بسم الله الرحمن الرحيم

Sampling

By

Dr. Nanees ghareeb

Definitions

Population:

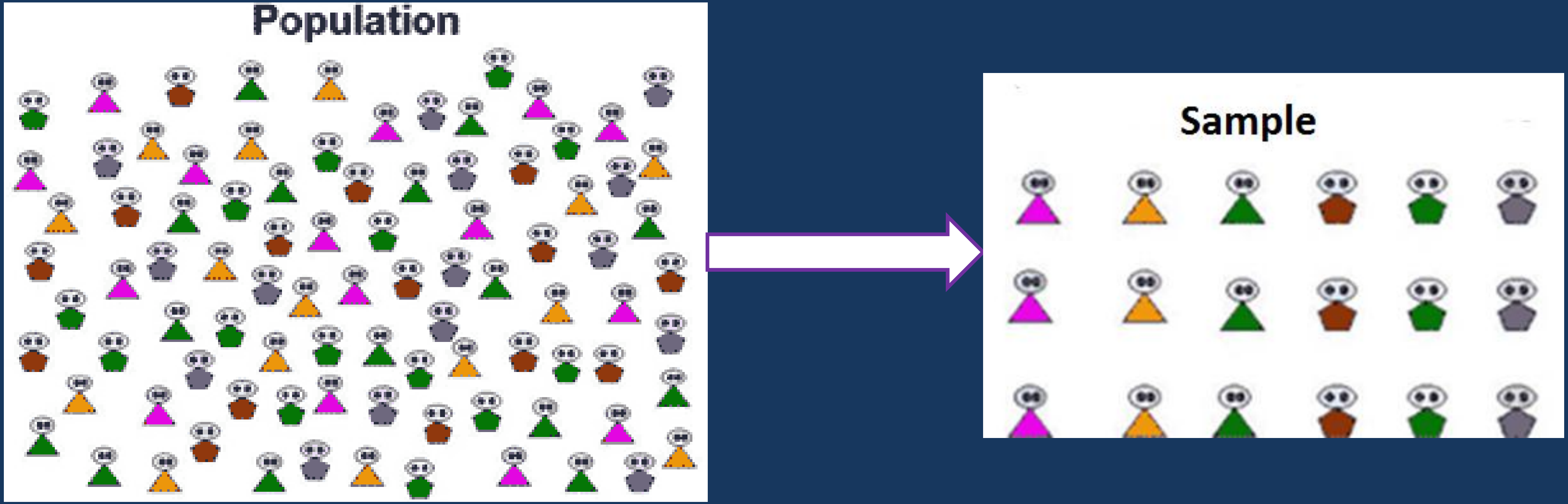
All people living in a place or any collection of individuals or things that we are interested in and their number may be finite or infinite e.g. Egyptians, students, blood cells, etc.

- Ideally to carry out an epidemiologic study we should examine the whole population,

- but since this is not always possible because it is:

  1. expensive
  2. time consuming and
  3. not feasible

- So, we have to select a group from the population → sample.


Population
Sample

Sample:

Group of individuals or things taken from a larger population and used to find certain information about this population.

Example: **examination of 5ml of blood can diagnose liver disease. We are not in need to examine all blood.**

The way that we follow in the selection of the sample will determine whether it is:

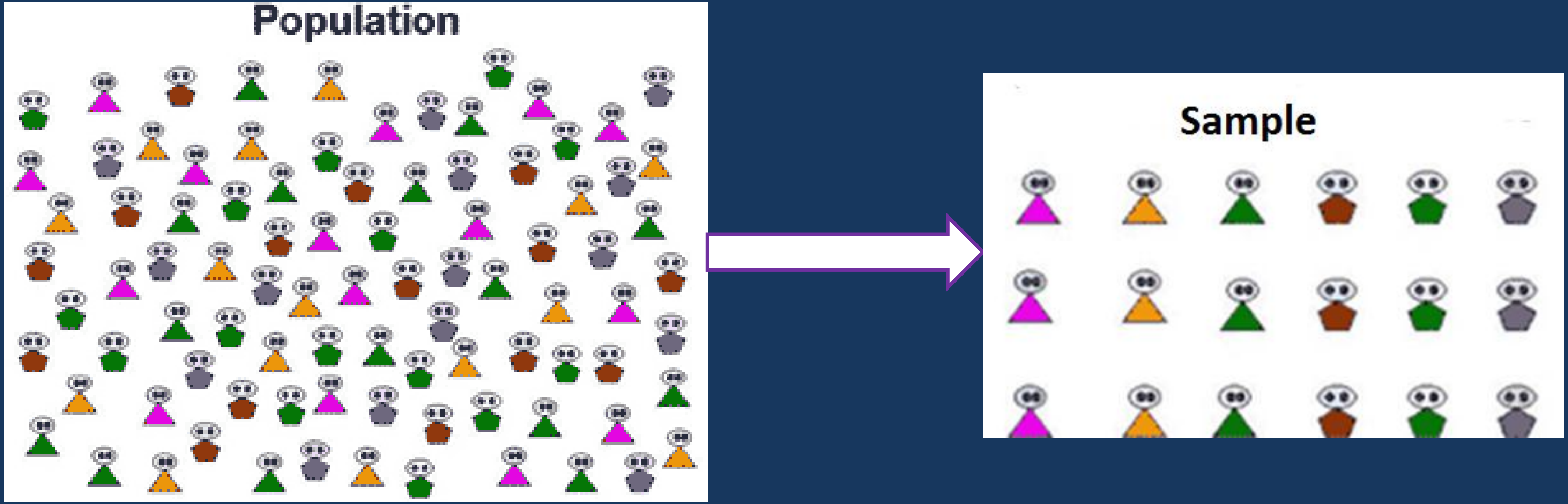
- A good representative sample → its result can be generalized on the whole population
- Not good representative sample → its result can not be generalized on the whole population

Sampling Units:

Each individual or thing of a population is called sampling unit.

Sampling frame:

All sampling units (all individuals of the population) are known and each of them can be identified by a number or mark.


Population
Sample

Why we use samples?

1. Cheaper than examining the whole population.

2. Less time consuming.

3. Feasible and can be repeated in other areas or times.

Types of samples

Probability samples

1. Every individual has an equal chance (probability) of being taken in the sample before the sample is drawn.
2. It is a good representation of the population.
3. Its results can be generalized.

Non-Probability samples

1. Chance of selection not equal for all individuals → it is biased.

Types of samples

Non probability

1- Accessibility sample
2- Quota sample

Probability

1- Simple random sample
2- Systematic random sample
3- Stratified random sample
4- Cluster sample
5- Multi-stage random sample

Non-Probability Samples

The non-probability sample doesn't allow us to get a true representation of the population from which it is drawn.

1. Accessibility sample:

- **The investigator chooses his sample by his opinion.**
- **The most convenient sample units are selected e.g. the nearest neighbors or relatives, volunteers, hospital cases, etc.**
- **The sample is completed when the desired number of population is reached.**

Advantages:

Cheap, quick, does not require sampling frame.

Disadvantages:

- Not representative of the whole population.
- It is biased due to subjective choice.
- Its findings could not be generalized. So, it has to be restricted in use in scientific medical research.
- Examples: sometimes we have to use this method e.g.:

1. Studying rare diseases which are available only in hospitals.
2. Studying occupational health hazards in workers exposed to that hazards.

2. Quota Sample

- The investigator will take a sample of a certain size and structure.
- The choice of the actual sampling units does not follow a special scheme but left to his choice.
- The sample is completed when the desired number of population is reached.

Advantages:

Cheap, quick, does not require sampling frame.

Disadvantages:

- Not a good representation of the population as it depends mainly on the investigator choice.
- It is biased due to subjective choice.
- Its findings could not be generalized, so seldom used in scientific medical research.

Examples:

1. **Interview of all persons passing in a certain street at a certain time.**
2. **In T.V. to know public opinion for the preferable programs.**

Probability Samples

- Every individual (or sample unit) has an equal chance (probability) of being taken in the sample before the sample is drawn.

- There is minimal role for the investigator in selection of individuals or sample units. So, bias of subjective (researcher) selection is minimal.

- Results obtained from researches based on probability sampling can be generalized on population with confidence.

Types of probability samples

1. Simple random sample
2. Systematic random sample
3. Stratified random sample
4. Cluster sample
5. Multi-stage random sample

Simple Random Sample:

- The **population** from which a simple random sample is drawn should be **uniform or homogeneous**.
- A **sample frame** must be present, to choose the needed units from it.
- The units are selected by using **random number tables** ** (either in statistical books or generated by the computer) or by **lottery** or **rotary** depending on the size of the sample.

Table of Random Numbers

36518	36777	89116	05542	29705	83775	21564	81639	27973	62413	85652	62817	57881
46132	81380	75635	19428	88048	08747	20092	12615	35046	67753	69630	10883	13683
31841	77367	40791	97402	27569	90184	02338	39318	54936	34641	95525	86316	87384
84180	93793	64953	51472	65358	23701	75230	47200	78176	85248	90589	74567	22633
78435	37586	07015	98729	76703	16224	97661	79907	06611	26501	93389	92725	68158
41859	94198	37182	61345	88857	53204	86721	59613	67494	17292	94457	89520	77771
13019	07274	51068	93129	40386	51731	44254	66685	72835	01270	42523	45323	63481
82448	72430	29041	59208	95266	33978	70958	60017	39723	00606	17956	19024	15819
25432	96593	83112	96997	55340	80312	78839	09815	16887	22228	06206	54272	83516
69226	38655	03811	08342	47863	02743	11547	38250	58140	98470	24364	99797	73498
25837	68821	66426	20496	84843	18360	91252	99134	48931	99538	21160	09411	44659
38914	82707	24769	72026	56813	49336	71767	04474	32909	74162	50404	68562	14088
04070	60681	64290	26905	65617	76039	91657	71362	32246	49595	50663	47459	57072
01674	14751	28637	86980	11951	10479	41454	48527	53868	37846	85912	15156	00865
70294	35450	39982	79503	34382	43186	69890	63222	30110	56004	04879	05138	57476
73903	98066	52136	89925	50000	96334	30773	80571	31178	52799	41050	76298	43995
87789	56408	77107	88452	80975	03406	36114	64549	79244	82044	00202	45727	35709
92320	95929	58545	70699	07679	23296	03002	63885	54677	55745	52540	62154	33314
46391	60276	92061	43591	42118	73094	53608	58949	42927	90993	46795	05947	01934
67090	45063	84584	66022	48268	74971	94861	61749	61085	81758	89640	39437	90044
11666	99916	35165	29420	73213	15275	62532	47319	39842	62273	94980	23415	64668
40910	59068	04594	94576	51187	54796	17411	56123	66545	82163	61868	22752	40101
41169	37965	47578	92180	05257	19143	77486	02457	00985	31960	39033	44374	28352
76418												

- **<u>So, simple random sample is used when:</u>**

1. **Population is uniform or homogeneous <u>and</u>**
2. **All sampling units are known and so sampling frame can be prepared**

Example:

Selection of 5 individuals out of 15.

- Give number for each individual (***sampling frame***).
- Randomly select the needed sample (5 units) by lottery from a box containing numbers from 1 to 15.

Systematic Random Sample:

Selection depends on an interval (K-interval) which is calculated from both the size of population and the size of the sample.

$$\text{K-interval} = \frac{\text{total population}}{\text{Sample size}}$$

Example:

- Suppose we have a population = 120 and it is required to take a sample of 12

- **K-interval =** $\frac{\textit{Total population}}{\textit{Sample size}} = \frac{120}{12} = 10$

- So, we have to select one out of each 10.

- Then randomly select one out of the 1st 10, say 2.

- Then repeatedly add the k-interval to the selected number.

- So, the sample will be the individuals number: 2, 12, 22, 32, 42, 52, 62, 72, 82, 92, 102 and 112.

- Patients can be selected from the outpatient clinic by a **modified** method of this sample.

- **Example:** select 8 persons from an outpatient clinic:
  - We take a random number from 1-10 (or 1-5 according to the rate), suppose the 3rd.
  - then we will take every 3rd person coming to the clinic i.e. 3rd, 6th, 9th, 12th, etc. till we reach the desired sample size (8 persons).

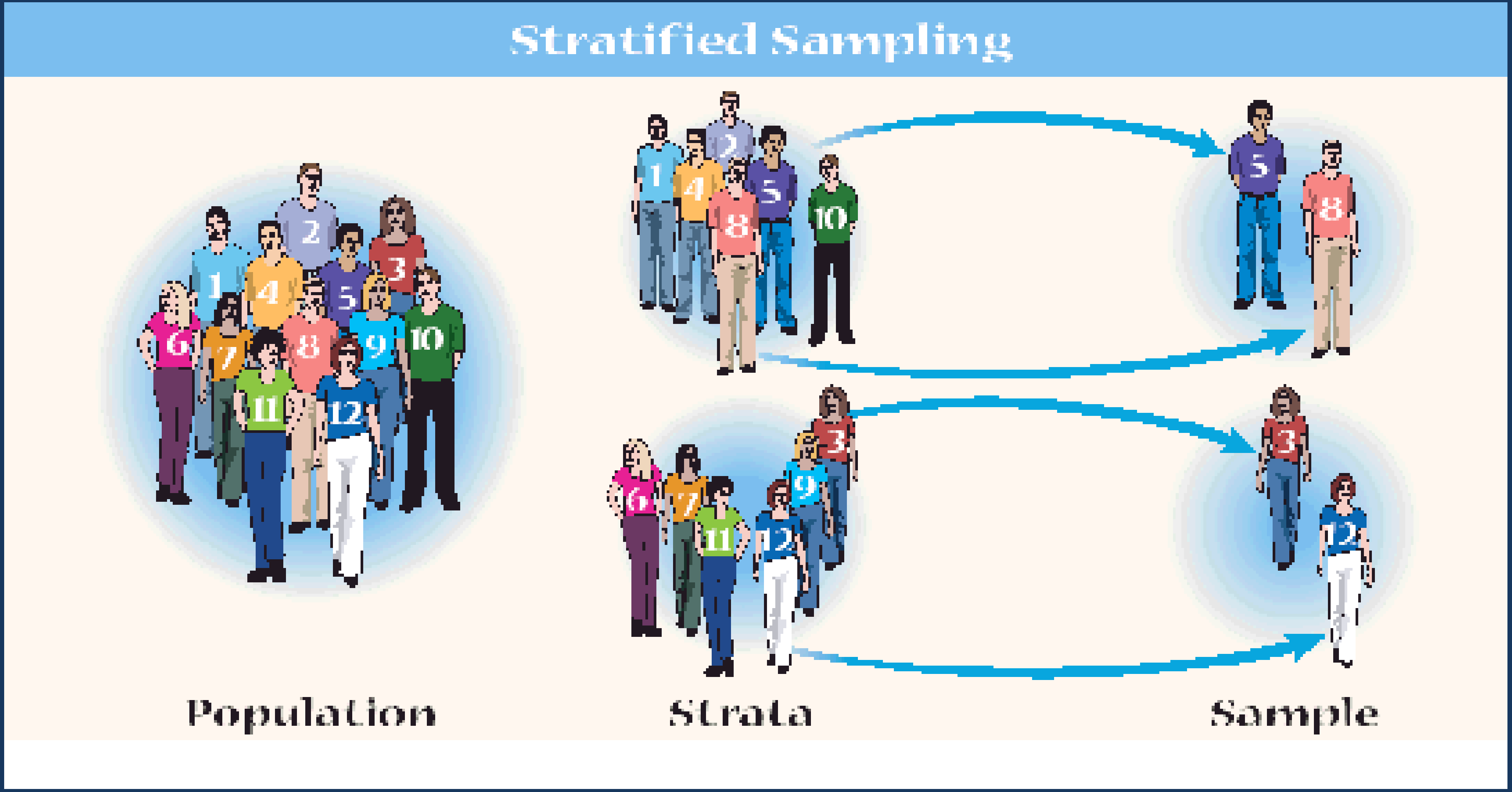
- By this way, there is **no bias** in selection (no subjective selection).

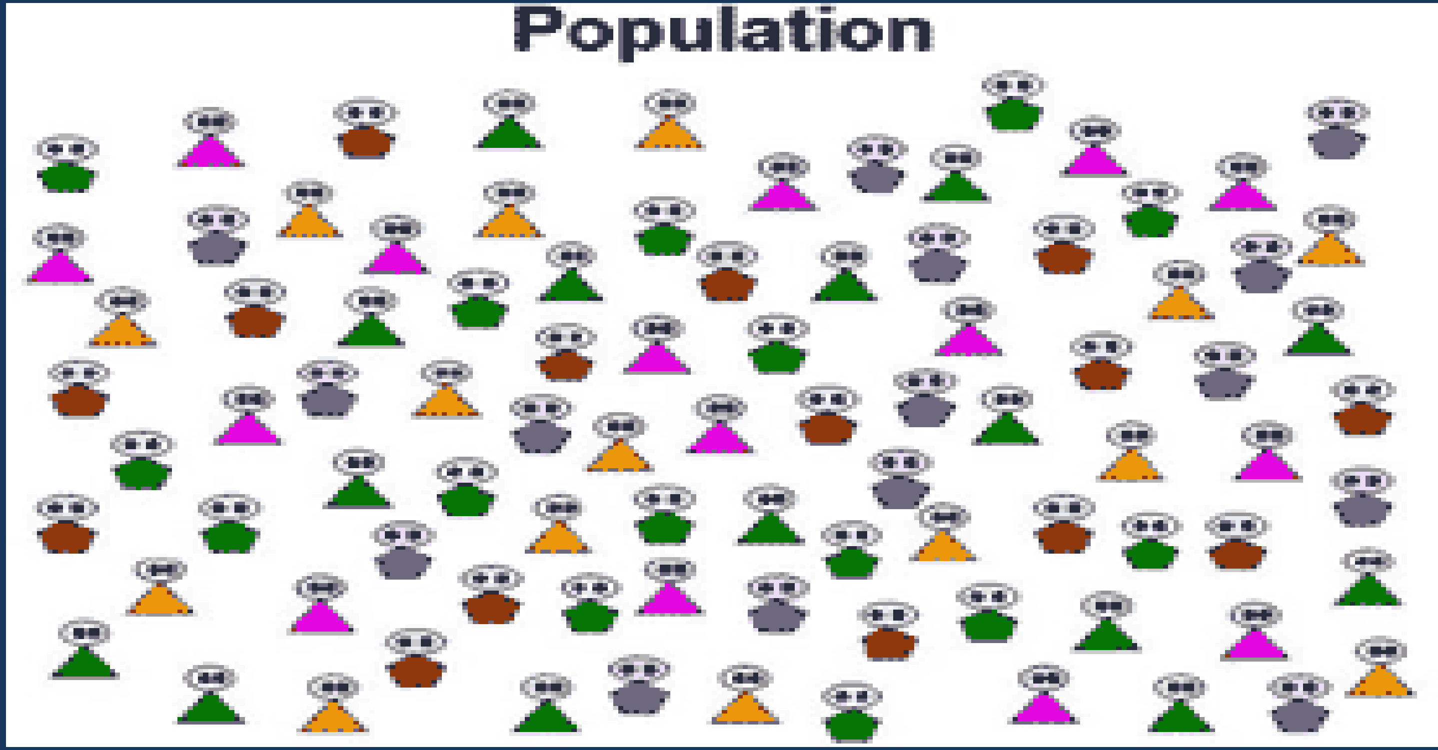
Advantages:

- Does not require sampling frame.
- No bias in selection.
- We can select sample from large scale population.

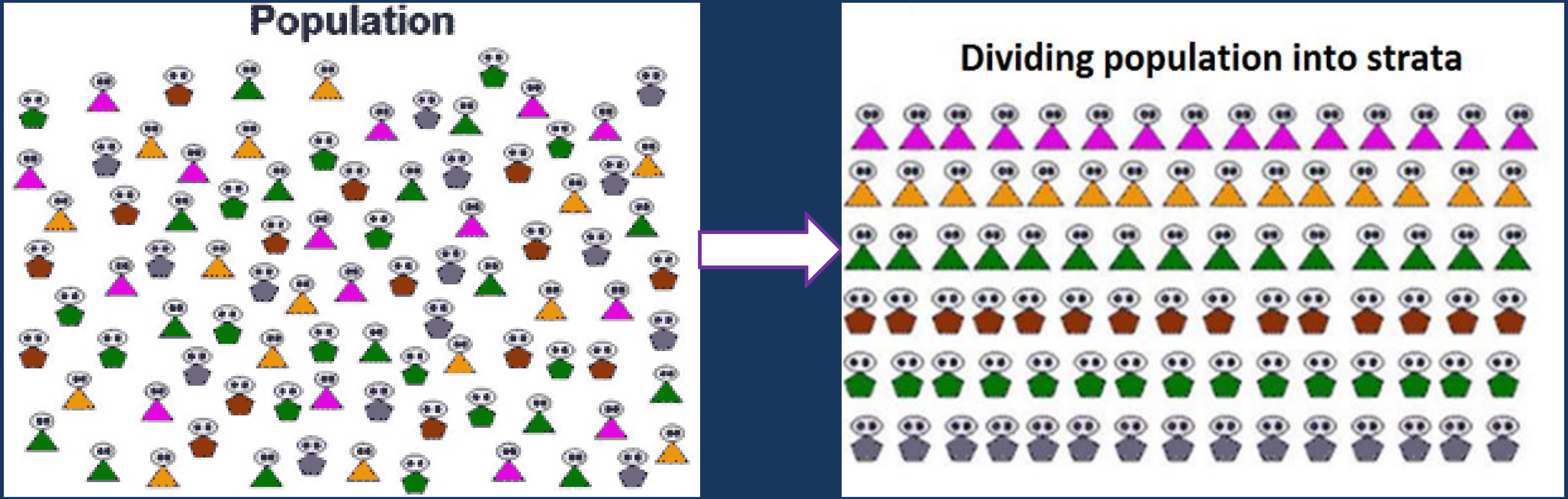
Stratified Random Sample:

It is used when the population is not homogeneous.

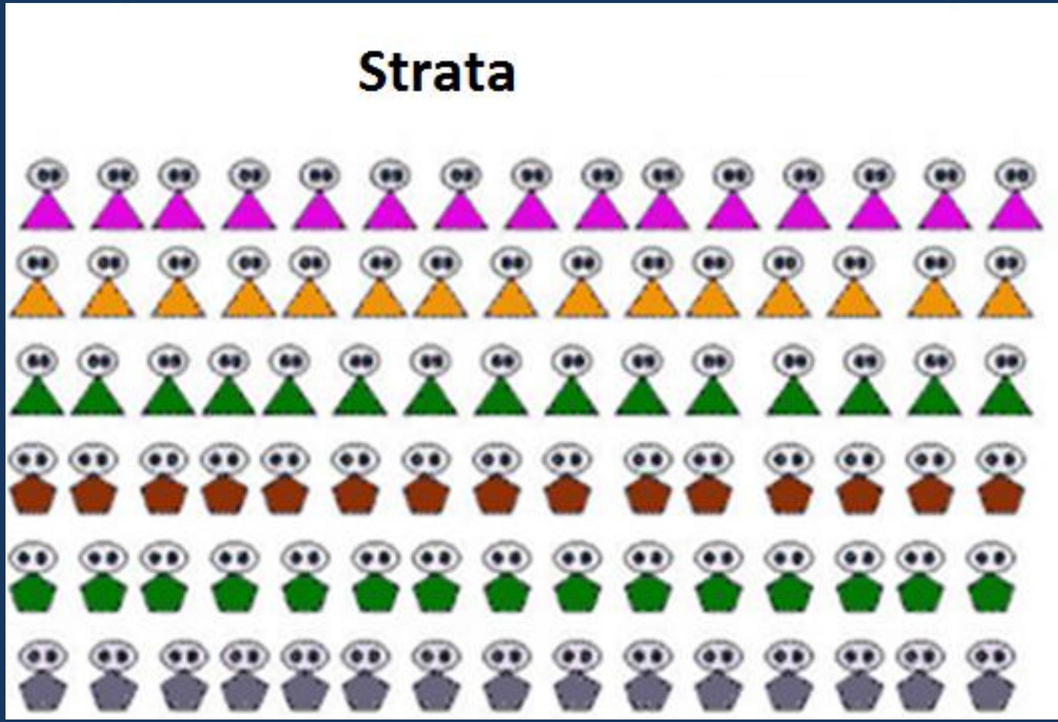


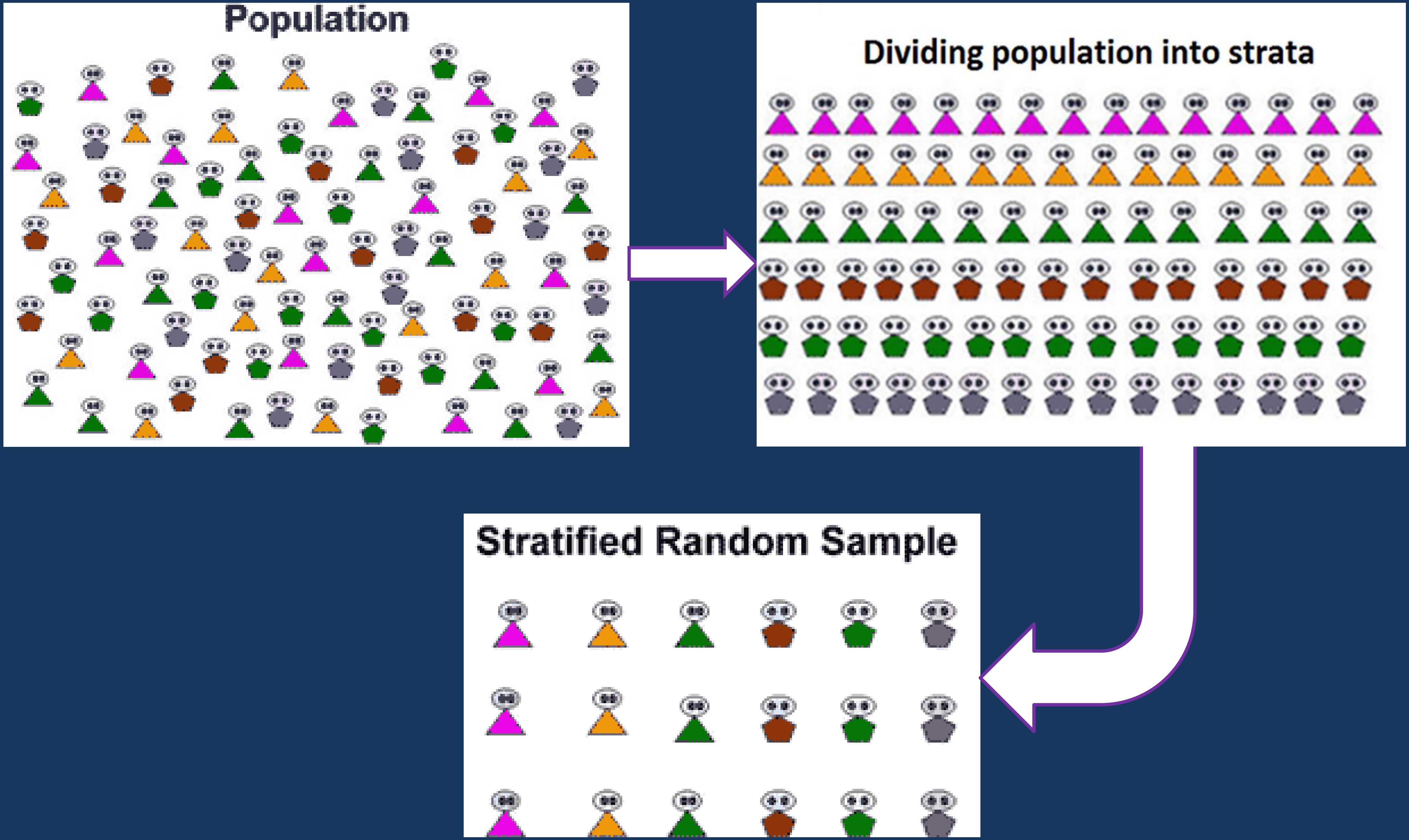


First: stratifying the population i.e. dividing the population into ***different strata*** each of which is as homogeneous as possible e.g. according to sex, age, residence, etc.


Population
Dividing population into strata

Second: selecting a simple random sample (or systematic random sample) from each stratum




Population
Dividing population into strata
Stratified Random Sample

How many will be taken from each stratum ?

Selection can be done using:

1- Equal allocation method:

$$\text{No. required from each stratum} = \frac{\text{Sample size}}{\text{No. of strata}}$$

2- Proportional allocation method:

$$\text{No. required from each stratum} = \text{Sample size} \times \frac{\text{Size of stratum}}{\text{Total population}}$$

العدد المطلوب من كل فئة = حجم العينة المطلوبة × (حجم الفئة / حجم المجتمع)

مثال: اذا أردنا أخذ عينة ممثلة لمدرسة عدد طلابها ٣٠٠، منهم ١٢٠ فى الصف الأول و ١٠٠ فى الصف الثانى و٨٠ فى الصف الثالث والعينة المطلوبة ٦٠ طالباً.

الطريقة الأولى: العدد المطلوب من كل صف = $\frac{٦٠}{٣}$ = ٢٠ طالباً

الطريقة الثانية:

العدد المطلوب من الصف الأول = ٦٠ × $\frac{١٢٠}{٣٠٠}$ = ٢٤ طالباً

العدد المطلوب من الصف الثانى = ٦٠ × $\frac{١٠٠}{٣٠٠}$ = ٢٠ طالباً

العدد المطلوب من الصف الثالث = ٦٠ × $\frac{٨٠}{٣٠٠}$ = ١٦ طالباً

مثال:

مصنع به ٨٠٠ عامل، منهم ٧٠٠ من الذكور و ١٠٠ من الإناث والعينة المطلوبة ٨٠

الطريقة الأولى: العدد المطلوب من الذكور وكذلك من الإناث = $\frac{٨٠}{٢}$ = ٤٠ عاملاً

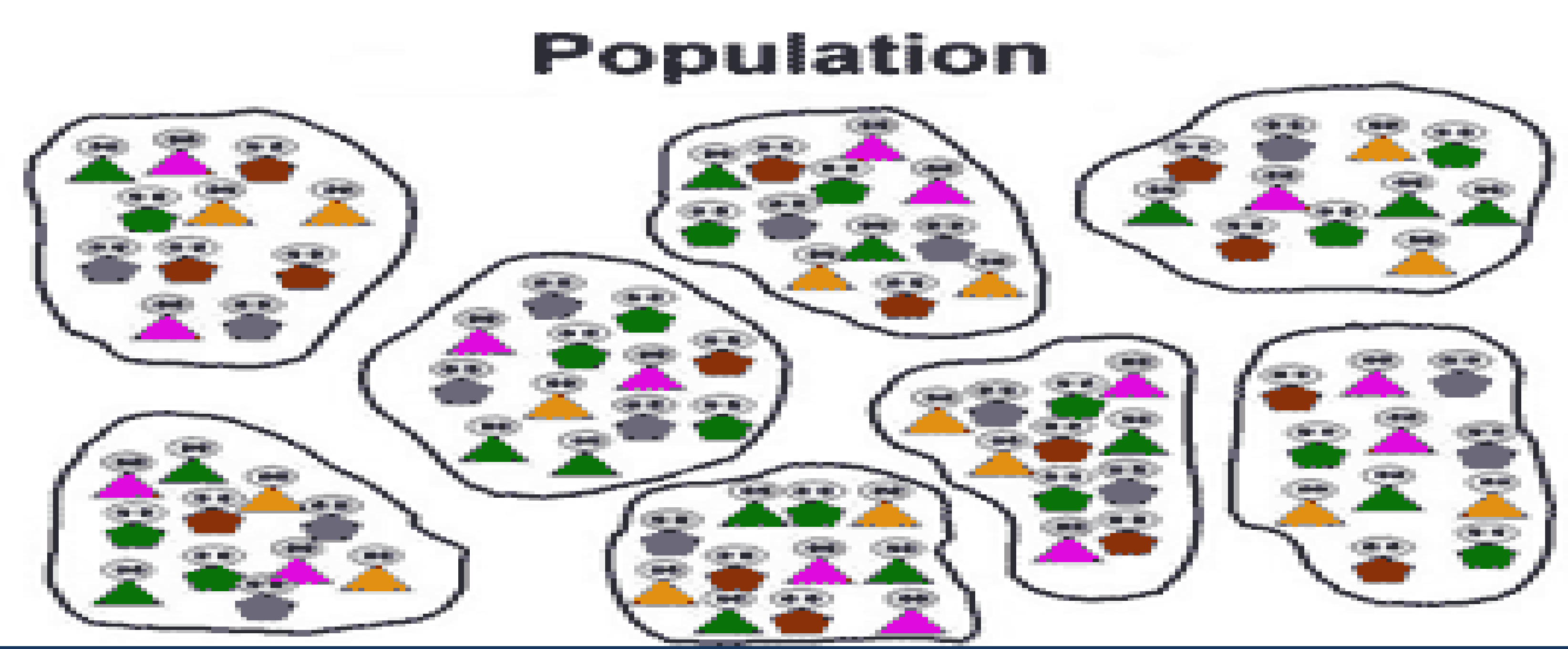
الطريقة الثانية:

العدد المطلوب من الذكور= ٨٠ × $\frac{٧٠٠}{٨٠٠}$ = ٧٠ عاملاً

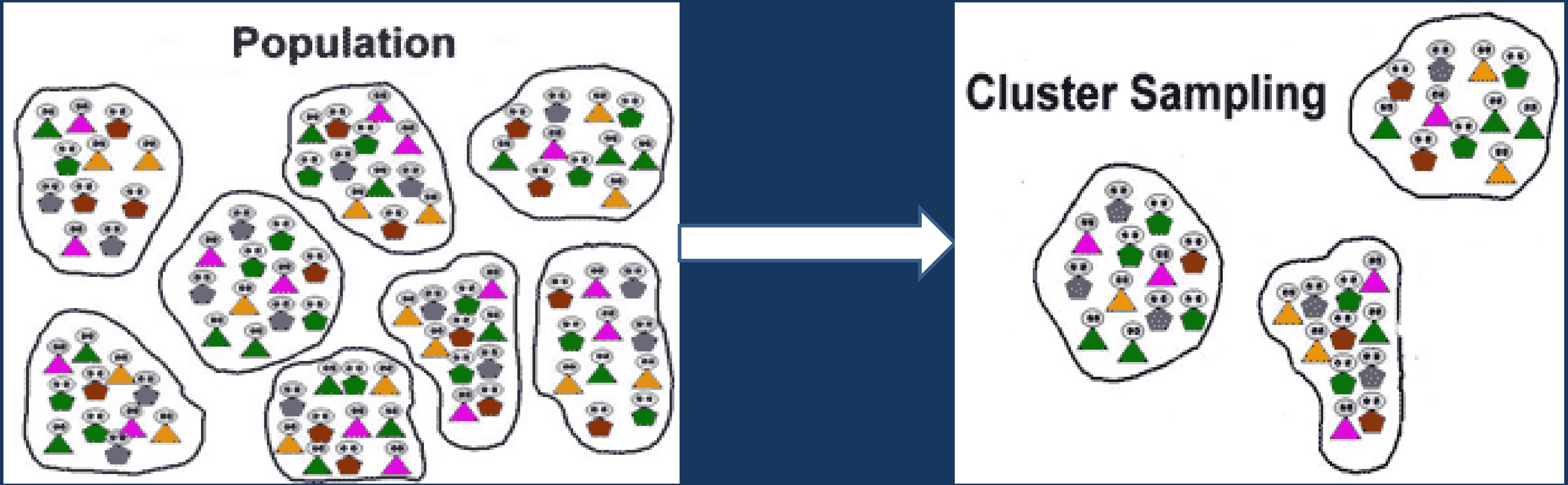
العدد المطلوب من الإناث = ٨٠ × $\frac{١٠٠}{٨٠٠}$ = ١٠ عاملات

Cluster Sample

- **A cluster:** is a group of individuals that is present in certain locality or geographical area e.g. village, school, classroom, etc.



- **First:** we select a random sample of clusters.



- **Then:** the clusters are taken as whole i.e. taking all individuals within the selected clusters.

Example:

- If we need to select 5 districts of Al-Zarqa Governorate :
- Prepare a list of all districts in AL-Zarqa Governorate
- Then select randomly 5 districts out of the total districts
- Then all people living in these 5 districts will be included in the study.

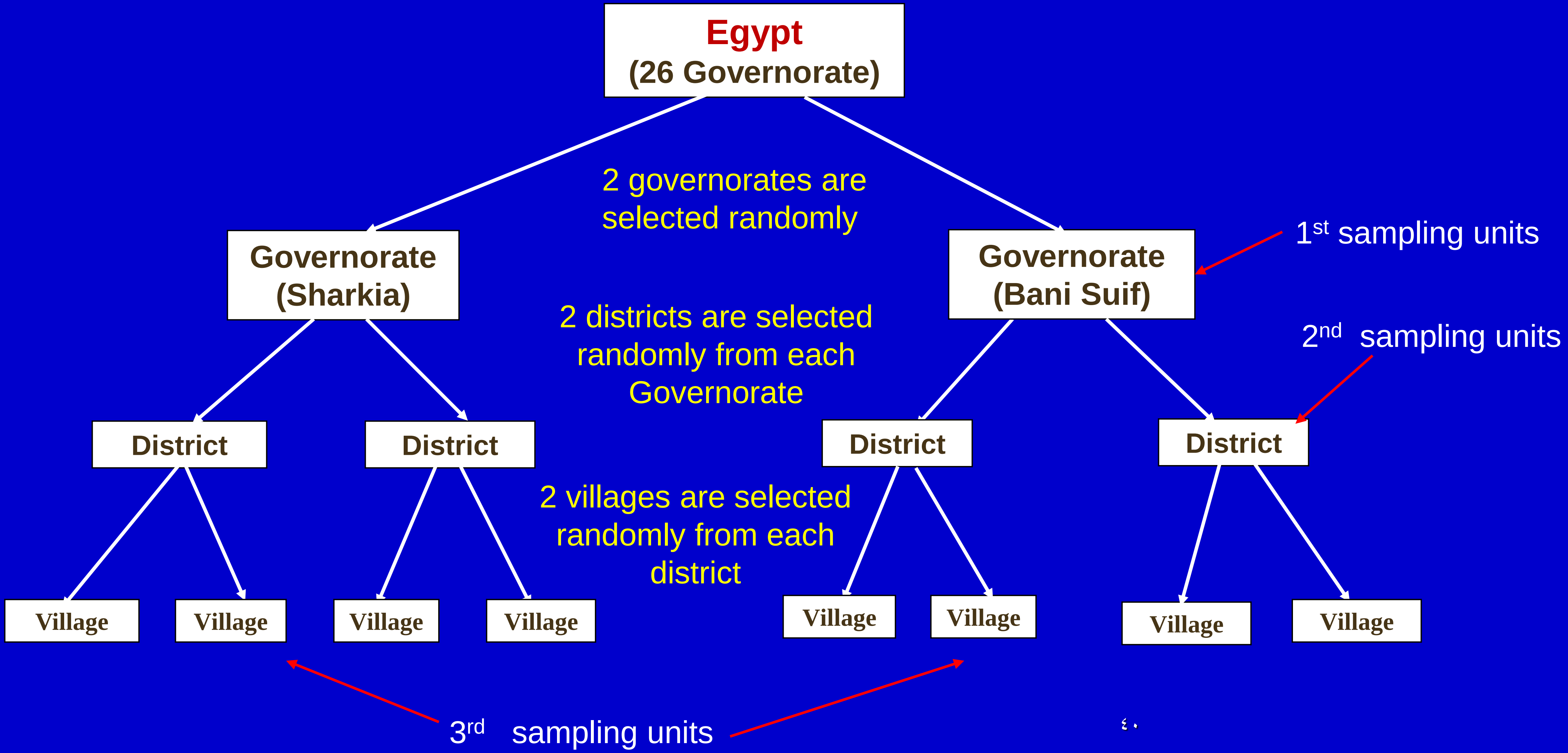
Example:

- We can obtain a random sample of primary school children in an area by:
- Starting with a list of schools
- Draw a simple random sample of schools
- Then all children within the selected schools form the sample of children.

Multistage Random sample

- **It is usually used in case of national or widespread studies.**
- **The field of work is arranged in levels or stages e.g. governorates, districts, villages, houses, families and individuals.**
- **From each stage we select randomly the desired sample.**

Example: Selection of sample of villages (8 for example) from Egypt for a morbidity Survey:



Example:

Selection of sample of villages (8 for example) from Egypt for a morbidity Survey:

- **First, we have 26 governorates**
- **Select 2 governorates randomly (Governorates are the 1st sampling units)**
- **Then from each governorate, select 2 administrative districts (Districts are the 2nd sampling units)**
  → two-stage sample.
- **Then from each district, select 2 villages randomly (Villages are th 3rd sampling units) → three-stage sample.**

Sample size

How many individuals (or things) will be included in the study.

Determinants of sample size:

1. **Available resources:** **man, money, materials and time.**

   ↑ resources → ↑ sample size and vice versa.

2. **Number of variables affecting the study.**

   ↑ no. of variables → ↑ sample size and vice versa.

3. **Prevalence of the problem or disease under study.**

↑ prevalence → ↓ sample size and vice versa.

4. **Power of statistical test:**

- It is the ability of the study to detect statistical significant relations.
- A power of 80% is suitable for most studies. It means that there is 20% error of missing a statistical significant difference in our selected sample.
- ↑ power → ↑ sample size and vice versa.

5. Level of significance

- It is the ability of the study to detect statistical insignificant relations.
- 95 % is usually the selected significant level.
- It means that 5% error can occur in the study for getting significant result although it is not truly significant.
- ↑ level of significance → ↑ sample size and vice versa.

6. Effect size:

It is the difference expected between treatment and control groups or the strength of association.

For example: **if the new treatment under study will produce percentage of cure 80% and the old treatment gives 70% cure rate, then the effect size is 10%.**

↑ effect size → decreases the sample size

- **Mean Value:** e.g. if we have mean value of 10±2 for Hb of normal population and we assume that the Hb of cases of lead poisoning will be 8±2.5 then the effect size will be the squared difference in the mean value divided by SD of the lead cases group = $\frac{(10-8)^2}{2.5}$

- If we have no mean value for cases, we can assume effect size of:
  - 0.2 for small suspected difference
  - 0.5 for moderate suspected difference
  - 0.8 for large suspected difference.

- We can get the mean value of the population from other previous studies or by doing a pilot study.

7. **Type of study:** usually cross-sectional and case control study need larger samples (one reading is needed from each person) than cohort or randomized studies which need follow up and many reading for the same person.

8. **Cost of each sample:** if the cost is expensive, we have to minimize the sample.

9. **Variability in the studied population:** if great, the sample size should be larger.

10. **Reliability and validity of the measurements**: The more valid and reliable method, the smaller is the sample.

- **Sample size is calculated simply by many computer statistical packages e.g. Open Epi , Epi 6, SPSS.**

- **But we have to fill some information in these statistical programs for calculation.**

- **The needed information is specific for each type of study.**

In cross sectional studies (population survey):

1. Population size: from which the sample will be chosen.
2. Prevalence of the disease or factor under study in the population: from records, previous studies, websites or pilot study.
3. Power of test (In Epi 6: Result farthest from the prevalence rate that you would accept in your sample, higher or lower): 80% is reasonable and common level.
4. Level of significance: 95% is reasonable and common level.

In cohort and randomized clinical trials:

1. Two sided confidence level (level of significance, 1-α): usually 95%.
2. Power of study (1-β): usually 80%.
3. Ratio of unexposed to exposed in the sample: for equal samples use 1.
4. Percent of disease or factor under study among unexposed (e.g. 5).
5. **One of the following:**
   a. Odds ratio (e.g. 2).
   b. Percent of disease or factor under study among exposed (e.g. 9.52)
   c. Risk ratio or prevalence ratio = $\frac{\%\ among\ exposed}{\%\ among\ non\ exposed}$ ($\frac{9.52}{5}$ = 1.904)
   d. Risk difference or prevalence difference = % among exposed − % among non exposed (9.52 − 5 = 4.52).

In unmatched case control study:

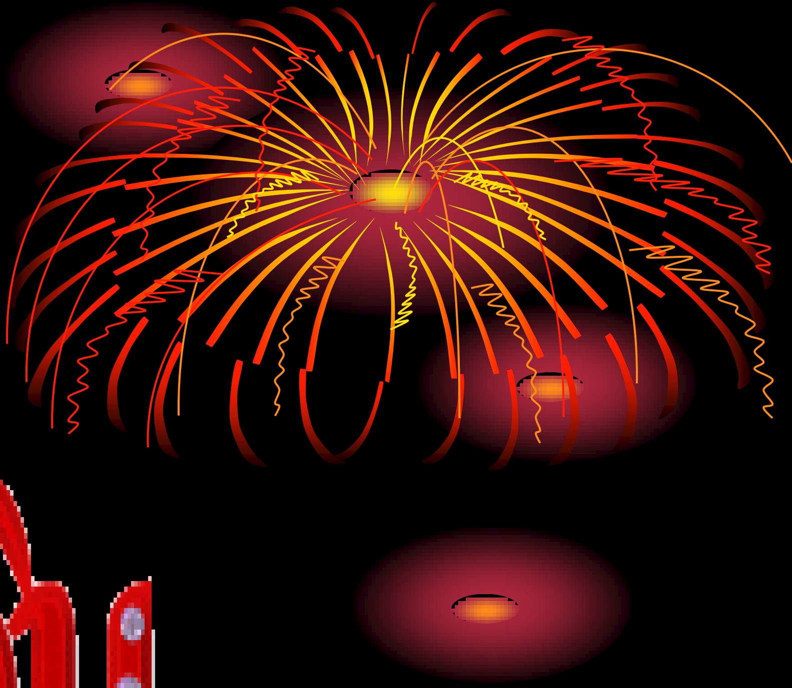
1. Two sided confidence level (level of significance, 1-α): usually 95%.
2. Power of study (1-β): usually 80%.
3. Ratio of control to cases in the sample: for equal samples use 1.
4. Expected frequency of exposure among controls (e.g. 40).
5. **One of the following:**
   a. Odds ratio (e.g. 2).
   b. Expected frequency of exposure among cases (e.g. 57.14).

Sample size for comparing two means (mean difference):

1. Two sided confidence level (level of significance, 1-α): usually 95%.
2. Power of study (1-β): usually 80%.
3. Ratio of sample size = $\frac{Group\ 2}{Group\ 1}$: for equal samples use 1.
4. Mean of group 1 and mean of group 2 (or difference between the 2 means, mean difference)
5. **One of the following:**
   a. Standard deviations of the 2 groups
   b. Variance of the 2 groups

https://www.openepi.com/SampleSize/

Sample Size for Cross-Sectional, Cohort, & Randomized Clinical Trial Studies		
Two-sided confidence level(%)	95	(1-alpha) usually 95%
Power (1-beta or % chance of detecting)	80	Usually 80%
Ratio of Unexposed to Exposed in sample	1.0	For equal samples, use 1.0
Percent of Unexposed with Outcome	5	Between 0.0 and 99.9
Please fill in 1 of the following. The others will be calculated.		
Odds ratio	2	
Percent of Exposed with Outcome	9.52	Between 0.0 and 99.9
Risk/Prevalence Ratio	1.90	
Risk/Prevalence difference	4.52	Between -99.99 and 99.99



Thank You