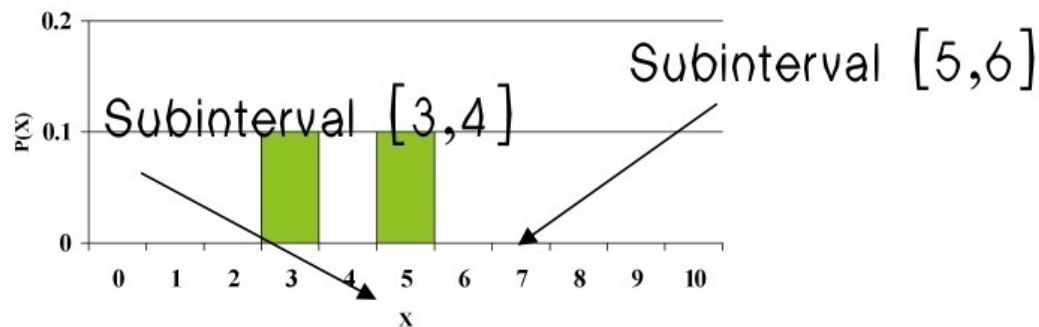


# UNIT-3

## UNIFORM DISTRIBUTION

- u The closed unit interval, which contains all numbers between 0 and 1, including the two end points 0 and 1



The probability  
density function

# The Expected Value of a continuous Random Variable

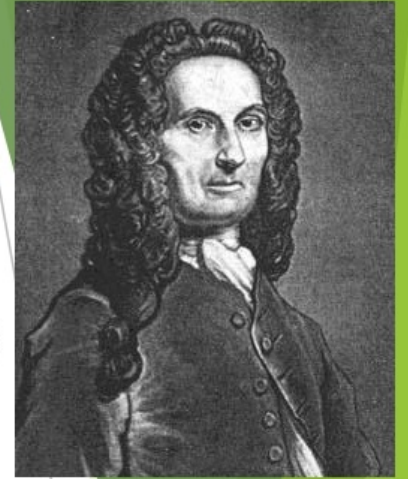
For an uniform random variable  $x$ , where  $f(x)$  is defined on the interval  $[a,b]$ , and where  $a < b$ ,

an  
d

# The Normal Distribution

## Overview

- Discovered in 1733 by de Moivre as an approximation to the binomial distribution when the number of trials is large
- Derived in 1809 by Gauss
- Importance lies in the Central Limit Theorem, which states that the sum of a large number of independent random variables (binomial, Poisson, etc.) will approximate a normal distribution
  - Example: Human height is determined by a large number of factors, both genetic and environmental, which are additive in their effects. Thus, it follows a normal distribution.



Abraham de Moivre (1667-1754)



Karl F. Gauss (1777-1855)

# The Normal Distribution

## Overview

- u A continuous random variable is said to be normally distributed with mean  $\mu$  and variance  $\sigma^2$  if its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x - \mu)^2/2\sigma^2}$$

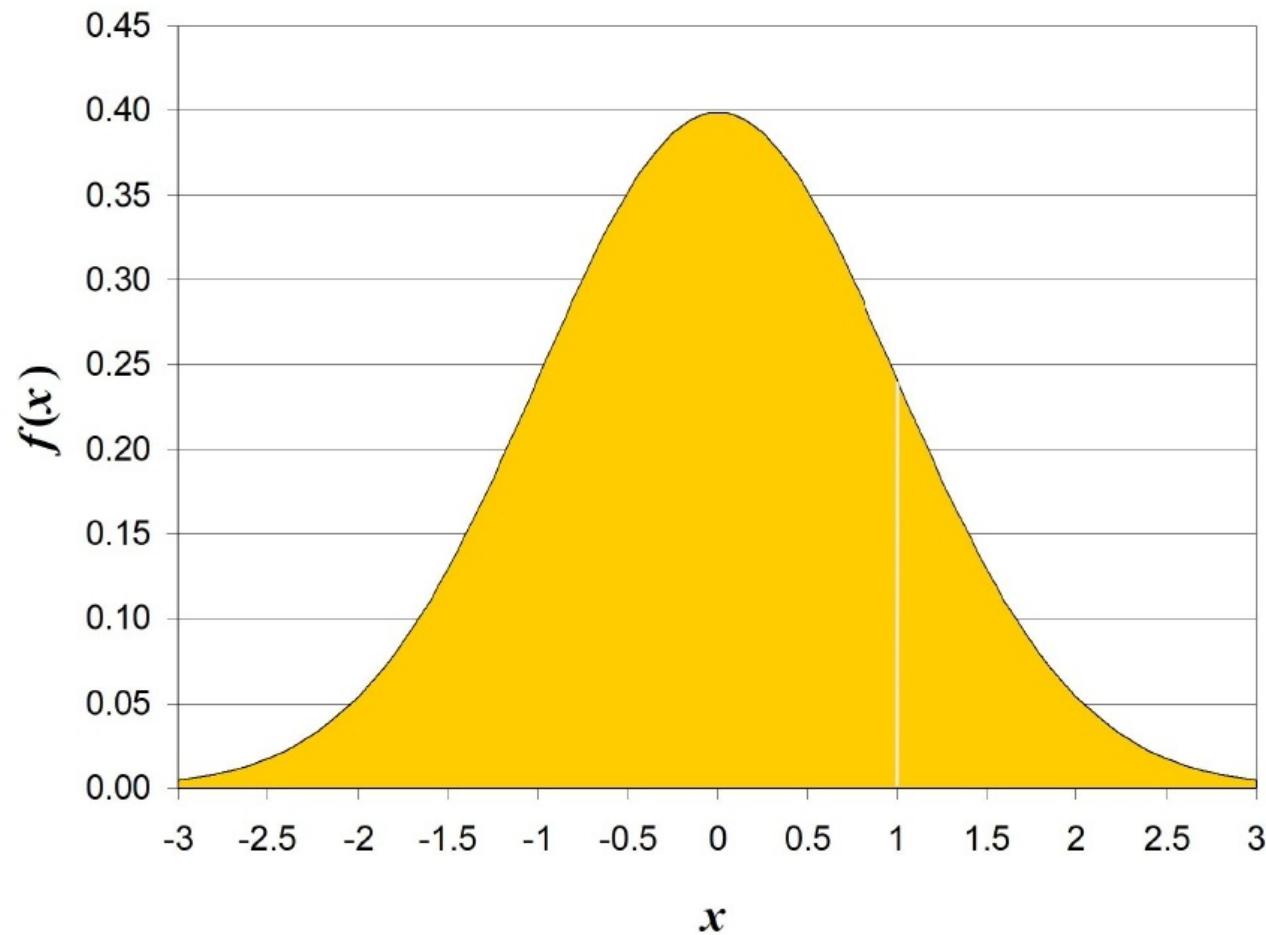
- u  $f(x)$  is not the same as  $P(x)$ 
  - u  $P(x)$  would be 0 for every  $x$  because the normal distribution is continuous

- u However,  $P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x) dx$



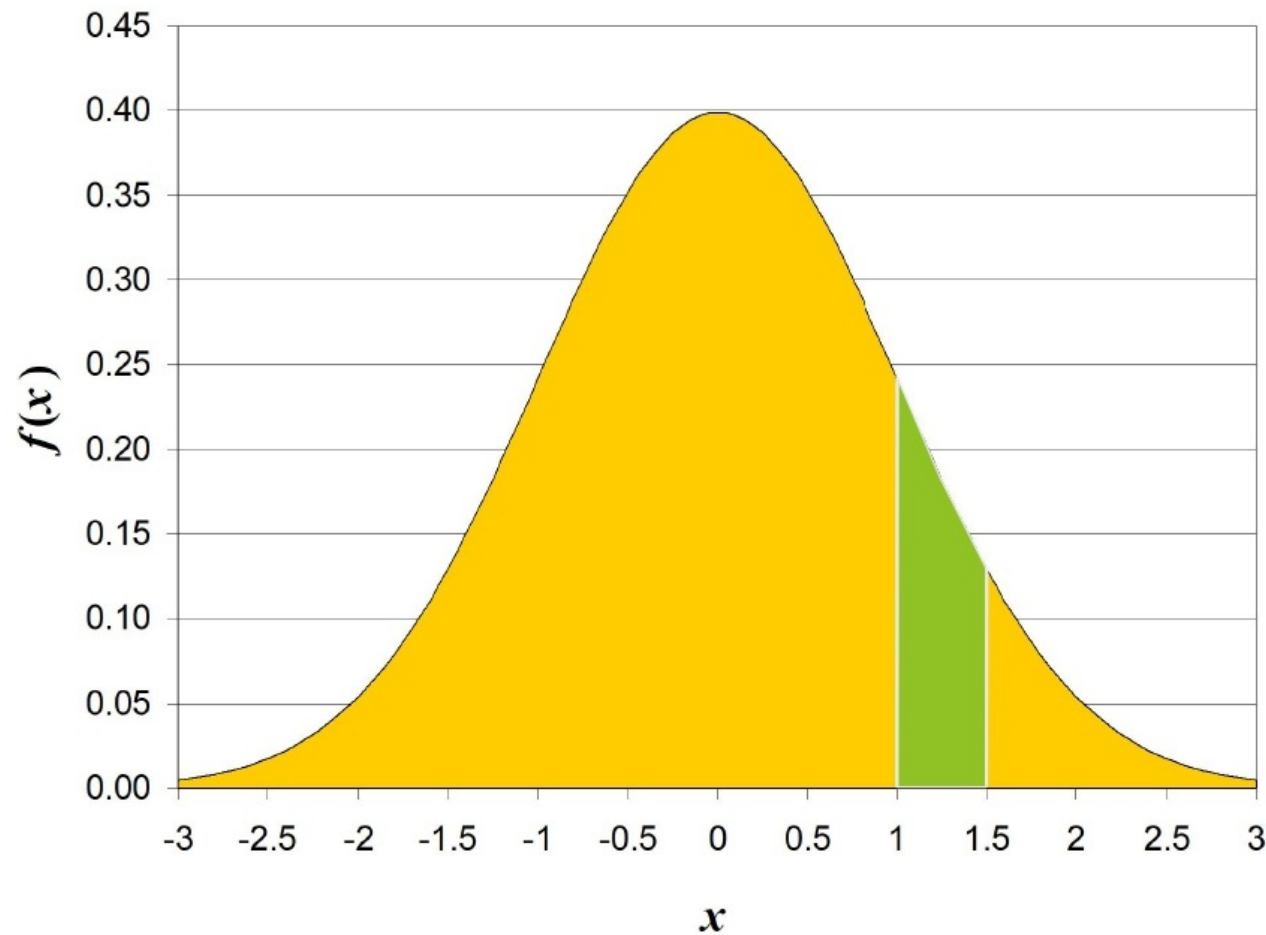
# The Normal Distribution

## Overview



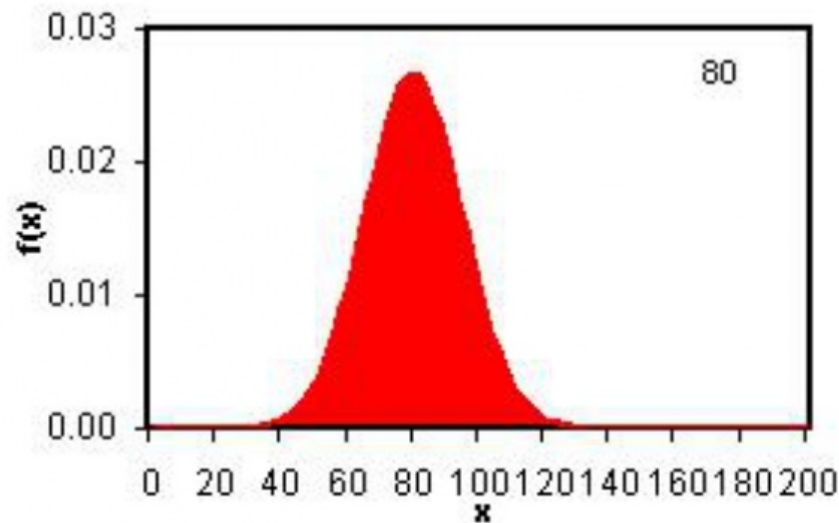
# The Normal Distribution

## Overview

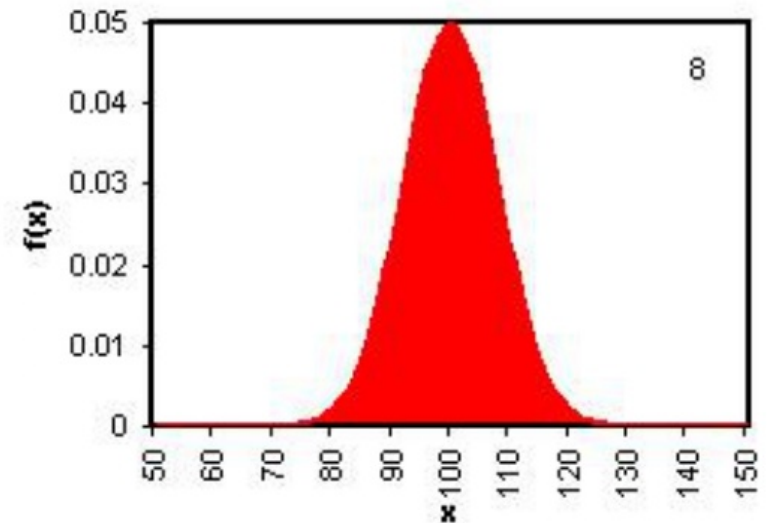


# The Normal Distribution

## Overview



Mean changes



Variance changes

# The Normal Distribution

## Length of Fish

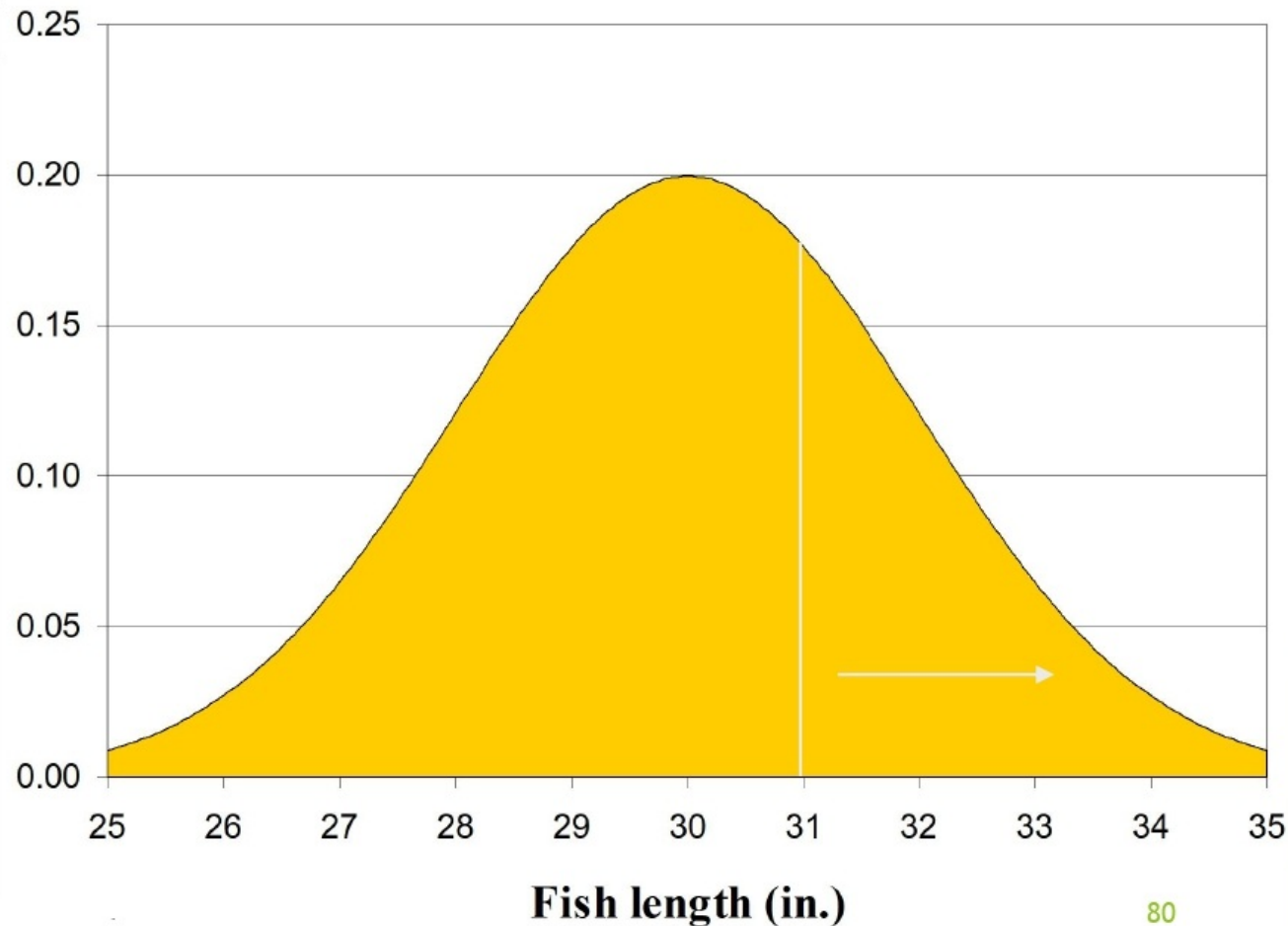
- u A sample of rock cod in Monterey Bay suggests that the mean length of these fish is  $\mu = 30$  in. and  $\sigma^2 = 4$  in.
- u Assume that the length of rock cod is a normal random variable
- u If we catch one of these fish in Monterey Bay,
  - u What is the probability that it will be at least 31 in. long?
  - u That it will be no more than 32 in. long?
  - u That its length will be between 26 and 29 inches?



# The Normal Distribution

## Length of Fish

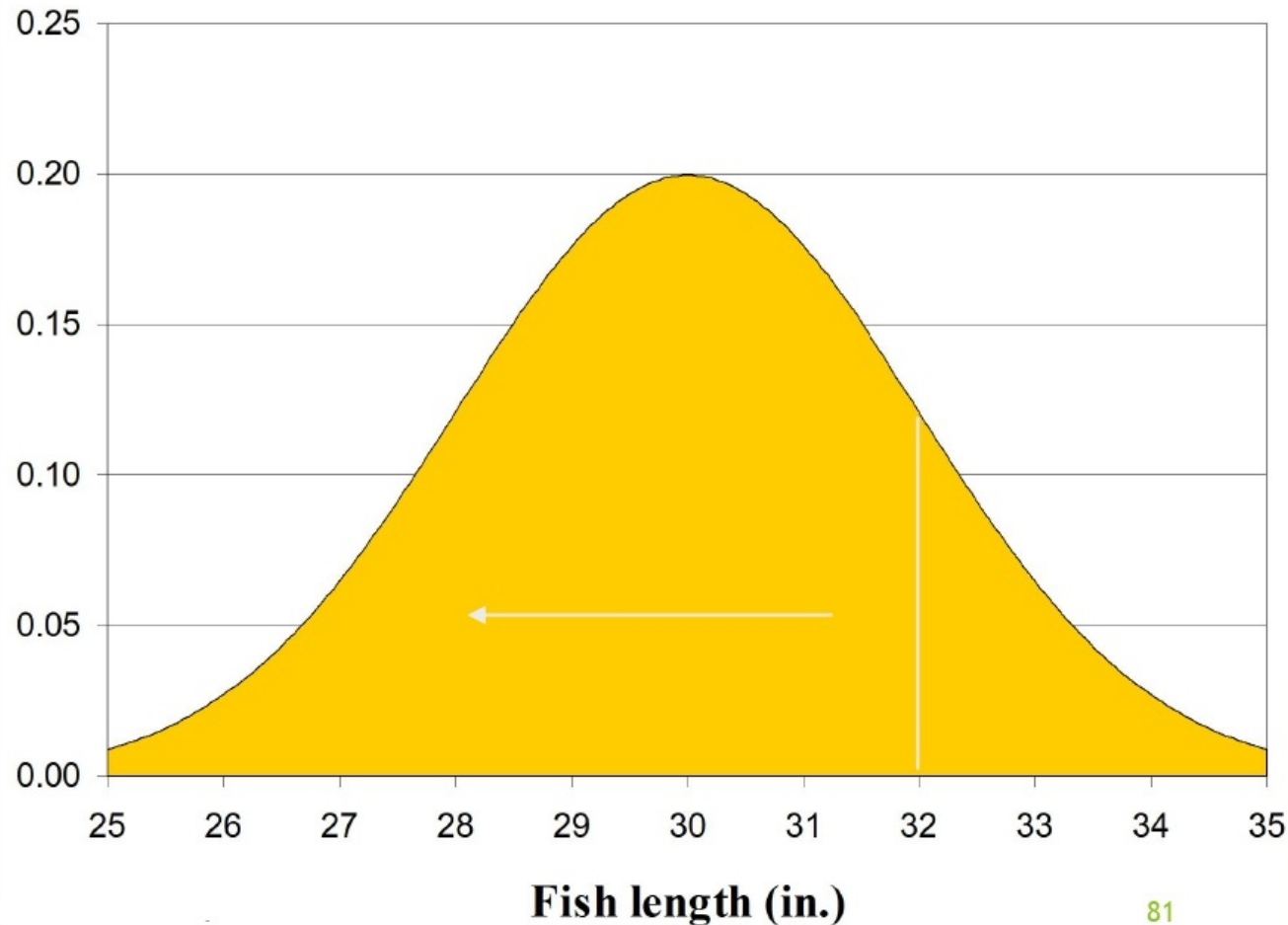
- What is the probability that it will be at least 31?



# The Normal Distribution

## Length of Fish

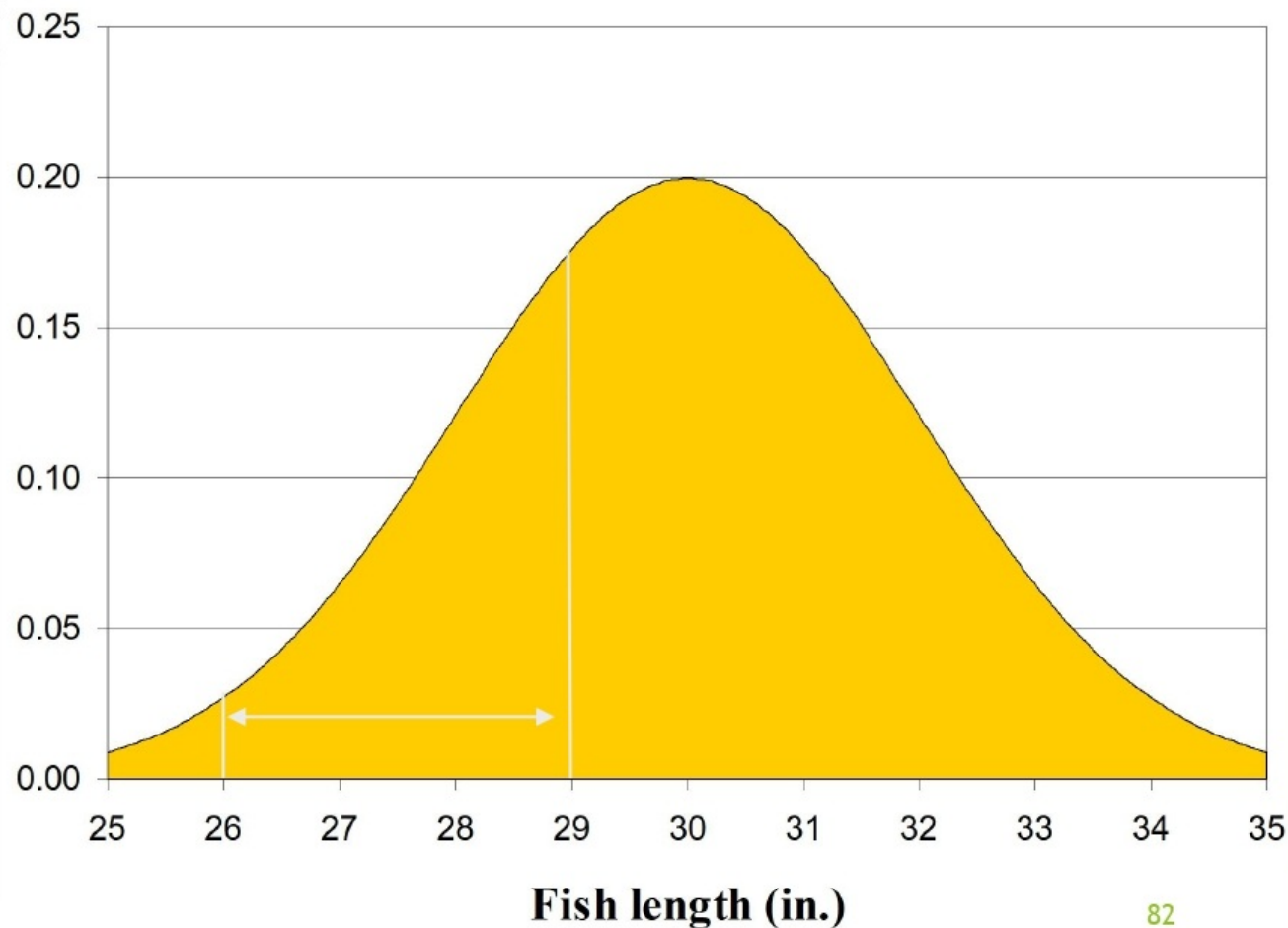
- u That it will be no more than 32 in. long?



# The Normal Distribution

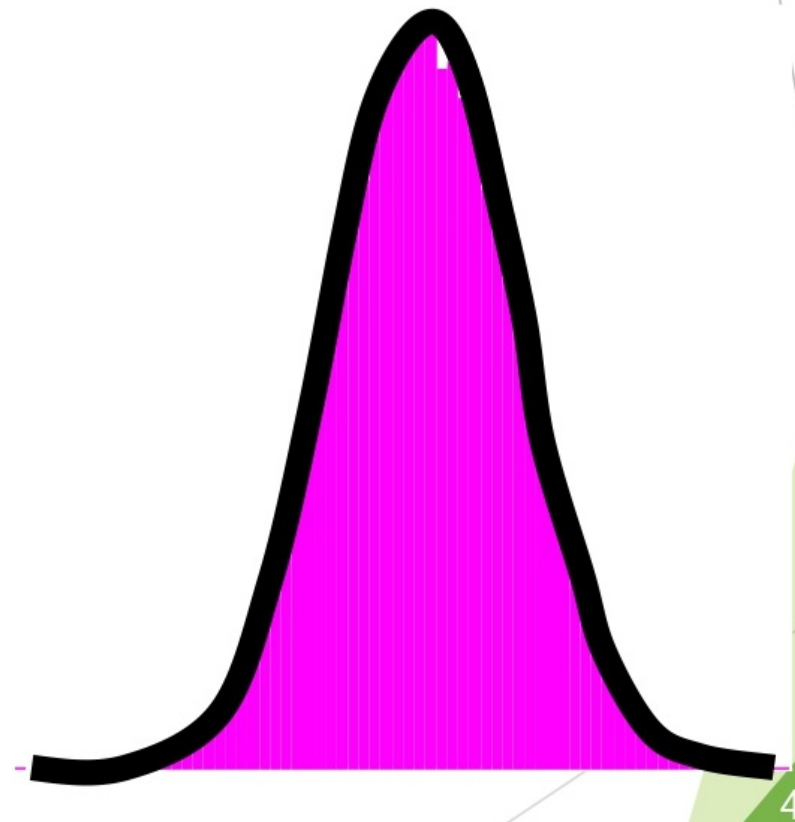
## Length of Fish

- u That its length will be between 26 and 29 inches?



# Standard Normal Distribution

•  $\mu=0$  and  $\sigma^2=1$





# Useful properties of the normal distribution

1. The normal distribution has useful properties:
  - u Can be added  $E(X+Y)=E(X)+E(Y)$  and  $\sigma^2(X+Y)=\sigma^2(X)+\sigma^2(Y)$
  - u Can be transformed with ***shift*** and ***change of scale*** operations

# Consider two random variables X and Y

Let  $X \sim N(\mu, \sigma)$  and let  $Y = aX + b$  where  $a$  and  $b$  are constants

*Change of scale* is the operation of multiplying  $X$  by a constant " $a$ " because one unit of  $X$  becomes " $a$ " units of  $Y$ .

*Shift* is the operation of adding a constant " $b$ " to  $X$  because we simply move our random variable  $X$  " $b$ " units along the x-axis.

If  $X$  is a normal random variable, then the new random variable  $Y$  created by these operations on  $X$  is also a normal random variable

# For $X \sim N(\mu, \sigma)$ and $Y = aX + b$

- $E(Y) = a\mu + b$
- $\sigma^2(Y) = a^2 \sigma^2$
- A special case of a change of scale and shift operation in which  $a = 1/\sigma$  and  $b = -1(\mu/\sigma)$
- $Y = (1/\sigma)X - \mu/\sigma$
- $Y = (X - \mu)/\sigma$  gives
- $E(Y) = 0$  and  $\sigma^2(Y) = 1$

# The Central Limit Theorem

- u That Standardizing any random variable that itself is a sum or average of a set of independent random variables results in a new random variable that is nearly the same as a standard normal one.
- u The only caveats are that the sample size must be large enough and that the observations themselves must be independent and all drawn from a distribution with common expectation and variance.



# SAMPLING TECHNIQUES AND SAMPLE SIZE DETERMINATION



# Sampling

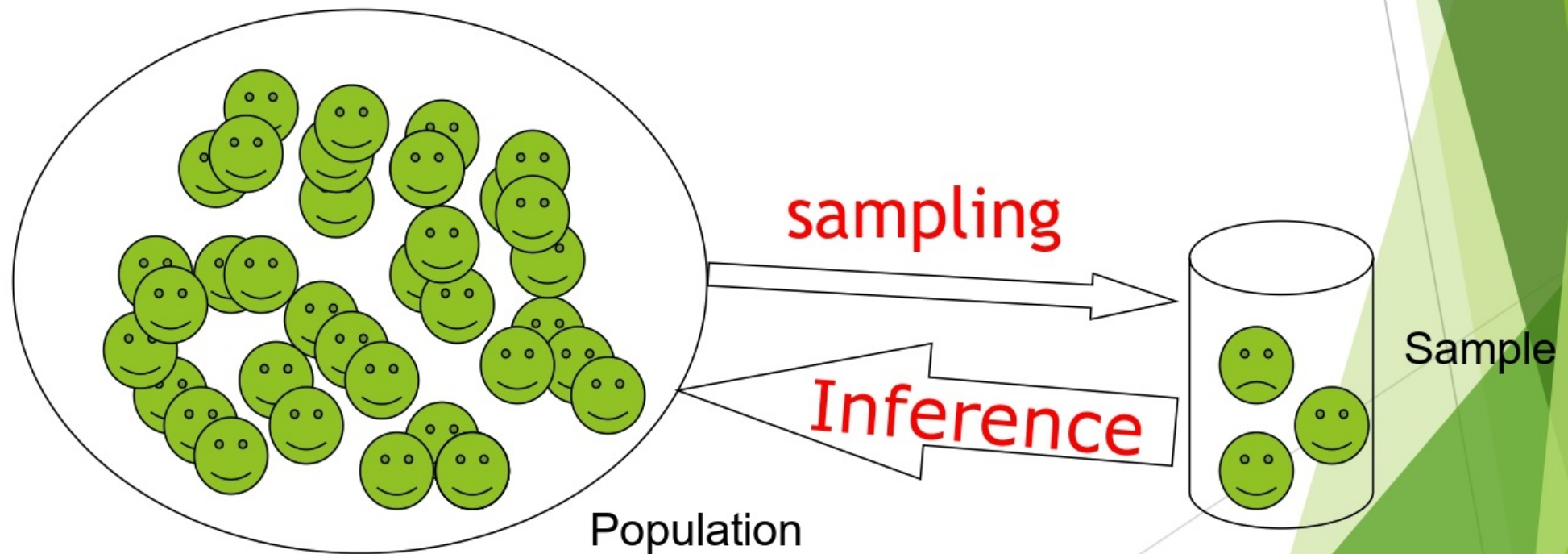
- A **sample** is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population” .
- Why sample?

# Why sample?

- u Cost in terms of money, time and manpower
- u Accessibility
- u Utility e.g. to do diagnostic laboratory test you don't draw the whole of patient's blood.
- A census is a sample consisting of the entire population.
- Even though a census is not full proof, it gives detailed information about every small area of the population.
- It has the following disadvantages:
  - u **Expensive**
  - u **Takes a long time**
  - u **Cumbersome & therefore inaccurately done ( a careful sample produces a more accurate data than a census.)**

# Sampling.....

- Sampling is the process of selecting a representative sample from populations.
- It Selecting cases (elements)—or locating people (or other units of analysis)—from a target population in order to study the population.

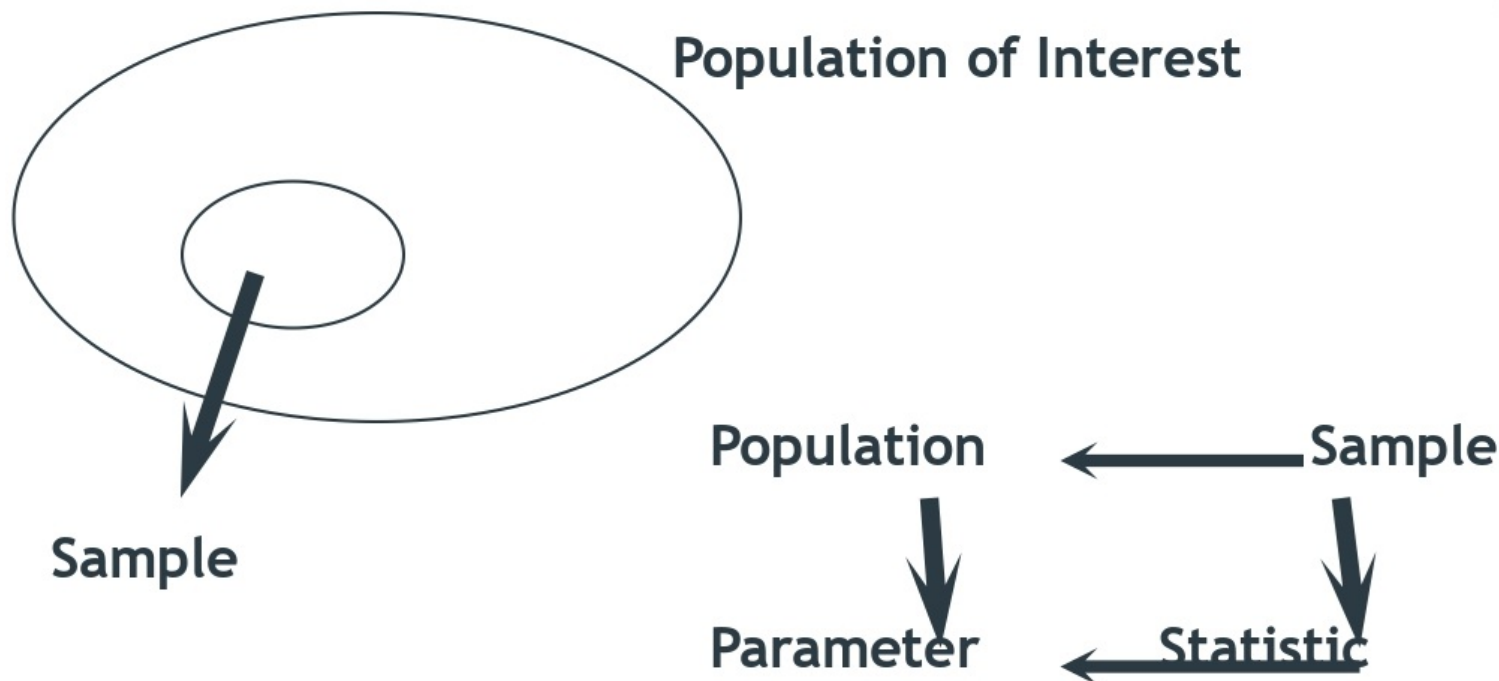


# Cont'd

- The process of obtaining information from a subset (sample) of a larger group (population)
- The results for the sample are then used to make estimates of the larger group
- Faster and cheaper than asking the entire population
- Two keys
  1. **Selecting the right people**
    - Have to be selected scientifically so that they are representative of the population
  2. **Selecting the right number of the right people**
    - To minimize sampling errors I.e. choosing the wrong people by chance



# Population Vs. Sample



We measure the sample using statistics in order to draw inferences about the population and its parameters.



# Characteristics of Good Samples

## ○ Representation

- Sample surveys are almost never conducted for the purposes of describing the particular sample under study. Rather they are conducted for purposes of **understanding the larger population** from which the sample was initially selected
- A great deal of work has been done over the years in developing sampling methods that provide representative samples for the general population.

E.g. international survey programs such as the DHS series, EPI coverage surveys have perfected the art of household sampling.

# Characteristics of Good Samples

- ✓ 3 factors that influence sample representativeness
  - υ Sampling procedure
  - υ Sample size
  - υ Participation (response)
- ✓ When might you sample the entire population?
  - υ When your population is very small
  - υ When you have extensive resources
  - υ When you don't expect a very high response
- **Accessible**
- **Low cost**

# Basic Terms

- **population (also called source population or target population):** is a group of individuals persons, objects, or items from which samples are taken for measurement.
  - It refers to the entire group of individuals or objects to which researchers are interested in generalizing the conclusions.

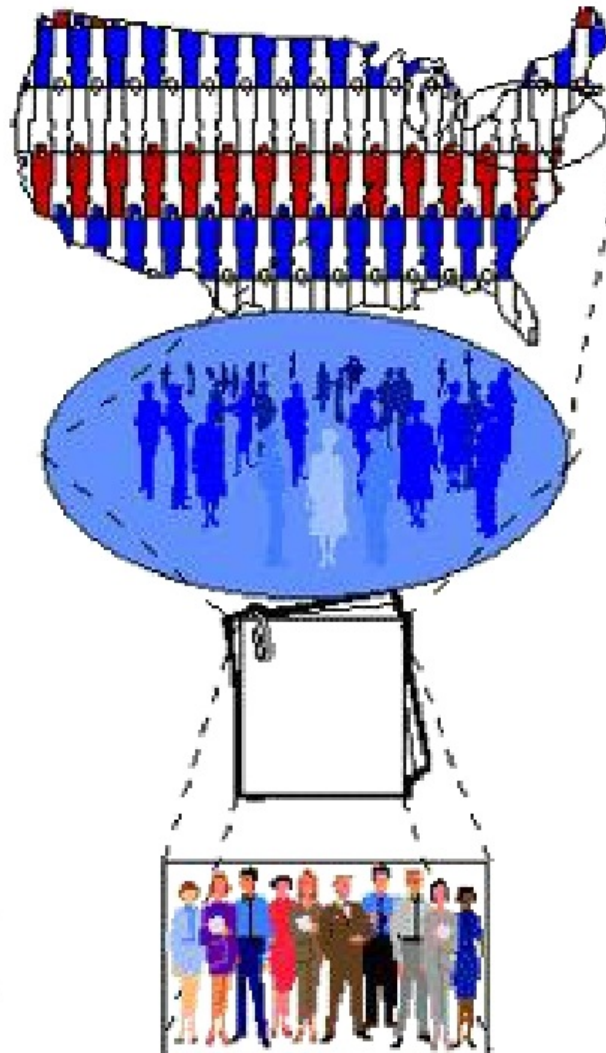
## Basic term cont'd....

Who do you want to generalize to?

What population can you get access to?

How can you get access to them?

Who is in your study?



The Theoretical Population

The Study Population

The Sampling Frame

The Sample



## Basic Terms cont'd...

- **Census:** Obtained by collecting information about each member of a population. Studying the whole population and requires a great deals of time, money and energy.
- **Sample survey:** study sample and draw conclusions about populations. It is cheaper in terms of cost, practical & convenient in terms of technicalities, saves time & energy.

## Basic Terms cont'd...

- **Sampling Frame:** is the list of people from which the sample is taken. It is the list from which the potential respondents are drawn.
  - ◆ It should be comprehensive, complete and up-to-date.
  - ◆ Examples of sampling frame: Electoral Register; Postcode Address File; telephone book and so on.
- **Probability samples:** With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample.

## Basic term cont'd....

- **Non-probability samples:** With non-probability sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen
- **Sampling unit** - the unit of selection in the sampling process
- **Study unit (study subjects)**- the unit on which information is collected or on which observations are made. E.g. Familiar examples are families, towns, litters, branches of a company, individual subjects or schools.



## Basic term cont'd....

- The **sampling unit** is not necessarily the same as the **study unit**.
  - If the objective is to determine the availability of latrine,
  - then the **study unit** would be the household;
  - If the objective is to determine the prevalence of trachoma, then the **study unit** would be the individual.
- ◆ **Sampling fraction (Sampling interval)** - the ratio of the number of units in the sample to the number of units in the reference population ( $N/n$ )



# Hierarchy of sampling

Study subjects

The actual participants in the study

Sample

Subjects who are selected

Sampling Frame

The list of potential subjects from which the sample is drawn

Source population

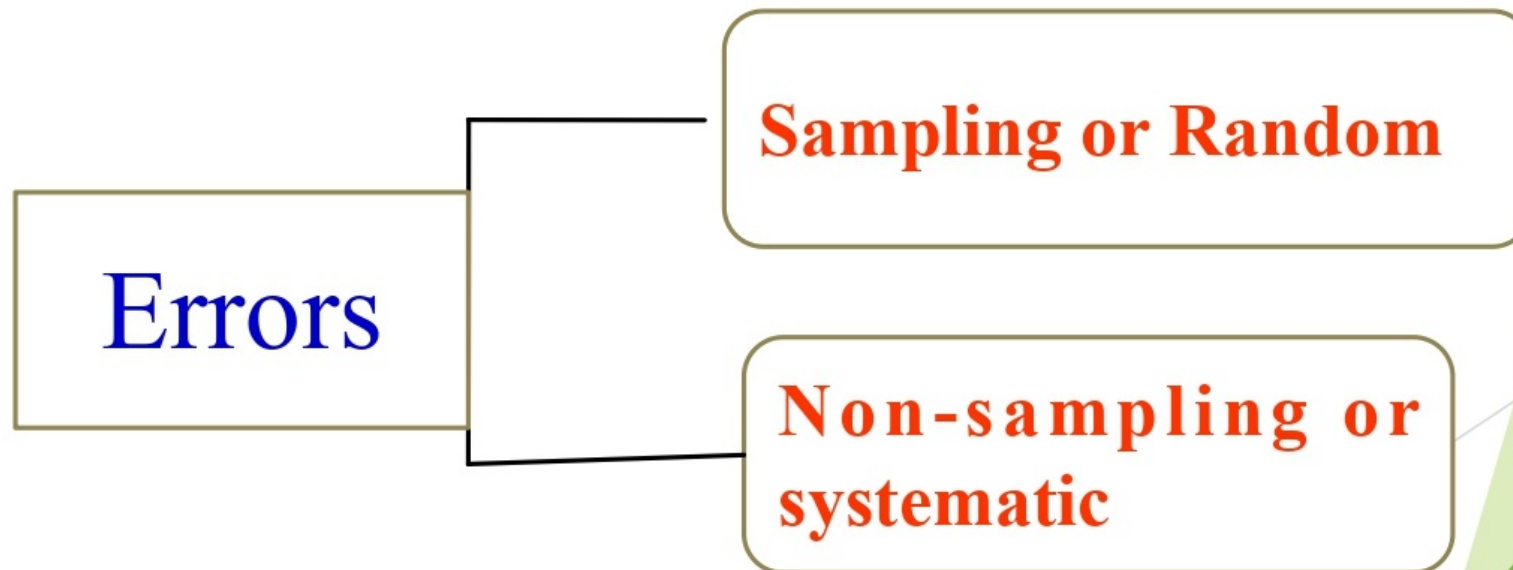
The Population from whom the study subjects would be obtained

Target population

The population to whom the results would be

# Errors in statistical Study

- ▶ A sample is expected to mirror the population from which it comes, however, there is no guarantee that any sample will be precisely representative of the population.
- ▶ No sample is the exact mirror image of the population .



# 1. Sampling error

- random error- the sample selected is not representative of the population due to chance
- The **uncertainty** associated with an estimate that is based on data gathered from a sample of the population rather than the full population is **known as sampling error**.
- Sampling errors are the **random variations** in the sample estimates around the true population parameters.

## Sampling error cont'd...

- ▶ **the level of it is controlled by sample size**
- ▶ a larger sample size leads to a smaller sampling error. it **decreases with the increase** in the size of the sample, and it happens to be of a smaller magnitude in case of homogeneous population.

When  $n = N \Rightarrow$  sampling error = 0

- ✓ Can not be avoided or totally eliminated



## Sampling error cont'd...

- ▶ why do **sample estimates have uncertainty** associated with them? There are two reasons.
  - Estimates of **characteristics from the sample** data can differ from those that would be obtained if the entire **population** were surveyed.
  - Estimates from one subset or sample of the population can differ from those based on a different sample from the same population (**sample to sample variations**).

# The cause of sampling error

- ▶ **Chance:** main cause of sampling error and is the error that occurs just because of **bad luck**.
- ▶ **Sampling bias:** Sampling bias is a tendency to favor the selection of participants that have particular **characteristics**.
- ▶ The chance component (sometimes called **random error**) exists no matter how carefully the selection procedures are implemented, and the only way to minimize chance-sampling errors is to select a **sufficiently large sample**.

## 2. Non Sampling Error

- ▶ It is a type of systematic error in the design or conduct of a sampling procedure which results in distortion of the sample, so that it is no longer representative of the reference population.
- ▶ We can eliminate or reduce the non-sampling error (bias) by **careful design of the sampling procedure** and not by increasing the sample size.
- ▶ It can occur whether the total study population or a sample is being used.



## ***Non-sampling Error.....***

- The basic types of non-sampling error
  - υ Non-response error
  - υ Response or data error
- A *non-response error* occurs when units selected as part of the sampling procedure do not respond in whole or in part
  - υ If non-respondents are not different from those that did respond, there is no non-response error
  - υ When non-respondents constitute a significant proportion of the sample (about 15% or more



## ***Non-sampling Error.....***

- A *response or data error* is any systematic bias that occurs during data collection, analysis or interpretation
  - Respondent error (e.g., lying, forgetting, etc.)
  - Interviewer bias
  - Recording errors
  - Poorly designed questionnaires

## Non-Sampling Error cont'd ...

- ▶ Systematic error makes survey results **unrepresentative** of the target population by distorting the survey estimates in one direction.
- ▶ Random error can distort the results in any given direction but tend to balance out on average
- ▶ Thus, the total survey error

sampling error + non-sampling error

## Advantage of sampling

- ▶ We obtain a sample rather than a complete enumeration (a census ) of the population for many reasons.
  - **Feasibility** it may be the only feasible method of collecting data
  - **Reduced cost** sampling reduces demands on resource such as finance, personal and material
  - **Greater accuracy** sampling may lead to better accuracy of collecting data.
  - **Greater speed** data can be collected and summarized more quickly

# Disadvantage of Sampling

- ▶ If sampling is biased, or not representative or too small the conclusion may not be **valid and reliable**
- ▶ If the population is very large and there are many sections and subsections, the sampling procedure becomes very complicated
- ▶ If the researcher does not possess the necessary skill and technical knowledge in sampling procedure, then the outcome will be devastated.



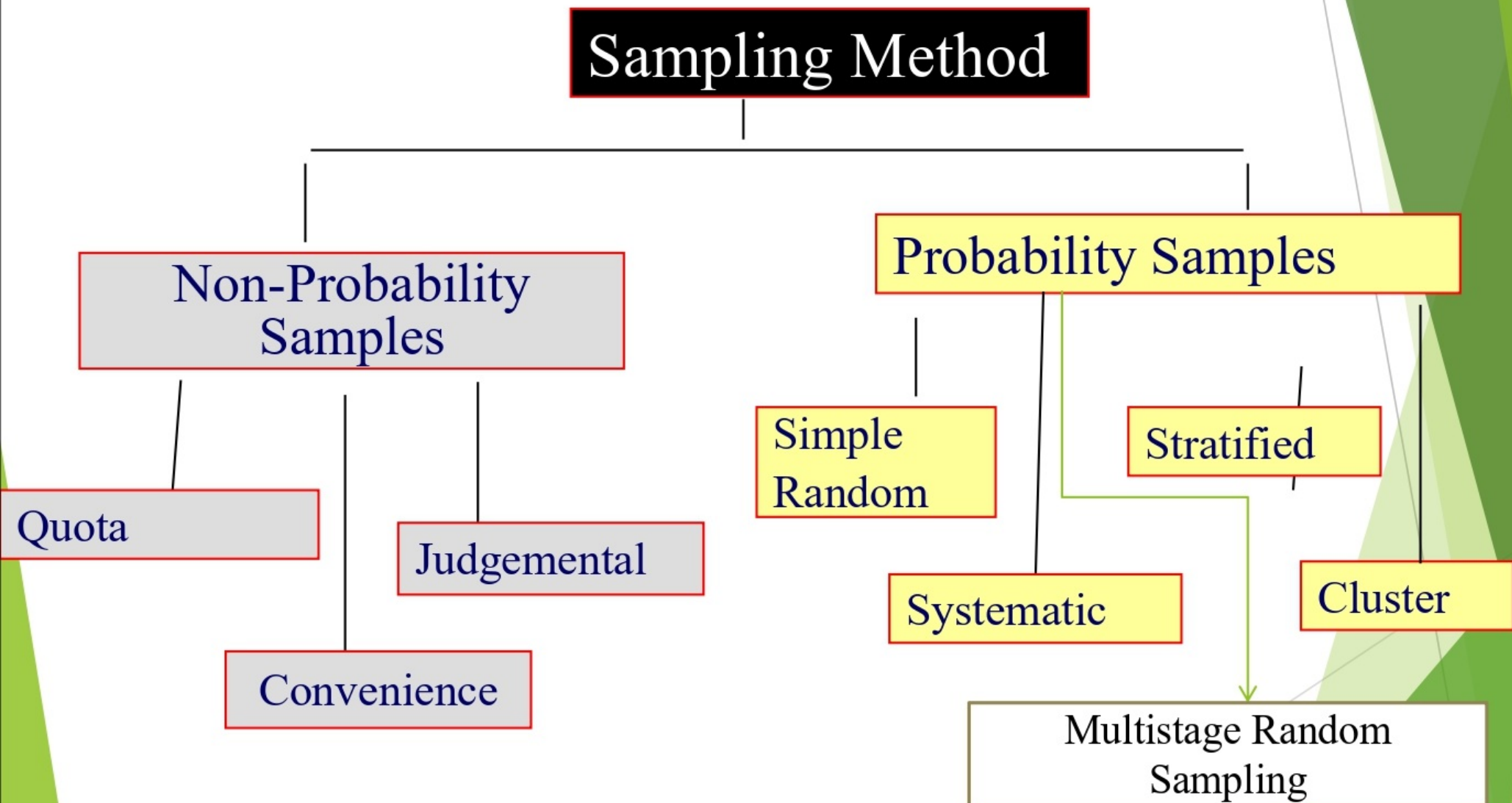
## Characteristics Of A Good Sample Design

- ▶ From what has been stated above, we can list down the characteristics of a good sample design as:
  - ➡ Sample design must result in a truly **representative sample**.
  - ➡ Sample design must be such which results in a **small sampling error**.
  - ➡ Sample design must be viable in the context of **funds available for the research study**.
  - ➡ Sample design must be such so that systematic bias can be controlled in a better way.
  - ➡ Sample should be such that the results of the sample study can be applied, **in general, for the universe with a reasonable level of confidence**.

# Types of Sampling

- ▶ How we Selecting the right subjects
  - The sample that we draw for our study determines the generalizability of our findings.
  - Sample should to have a good representation of the population.

# Types of Sampling Methods



# Probability Sampling Method ...

- ▶ The random ("equal chance") and "independent" components of random sampling are what makes us confident that the sample has a reasonable chance of representing the population
- ▶ What does it mean to be independent? The researchers select each person for the study separately.
- ▶ Let us say you were asked to participate in an experiment, enjoyed it, and told your friends to contact the researcher to volunteer for the study.
  - This would be an example of non-independent sampling.



# Probability Sampling Method cont'd ...

## ➤ In probability sampling

- ▶ A sampling **frame exists** or can be compiled.
- ▶ should have an equal or at least a known or **nonzero** chance of being included in the sample.
- ▶ Generalization is possible (from sample to population)
  - Simple Random Sampling,
  - Systematic Sampling,
  - Stratified Random Sampling,
  - Cluster Sampling
  - Multistage Sampling.

# 1. Simple Random Sampling(SRS)

- ▶ Simple random sampling is the **most straightforward** of the random sampling strategies.
- ▶ To use SRS there should be
  - sampling **frame** for the population
  - All possible samples of “**n**” subjects are equally likely (  $\frac{1}{n}$  ) to occur.
  - population is small, relatively homogeneous & readily available

# Simple Random Sampling cont'd ...

## ► Procedures to select the sample

- ✓ The specific procedures that you follow may vary depending on your resources, but all involve some type of random process. Depending on the complexity of the population, we can use different **tools to select “n” samples** from the given sampling frame.

- These are lottery method,
- **table of random number** (they are available in the appendix of many research methods and statistics textbooks) or
- computer generated random number.



# Simple Random Sampling cont'd ...

- ▶ Lottery method is appropriate if the total population is **not too large**, otherwise if the population is too large then it will be very difficult to use lottery method.

Thus, table of **random number or computer generated random number** is the feasible method to be used.

- ▶ Sampling schemes may be
  - **without replacement**- no element can be selected more than once in the same sample,  $\binom{N}{n}$  possible samples.
  - **with replacement**- an element may appear multiple times in the one sample  $N^n$  possible samples.



## 2. Systematic Random Sampling

- ⊕ Systematic sampling is thought as random, as long as the periodic interval is determined beforehand and **the starting point is random**
- ⊕ A method of selecting sample members from a larger population according to a **random starting point** and a fixed, periodic interval.
- ⊕ Typically, every  $n^{\text{th}}$  member is selected from the total population for inclusion in the sample population.
- ⊕ It is frequently chosen by researchers for its simplicity and its periodic quality.
- ⊕ it needs the population to be **homogeneous, however the method does not require frame.**

## Steps in systematic sampling:

- ▶ Define the population
- ▶ Determine the desired sample size ( $n$ )
- ▶ List the population from 1 to  $N$
- ▶ Determine  $K$ , where  $k=N/n$
- ▶ Select a random number between 1 and  $k$ , let us denote this number by “ $a$ ”  
Starting at  $a$ , take every  $K^{\text{th}}$  number on the list until the desired sample is obtained.
- ▶ Then the selected list will be

$a, a+k, a+2k, a+3k, \dots, a+(n-1)k$   
**Note: Systematic sampling should not be used when a cyclic repetition is inherent in the sampling frame**

## E.g. systematic sampling

- $N = 1200$ , and  $n = 60$   
→ sampling fraction =  $1200/60 = 20$
- List persons from 1 to 1200
- Randomly select a number between 1 and 20 (e.g. 8)
- 1<sup>st</sup> person selected = the 8<sup>th</sup> on the list
- 2<sup>nd</sup> person =  $8 + 20 = 28^{\text{th}}$  list e.t.c.

## Systematic sampling....

- It relies on arranging the target population according to some **ordering** scheme and then selecting elements at regular intervals through that ordered list.
- Systematic sampling involves a random start and then proceeds with the selection of every ***k*th** element from then onwards. In this case,  $k = (\text{population size} / \text{sample size})$ .
- It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the ***k*th** element in the list.



### 3. Stratified Random Sampling

- ▶ Stratified random sampling is used when we have **subgroups in our population** that are likely to differ substantially in their responses or behavior (i.e. **if the population is heterogeneous**).
- ▶ In stratified random sampling, the population is first divided into a number of **parts or 'strata'** according to some characteristic, chosen to be related to the major variables being studied.
- ▶ For example, you are interested in visual-spatial reasoning and previous research suggests that men and women will perform **differently** on these types of task

## Stratified Random Sampling cont'd...

- So, you divide your sample into male and female members and randomly select the required sample size within each subgroup (or "stratum")
- With this technique, you are guaranteed to have enough of each subgroup for meaningful analysis.
- Often we used simple random sampling to select a sample from each strata after stratification.

## Steps involve in stratified sampling method:

- ✦ Define the population
- ✦ Determine the desired sample size
- ✦ Identify the variable and subgroups (strata) for which you want to guarantee appropriate representation (either proportional or equal)
- ✦ Classify all members of the population as a member of one of the identified subgroups
- ✦ Randomly select (using simple random sampling or others) an appropriate number of individuals from each subgroup.
- ✦ Then the total sample size will be the sum of all samples from each subgroup.



- There are two methods to get the study subject from each subgroup,
  - proportional allocation or
  - equal allocation.
- We use proportional allocation technique when our subgroups vary dramatically in size in our population
- Let  $N$  be total population and  $N_1, N_2, \dots, N_k$  be the subtotal population for strata 1, 2, ...,  $K$  respectively. Moreover let  $n$  be the total sample size and  $n_1, n_2, \dots, n_k$  be the subsample for strata 1, 2, ...,  $k$  respectively in which  $N = N_1 + N_2 + \dots + N_K$  and  $n = n_1 + n_2 + \dots + n_k$
- Then the subsample " $n_i$ " which will be selected from subgroup  $N_i$  can be computed by

$$n_i = \frac{n * N_i}{N} \text{ where } i = 1, 2, 3, \dots, k$$



- ➡ The higher the population in the subgroup, the higher the sample size will be.
- ➡ However, **equal allocation** will be used if the total population from each subgroup is **approximately equal**.

# Advantage of stratified sampling over simple random sampling

- ➡ The representativeness of the sample is improved. That is, adequate representation of **minority subgroups** of interest can be ensured by stratification and by varying the sampling fraction between strata as required.

## ◆ **DEMERIT**

- ➡ Sampling frame for the entire population has to be prepared separately for each stratum.

## 4. Cluster Random Sampling

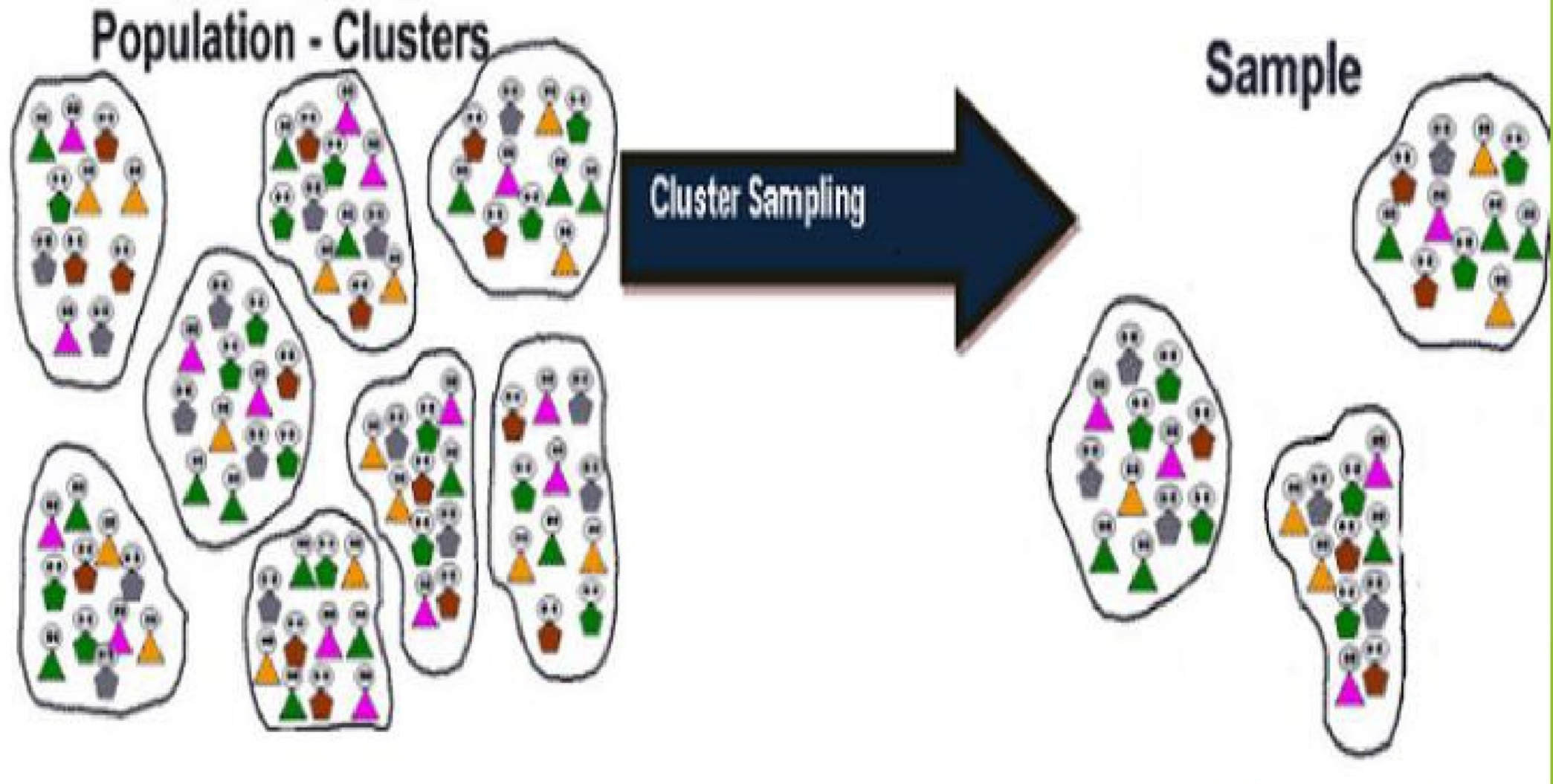
- ▶ In this sampling scheme, selection of the required sample is done on **groups of study units (clusters)** instead of each study unit individually.
- ▶ The sampling **unit is a cluster**, and the **sampling frame** is a list of these clusters.
- ▶ If the study covers wide geographical area, using the other methods will be too costly.
- ▶ The idea is, divided the total population in to different clusters and then the unit of selection will be cluster.
- ▶ Therefore, total population in the selected cluster will be taken as the sample.

## Steps in cluster sampling are:

- ✓ Define the population
- ✓ Determine the desired sample size
- ✓ Identify and define a **logical cluster** (can be kebele, Got, residence, and so on)
- ✓ Make a list of all clusters in the population
- ✓ Estimate the average number of population number per cluster
- ✓ Determine the number of clusters needed by dividing the **sample size by the estimated size of the cluster**
- ✓ Randomly select the required number of clusters (using table of random number as the total number of clusters is manageable)
- ✓ Include in the sample all population in the selected cluster.



- Consider the following graphical display:



## 5. Multistage Random Sampling

- ▶ This is the most complex sampling strategy.
- ▶ The researcher combines simpler sampling methods to address sampling needs in the most effective way of possible.
- ▶ **Example 1,**
  - ✓ The administrator might begin with a cluster sample of all **schools in the district.**
  - ✓ Then he might set up a **stratified sampling** process within clusters.
  - ✓ Within schools, the administrator could conduct a **simple random sample of classes or grades.**
  - ✓ By combining various methods, researchers achieve a rich variety of results useful in different contexts.

# Non-Probability Sampling Method

- ④ In the presence of constraints to use probability sampling strategies, the alternative sampling method is **non-probability sampling method**.
- ④ Non-probability sampling strategies are used when it is practically impossible to use probability sampling strategies.
- ④ **Non-probability** sampling is sampling procedure which does not afford any basis for **estimating the probability** that each item in the population has of being included in the sample.



## Cont'd.....

- Subjective units of population have a zero or unknown probability of selection before drawing the as sample. Hence obtained a non-representative samples.
- Sampling error can not be computed
- Survey results cannot be projected to the population

### Advantages

- Cheaper and faster than probability
- Reasonably representative if collected in a thorough manner



# 1. Judgment Sampling/ Purposive sampling

- ▶ The researcher selects the sample based on judgment. **A researcher exerts some effort in selecting a sample that seems to be most appropriate for the study.**
- ▶ This is used primarily when there is a limited number of people that have expertise in the area being researched

## 2. Convenience Sampling

- ▶ Convenience sampling selects a particular group of people but it does not come close to sampling all of a population.
- ▶ The sample would generalize only to similar programs in similar cities.
- ▶ It looks just like **cluster sampling**.
- ▶ The major difference is that the clusters of research participants are selected by convenience rather than by a random process.

Cont'd.....

- u Sometimes known as grab or opportunity sampling or accidental or haphazard sampling.
- u A type of non probability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, **readily available and convenient**.
- u The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough.
- u This type of sampling is most useful for pilot testing.

### 3. Quota sampling

- ▶ It is a method that ensures a certain number of sample units from different categories with specific characteristics are represented. The investigator interviews as many people in each category of study unit as he can find until he has filled his quota.
- ▶ It is the non-probability equivalent of **stratified sampling**. This differs from stratified sampling, where the strata are filled by random sampling.
- ▶ The population is first segmented into mutually exclusive sub-groups, just as in stratified sampling.



## Cont'd

- u Then judgment used to select subjects or units from each segment based on a specified proportion.
- u For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.
- u It is this second step which makes the technique one of non-probability sampling.

# Cont'd

- u In quota sampling the selection of the sample is non-random.
- u For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection.
- u This random element is its greatest weakness and quota versus probability has been a matter of controversy for many years

## 4. Snowball sampling

- ▶ It is a special non-probability method used when the desired sample characteristic is rare.
- ▶ Snowball sampling relies on referrals from **initial subjects to generate additional subjects.**
- ▶ What we need to do in case of snowball sampling is that first identify someone **who meets the criteria and then let him/her bring the other he/she knew.**

## Cont'd

- u While this technique can dramatically lower search costs, it comes at the expense of **introducing bias because** the technique itself reduces the likelihood that the sample will represent a good cross section from the population.
- u **Selection of additional respondents is based on referrals from the initial respondents.**
  - u friends of friends
- u **Used to sample from low incidence or rare populations.**



# Sample Size Determination

- ④ Determining the sample size for a study is a crucial component of study to include sufficient numbers of subjects so that statistically significant results can be detected.
- ④ "How large a sample do I need?"
- ④ The answer will depend on the **aims, nature and scope of the study and on the expected result**. All of which should be carefully considered at the planning stage.

# Sample.....

- If sample (“**n**”) is
  - **Large**
    - Increase accuracy
    - Costly / complex
  - **Small**
    - Decrease accuracy
    - Less costly



**Optimum  
sample**



**How ?**

# Factors to determine sample size

- u Size of population
- u Resources – subjects, financial, manpower
- u Method of Sampling- random, stratified
- u **Degree of difference** to be detected
- u Variability (**S.D.**) – pilot study, historical
- u Degree of Accuracy (or errors)
  - Type I error (**alpha**)  $p < 0.05$
  - Type II error (**beta**) less than 0.2 (20%)
  - Power of the test : more than 0.8 (80%)
- u Statistical Formulae
- u Dropout rate, non-compliance to Rx



- Sample size determination depending on outcome variables.
- ▶ There are **three** possible categories of **outcome variables**.
  - u **The first** is where the variable of interest has only two alternatives response: yes/no, dead/alive, vaccinated/not vaccinated and so on.
  - u **The second category** covers those outcome variable with multiple, mutually exclusive alternatives responses, such as marital status, religion, blood group and so on.
  - u For these two categories of outcome variables, the data are generally express as percentages or rates.
  - u So we can use percentage to compute the sample size.



- u The third category covers **continuous response variables** such as birth weight, age at first marriage, blood pressure and serum uric acid level, for which numerical measurement are usually made.
- u In this case the data are summarize in the form of **means and standard deviations or their derivatives.**

# Sample Size.....

- ▶ There are several approaches to determining the sample size.
- ▶ Depending on the type of response variable, whether it is categorical or continuous, we will have two sets of formulas.
- ▶ The sample size determination formulas come from the formulas for the **maximum error of the** estimates and is derived by solving for  $n$ .

# Sample for Single population

To estimate sample size for single survey using simple or systematic random sampling, need to know:

- Estimate of the prevalence of the outcome
- Precision desired
- Design effect
- Size of total population
- Level of confidence (always use 95%)

## Sample size for single population mean

- ▶ This is the condition in which the research question is about mean.
- ▶ **Standard deviation ( $\sigma$ )** of the population: It is rare that a researcher knows the exact standard deviation of the population.
- ▶ Typically, the standard deviation of the population is estimated:
  - from the results of a previous survey,
  - from a pilot study,
  - from secondary data,
  - from judgment of the researcher.



- ▶ **Maximum acceptable difference (w):** This is the maximum amount of error that you are willing to accept.
- ▶ **Desired confidence level ( $Z_{\alpha/2}$ ) :** is your level of certainty that the sample mean does not differ from the true population mean by more than the maximum acceptable difference. Commonly we use a 95% confidence level.
- ▶ Then the sample size determination formula for single population mean is defined by:

$$n = \frac{Z_{\alpha/2}^2 * \sigma^2}{w^2}$$

### ➤ Where

- υ  $\alpha$  = The level of significance which can be obtain as 1-confidence level.
- υ  $\sigma$  = Standard deviation of the population
- υ  $w$  = Maximum acceptable difference
- υ  $z_{\alpha/2}$  = The value under standard normal table for the given value of confidence level

## ► Sample Size for Single Population Proportion

- ➡ This is the situation in which the variable of interest is categorical.
- ➡ Three questions must be answered to determine the sample size for single population proportion:
- ➡ **Best estimate of population proportion of the variable of interest** : Make your best estimate of what the actual percent of the survey characteristic is.
- ➡ The possible source of this proportion are:
  - ✓ from the results of a previous study,
  - ✓ item from a pilot study,
  - ✓ item judgment of the researcher.
  - ✓ item Simply taking 50%

- ▶ Then the formula for the **sample size of single population proportion** is defined as:

$$n = \frac{z_{\alpha/2}^2 * p(1 - p)}{w^2}$$

- ▶ Where  $\alpha$  = the level of significance which can be obtained as 1- confidence level.
- ▶  $P$  = best estimate of population proportions
- ▶  $W$  = maximum acceptable difference
- ▶ the value under standard normal table for the given value of confidence level



# Example 1

- One of MPH student want to conduct a research on the prevalence of ANC utilization of mothers in DABAT district. Given that the prevalence from the previous study found to be 45.7% , what will be the sample size he should take to address his objective?

**Solution:**

- Margin of error  $d = 5\%$
- A confidence level of 95% will give the value of as  $Z_{\alpha/2} = 1.96$ .
- Then using the formula :

$$\begin{aligned} n &= \frac{\left( Z_{\alpha/2} \right)^2 P (1 - P )}{W^2} = \frac{\left( Z_{0.05/2} \right)^2 0.457 (1 - 0.457 )}{0.05^2} \\ &= \frac{(1.96)^2 0.457 (0.543 )}{0.05^2} \\ &= 382 \end{aligned}$$

# Some Considerations

- u The final sample size will be corrected for
  - Nonresponse, lost to follow up, lack of compliance and so on
  - Consider the total size of the population (N): if  $N < 10000$  then we need correction the formula which is defined by

$$n_f = \frac{n_o}{1 + \frac{n_o}{N}}$$

- u Where  $n_f$  = final sample size,  $n_o$  = sample size from the above formula and N total population.
  - Take the design effect in to account if needed

# Incorrect sample size will lead to

- Wrong conclusions
- Poor quality research (Errors)
  - Type II error can be minimized by increasing the sample size
- Waste of resources
- Loss of money
- Ethical problems
- Delay in completion

# Chi Square Test for Independence (Example)

|        | Republican | Democrat | Independent | Total |
|--------|------------|----------|-------------|-------|
| Male   | 200        | 150      | 50          | 400   |
| Female | 250        | 300      | 50          | 600   |
| Total  | 450        | 450      | 100         | 1000  |

- **State the hypotheses**

$H_0$ : Gender and voting preferences are independent.

$H_a$ : Gender and voting preferences are not independent

- **Choose significance level**

Say, 0.05



# Chi Square Test for Independence

## u Analyze sample data

- Degrees of freedom =

$$|g|-1 * |v|-1 = (2-1) * (3-1) = 2$$

- Expected frequency count =

$$E_{g,v} = (n_g * n_v) / n$$

$$E_{m,r} = (400 * 450) / 1000 = 180000/1000 = 180$$

$$E_{m,d} = (400 * 450) / 1000 = 180000/1000 = 180$$

$$E_{m,i} = (400 * 100) / 1000 = 40000/1000 = 40$$

$$E_{f,r} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{f,d} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{f,i} = (600 * 100) / 1000 = 60000/1000 = 60$$

|        | Republican | Democrat | Independent | Total |
|--------|------------|----------|-------------|-------|
| Male   | 200        | 150      | 50          | 400   |
| Female | 250        | 300      | 50          | 600   |
| Total  | 450        | 450      | 100         | 1000  |

# Chi Square Test for Independence

|        | Republican | Democrat | Independent | Total |
|--------|------------|----------|-------------|-------|
| Male   | 200        | 150      | 50          | 400   |
| Female | 250        | 300      | 50          | 600   |
| Total  | 450        | 450      | 100         | 1000  |

- Chi-square test statistic

$$X^2 = \left[ \sum \frac{(O_{g,v} - E_{g,v})^2}{E_{g,v}} \right]$$

- $X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40 +$

$$(250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/40$$

- $X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$

- $X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$

# Chi Square Test for Independence

- ▮ **P-value**

- ▮ Probability of observing a sample statistic as extreme as the test statistic
  - ▮  $P(X^2 \geq 16.2) = 0.0003$
- ▮ Since **P-value** (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis
- ▮ There is a relationship between gender and voting preference