From a few to the whole crowd: a practical introduction to sampling and survey data analysis

Annibale Cois, Meng, MPH, PhD
Burden of Disease Research Unit, South African Medical Council &
Division of Epidemiology & Biostatistics,
School of Public Health, University of Cape Town

Email: annibale.cois@mrc.ac.za
ORCID: 0000-0002-7014-6510
WEB: annibalecois.github.io

Cape Town International Convention Centre, 24 September 2024

Sampling process

Population

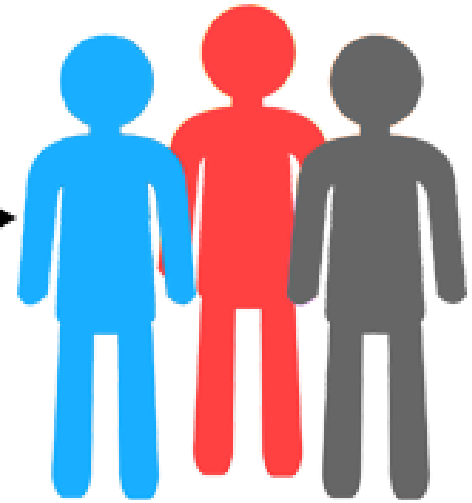
Sampling Process

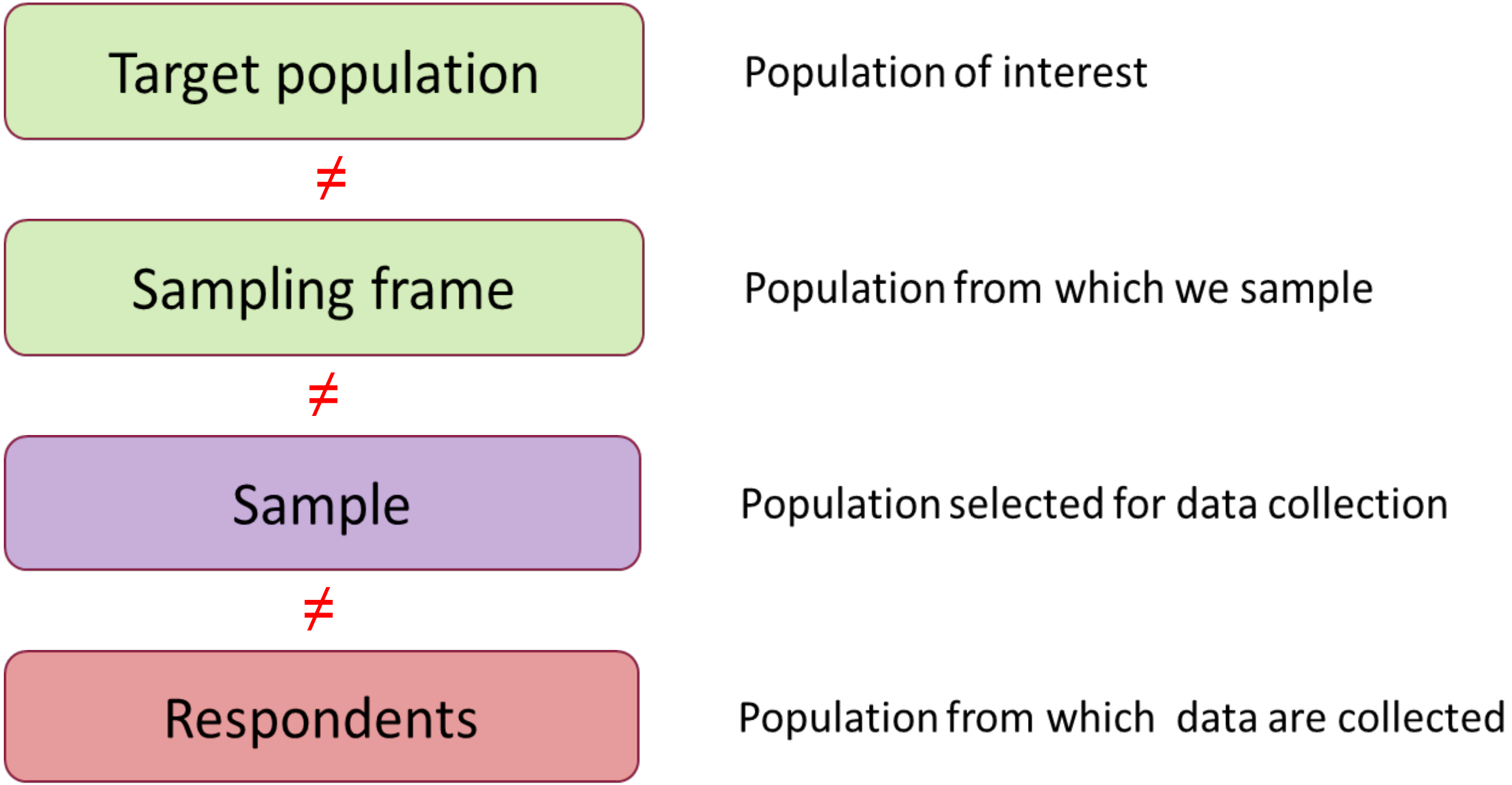
Sampling

IDENTIFY THE POPULATION

DEFINE A STRATEGY

DEFINE A SAMPLE SIZE




Target population
≠
Sampling frame
≠
Sample
≠
Respondents
Population of interest
Population from which we sample
Population selected for data collection
Population from which data are collected

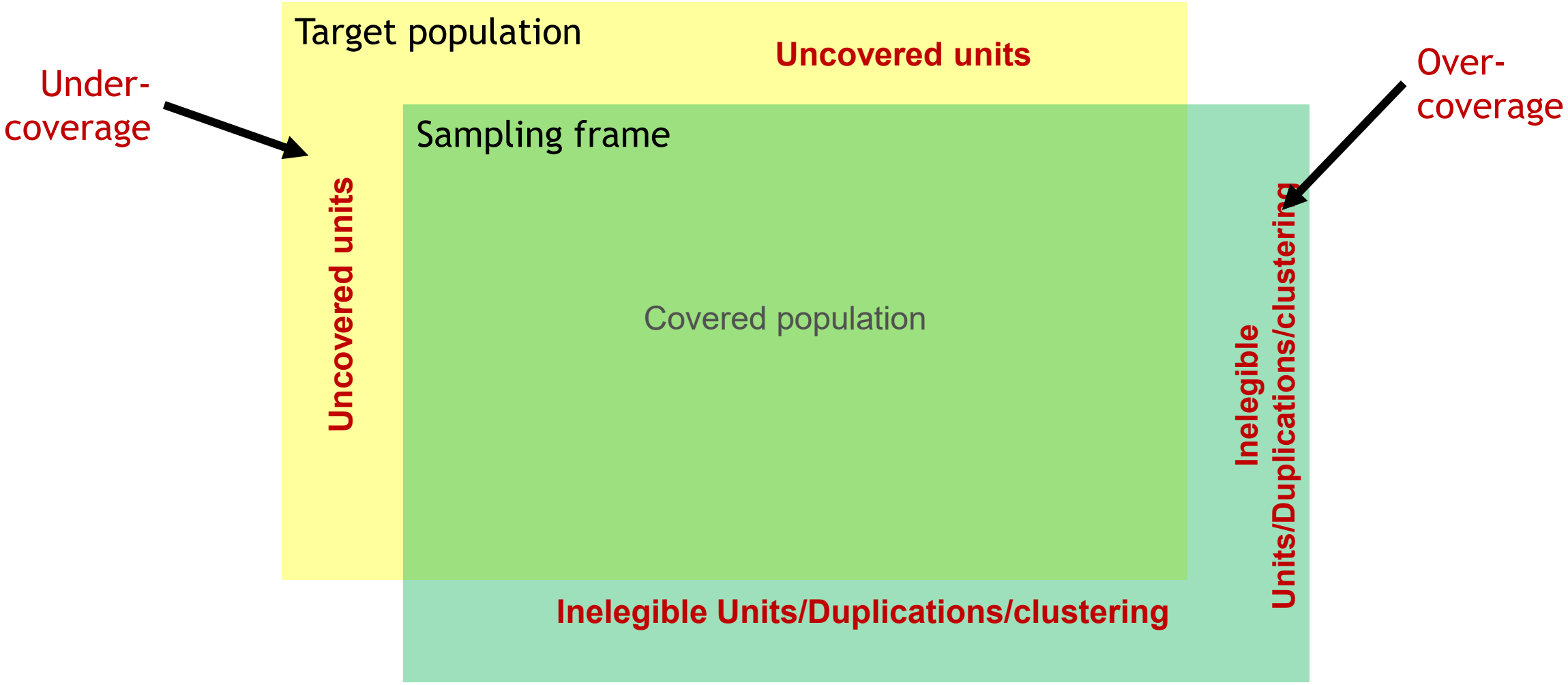
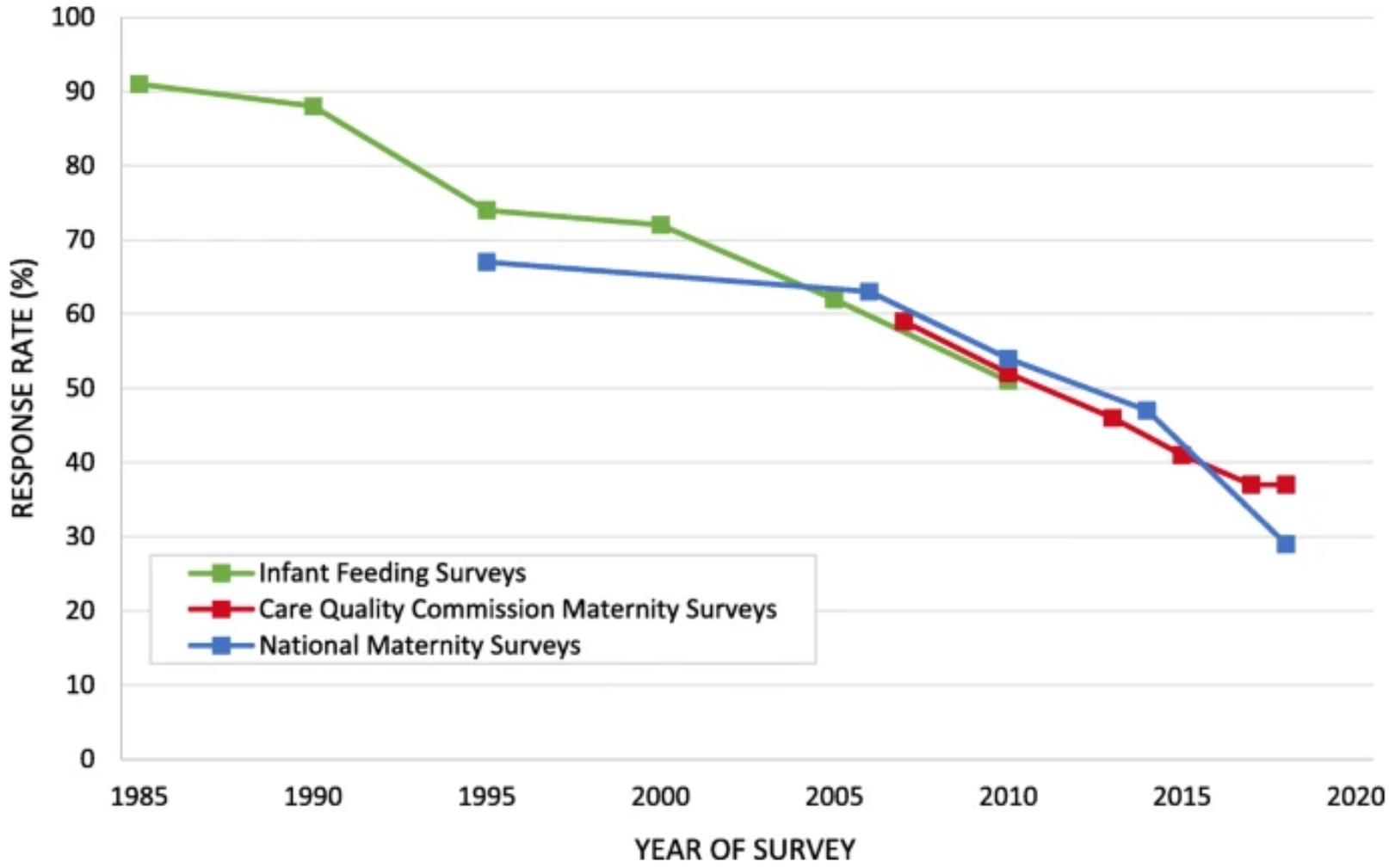

Target population
Uncovered units
Under-coverage
Over-coverage
Sampling frame
Uncovered units
Covered population
Inelegible Units/Duplications/clustering
Inelegible Units/Duplications/clustering

Fig. 1



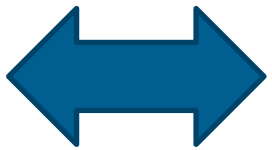
Response rates to the IFS, CQC Maternity Surveys, and NMS (1985–2018)

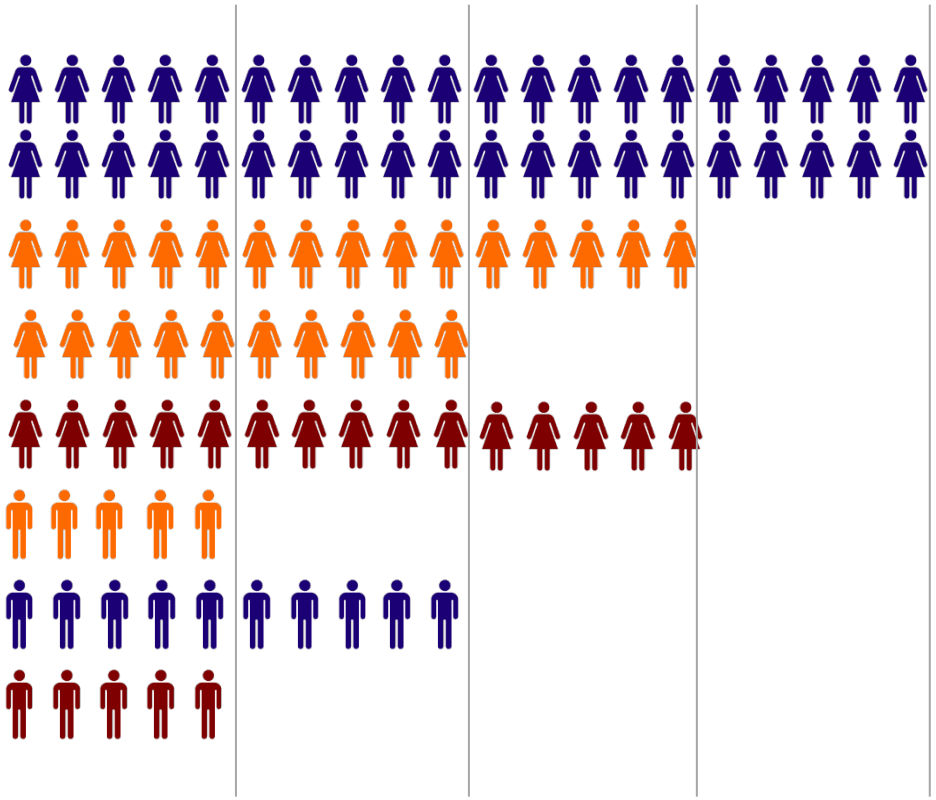
Harrison et al. Trends in response rates and respondent characteristics in five National Maternity Surveys in England during 1995–2018. Arch Public Health 78, 46 (2020). https://doi.org/10.1186/s13690-020-00427-w]

Representativeness

Representativeness

The sample is similar to the target population in all characteristics of interest





Target population
N=100

n=10

Sample A

Sampling

Representative?

Sample B

Sampling

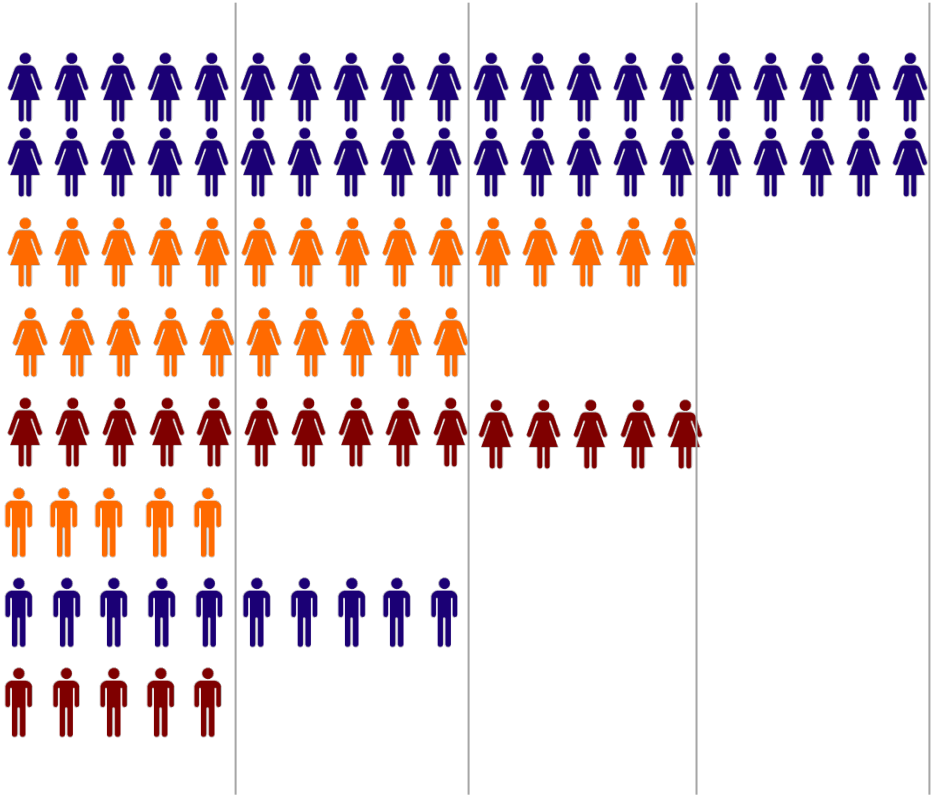
Representative?

Sampling

Sample C

Representative?

Characteristics of interest: Gender & hair colour

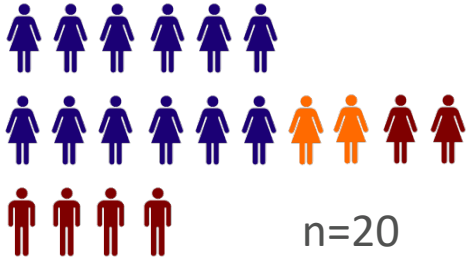


Target population
N=100

Sampling

Sampling

Sample D



n=20

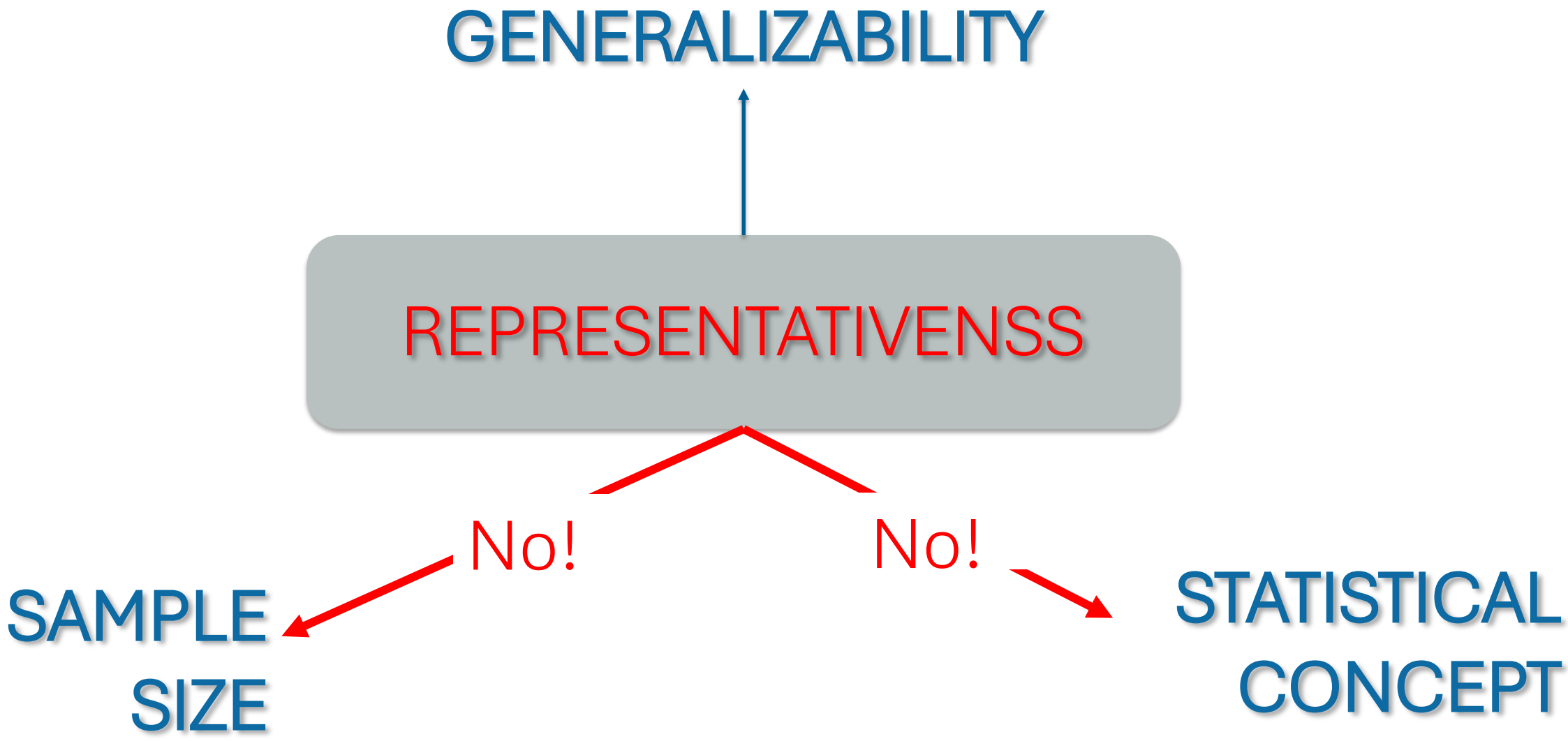
Representative ?

Sample E

n=40

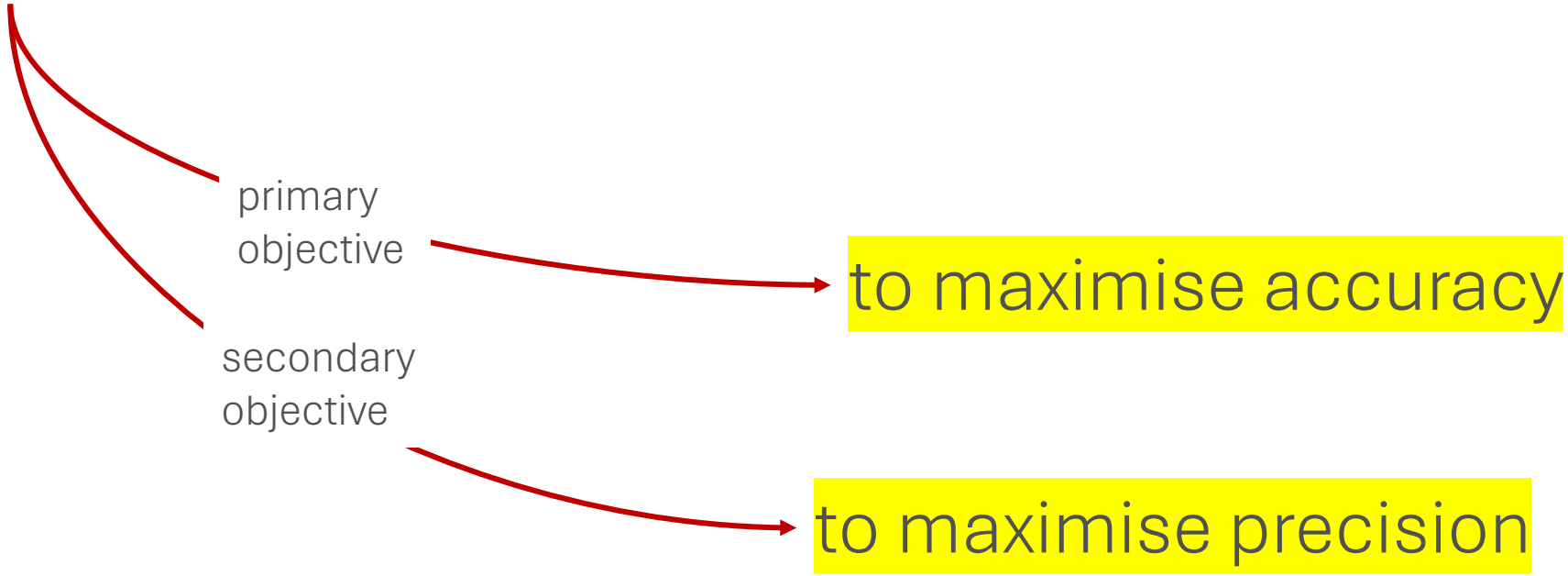
Representative?

Characteristics of interest: Gender & hair colour


GENERALIZABILITY
REPRESENTATIVENSS
No!
No!
SAMPLE SIZE
STATISTICAL CONCEPT

Sampling Strategies

SAMPLING STRATEGY



Probability vs non-probability samples

A probability sampling strategy is one in which every unit in the population has a chance of being selected in the sample, and this probability can be accurately determined.

probability of selection p >0

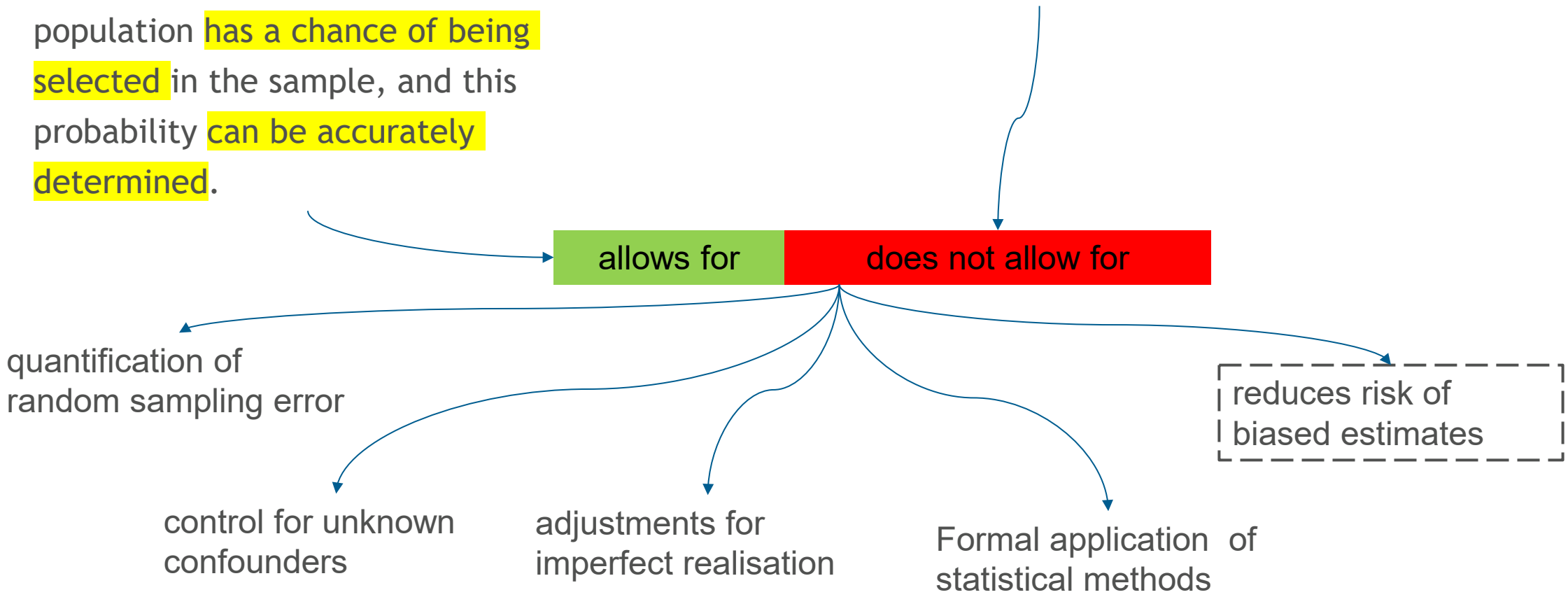
p is known

Everything else is non-probability sampling

Advantages of probability sampling

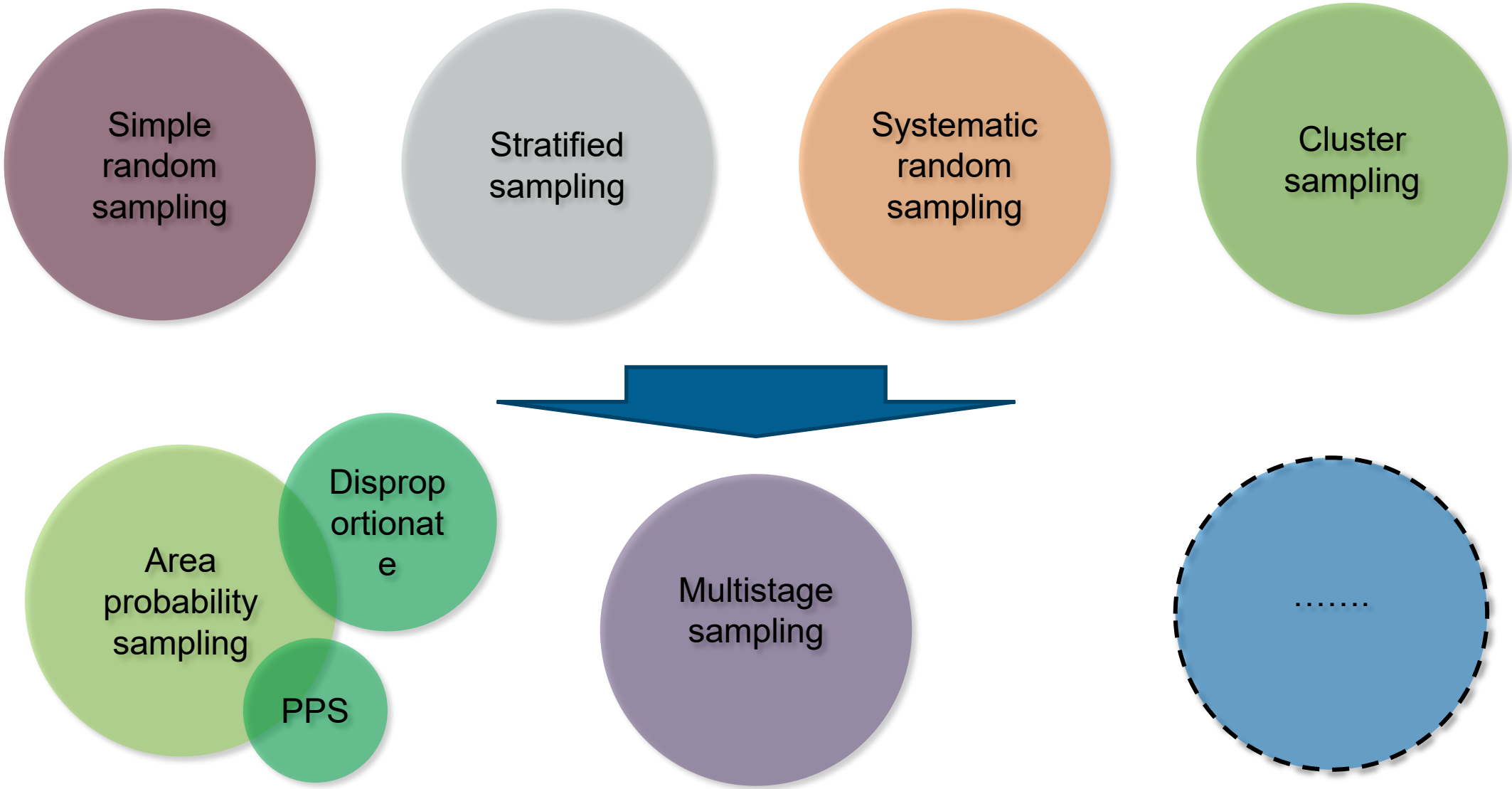
A probability sampling strategy is one in which every unit in the population has a chance of being selected in the sample, and this probability can be accurately determined.

Everything else is non-probability sampling

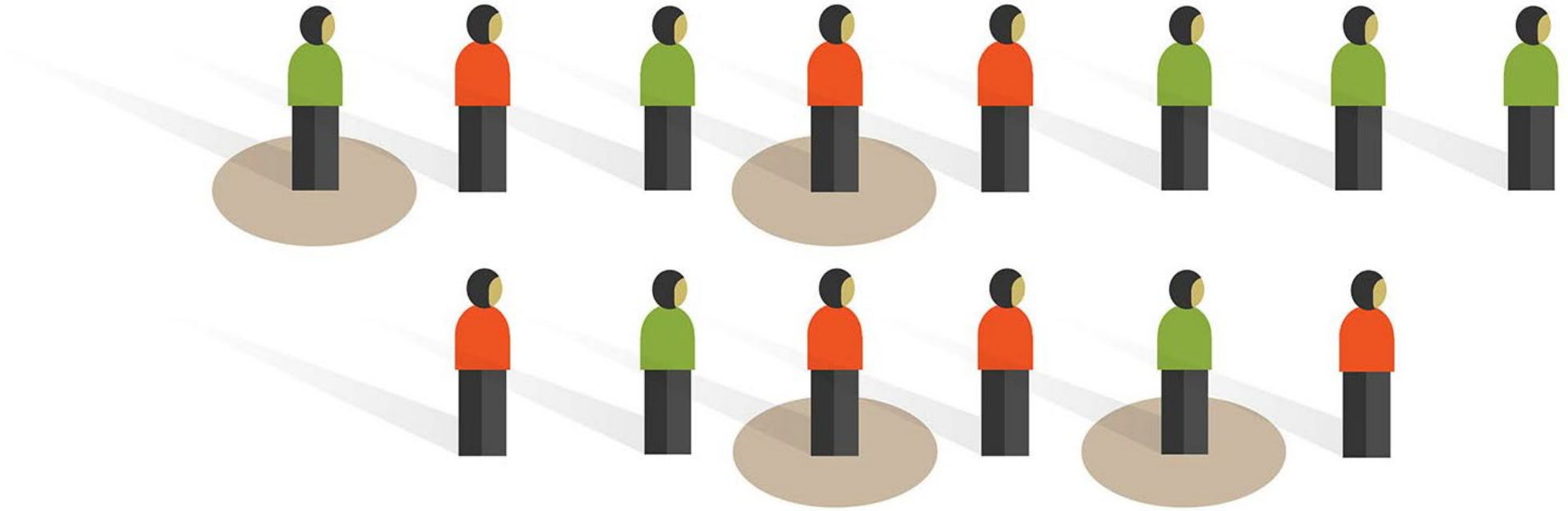


Advantages of non-probability sampling

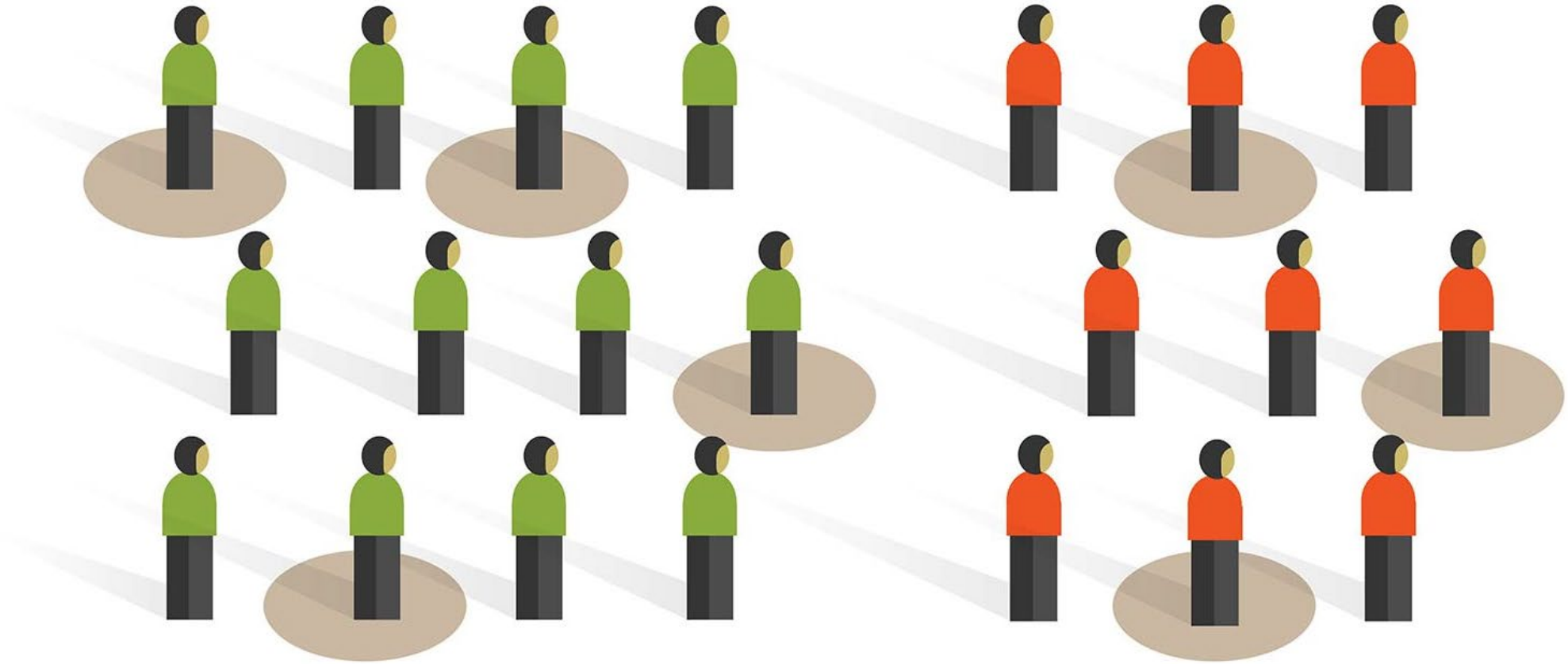
Probability sampling strategies



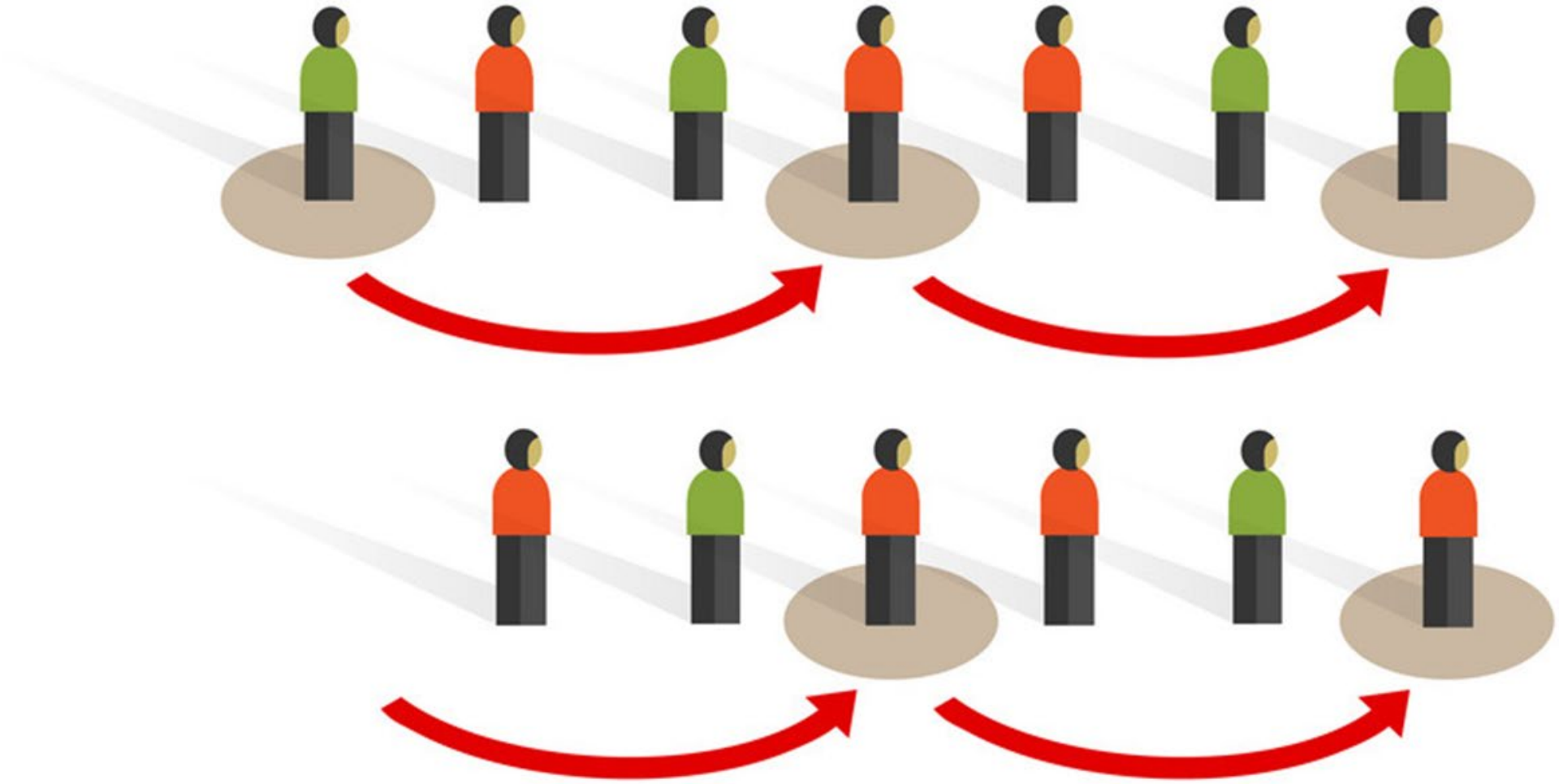
Simple random sampling



Stratified random sampling

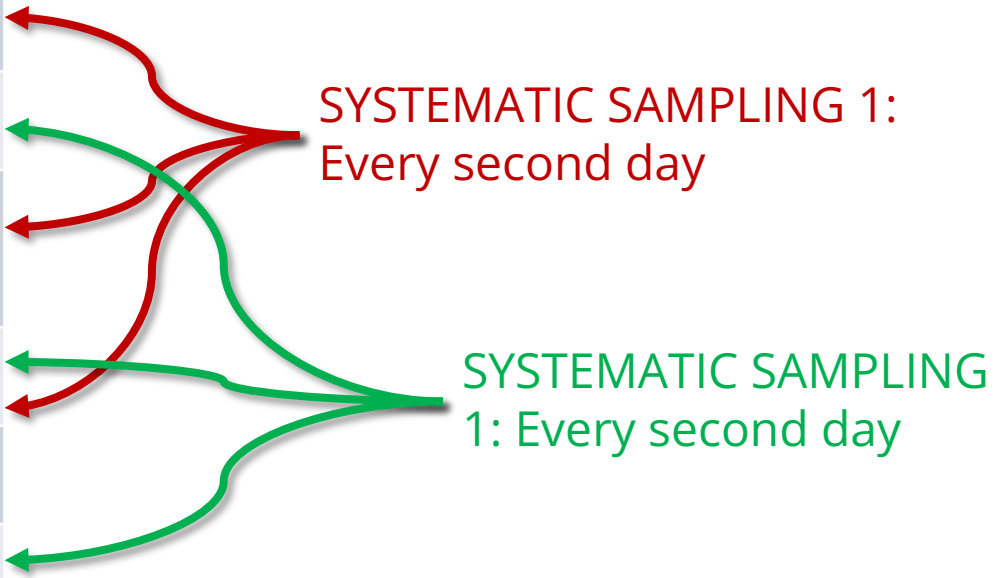


Systematic random sampling

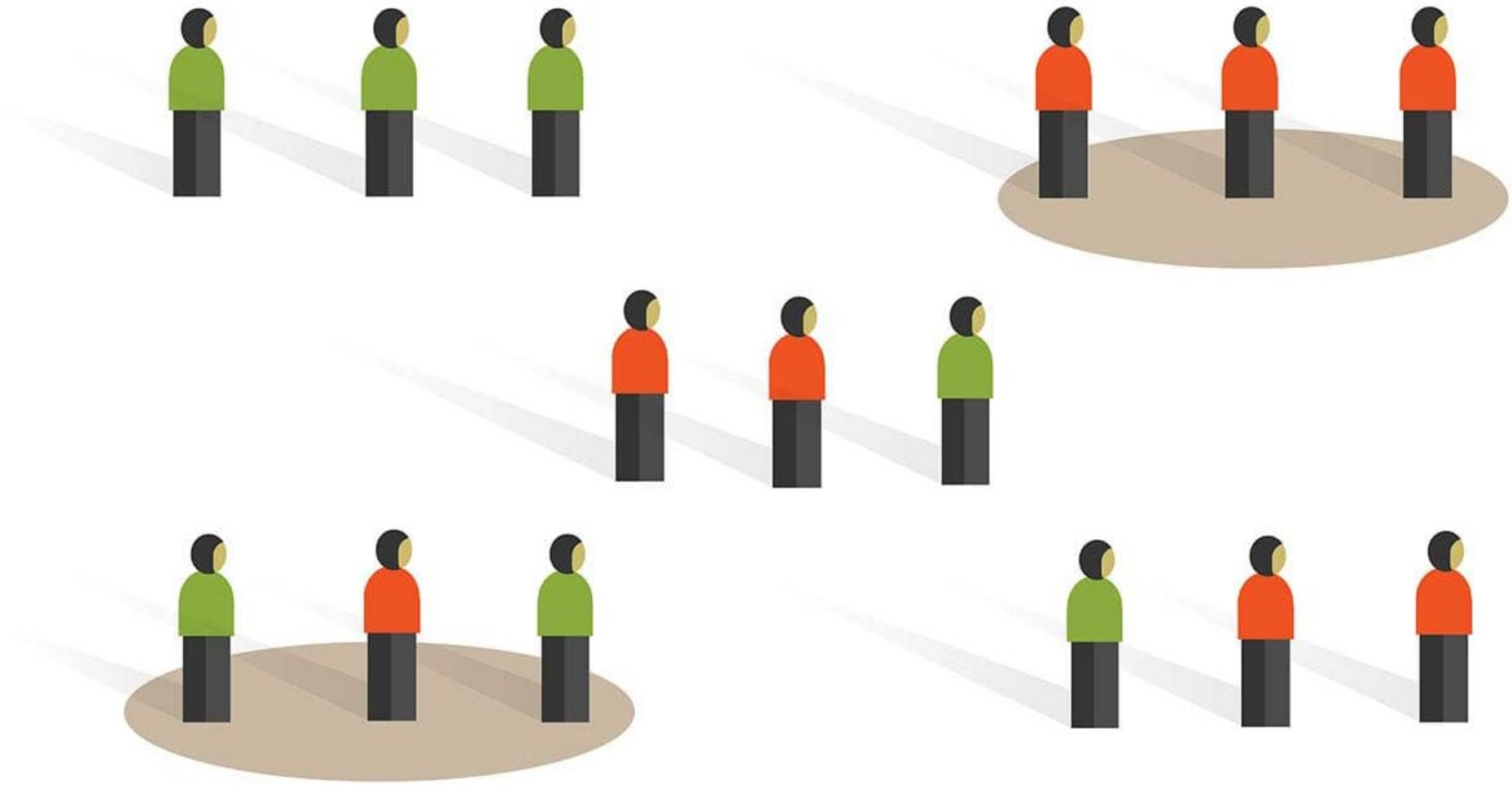


Diabetic clinic

Day	Type of patients
Monday	New patients
Tuesday	Returning patients
Wednesday	Returning patients
Thursday	Returning patients
Friday	Returning patients
Saturday	Returning patients

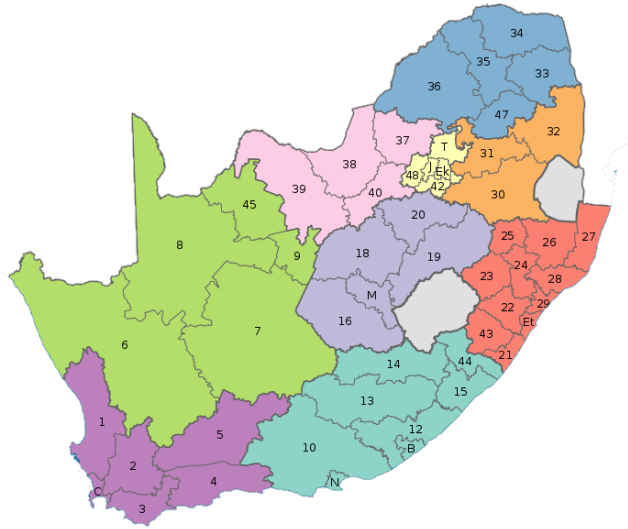


Cluster random sampling



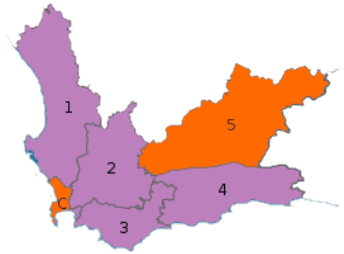
Multistage

Stratification by province



Randomly select 2 districts in each province

Stage 1



Randomly select 3 clinics in each district

Stage 2

Randomly select 4 patients in each clinic

Stage 3

Random selection
(not all clusters are included in the sample)

Large number

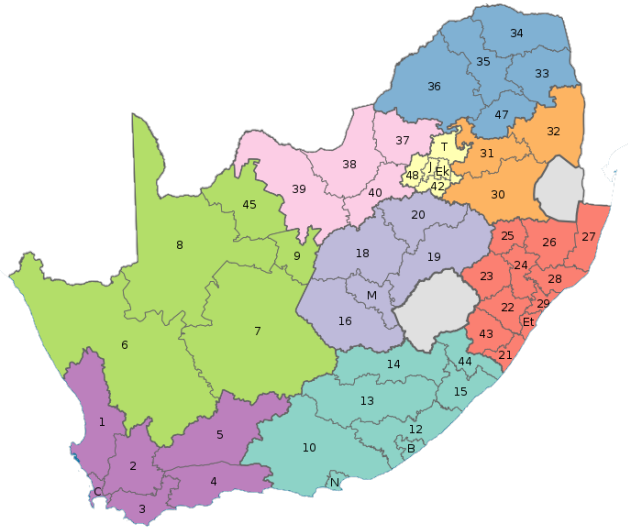
Cluster vs stratum

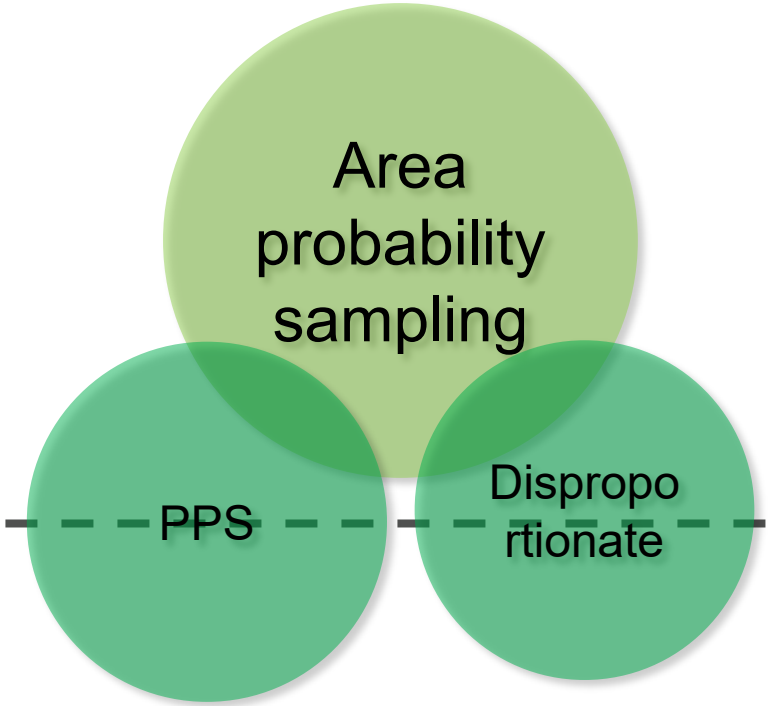
Non-random selection
(all strata are included in the sample)

Small number

Multistage

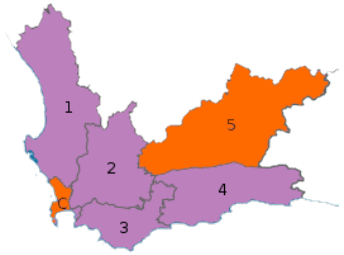
Stratification by province





Randomly select 2 districts in each province

Stage 1



Randomly select 3 clinics in each district

Stage 2

Randomly select 4 patients in each clinic

Stage 3

Advantages and disadvantages

Design	Advantages	Disadvantages
Simple random	• Requires little knowledge of population in advance.	• May not capture certain groups of interest. • May not be very efficient.
Systematic	• Easy to analyze data and compute sampling (standard) errors. • High precision.	• Periodic ordering of elements in sample frame may create biases in the data. • May not capture certain groups of interest. • May not be very efficient.
Stratified	• Enables certain groups of interest to be captured. • Enables disproportionate sampling and optimal allocation within strata. • Highest precision.	• Requires knowledge of population in advance. • May introduce more complexity in analyzing data and computing sampling (standard) errors.
Cluster	• Lowers field costs. • Enables sampling of *groups* of individuals for which detail on individuals themselves may not be available.	• Introduces more complexity in analyzing data and computing sampling (standard) errors. • Lowest precision.

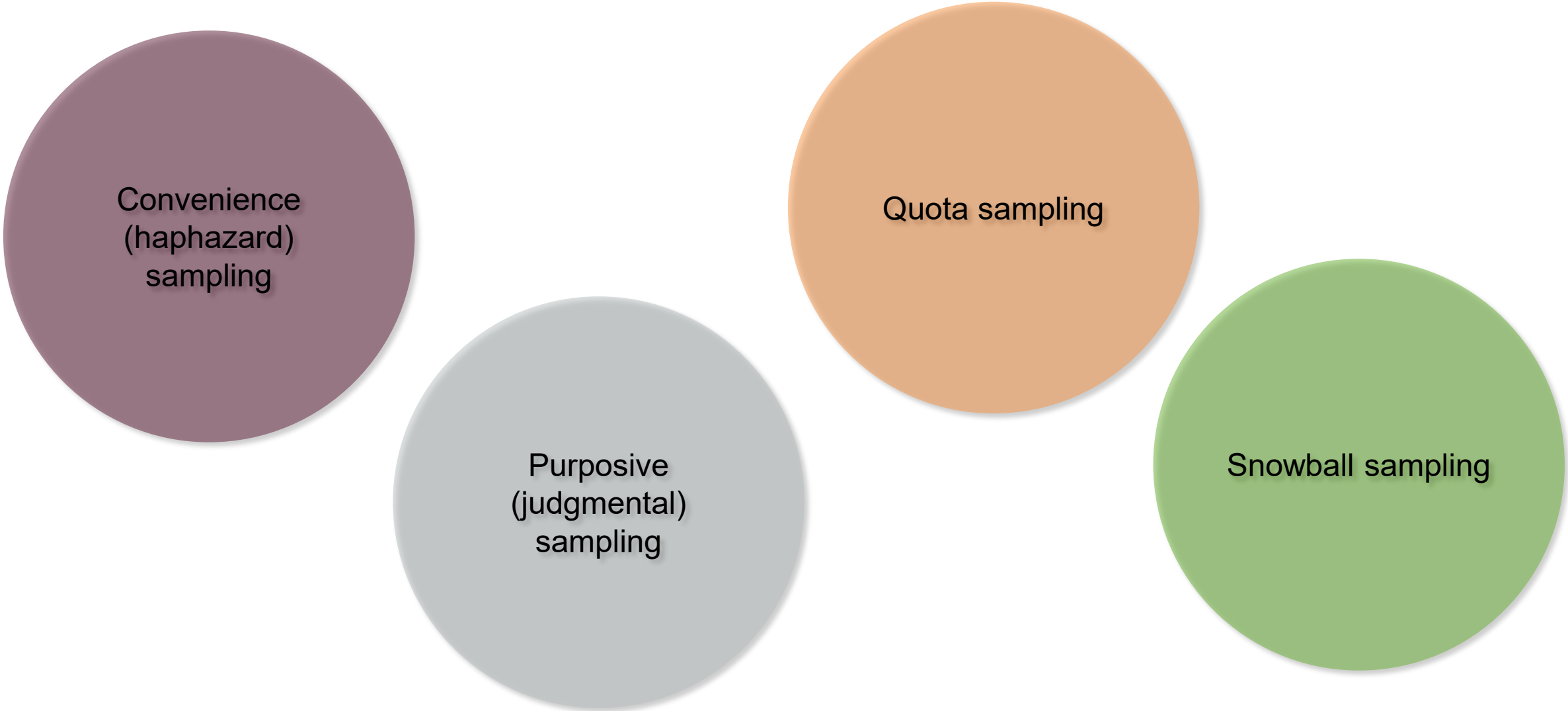
(Table form Aday and Cornelius 2006:135)

Efficiency

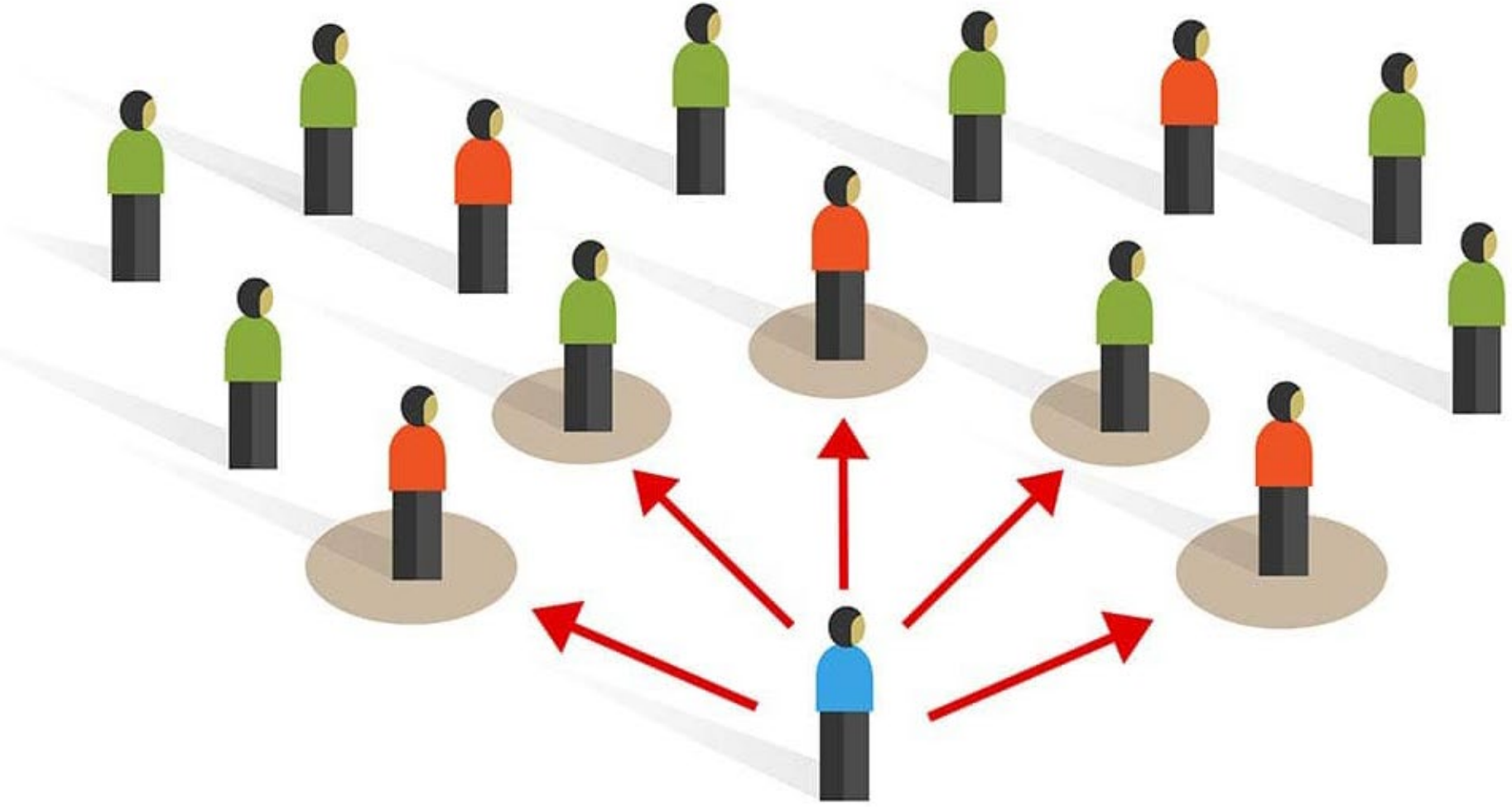
Risk of bias

Type of analysis

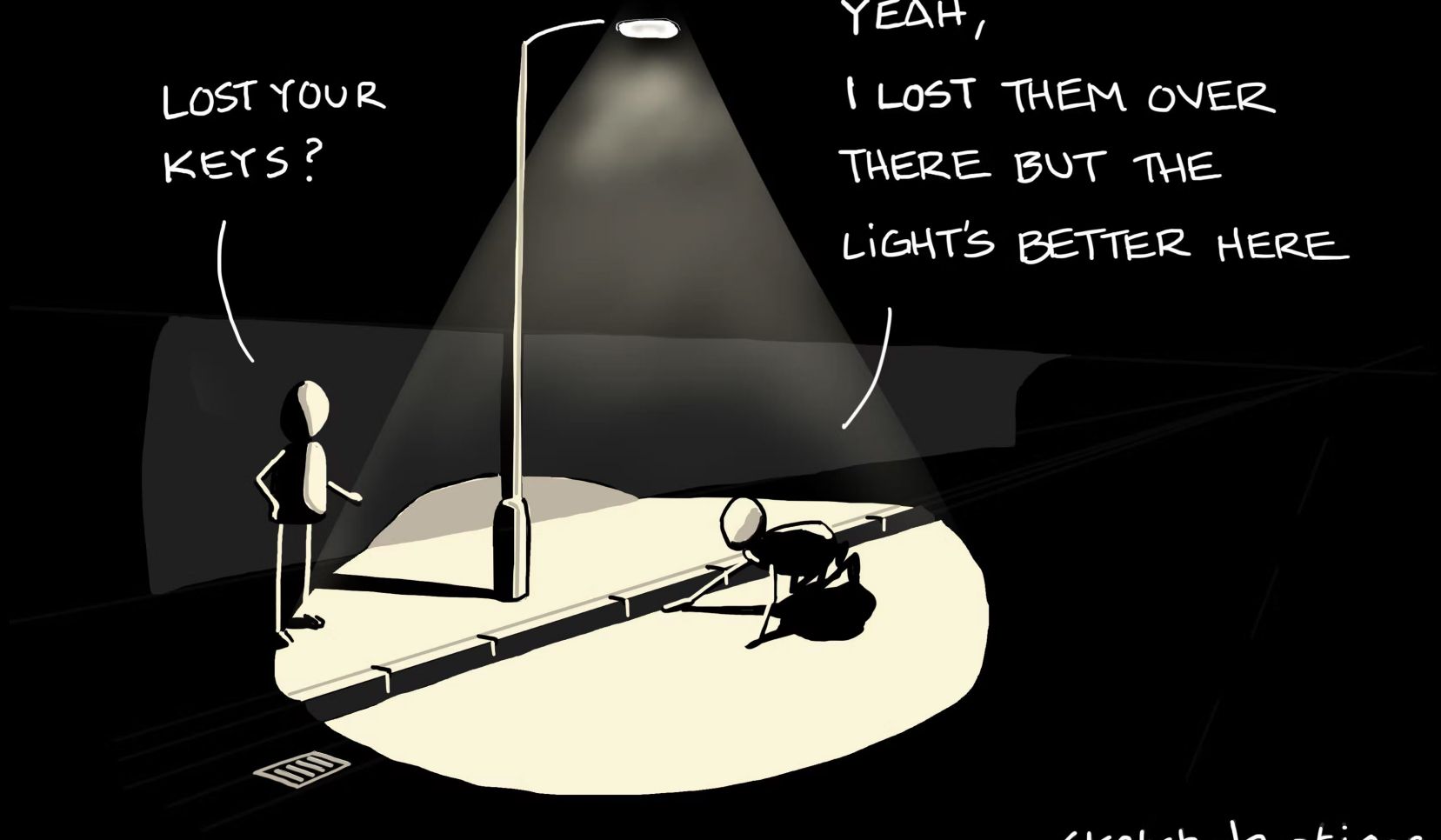
Non-probability sampling strategies



Convenience sampling

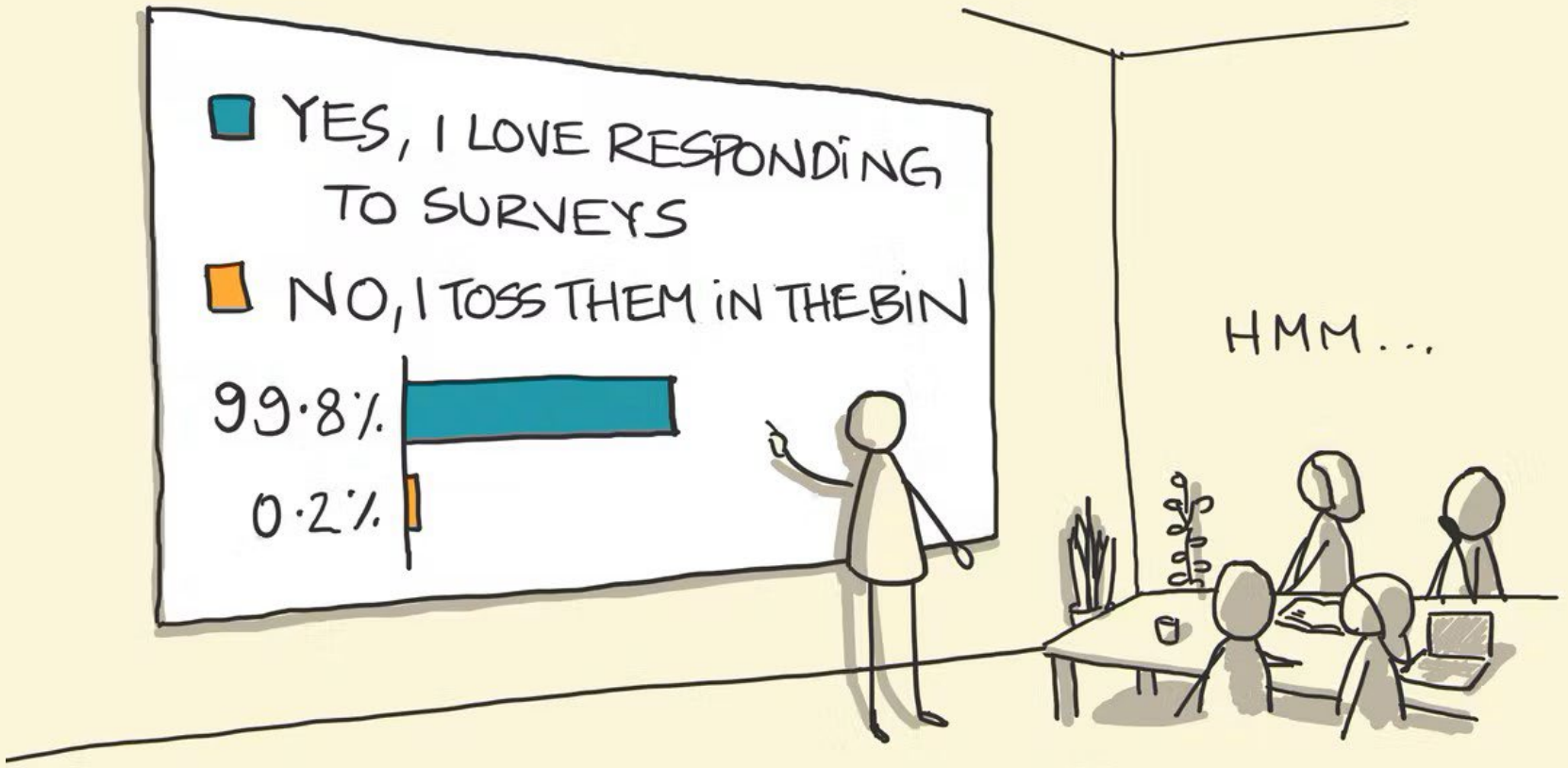


LOOKING UNDER THE LAMPPOST



sketchplanations

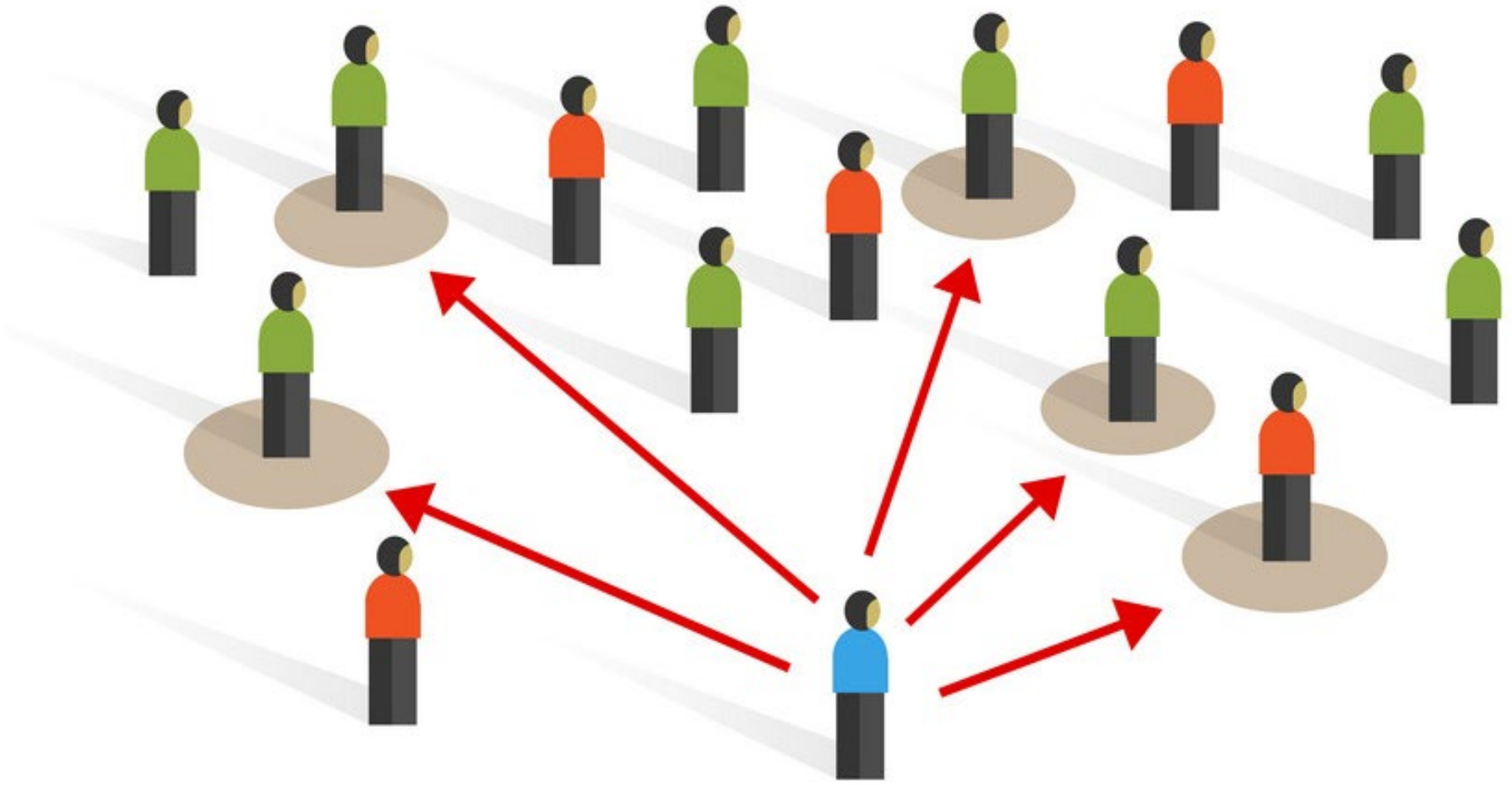
SAMPLING BIAS

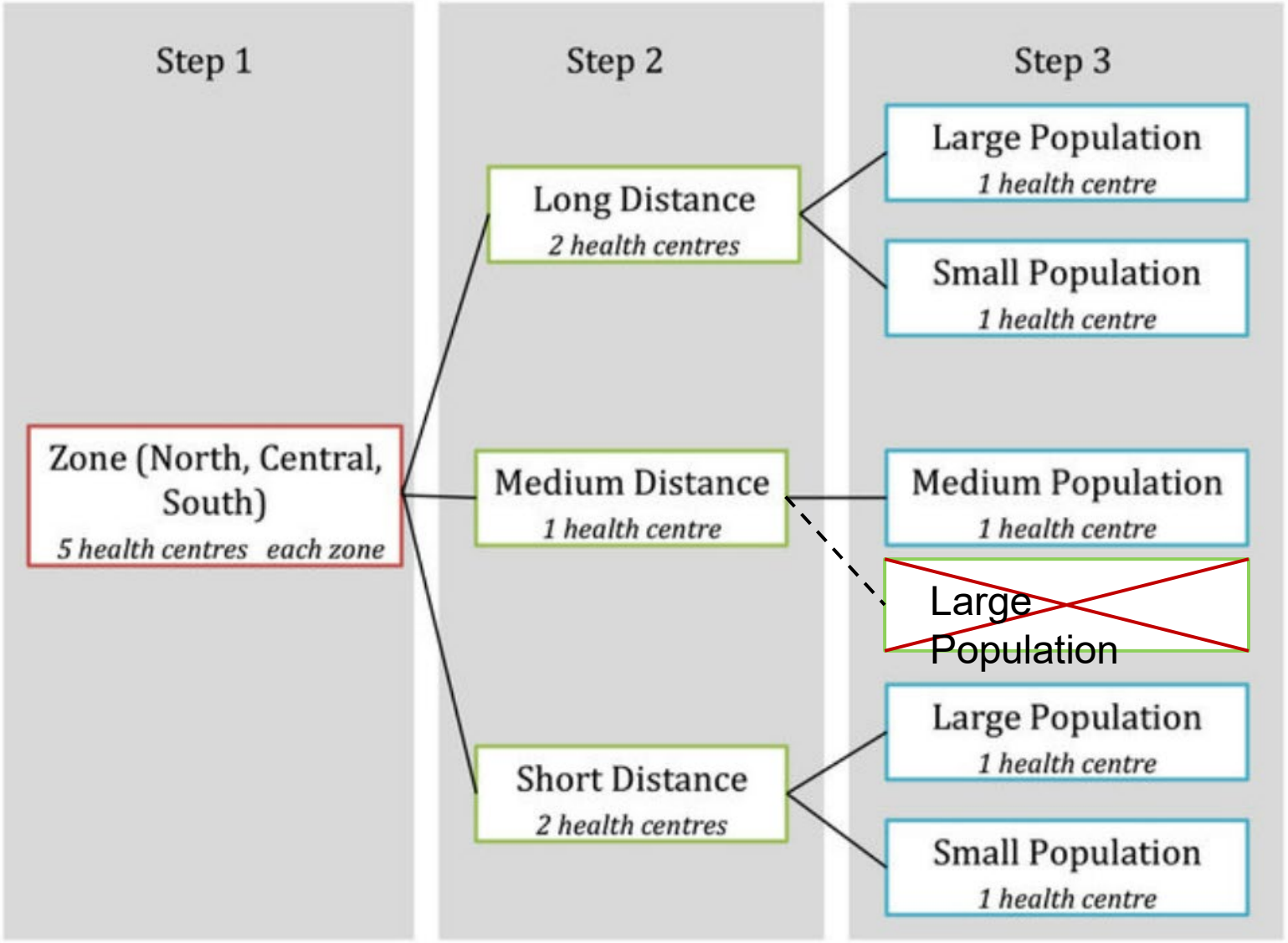


"WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS"

sketchplanations

Purposive sampling

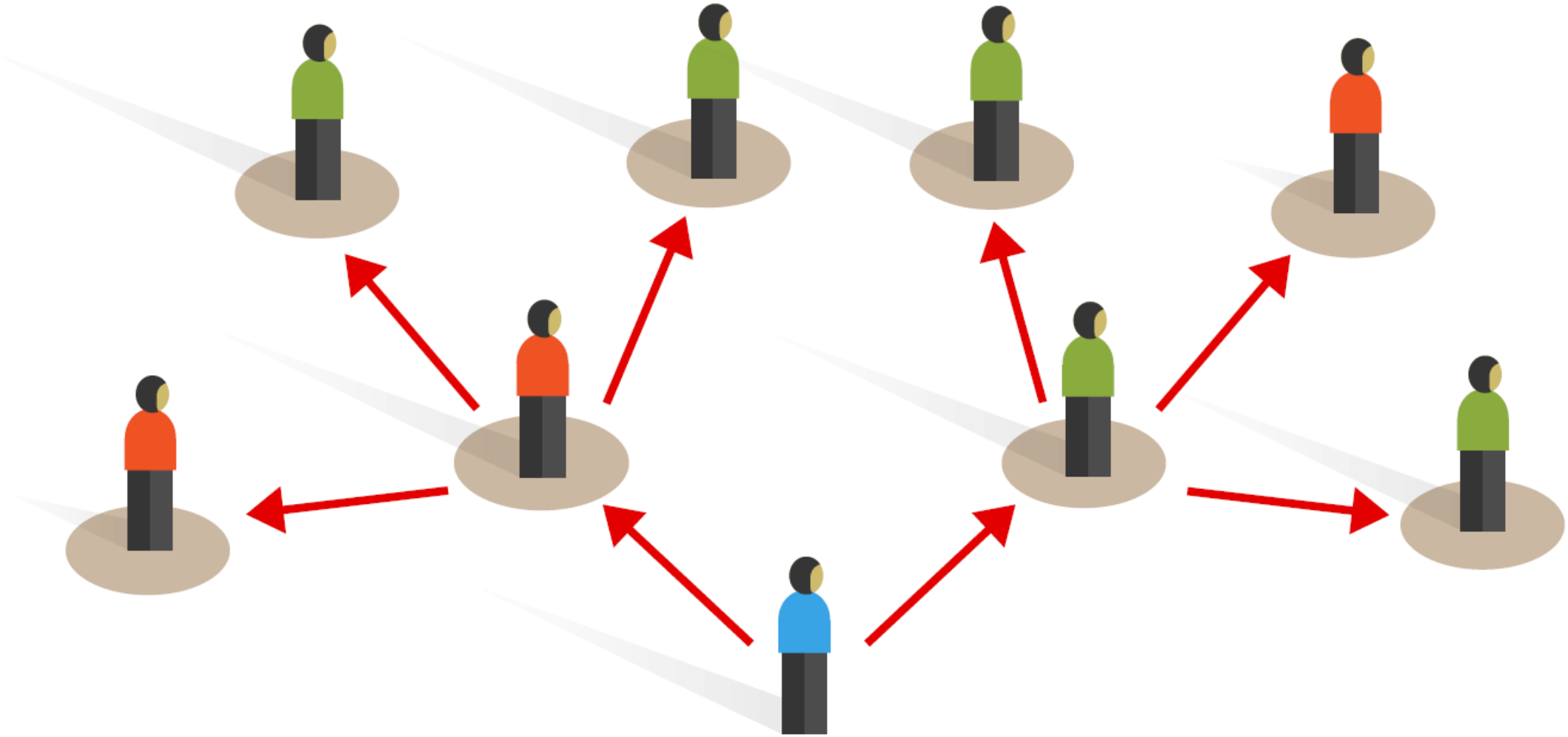




Some elements of the population have no chance of selection (p =0)

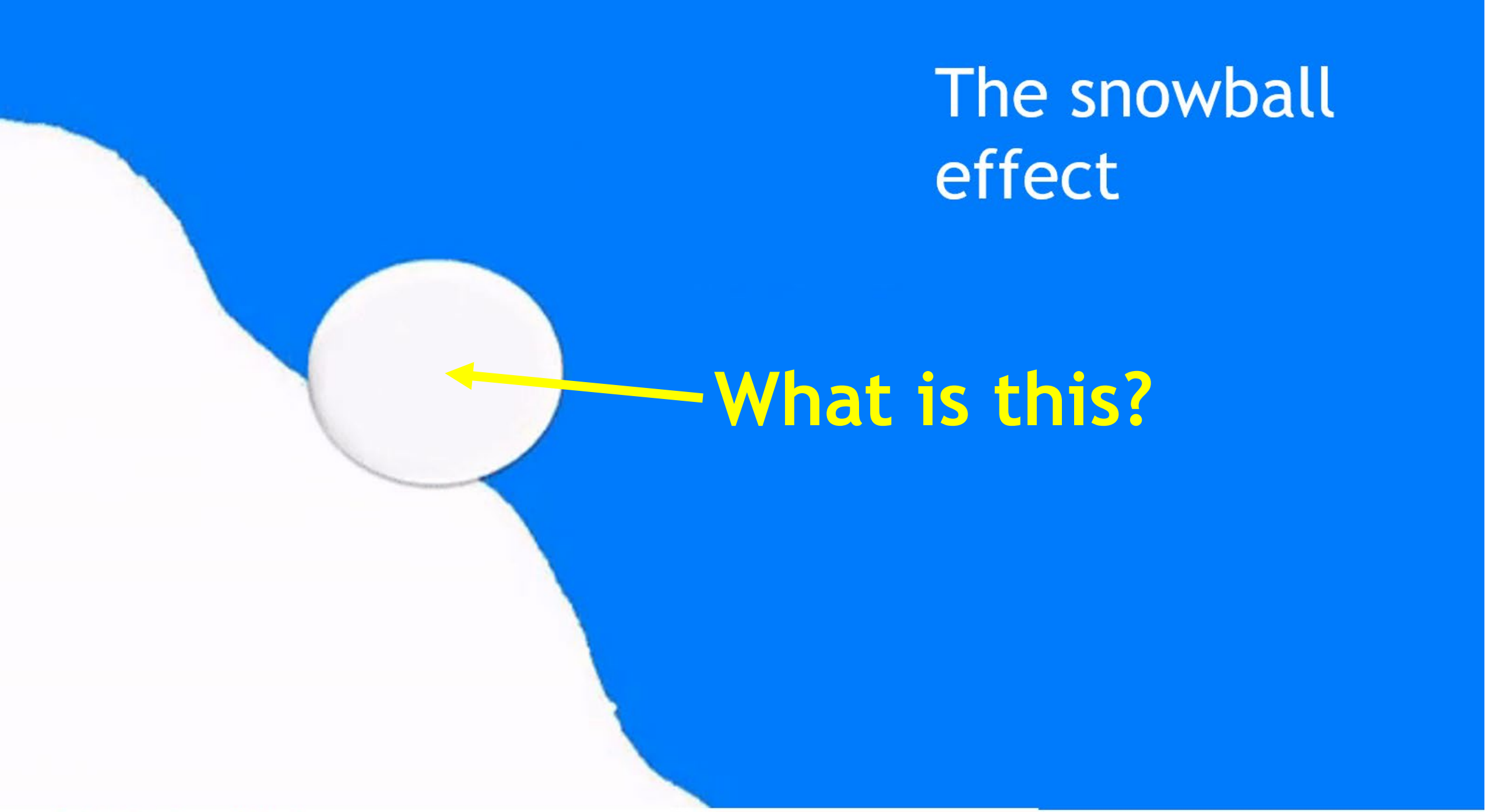
Hasselback L et al. Malar J 13, 295 (2014). https://doi.org/10.1186/1475-2875-13-295

Snowball sampling

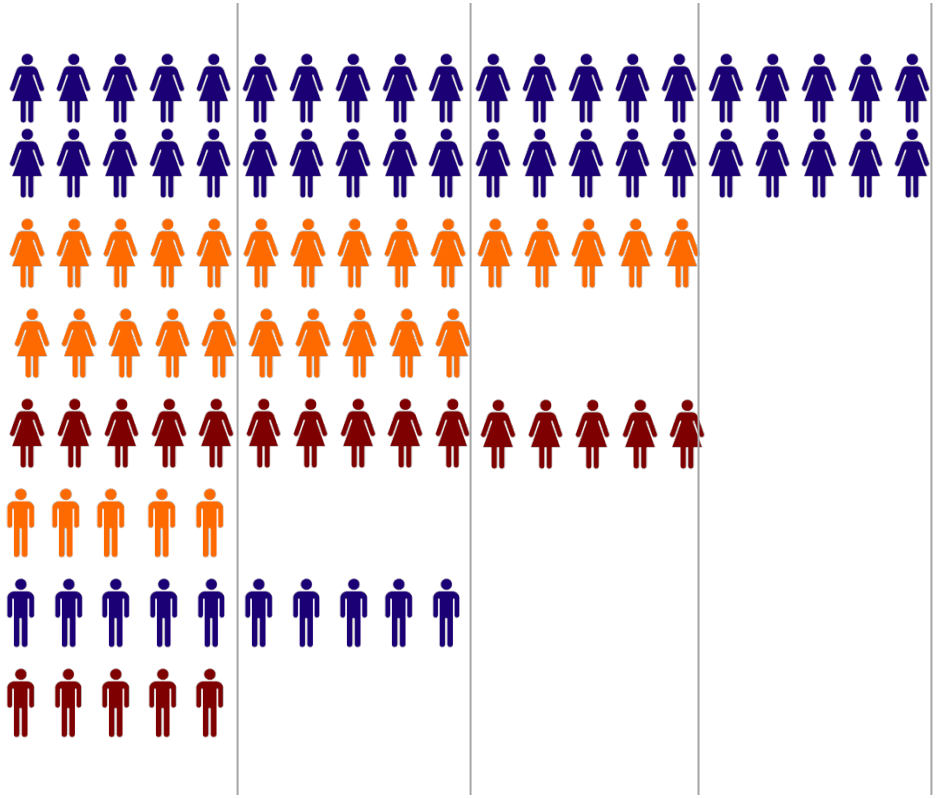


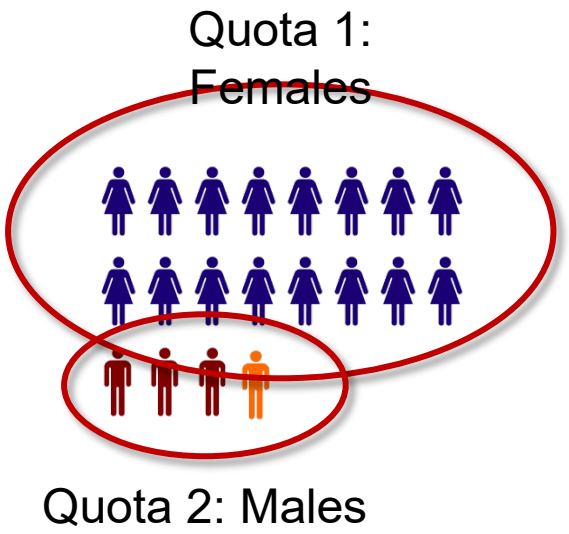
The snowball effect

The snowball effect



Quota sampling





1. Only the **selected** traits of the population are taken into account in forming the subgroups
2. Selection within subgroups is **not random**

Power, Precision, Sample size

Power and precision

Power

Ability of a study to conclude that two population differ in some specific characteristics if they really are different

↓

Hypothesis testing

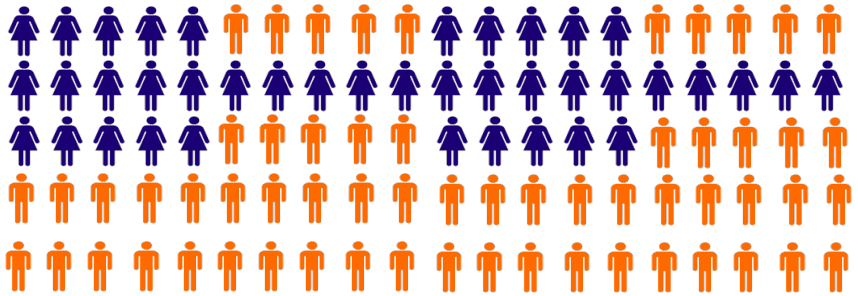
Precision

Width of the interval of uncertainty around the estimate of some population characteristic

↓

Uncertainty around point estimates

Population 1



Population 2

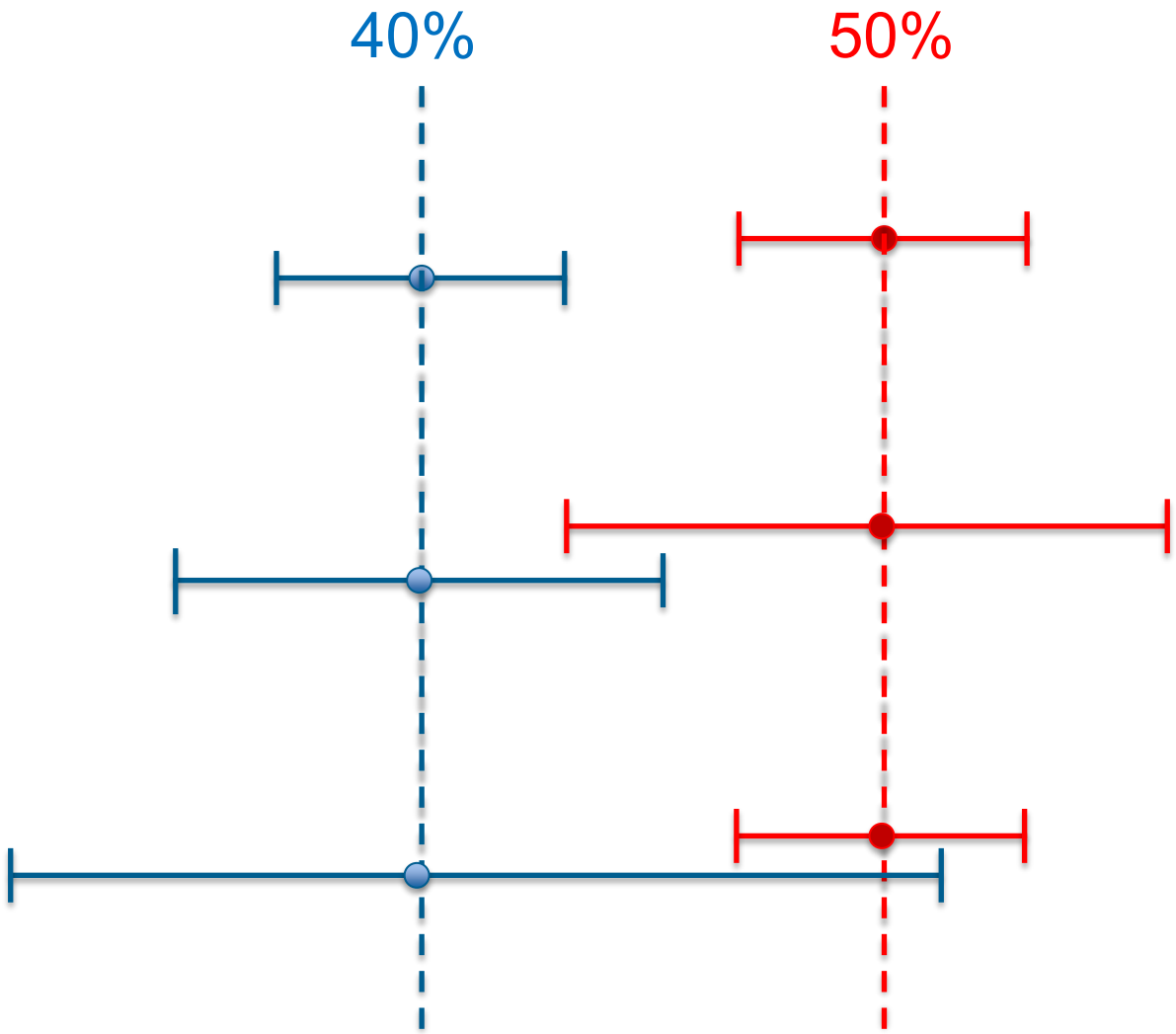
Research question: is the sex distribution the same in these two populations?

Sample 1

Sample 2

5/10 – 4/10 = 1/10 <> 0

Answer: No

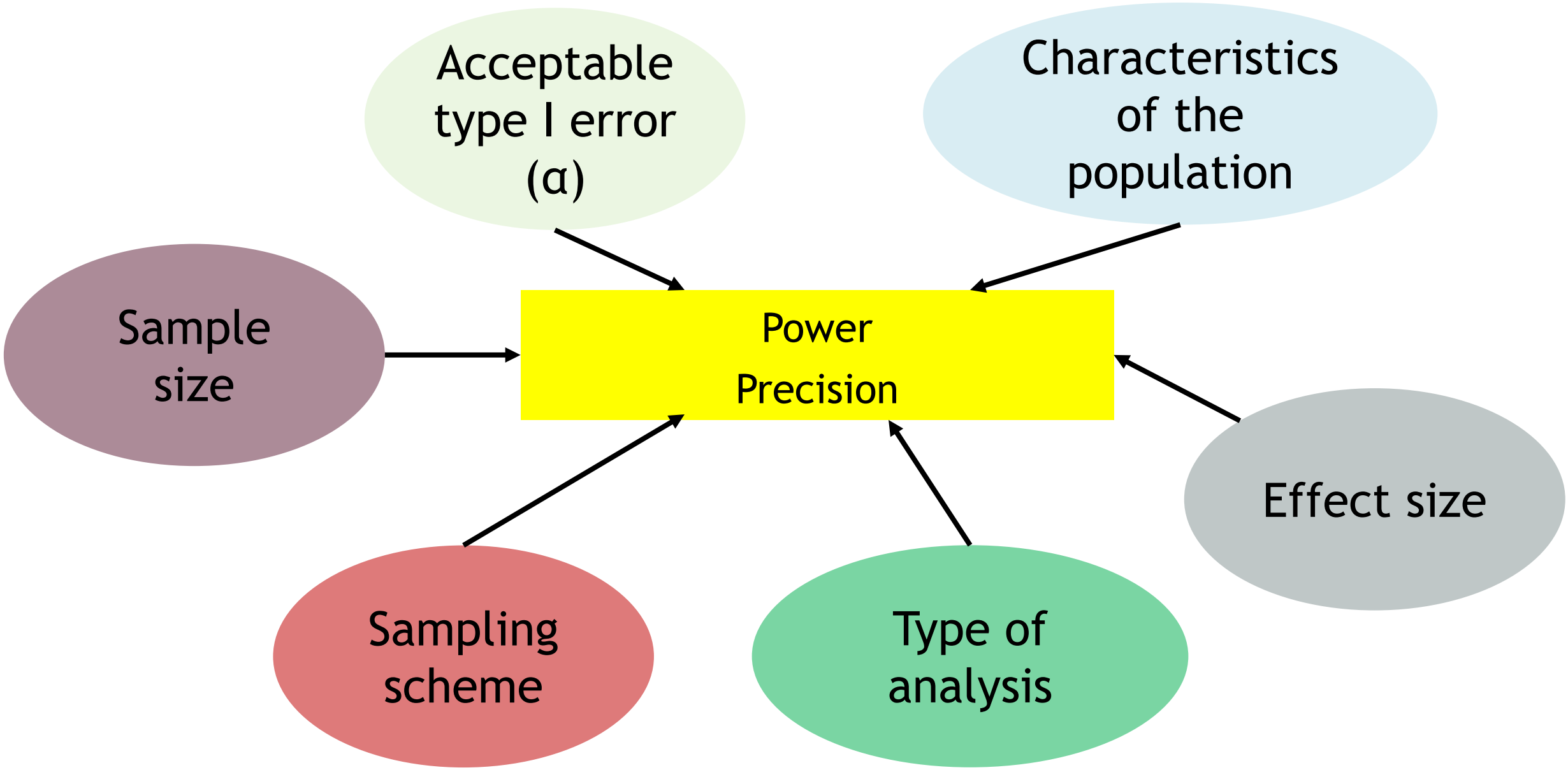


Estimates 1

Estimates 2

Estimates 3

What affects the power of a study?



Sample size for single proportion (descriptive studies)

$$N = \frac{Z^2_{\alpha/2} \times P \times (1 - p) \times D}{E^2}$$

N - sample size

P - prevalence or proportion of event

E - precision (or margin of error) with which a researcher want to measure something

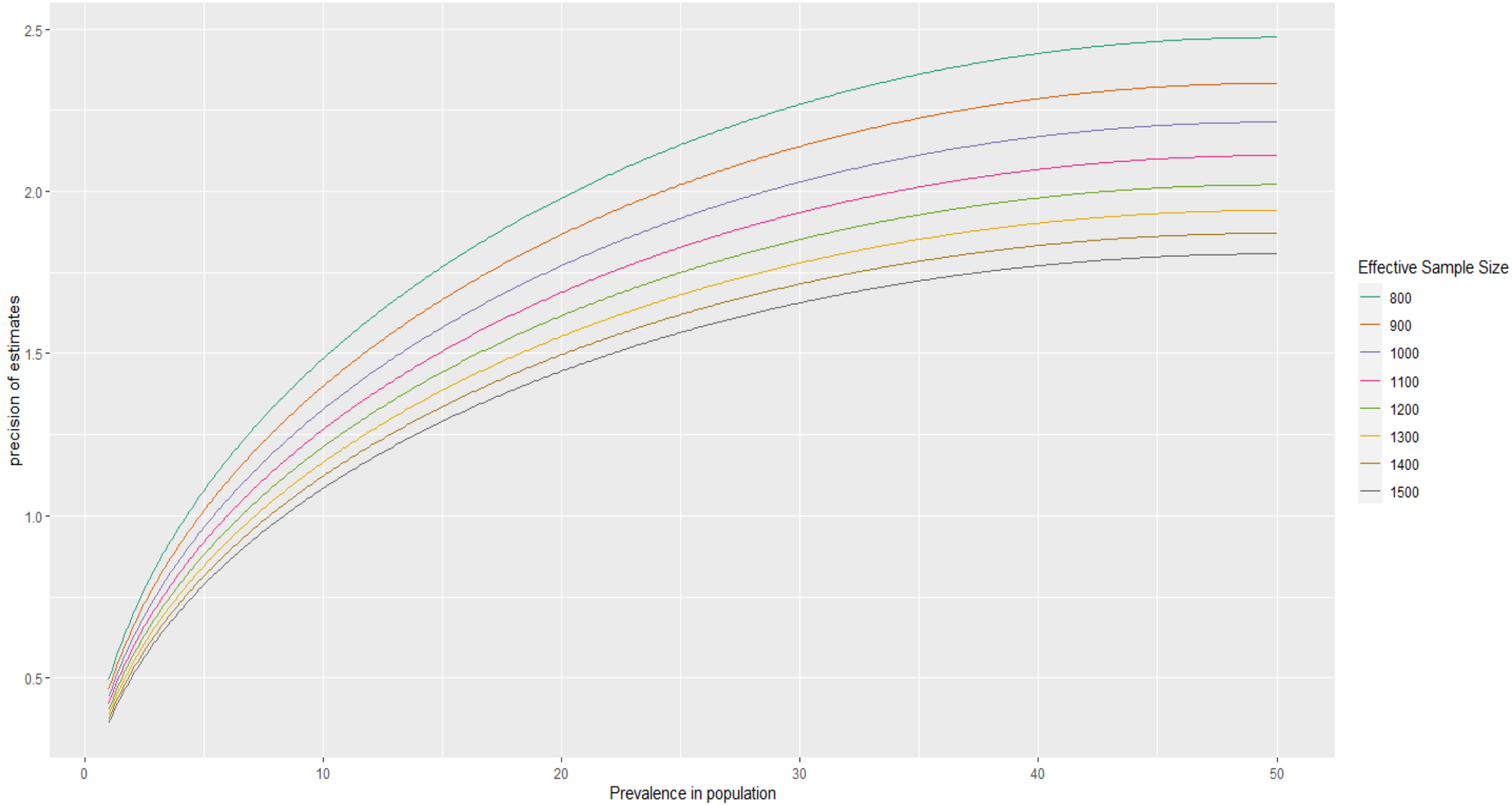
D - design effect reflects the sampling design used in the survey type of study. This is 1 for simple random sampling and higher values (usually 1 to 2) for other designs such as stratified, systematic, cluster random sampling

$\mathbf{Z_{\alpha/2}}$ - 1.96 for alpha 0.05

Sample size for single mean (descriptive studies)

$$n = (Z_{\alpha/2})^2 \, s^2 / d^2$$

s = population standard deviation
d is the require precision
. $Z_{\alpha/2}$ is normal deviate for two- tailed alternative hypothesis at a level of significance α.


precision of estimates
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
Prevalence in population
Effective Sample Size
800
900
1000
1100
1200
1300
1400
1500

SAMPLE SIZE BASIC CALCULATOR

PREVALENCE STUDY - BINOMIAL APPROXIMATION

A) SAMPLE SIZE CALCULATION

REQUIRED **ABSOLUTE PRECISION** (±%): d =	5.00E+00	%	
ESTIMATED **PREVALENCE IN POPULATION** (%): p =	5.00E+01	%	
ESTIMATED **NON-RESPONSE RATE** (%): nr =	0	%	
DESIGN EFFECT: df =	1		
EFFECTIVE SAMPLE SIZE: n1 =	384	(	384.16)
SAMPLE SIZE **NEEDED**: n2 =	384	(	384.16)

B) CALCULATION OF PRECISION

AVAILABLE **SAMPLE SIZE**: n2 =	2743		
ESTIMATED **PREVALENCE IN POPULATION** (%): p =	30	%	
ESTIMATED **NON-RESPONSE RATE** (%): nr =	20	%	
DESIGN EFFECT: df =	1.7		
EFFECTIVE SAMPLE SIZE: n1 =	1291	(	1290.82)
ABSOLUTE PRECISION (±%): d =	2.5	(	2.50)