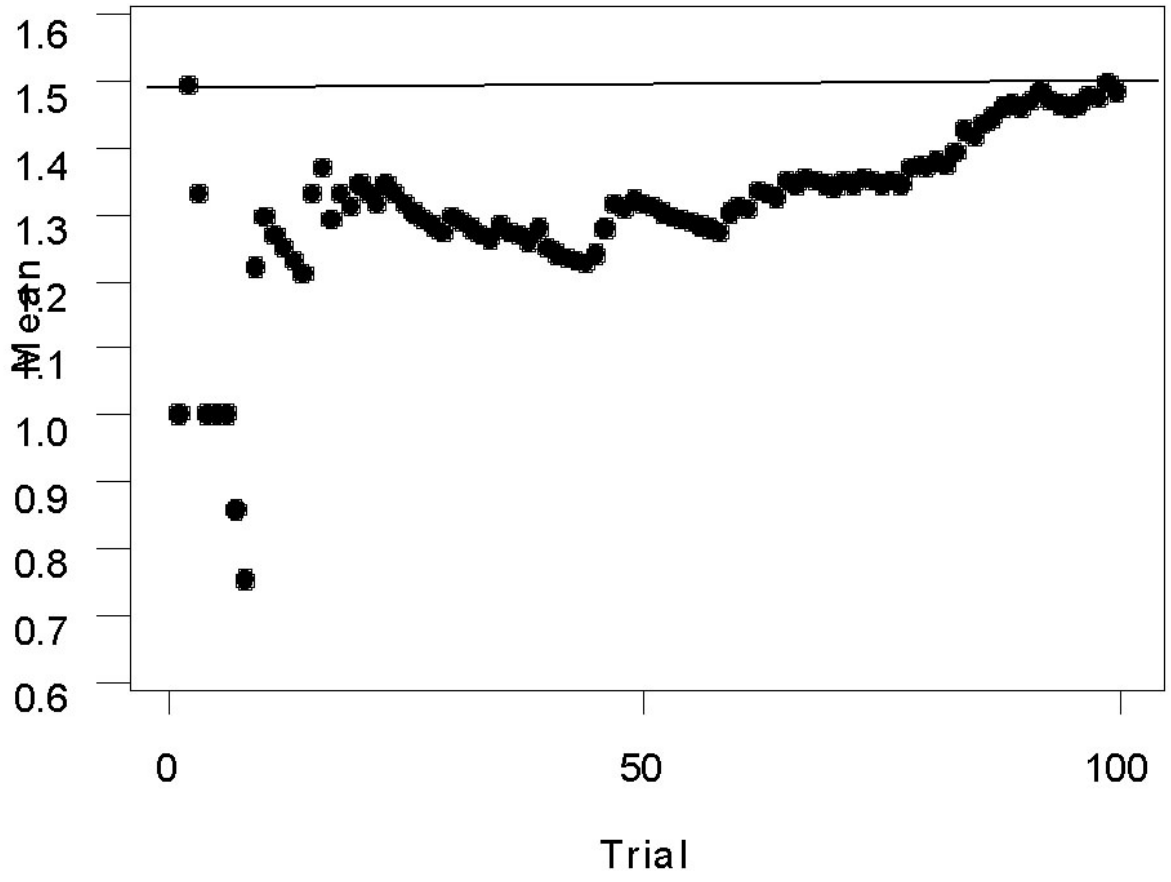
x	$P(x)$
0	0.06
1	0.58
2	0.22
3	0.10
4	0.03
5	0.01



The following data represent the number of DVDs rented by 100 randomly selected customers in a single visit. Compute the mean number of DVDs rented.

1	1	1	1	1	1	1	2	2	2
2	1	1	1	1	1	3	1	1	3
1	1	2	1	1	1	1	2	3	0
0	1	1	1	1	1	1	1	4	1
1	3	1	2	2	1	3	1	1	1
1	2	1	1	3	1	1	2	3	2
0	0	1	1	3	1	2	1	2	3
0	2	1	1	1	1	1	3	3	1
5	1	1	2	2	3	1	2	2	4
2	2	2	0	1	2	1	1	1	0

As the number of trials of the experiment increases, the mean number of rentals approaches the mean of the probability distribution.



Because the mean of a random variable represents what we would expect to happen in the long run, it is also called the **expected value**, $E(X)$, of the random variable.

EXAMPLE Computing the Expected Value of a Discrete Random Variable

A term life insurance policy will pay a beneficiary a certain sum of money upon the death of the policy holder. These policies have premiums that must be paid annually. Suppose a life insurance company sells a $250,000 one year term life insurance policy to a 49-year-old female for $530. According to the National Vital Statistics Report, Vol. 47, No. 28, the probability the female will survive the year is 0.99791. Compute the expected value of this policy to the insurance company.

	x	$P(x)$
Survives	530	0.99791
Does not survive	530 – 250,000 = -249,470	0.00209

$$E(X) = 530(0.99791) + (-249{,}470)(0.00209) = \$7.50$$

The company expect to make $7.50 for each 49-year old female client it insures. The $7.50 is a long term result. It's the average profit per 49-year old female insured is $7.50.

Standard Deviation of a Discrete Random Variable

The standard deviation of a discrete random variable X is given by

$$\sigma_x = \sqrt{\sum\left[(x-\mu_x)^2 \cdot P(x)\right]}$$
$$= \sqrt{\sum\left[x^2 \cdot P(x)\right] - \mu_x^2}$$

where x is the value of the random variable, μ_X is the mean of the random variable, and $P(x)$ is the probability of observing a value of the random variable.

EXAMPLE Computing the Standard Deviation of a Discrete Random Variable

Compute the variance and standard deviation of the following probability distribution which represents the number of DVDs a person rents from a video store during a single visit. Remember, the mean that we found was 1.49.

x	$P(x)$
0	0.06
1	0.58
2	0.22
3	0.10
4	0.03
5	0.01

EXAMPLE Computing the Variance and Standard Deviation of a Discrete Random Variable

x	μ	$x-\mu_x$	$(x-\mu_x)^2$	$P(x)$	$(x-\mu_x)^2 P(x)$
0	1.49	–1.49	2.2201	0.06	0.133206
1	1.49	–0.49	0.2401	0.58	0.139258
2	1.49	0.51	0.2601	0.22	0.057222
3	1.49	1.51	2.2801	0.1	0.22801
4	1.49	2.51	6.3001	0.03	0.189003
5	1.49	3.51	12.3201	0.01	0.123201

$$\sigma_x^2 = \sum (x-\mu_x)^2 \cdot P(x) = 0.8699$$

$$\sigma_x = \sqrt{0.8699} = 0.9327$$

Continuous probability distribution

Definition: a non negative function **f(x)** is called probability distribution of continuous Random Variable X .

- if the total area bounded by the curve along the X-axis is 1 and if the sub area under the curve bounded by the curve & X-axis and perpendicularly erected at any points a and b give the probability that X is between a and b.

Properties of continuous probability distribution

a) The total area under the curve is one i.e. $\int_{-\infty}^{\infty} f(x) = 1$

b) $P(a \le X \le b)$ = the area under the curve between the point a and b.

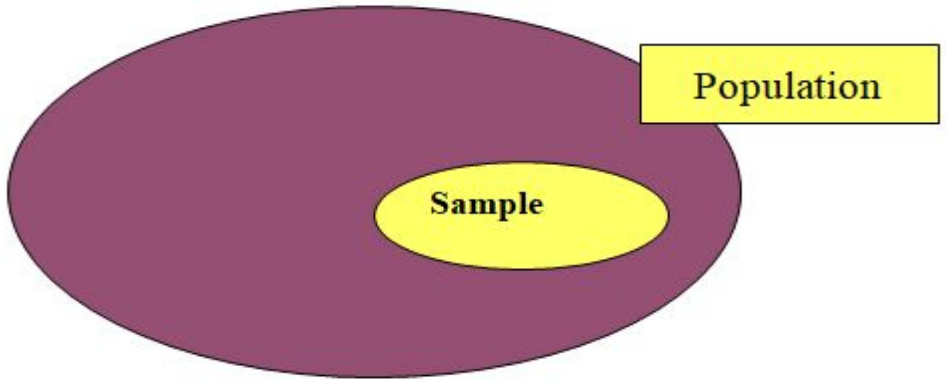
c) $P(X) \ge 0$

d) $P(X = a) = 0$

e) $P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$

What is sampling?

- A sample is some part of a larger body specially selected to represent the whole.
- Sampling 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

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

Example

- 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

- Sampling unit is the unit about which information is collected.
- Unit of analysis is the unit 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.

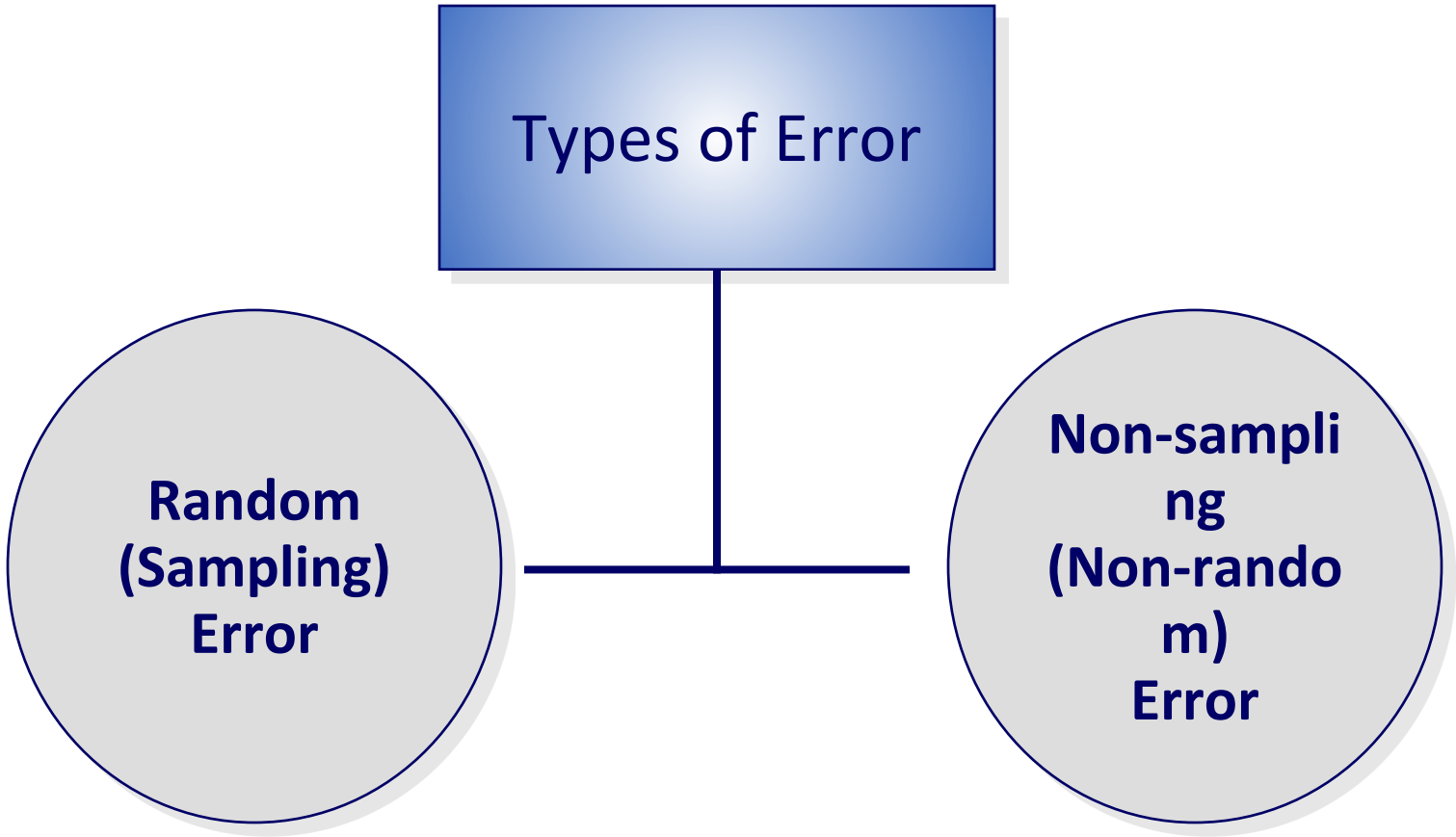
Sampling Frame

- 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

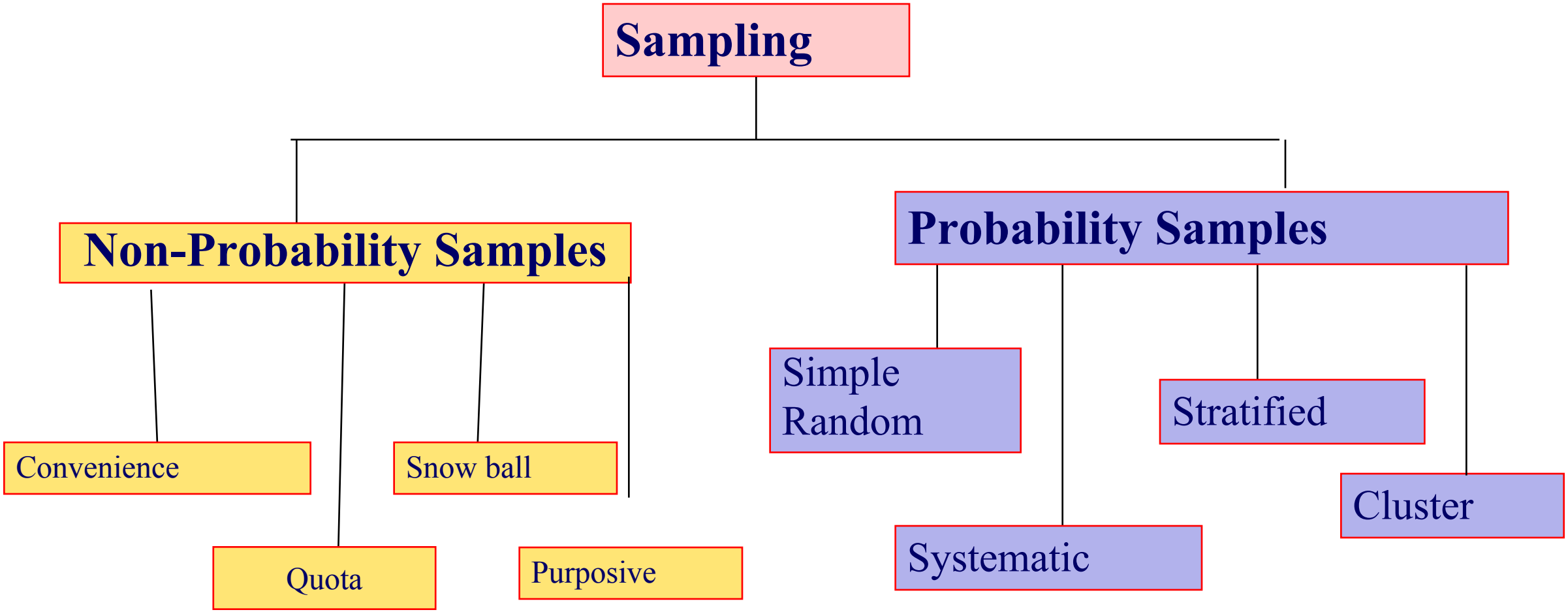
- A set of rules or procedures that specify how a sample is to be selected
- This can either be probability or non-probability
- Sample size: The number of elements in the obtained sample


Types of Error
Random (Sampling) Error
Non-sampling (Non-random) Error

Sampling Error: any type of bias that results from mistakes in either the selection process for prospective sampling units or in determining the sample size.

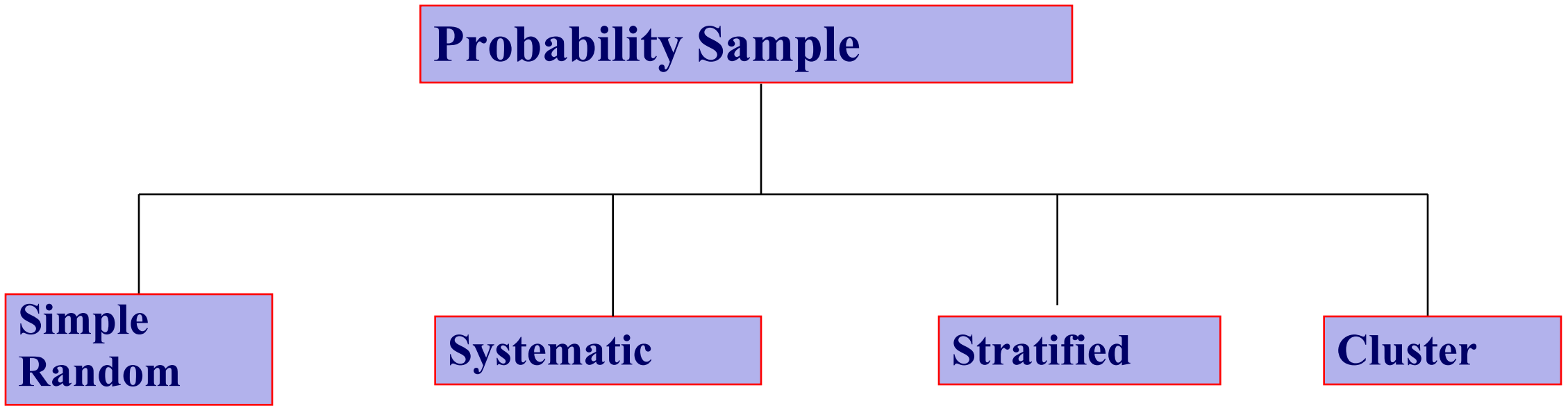
Non sampling Error: bias that occurs in a research study regardless of whether a sample or census is used; e.g., bias caused by measurement errors, response errors, coding errors, etc.

Types of Sampling Methods



Probability Sampling

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

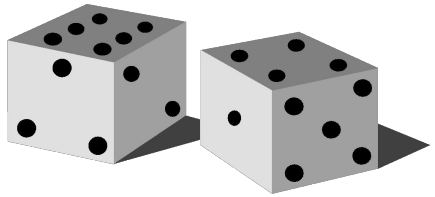


Types of Probability Sampling

- 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

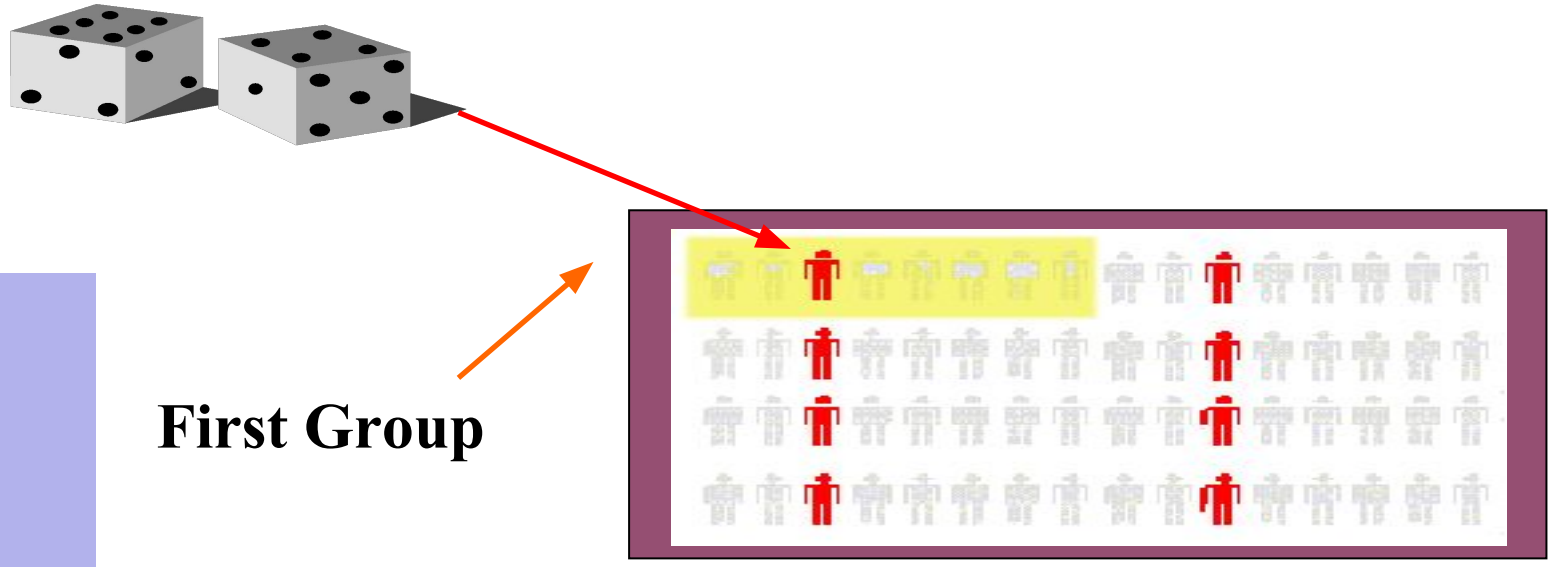


Systematic sample

- This method is referred to as a systematic sample with a random start.
- This is done by picking every 5^{th} or 10^{th} 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.

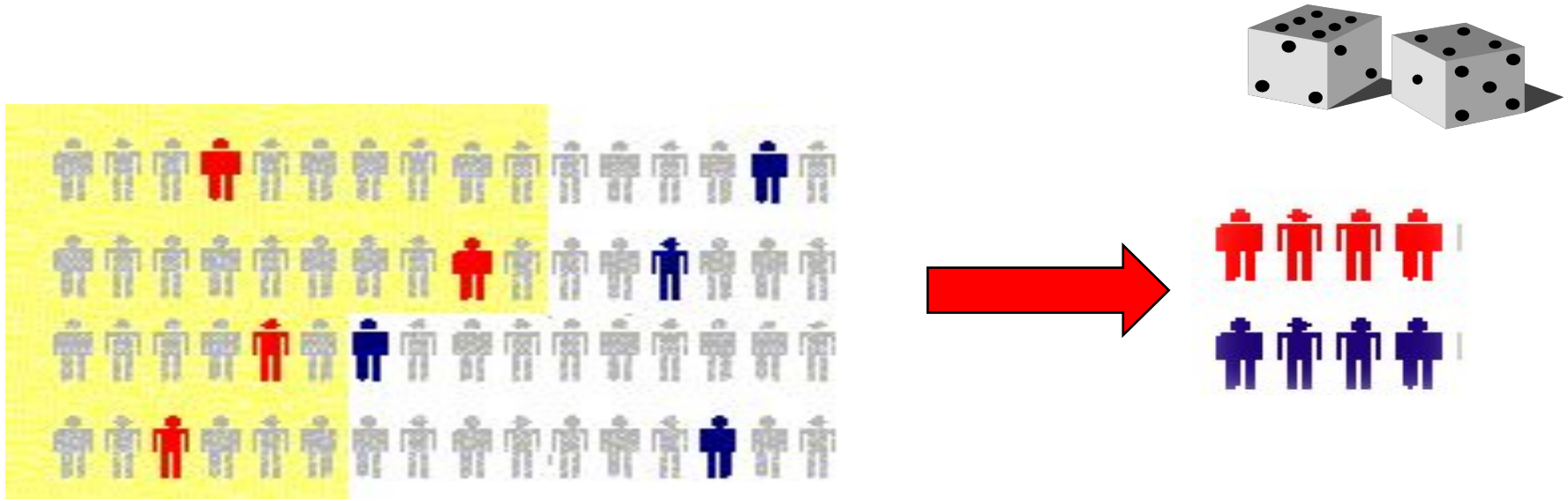


Stratified Random sample

- 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.
- 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.

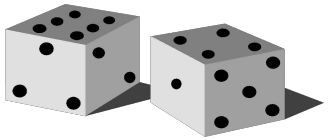


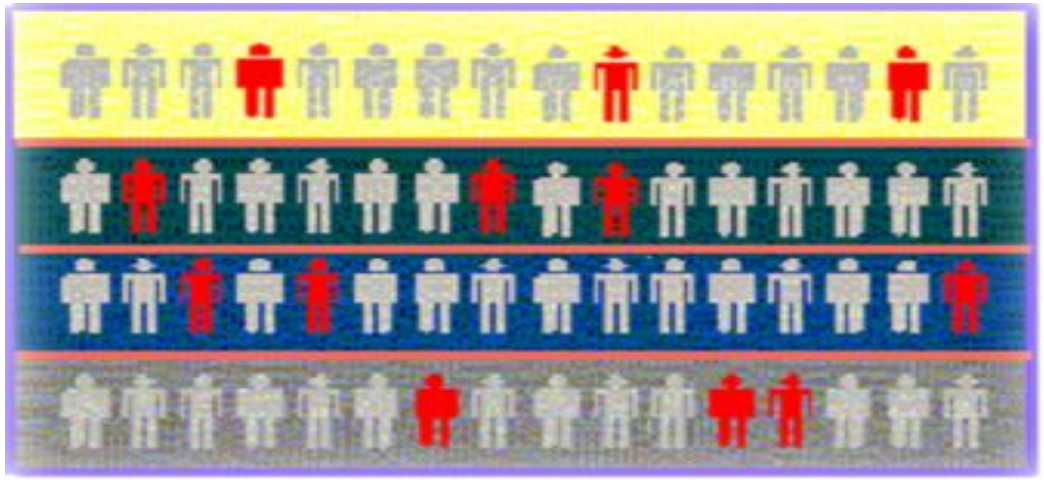
Cluster sample

- 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. It 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.

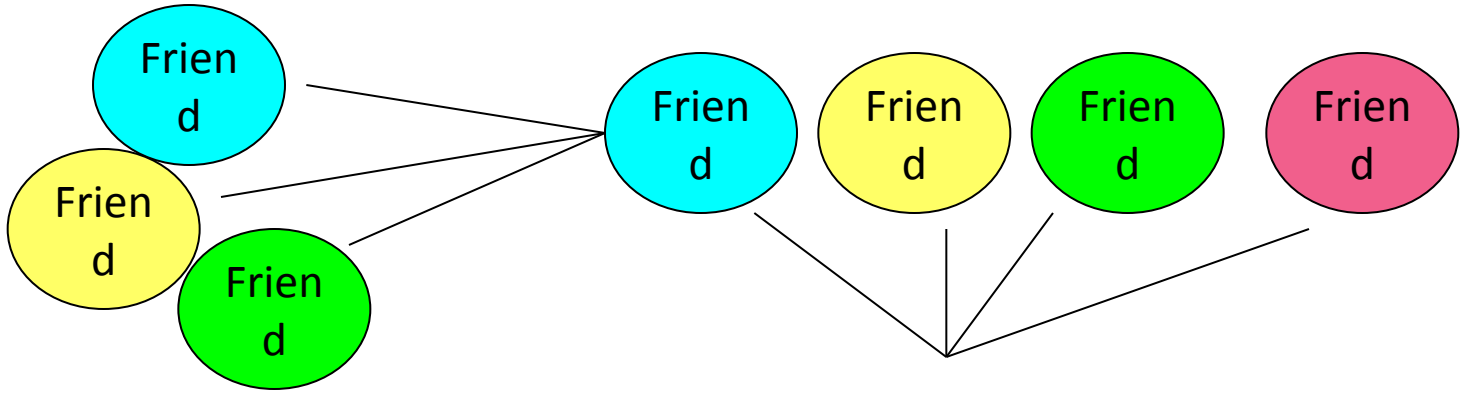
Quota

- 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

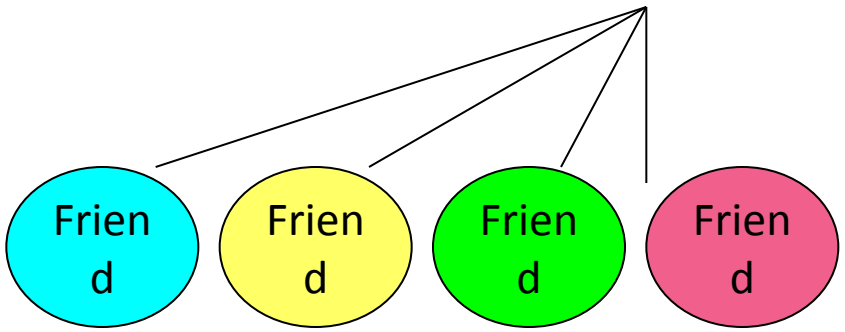
- Selecting sample on the basis of knowledge of the research problem to allow selection of appropriate persons for inclusion in the sample
- Expert judgment picks useful cases for study
- Good for exploratory, qualitative work, and for pre-testing a questionnaire.

Snowball



- Recruiting people based on recommendation of people you have just interviewed

- Useful for studying invisible/illegal populations, such as drug addicts

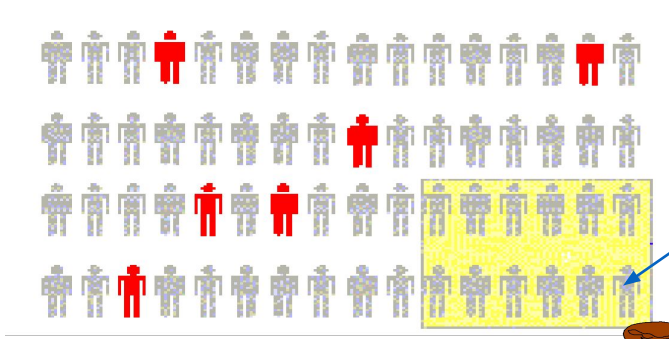


Non-sampling Errors

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

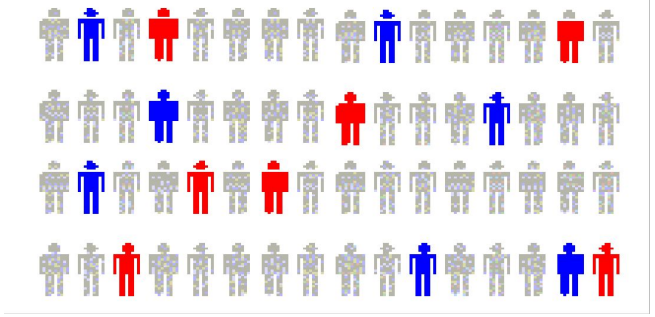
Types of Survey Errors

- Coverage error
- Non response error
- Sampling error
- Measurement error

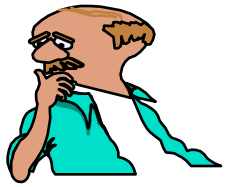


Excluded from frame.

Follow up on non responses.



Chance differences from sample to sample.



Bad Question!

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 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

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

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

$$n = 384$$

Sampling distribution

Sampling distribution of the mean

A theoretical probability distribution of sample means that would be obtained by drawing from the population all possible samples of the same size.

No matter what we are measuring, the distribution of any measure across all possible samples we could take approximates a normal distribution, as long as the number of cases in each sample is about 30 or larger.

Problem 1

A study is to be performed to determine a certain parameter in a community. From a previous study a sd of 46 was obtained.

If a sample error of up to 4 is to be accepted. How many subjects should be included in this study at 99% level of confidence?

Problem 2

It was desired to estimate proportion of anaemic children in a certain preparatory school. In a similar study at another school a proportion of 30 % was detected.

Compute the minimal sample size required at a confidence limit of 95% and accepting a difference of up to 4% of the true population.