

Sampling and Sample size

Wondwosen Teshager(MPH)

Learning objectives

After completing this chapter, the student should be able to:

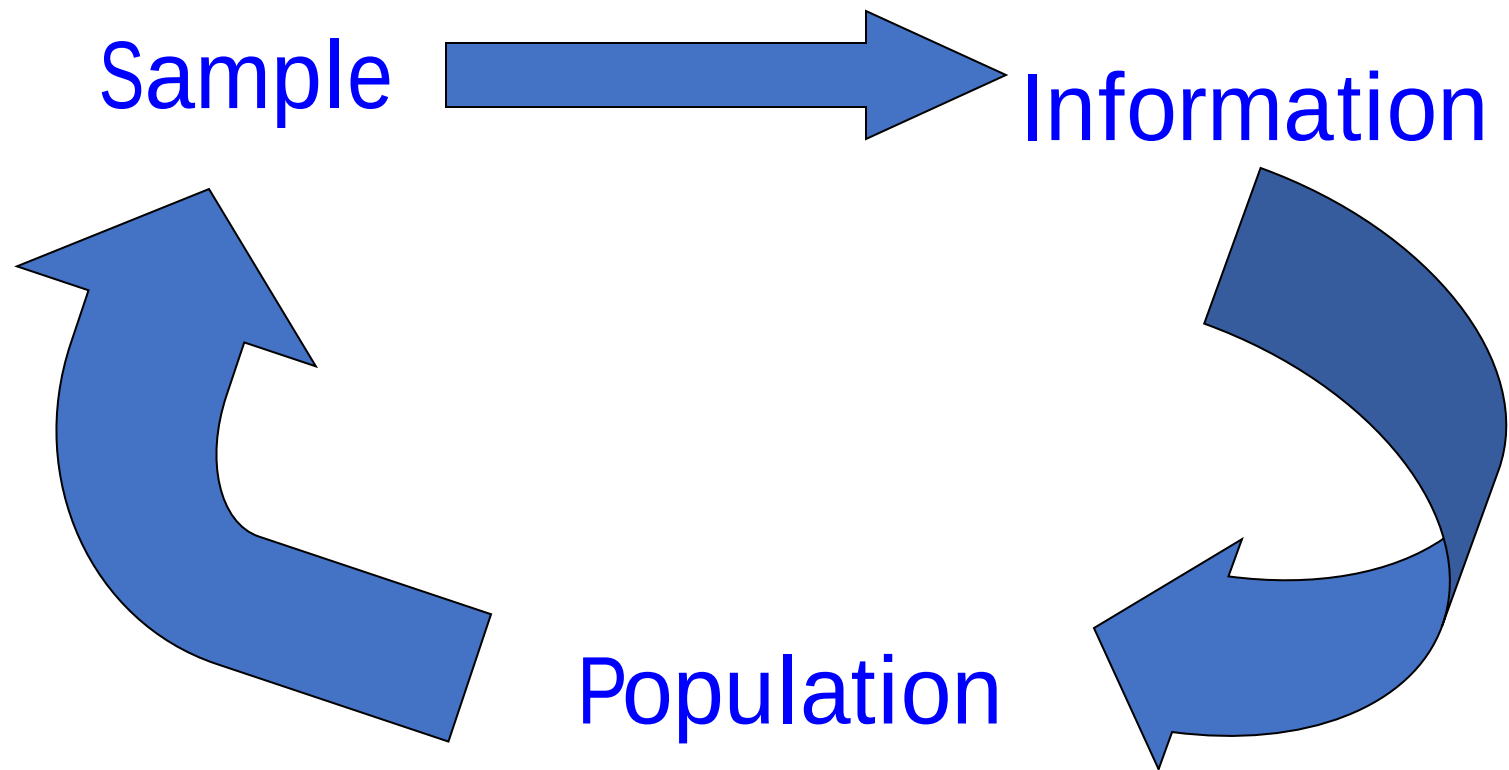
- Identify probability and non probability sampling techniques
- Be able to choose appropriate sampling technique
- Identify the advantages and limitations of each sampling techniques
- Calculate sample size for your research proposal

Sampling

- Some studies involve only small numbers of people and thus all of them can be included.
- however, research focuses on a large population is only possible to include some of its members in the investigation due to practical reasons, time, cost factor and other constraints.
- Thus we need sample.

Sampling

- ❑ Sampling is the process of selecting a number of study units from a defined study population.
- ❑ Sample is a finite subset of statistical individuals in a population and the number of individuals in a sample is called the sample size.
- ❑ To draw conclusions which are valid for the whole study population, the investigator should take care to draw a sample in such a way that it is representative of that population.



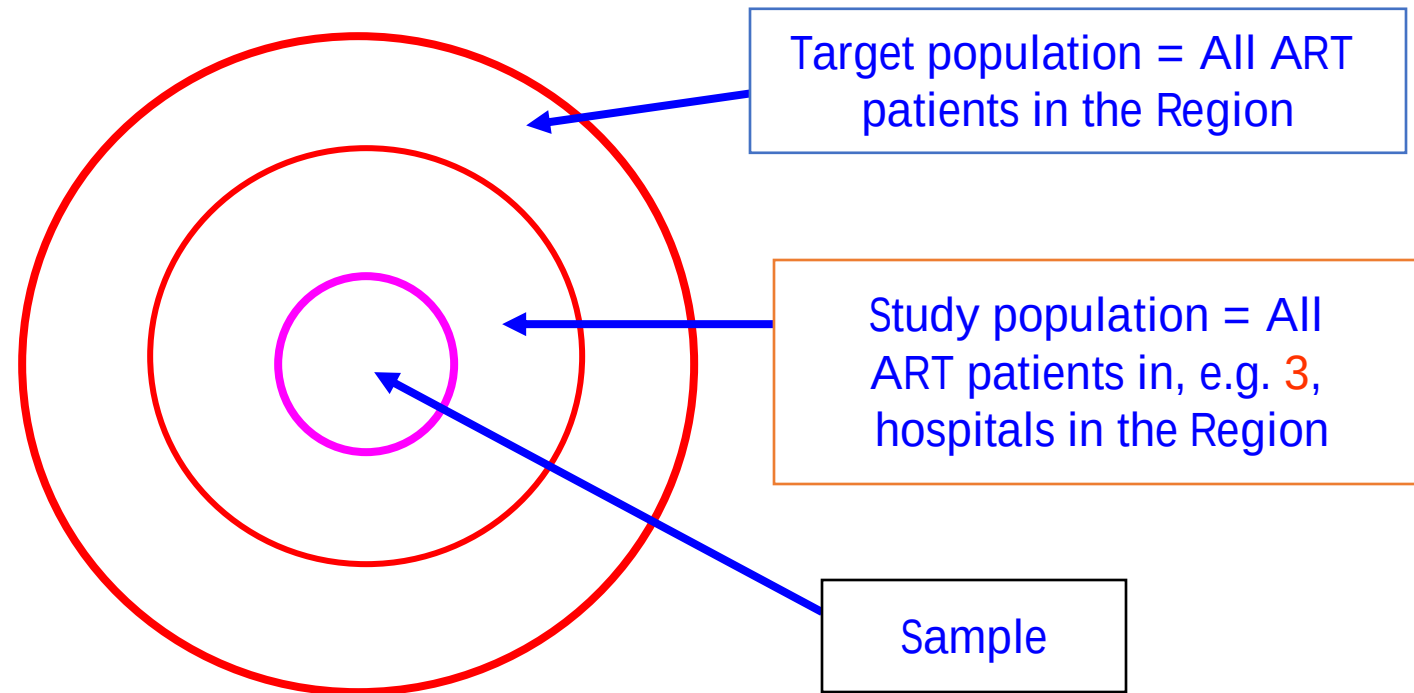
Common terms used in sampling

- **Population**: it is the collection of all items of interest.
- **Sampling**: It is the method by which we select a sample from the population
- **Reference population (or target population)**: the population of interest to whom the researchers would like to make generalizations.
- **Sample population**: the actual group in which the study is conducted = **Sample**

Common terms ----

- **Study population:** the subset of the target population from which a sample will be drawn.
- **Study unit:** the population on which information will be collected (measurement is done).
- **Sampling unit** - the unit of selection in the sampling process
- **Sampling frame:** It is the list of all the sampling units in the source population and from which a random sample is to be drawn.

Researchers are interested to identify factors associated with ART use among HIV/AIDS patients attending certain hospitals in a given Region



Advantages of sampling:

- **Feasibility**: Sampling may be the only feasible method of collecting information.
- **Reduced cost**: Sampling reduces demands on resource such as finance, personnel, and material.
- **Greater accuracy**: Sampling may lead to better accuracy of collecting data
- **Sampling error**: Precise allowance can be made for sampling error
- **Greater speed**: Data can be collected and summarized more quickly.

Disadvantage

Disadvantage

- There is always a sampling error
- Sampling may create a feeling of discrimination within the population
- Sampling may be inadvisable where every unit in the population is legally required to have a record.

Errors in sampling

1. Sampling error/ Random error

- 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.
- 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**.

1. Sampling error/ Random error

- It is the random variations in the sample estimates around the true population parameters
- No sample is the exact mirror image of the population
 - Sampling error (chance)
 - ✓ Can not be avoided or totally eliminated

Error sampling cont...

- Sampling error **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

Error in Sampling cont---

2. Non Sampling Error (Measurement 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.

Error in Sampling cont-

- It can occur whether the total study population or a sample is being used.
- It may either be produced by participants in the study or be an innocent by product of the **sampling plans and procedures**
- These biased observations can be innocent but very **devastating** to the findings of the study

Error in sampling cont'd...

- **Example:** If you take male students only from a student dormitory in Ethiopia in order to determine the proportion of smokers, you would result in an overestimate, since females are less likely to smoke.
 - Increasing the number of male students would not remove the bias.

sampling

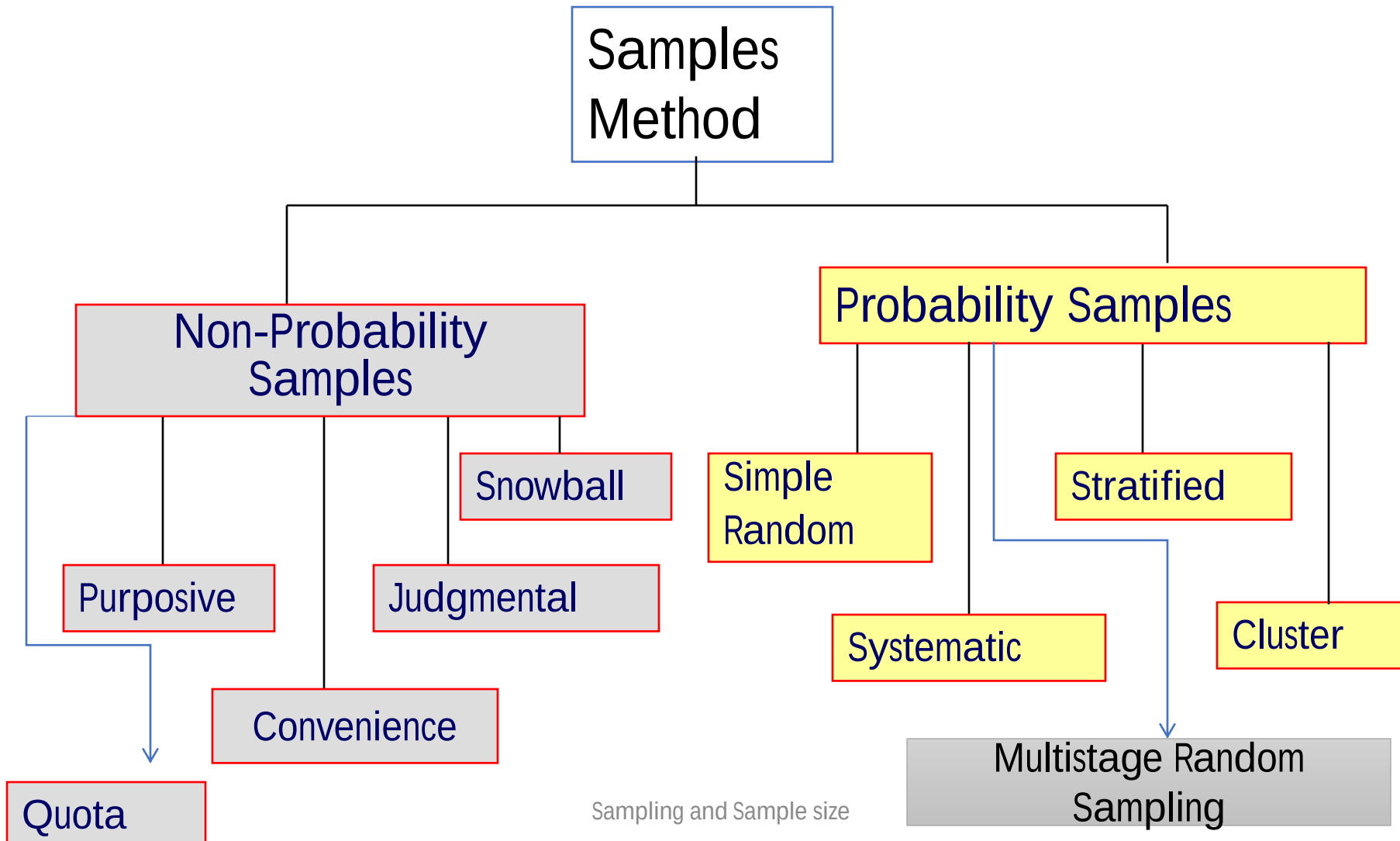
- While selecting a SAMPLE, there are basic questions:
 - What is the group of people (STUDY POPULATION) from which we want to draw a sample?
 - How many people do we need in our sample?
 - How will these people be selected?

Sampling Methods

Two broad divisions:

- A. Probability sampling methods
- B. Non-probability sampling methods

Types of Sampling Methods



A. Probability sampling methods

- Involves random selection of a sample
- Every sampling unit should have a known and non-zero probability of selection into the sample.
- Involves the selection of a sample from a population, based on chance.

1. Simple random sampling

- The required number of individuals are selected at random from the sampling frame, a list or a database of all individuals in the population.
- Each member of a population has an equal chance of being included in the sample.
- To use a SRS method:
 - Make a numbered list of all the units in the population i.e. Sampling frame
 - Each unit should be numbered from 1 to N (where N is the size of the population)
 - Select the required number.

1. Simple random sampling

- The randomness of the sample is ensured by:
 - Use of “lottery” methods
 - Table of random numbers
 - Computer programs

SRS has certain limitations:

- Difficult if the reference population is dispersed.
- Minority subgroups of interest may not be selected.

2. Systematic random sampling

- Sometimes called interval sampling
- Selection of individuals from the sampling frame systematically rather than randomly
- Individuals are taken at regular intervals down the list and The starting point is chosen at random

2. Systematic random sampling

- Important if the reference population is arranged in some order:
 - Order of registration of patients
 - Numerical number of house numbers
 - Student's registration books
- Taking individuals at fixed intervals (every k^{th}) based on the sampling fraction.

Steps in systematic random sampling

1. Number the units on your frame from 1 to N (where N is the total population size).
2. Determine the sampling interval (K) by dividing the number of **units** in the population by the desired sample size.
3. Select a number between one and K at random. This number is called the random start and would be the first number included in your sample.
4. Select every K^{th} unit after that first number

Note: Systematic sampling should not be used when a cyclic repetition is inherent in the sampling frame.

systematic random sampling

systematic random sample

number the units in the population from 1 to N

N = 100

decide on the n (sample size) that you want or need

want n = 20

$k = N/n =$ the interval size

N/n = 5

randomly select an integer between 1 to k

**select a random number from 1-5:
chose 4**

then take every kth unit

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

3. Stratified random sampling

- It is done when the population is known to be have **heterogeneity** with regard to some factors and those factors are used for stratification.
- Using **stratified sampling**, the population is divided into **homogeneous**, mutually exclusive groups called **strata**, and
- A population can be stratified by any variable that is available for all units prior to sampling (e.g., age, sex, province of residence, income, etc.).

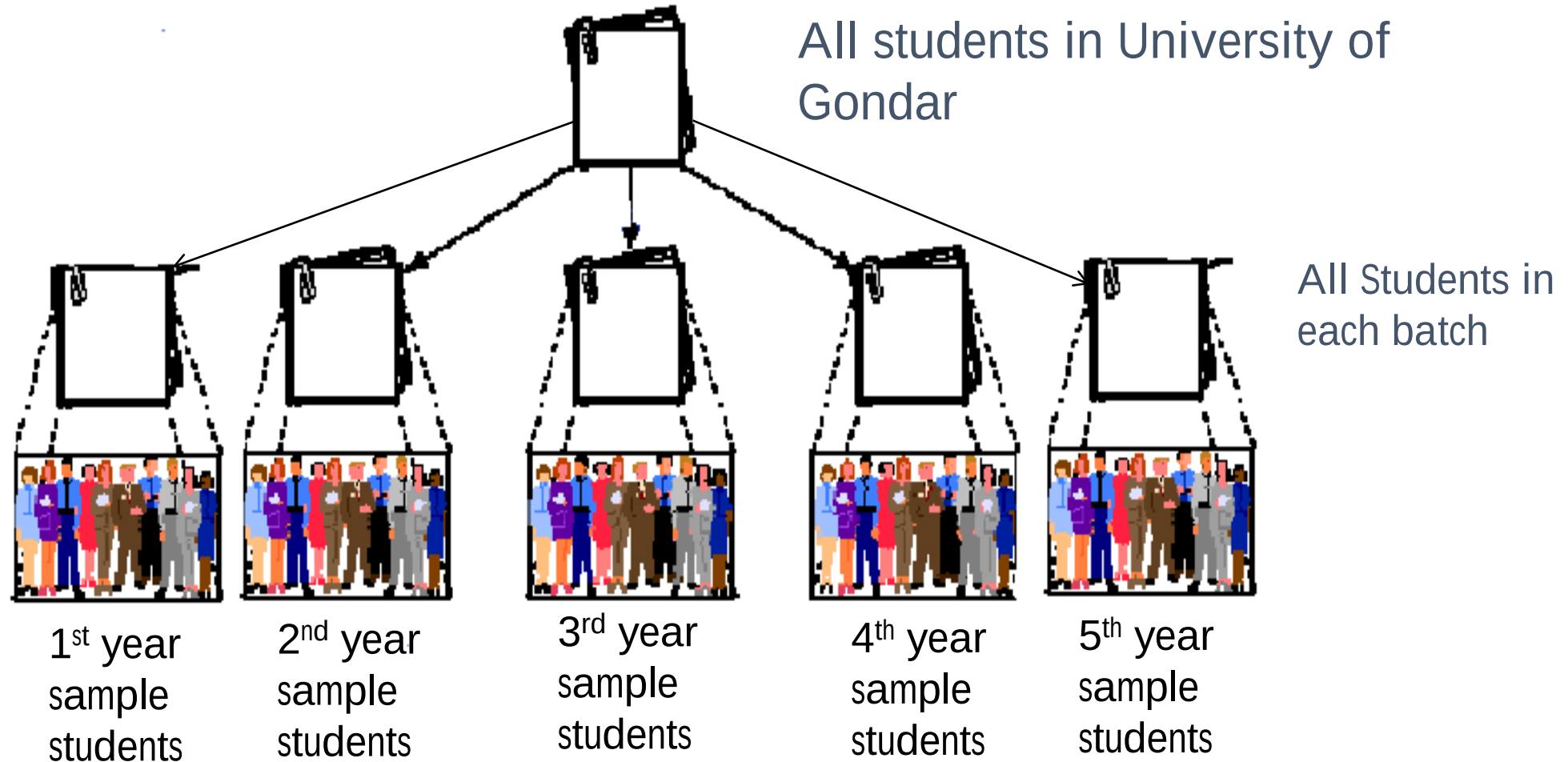
3. Stratified random sampling

- A separate sample is taken independently from each stratum.
- Any of the sampling methods mentioned in this section (and others that exist) can be used to sample within each stratum.
- Equal allocation:
 - Allocate equal sample size to each stratum
- Proportionate allocation:

Why do we need to create strata?

- It can make the sampling strategy more efficient.
- A larger sample is required to get a more accurate estimation if a characteristic varies greatly from one unit to the other.
- For example, if every person in a population had the same salary, then a sample of one individual would be enough to get a precise estimate of the average salary.

Stratified Random Sampling



4. Cluster sampling

- Sometimes it is too expensive to carry out SRS
 - Population may be **large and scattered**.
 - Complete list of the study population unavailable
 - Travel costs can become **expensive** if interviewers have to survey people from one end of the country to the other.

4. Cluster sampling

- Cluster sampling is the most widely used to reduce the cost
- The clusters should be homogeneous, unlike stratified sampling where the strata are heterogeneous
- It is the selection of groups of study units (clusters) instead of the selection of study units individually.

Steps in cluster sampling

- Cluster sampling divides the population into groups or clusters.
- A number of clusters are selected randomly to represent the total population, and then all units within selected clusters are included in the sample.
- No units from non-selected clusters are included in the sample — they are represented by those from selected clusters.
- This differs from stratified sampling, where some units are selected from each group.

Example

- In a school based study, we assume students of the same school are homogeneous.
- We can select randomly sections and include all students of the selected sections only

cluster sampling

Advantage

- Cost reduction
- It creates 'pockets' of sampled units instead of spreading the sample over the whole territory.
- Sometimes a list of all units in the population is not available, while a list of all clusters is either available or easy to create.

Disadvantage

- Creates a loss of efficiency when compared with SRS.
- Another drawback to cluster sampling is that you do not have total control over the final sample size.

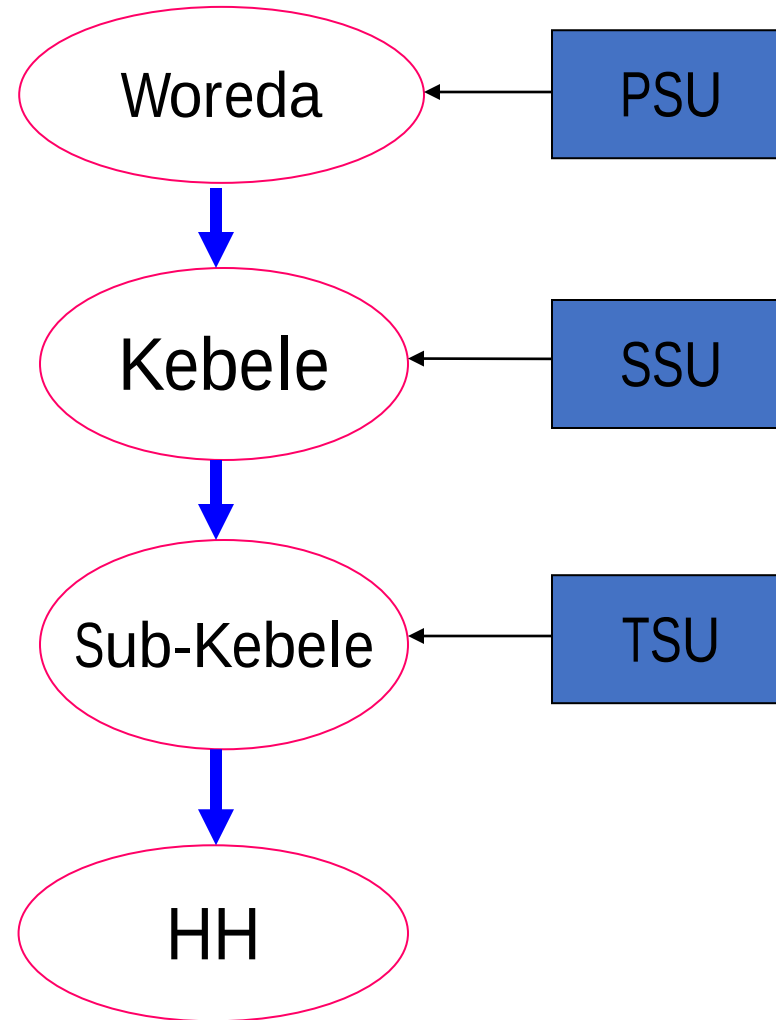
5. Multi-stage sampling

- Similar to the cluster sampling, except that it involves picking a sample from within each chosen cluster, rather than including all units in the cluster.
- This type of sampling requires at least two stages.
- The procedure is carried out in phases and it usually involves more than one sampling method.

5. Multi-stage sampling

- The primary sampling unit (PSU) is the sampling unit in the first sampling stage.
- The secondary sampling unit (SSU) is the sampling unit in the second sampling stage, etc.
- In the first stage, large groups or clusters are identified and selected. These clusters contain more population units than are needed for the final sample.

Sampling



5. Multi-stage sampling

- In the second stage, population units are picked from within the selected clusters (using any of the possible probability sampling methods) for a final sample.
- If more than two stages are used, the process of choosing population units within clusters continues until there is a final sample.
- With multi-stage sampling, you still have the benefit of a more concentrated sample for cost reduction.

B. Non-probability sampling

- In non-probability sampling, every item has an **unknown chance of being selected**.
- There is an assumption that **there is an even distribution of a characteristic of interest within the population**.
- But in probability sampling, **random is a feature of the selection process**.
- This is what makes the researcher believe that any sample would be representative and because of that, results will be accurate.

The most common types of non-probability sampling

1. Convenience or haphazard sampling
2. Volunteer sampling
3. Judgment sampling
4. Quota sampling
5. Snowball sampling technique

1. Convenience or haphazard sampling

- is sometimes referred to as haphazard or accidental sampling.
- It is a method in which for convenience sake the study units that happen to be available at the time of data collection are selected in the sample
- It is not normally representative of the target population because sample units are only selected if they can be accessed easily and conveniently.

1. Convenience or haphazard sampling

- The obvious advantage is that the method is easy to use, but that advantage is greatly offset by the presence of bias.
- Although useful applications of the technique are limited, it can deliver accurate results when the population is homogeneous.

2. Volunteer sampling

- As the term implies, this type of sampling occurs when **people volunteer to be involved in the study.**
- In psychological experiments or pharmaceutical trials (drug testing), for example, it would be difficult and unethical to enlist random participants from the general public.
- In these instances, the sample is taken from a group of volunteers.
- Sometimes, the researcher offers payment to attract respondents.

Volunteer sampling

- In exchange, the volunteers accept the possibility of a lengthy, demanding or sometimes unpleasant process.
- Sampling voluntary participants as opposed to the general population may introduce **strong biases**.
- Often in opinion polling, only the people who care strongly enough about the subject tend to respond.
- **The silent majority** does not typically respond, resulting in large **selection bias**.

3. Judgment sampling

- This approach is used when a sample is taken based on certain judgments about the overall population.
- The underlying assumption is that the investigator will select units that are characteristic of the population.
- The critical issue here is objectivity: how much can judgment be relied upon to arrive at a typical sample?
- Judgment sampling is subject to the researcher's biases.
- One advantage of judgment sampling is the reduced cost and time involved in acquiring the sample.

4. Quota sampling

- It is the non probability equivalent of stratified sampling.
- Like stratified sampling, the researcher first identifies the strata and their proportions as they are represented in the population.
- Then **convenience or judgment** sampling is used to select the **required number of subjects** from each stratum.
- This differs from stratified sampling, where the strata are filled by random sampling.

Quota sampling...

- This is one of the most common forms of non-probability sampling.
- Sampling is done until a specific/fixed number of units (quotas) for various sub-populations have been selected.
- The main argument against quota sampling is that **it does not meet the basic requirement of randomness.**
- Some units may have no chance of selection or the chance of selection may be unknown.
- Therefore, the sample may be **biased.**

Quota sampling...

- Quota sampling is generally less expensive than random sampling.
- It is also easy to administer, especially considering the tasks of listing the whole population, randomly selecting the sample and following-up on non-respondents can be omitted from the procedure.
- Quota sampling is an effective sampling method when information is urgently required and can be conducted without sampling frames.
- In many cases where the population has no suitable frame, quota sampling may be the only appropriate sampling method.

5. Snowball sampling

- A technique for selecting a research sample where existing study subjects recruit future subjects from among their friends.
- Thus the sample group appears to grow like a rolling snowball.
- This sampling technique is often used in hidden populations which are difficult for researchers to access; example populations would be drug users or commercial sex workers.
- Because sample members are not selected from a sampling frame, snowball samples are subject to numerous biases. For example, people who have many friends are more likely to be recruited into the sample.

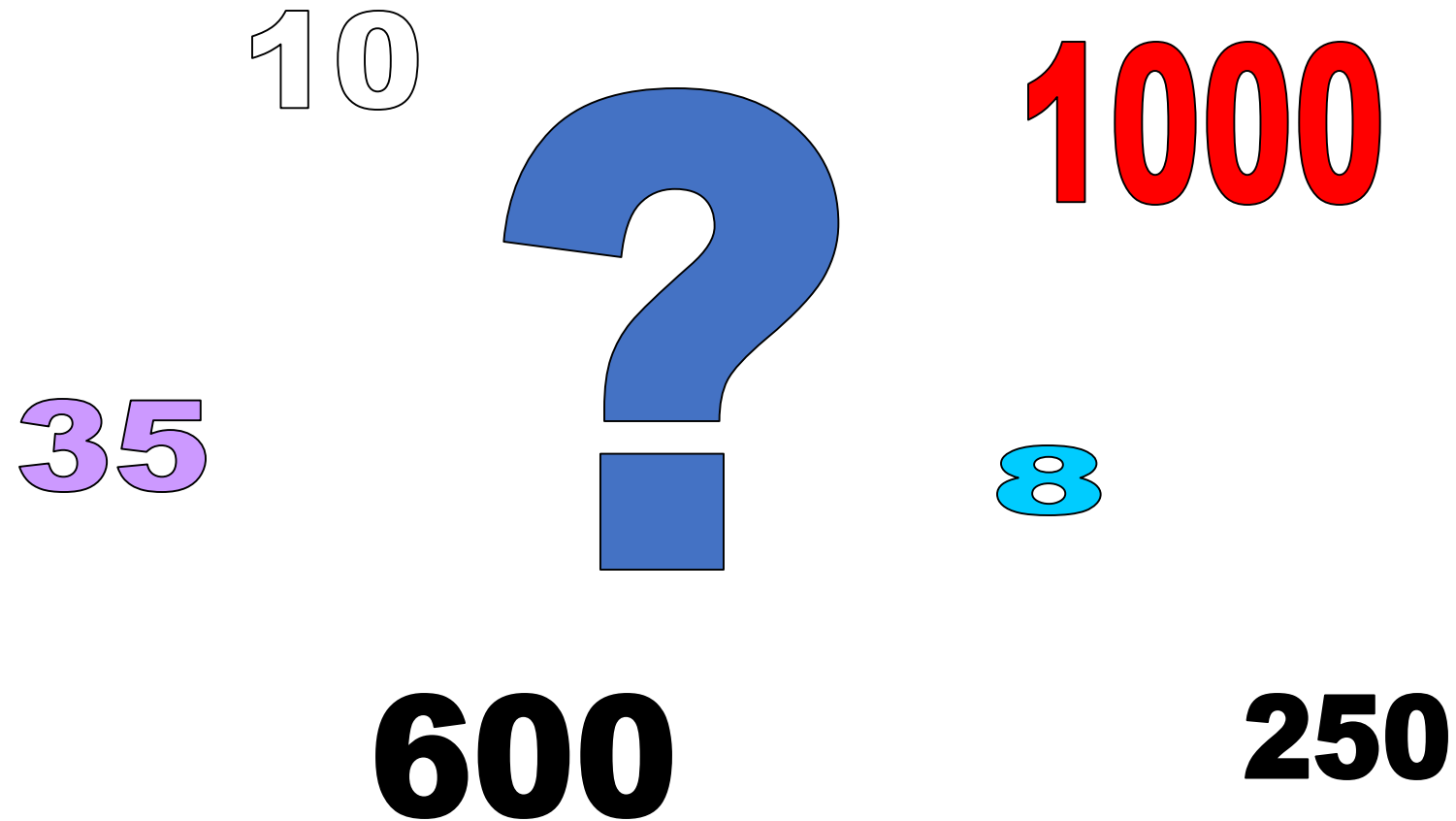
Sample size Determination

- In planning any investigation we must decide how many people need to be studied in order to answer the study objectives.
- The eventual sample size is usually a compromise between what is desirable and what is feasible
- The feasible sample size is determined by the availability of resources.
- It is also important to remember that resources are not only needed to collect the information, but also to analyze it.

Sample size Determination

- It is a widespread belief among researchers that the bigger the sample, the better the study becomes. This is not necessarily true.
- In general it is much better to increase the accuracy of data collection (improving by training of data collectors and pre testing data collection tools) than to increase sample size after a certain point.
- It is better to make extra efforts to get a representative sample rather than to get a very large sample.

How many people to study?



If too many....

- Waste of resources!



If too few....

- May fail to detect an important effect
- Estimates of effect may be too imprecise (wide CI's)

Sample size ...

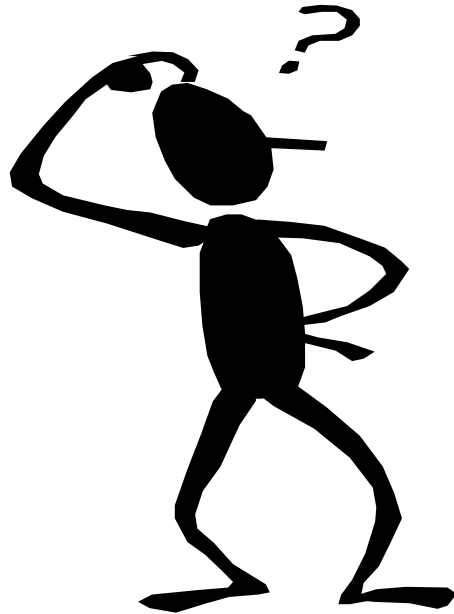
- Which variables should be included in sample size calculation?
 - It should relate to the study's primary outcome variable
 - If the study have secondary outcome variables which are considered important, the sample size should also be sufficient for the analysis of these variables.
- Answer depends on:
 - How different or dispersed the population is.
 - Desired level of confidence.
 - Desired degree of accuracy.

How to do we calculate a sample size

Formulae and software commands in notes

or

ask statistician



- Rules of thumb approach
- Confidence interval approach
- Hypothesis testing approach

In order to calculate the required sample size, you need to know the following facts:

a) The reasonable estimate of the key proportion to be studied.

- If you cannot guess the proportion, take it as 50%.

b) The degree of accuracy required.

- the allowed deviation from the true proportion in the population as a whole.
- It can be within 1% or 5%, etc.

Cont...

c) The confidence level required, usually specified as 95%.

d) The size of the population that the sample is to represent.

- If it is more than 10,000 the precise magnitude is not likely to be very important;
- but if the population is less than 10,000 then a smaller sample size may be required

Estimating a proportion

- To construct a confidence interval about a proportion, you must use the maximum error of the estimate, which is

$$E = z_{\alpha/2} \sqrt{\frac{pq}{n}}$$

- Confidence intervals about proportions must meet the criteria that $np\hat{p} \geq 5$ and $nq\hat{p} \geq 5$.

Estimating a proportion

- Estimate how big the proportion might be (P)
- Choose the margin of error you will allow in the estimate of the proportion (say $\pm w$).
- Choose the level of confidence that the proportion in the whole population is indeed between $(p-w)$ and $(p+w)$. We can never be 100% sure
- The minimum sample size required, for a very large population ($N > 10,000$)

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

Some Considerations

- 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 formul which is defined by:

$$n_f = \frac{n_0}{1 + \frac{n_0}{N}}$$

Where n_f =final sample size, n_0 =sample size from the above formula and N =total population.

- take the design effect in to account if needed

Example:

1. A hospital administrator wishes to know what proportion of discharged patients are unhappy with the care received during hospitalization. If 95% Confidence interval is desired to estimate the proportion within 5%, how large a sample should be drawn?

- $n = Z^2 p(1-p)/w^2 = (1.96)^2 (.5 \times .5) / (.05)^2 = 384.2 \approx 385$ patients

N.B. If you don't have any information about P, take it as 50% and get the maximum value of PQ which is 1/4 (i.e., 25%).

2. Suppose that you are interested to know the proportion of infants who breastfed >18 months of age in a rural area. Suppose that in a similar area, the proportion (p) of breastfed infants was found to be 0.20. What sample size is required to estimate the true proportion within $\pm 3\%$ points with 95% confidence. Let $p=0.20$, $d=0.03$, $\alpha = 5\%$

$$n = \frac{(z_{\alpha/2})^2 \cdot pq}{d^2} = \frac{(1.96)^2 (0.2)(0.8)}{(0.03)^2} = 683$$

Estimating a mean

□ the formula for determining a confidence interval for a mean will be

$$E = z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

E = maximum error of the estimate (also called the margin of error), σ = standard errors of the population

□ The required (minimum) sample size for a very large population is given by :

$$n = \frac{z_{\alpha/2}^2 \cdot \sigma^2}{E^2}$$

Examples

1: A health official wishes to estimate the mean serum cholesterol in a population of men. From previous similar studies a standard deviation of 40 mg/100ml was reported. If he is willing to tolerate a marginal error of up to 5 mg/100ml in his estimate, how many subjects should be included in his study? ($\alpha = 5\%$, two sided)

- a) If the population size is assumed to be very large, the required sample size would be:

$$n = (1.96)^2 (40)^2 / (5)^2 = 245.86 \approx 246 \text{ persons}$$

2. Find the minimum sample size needed to estimate the drop in heart rate (μ) for a new study using a higher dose of propranolol than the standard one. We require that the two-sided 95% CI for μ be no wider than 5 beats per minute and the sample sd for change in heart rate equals 10 beats per minute.

$$n = (1.96)^2 10^2 / (2.5)^2 = 62 \text{ patients}$$

3. Suppose that for a certain group of cancer patients, we are interested in estimating the mean age at diagnosis. We would like a 95% CI of 5 years wide. If the population SD is 12 years, how large should our sample be?

$$n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{d^2} = \frac{(1.96)^2 (144)}{(2.5)^2} = 88.5 \approx 89$$

Suppose $d=1$ Then the sample size increases

$$n = \frac{z^2 \cdot \sigma^2}{d^2} = \frac{(1.96)^2 (144)}{(1)^2} = 553.2 \approx 554$$