From a few to the whole crowd

Dr Annibale Cois, PhD

Burden of Disease Research Unit, South African Medical Research Council &
School of Public Health, University of Cape Town

Dr Makandwe Nyirenda, PhD
Burden of Disease Research Unit, South African Medical Research Council &
School of Nursing and Public Health, University of KwaZulu-Natal

Cape Town, SAMRC Conference Centre, 2-3 October 2025

The South African Medical Research Council
recognizes the catastrophic and persisting consequences of colonialism and apartheid, including land dispossession and the intentional imposition of educational and health inequities.

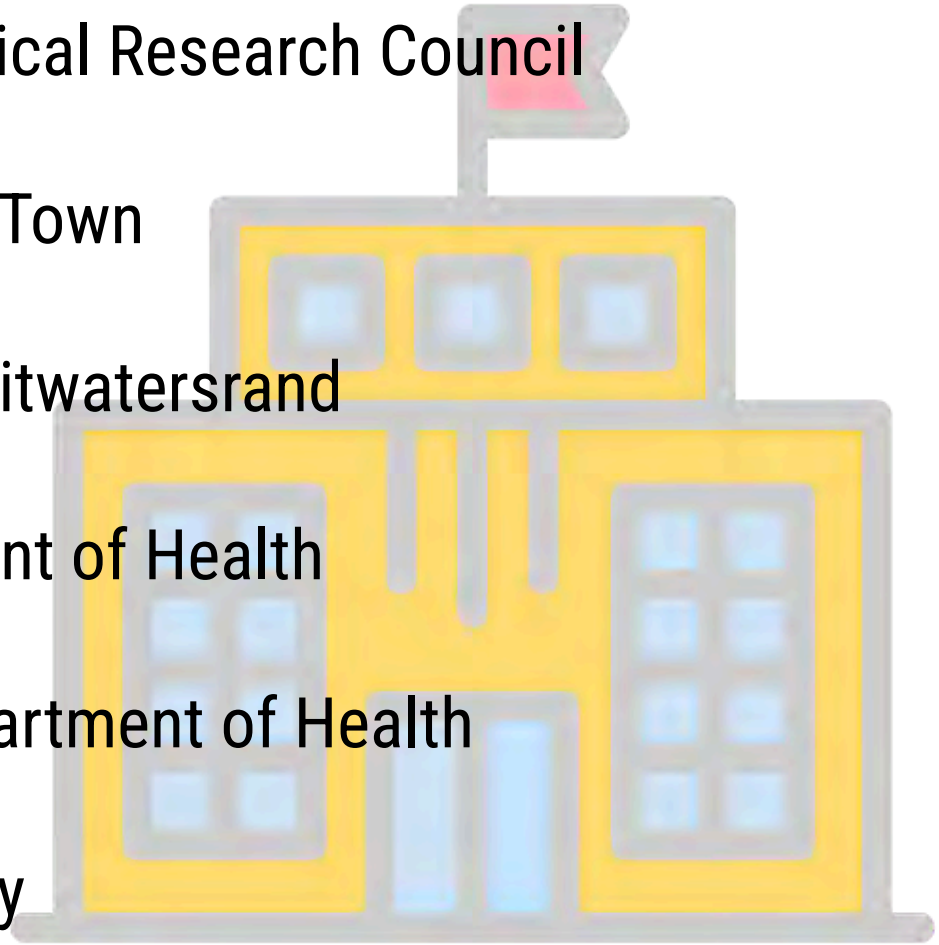
Acknowledging the SAMRC's historical role and silence during apartheid,
we commit our capacities and resources to the continued promotion of justice and dignity in health research in South Africa.

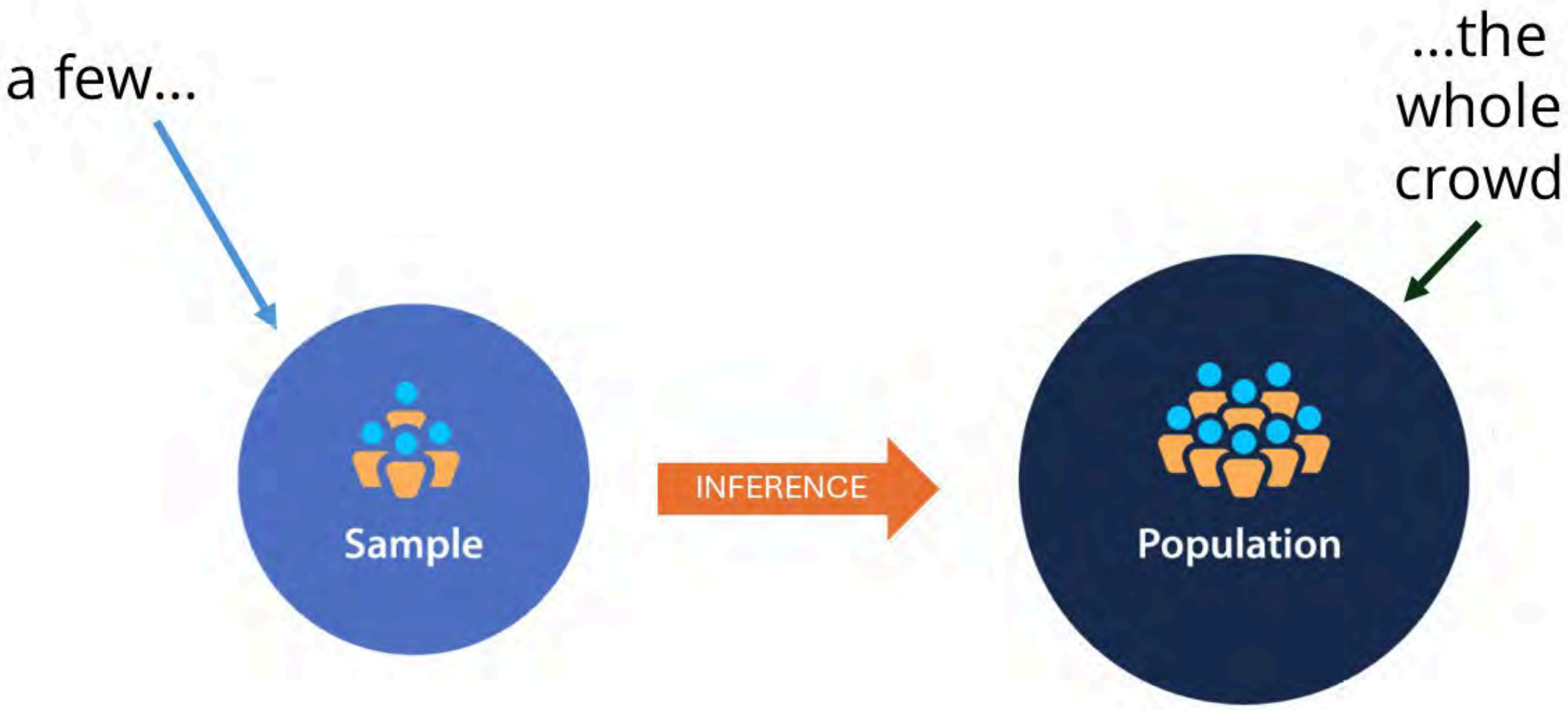
transformation

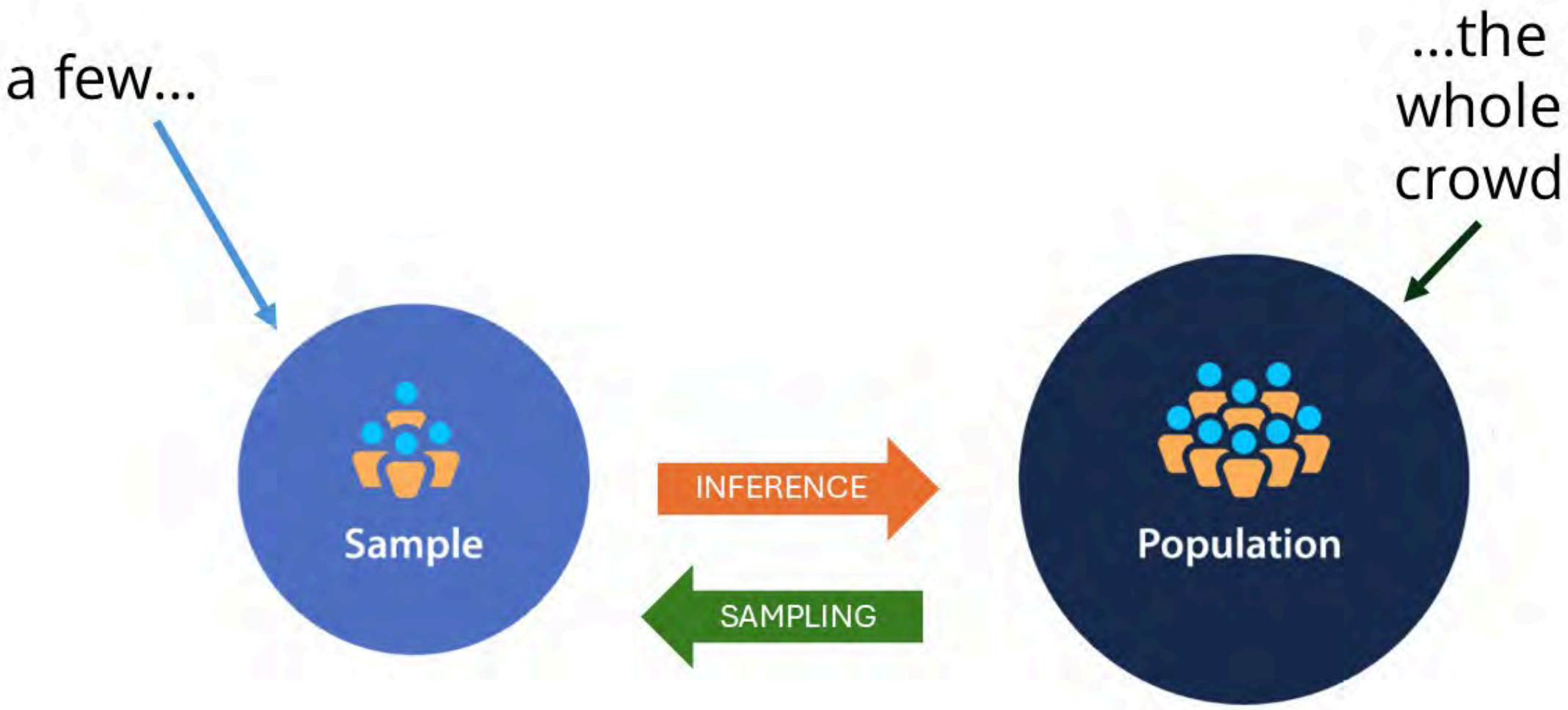
Introduction

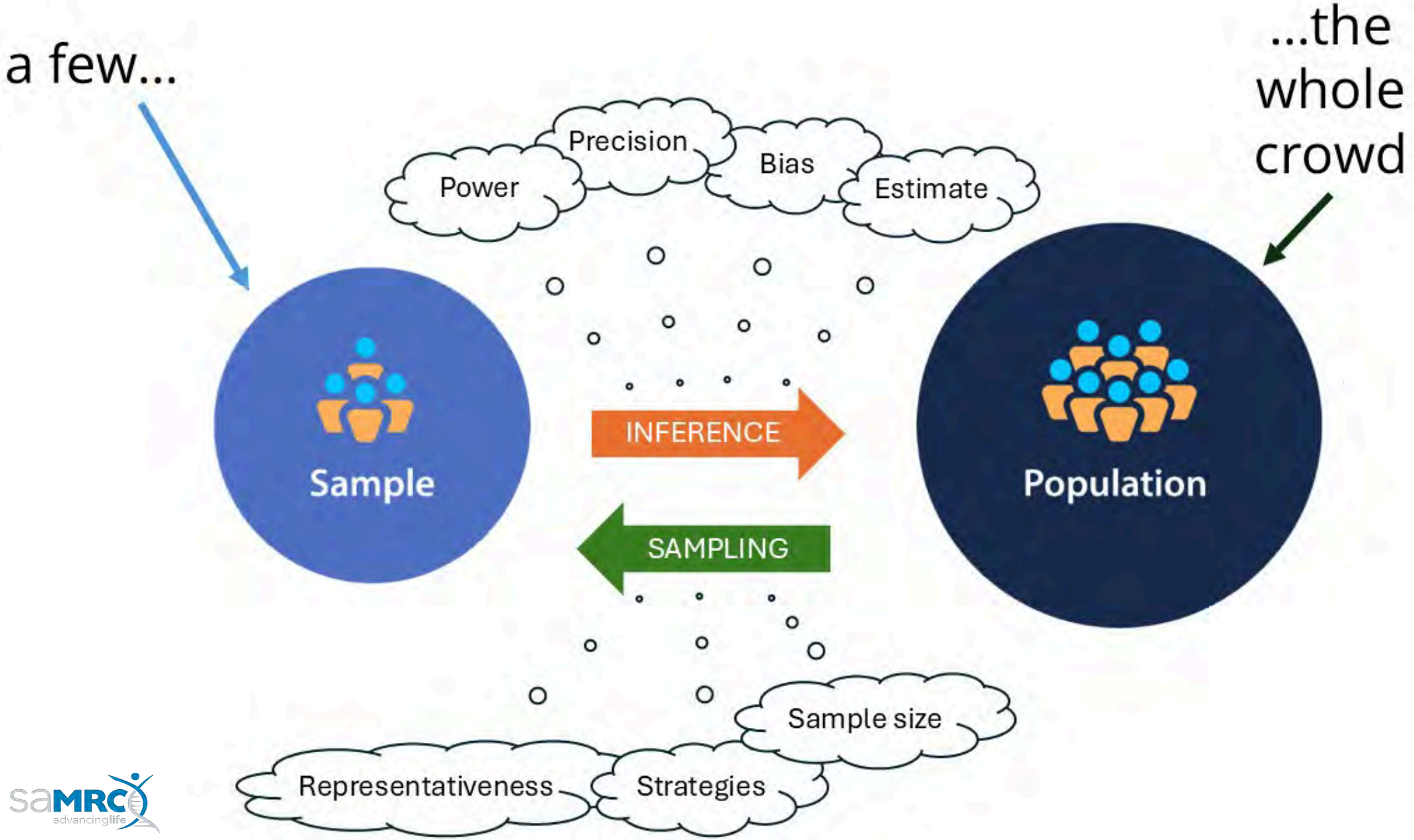
Learning objectives, methods

- South African Medical Research Council
- University of Cape Town
- University of the Witwatersrand
- National Department of Health
- Western Cape Department of Health
- Makerere University




a few...
...the whole crowd
Sample
INFERENCE
Population


a few...
...the whole crowd
Sample
INFERENCE
SAMPLING
Population


a few...
...the whole crowd
Precision
Power
Bias
Estimate
Sample
INFERENCE
Population
SAMPLING
Sample size
Representativeness
Strategies

https://github.com/AnnibaleCois/SurveyLab


Thaye
Perayea
Mantsocha
Otjimkos
Mmabani
Rakokaa
Toraseng
Peninsula
Luredjera
Mowebane
Virmelo
Durtis
Matshane
Nojea
Sekuwa
Sidvoyale
Hluti
Meyofebeng
Theyasen
Rashakos
Goritz
Kuronga
Kasapa
Vamko
Timoyulu
Lettengwe
Beria
Hoachapa
Gabarowe
Keetya
Gwesebjwe
Dukronga
Wandu
Tobalane
Kroonla
Meyaneng
Tabateki
Ngozini
Ondadjera
Ceredock
Dulekume
Wachanas
Moita
Alikal
Empane
Lurebib
Phane
Central Region
Tobabala
Moleronga
Thohoni
Metsina
Reliseru
Khudinare
Buleya
Rayatga
Gohohe
Malutikulu
South East
Bridford
Nokhokane
Sewi
Omuthinie
Khorihao
Konekal
Khoriva
Thahane
Ngobalunye
Eenranos
Alinoni
Mogodika
Rayafone
Kwalunya
Senenan
Quta
Karigwa
Nkudis
Grahaloanong
Gobamun
Quthing
0
60
80
100 km

The research questions

1. What is the prevalence of COPD in our population?

2. Does the presence of a coal-fired powerstations in the town affects the prevalence of COPD in the area?


Thaye
Perayea
Mantsocha
Sidvoyale
Otjimkos
Mmabani
Rakokaa
Hluti
Meyofebeng
Toraseng
Luredjera
Northern Peninsula
Theyaseng
Rashakos
Durtis
Mowebane
Virmelo
Matshane
Nojea
Goritz
Kuronga
Mandunnaka
Vamko
Pretokuwa
Kasapa
Mbalume
Timoyulu
Ramaputsoe
Beria
Hoachapa
Gabarowe
Gwesebjwe
Wesselha
Keetya
Dukronga
Wandu
Berion
Kroonla
Tobalane
Midanha
Taratena
Meyaneng
Ngozini
Mapufong
Nayeng
Kuteni
Warutjo
Wachanas
Ceredock
Temre
Dulekume
Alikal
Empaney
Lurebib
Maltaties
Tshabina
Ngobakulu
Bhulembu
Tobabala
Metsina
Wanbis
Central Region
Saldamanus
Nkadinare
Buhena
Mmawa
Khudinare
Stiltoria
Thohoni
Reliseru
Buleya
Rayatga
Moferabe
Nhlamelunye
Mandunwane
Shotshane
Malutikulu
Tamatleng
Maunatau
South East
Bridford
Nokhokane
Nakojibe
Sewi
Knygel
Omuthinie
Khorihao
Konekal
Rayafone
Khoriva
Thahane
Ngobalunye
Eenranos
Alinoni
Tshatulu
Mogodika
Kwalunya
Senenan
Lavugano
Quta
Karigwa
Nkudis
Mandunhaba
Grahaloanong
Gobamun
Outhing

Learning objectives

- Develop a deep understanding of the links between sampling strategy, sample size, analysis technique, and accuracy of population estimates
- Understand the statistical and practical implications of using complex sampling schemes including clustering and stratification
- Learn alternative approaches for analysing survey data
- Discuss the pros and cons of various sampling strategies and analysis methods to avoid misleading results and waste of resources
- Gain practical insights on collecting and analysing sample data through virtual surveys on a simulated human population

Outline

Day 1

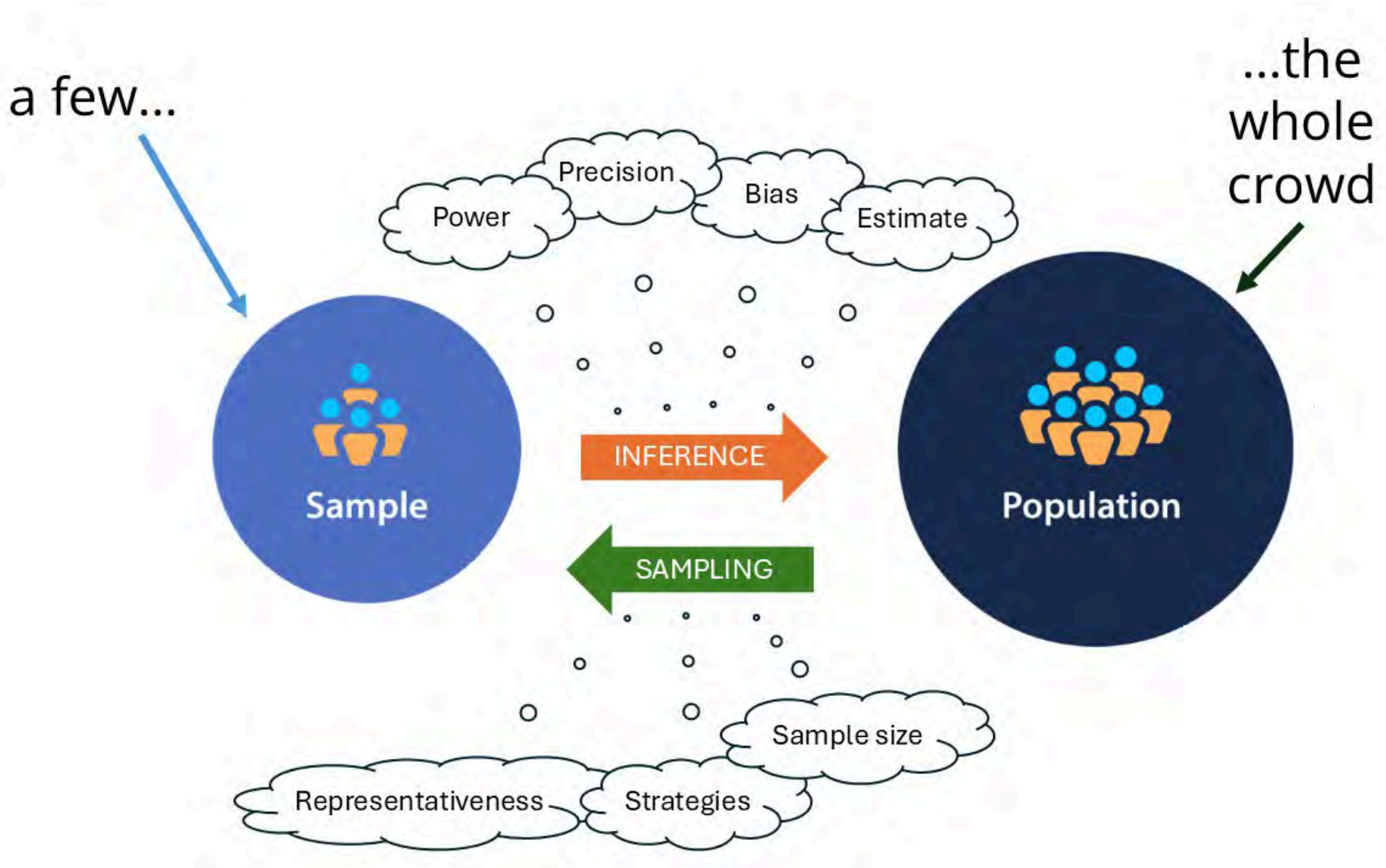
- Fundamentals: concepts and terminology
- Sampling strategies
- The *SurveyLab* environment
- Conducting a survey
- Discussion & recap

Day 2

- Analysing survey data: principles and approaches
- Survey data analysis with R
- Answering our research questions
- From questions, to sampling to estimates: a summary
- Conclusions

Fundamentals

Concepts, terminology, sampling strategies

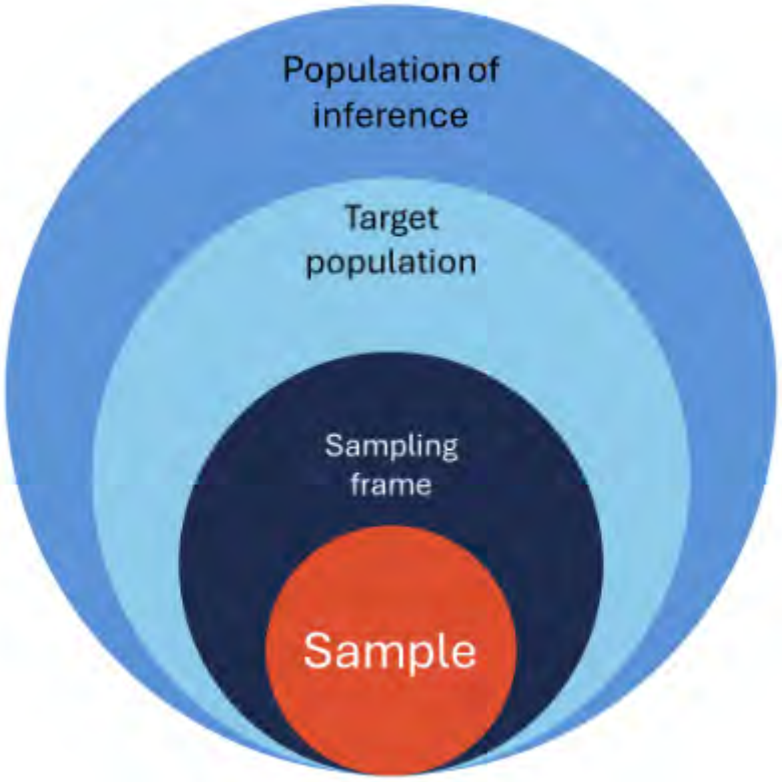


Groves, Fowler Jr, Couper, Lepkowski, Singer, and Tourangeau (2009); Groves and Lyberg (2010)

Populations

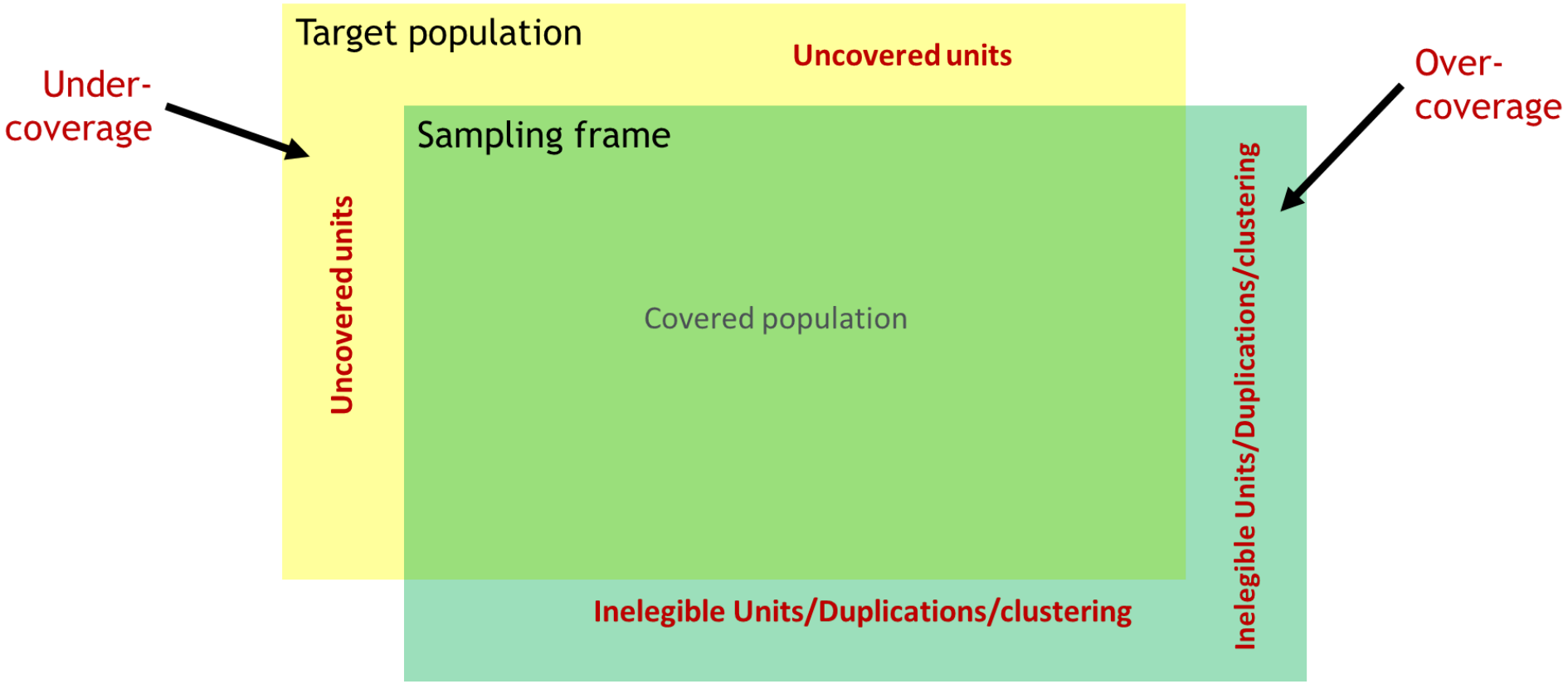
The population of inference is the broad group to whom we want to generalize our findings.

The target population is the specific, finite group of elements that the study aims to represent and is practically accessible for data collection.

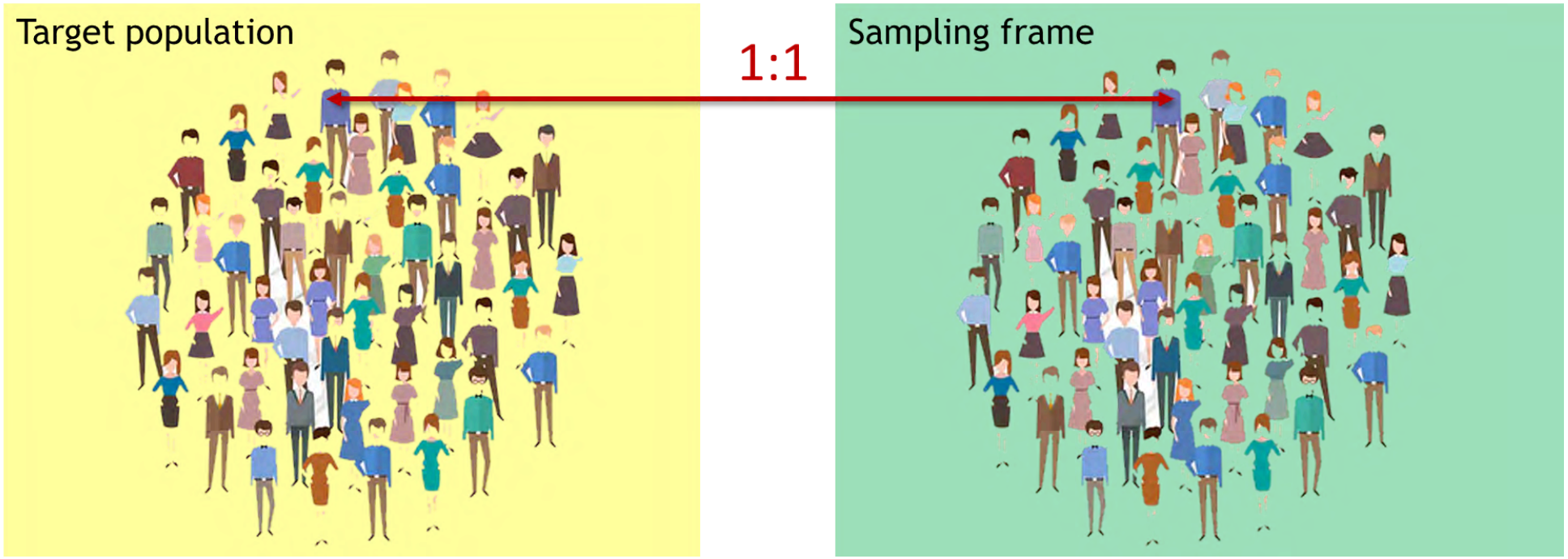


The sampling frame is a "list" of all those within a population who can be sampled

Lee (2008); Raymond and Darsaut (2025)

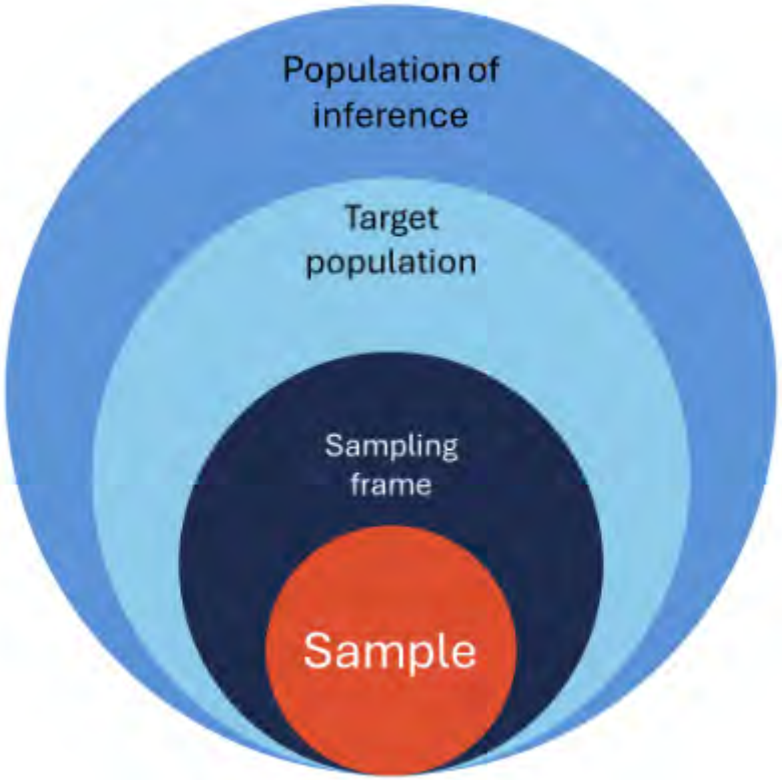

Target population
Uncovered units
Under-
coverage
Over-
coverage
Sampling frame
Uncovered units
Covered population
Inelegible Units/Duplications/clustering
Inelegible Units/Duplications/clustering

OXFAM (2024)



OXFAM (2024)

Sample



Tyrer and Heyman (2016); Banerjee and Chaudhury (2010)

What is a sample?

How do we choose a sample?

Statistical rigour

"We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression."

— Sir R. A. Fisher

Types of Sampling Approaches

- Non-probability sampling
  - Selection is based on some arbitrary criteria of the person picking the sample

- *Probability sampling*
  - The probability of being included in the sample is known for each subject in the population

Types of nonprobability sampling

- Convenience sampling
- Quota sampling
- Snowball sampling

Types of probability sampling

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Cluster sampling

Probability Sampling

- *Simple random sampling (SRS)*
  - Every subject has an equal probability of being selected
  - Use a random number table with a random start or a computer-generated random number

- *Systematic sampling*
  - Every k^{th} item is selected; k is determined by dividing the number of items in the sampling frame by the desired sample size. Using a random starting point
  - Do not use when there is cyclic repetition

Probability Sampling

- *Stratified sampling*
  - Population is first divided into relevant strata (subgroups) and a random sample is taken from each stratum
  - Characteristics used to stratify should be related to the measurement of interest
  - Can be the most efficient

- *Cluster sampling*
  - Multistage process in which the population is divided into clusters and a subset of clusters is randomly selected. Either all subjects in the cluster sampled or SRS used to select a subsample
  - Requires larger sample size, but may be most practical

What biases can result from sampling?

Statistical rigour

"We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression."

— Sir R. A. Fisher

Biases

- Sample not representative of the population
  - Non-response
  - Geographical/seasonal/volunteer bias
- Instrument/Interviewer reliability

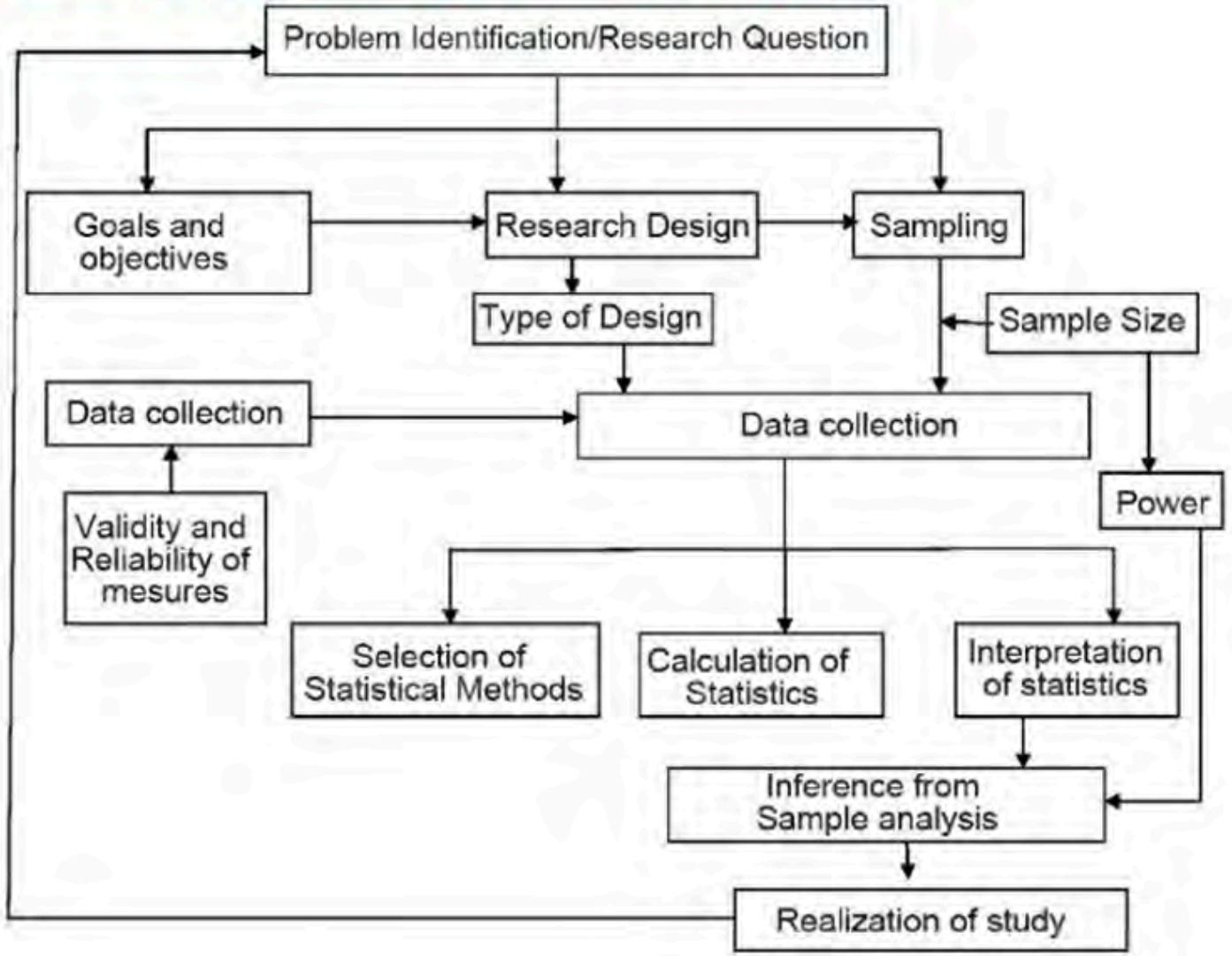
- *Bias is the result*
  - Statistical method used estimating measure thought to be estimated
  - Often results from poor study design
    - Experimental and control groups exposed to varying environmental factors
    - Two groups being compared are clinical/demographically different

  - A good statistical design minimizes bias, ensures adequate sample size (not too big/small), all study questions are answered

"The most elegant design of a clinical study will not overcome the damage caused by unreliable or imprecise measurement."

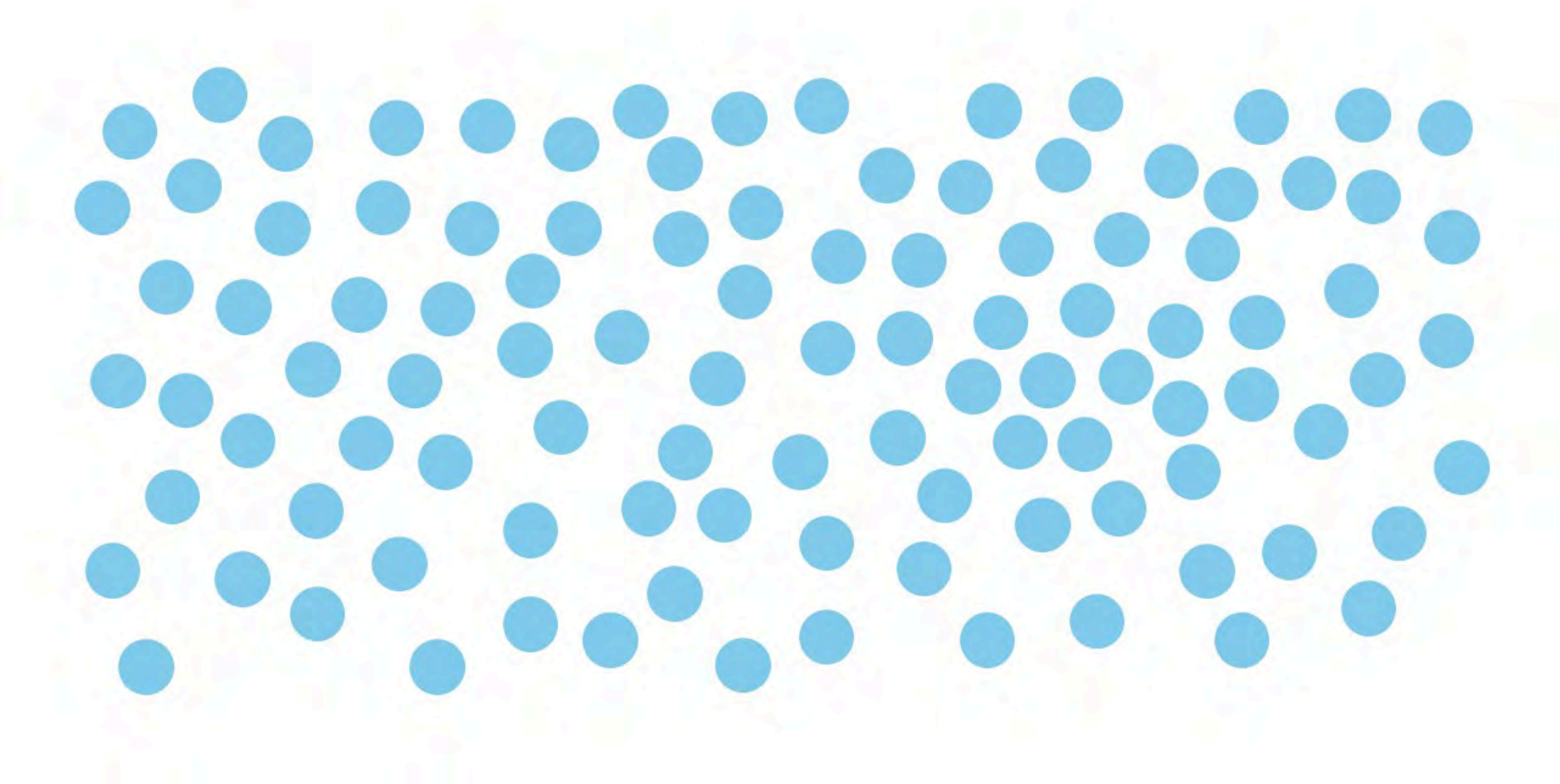
Fleiss, 1986

Research design framework



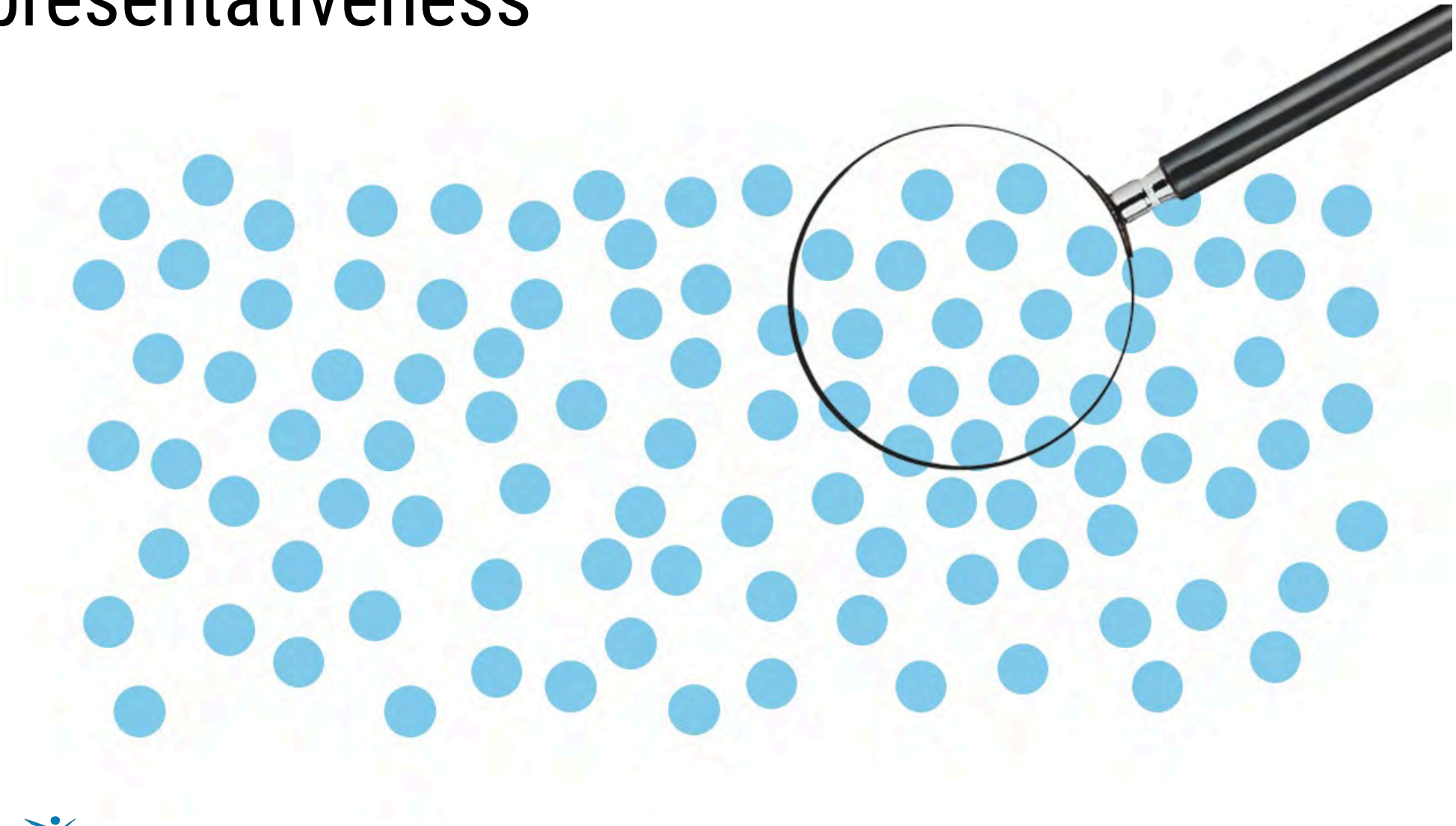
Suresh, Thomas, and Suresh (2011b)

Representativeness

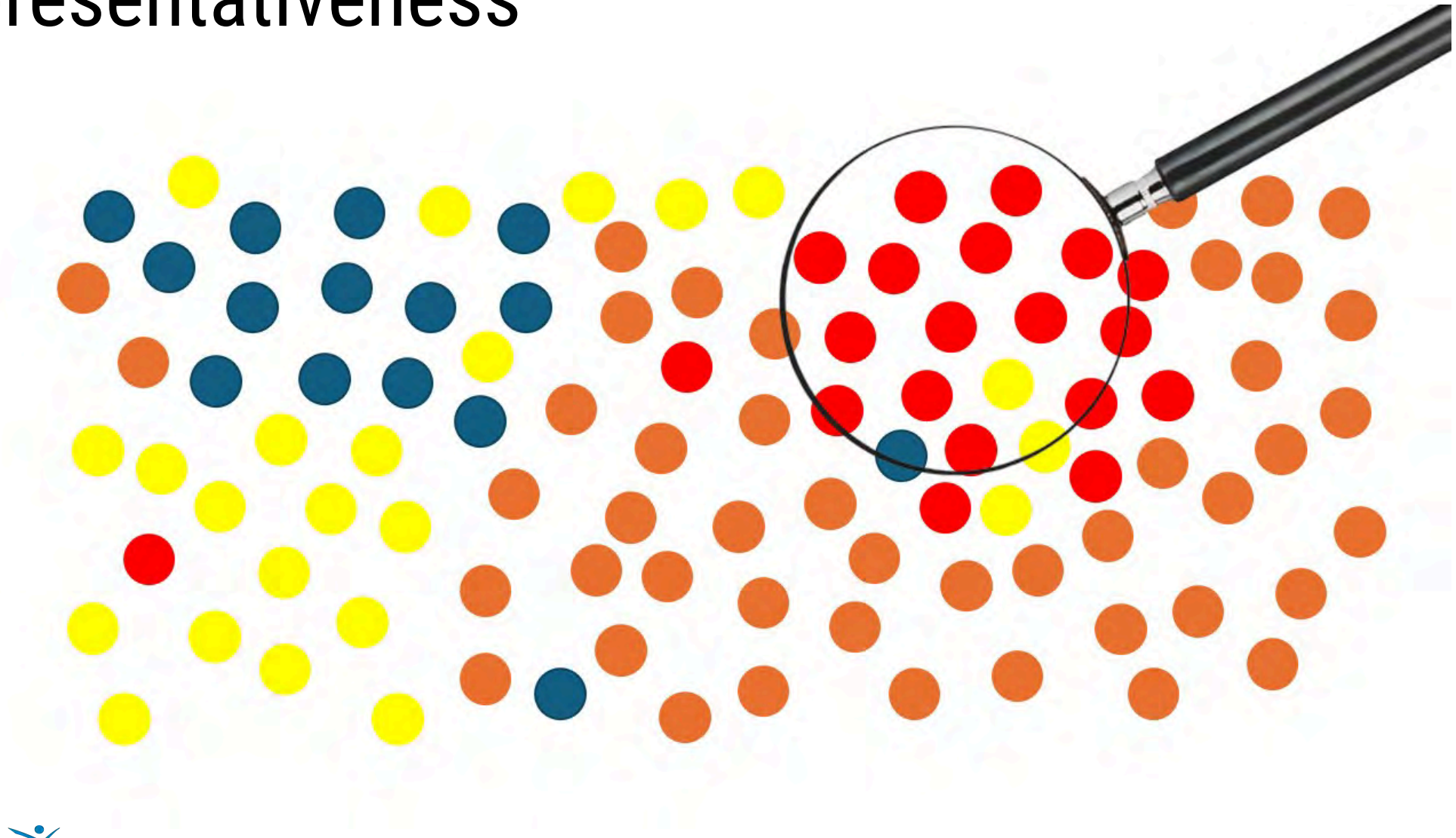


Rudolph, Zhong, Duggal, Mehta, and Lau (2023); Cornesse (2018)

Representativeness



Representativeness



Representativeness

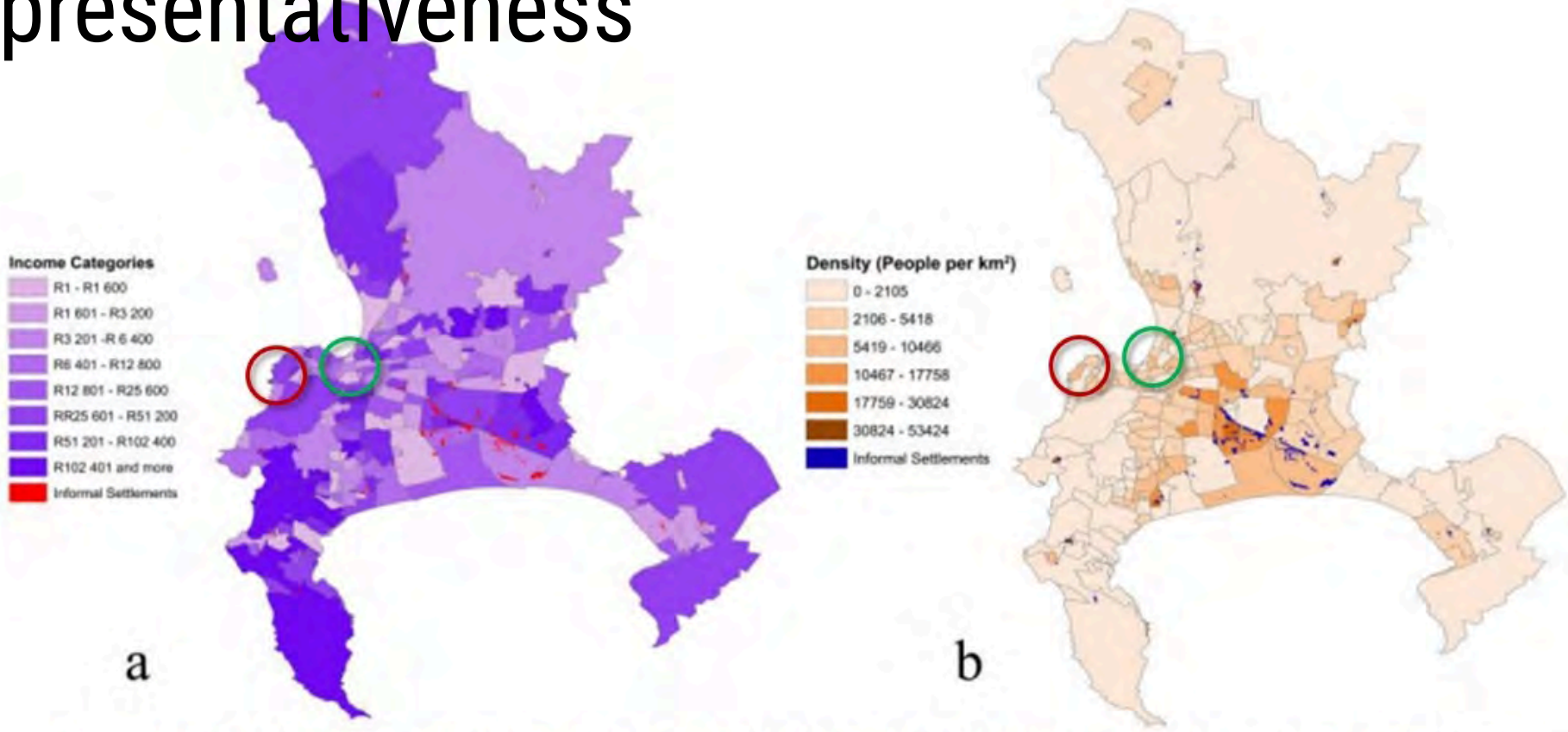
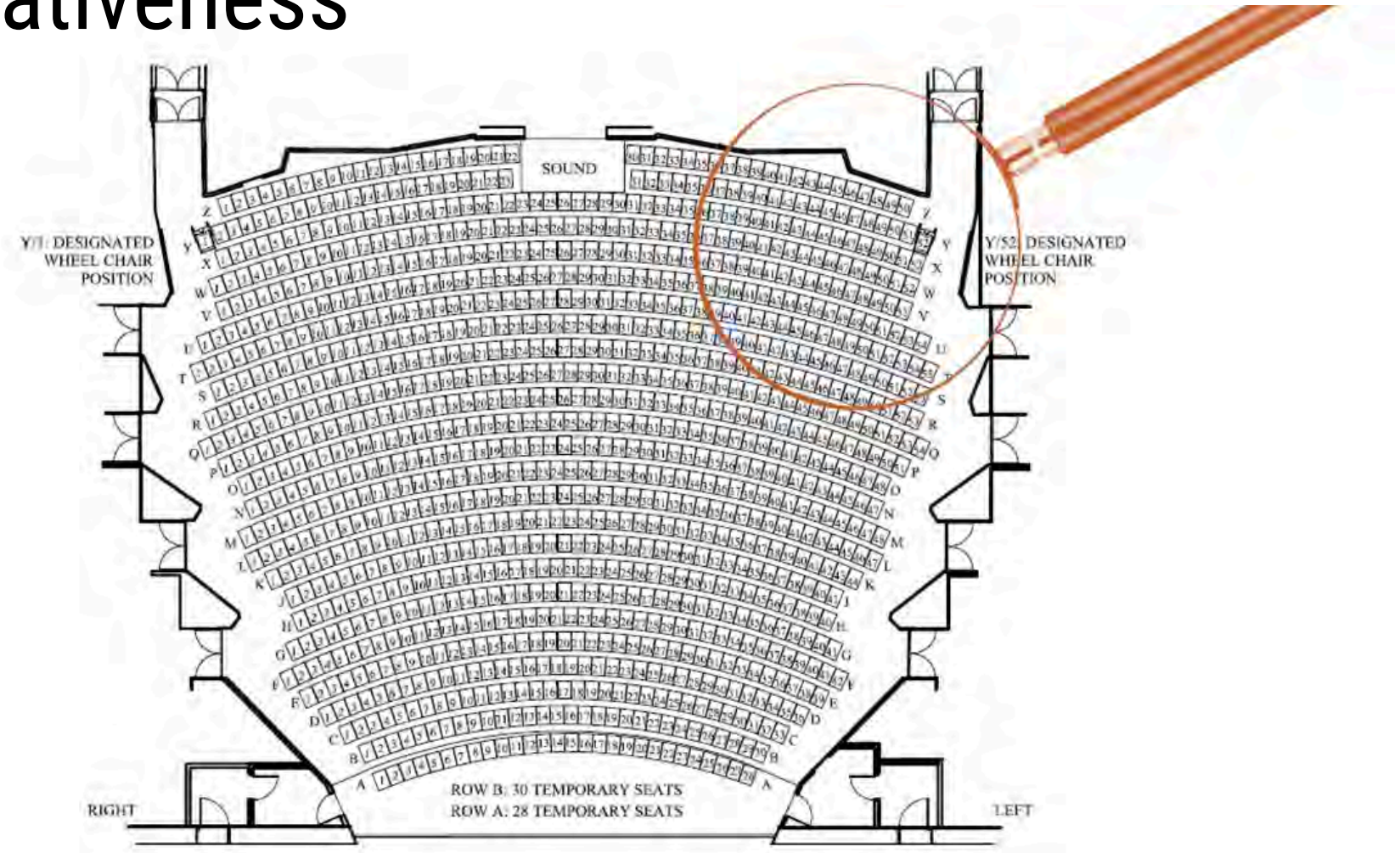


Figure 1: Cape Town suburbs delineated by monthly household income categories (a) and population density (b), overlaid with presence of informal settlements.

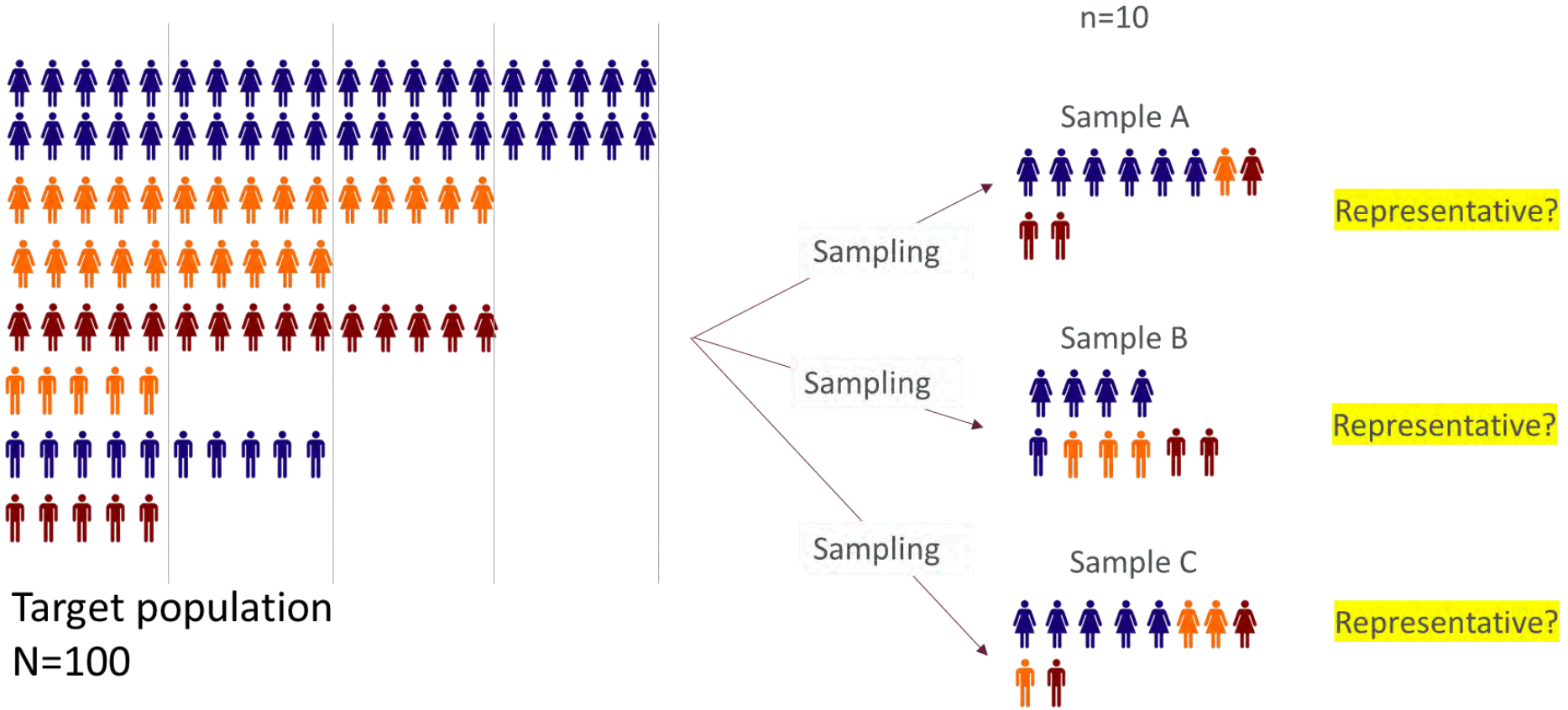
Data sourced from City of Cape Town (2016a)

Currie, Musango, and May (2017)

Representativeness

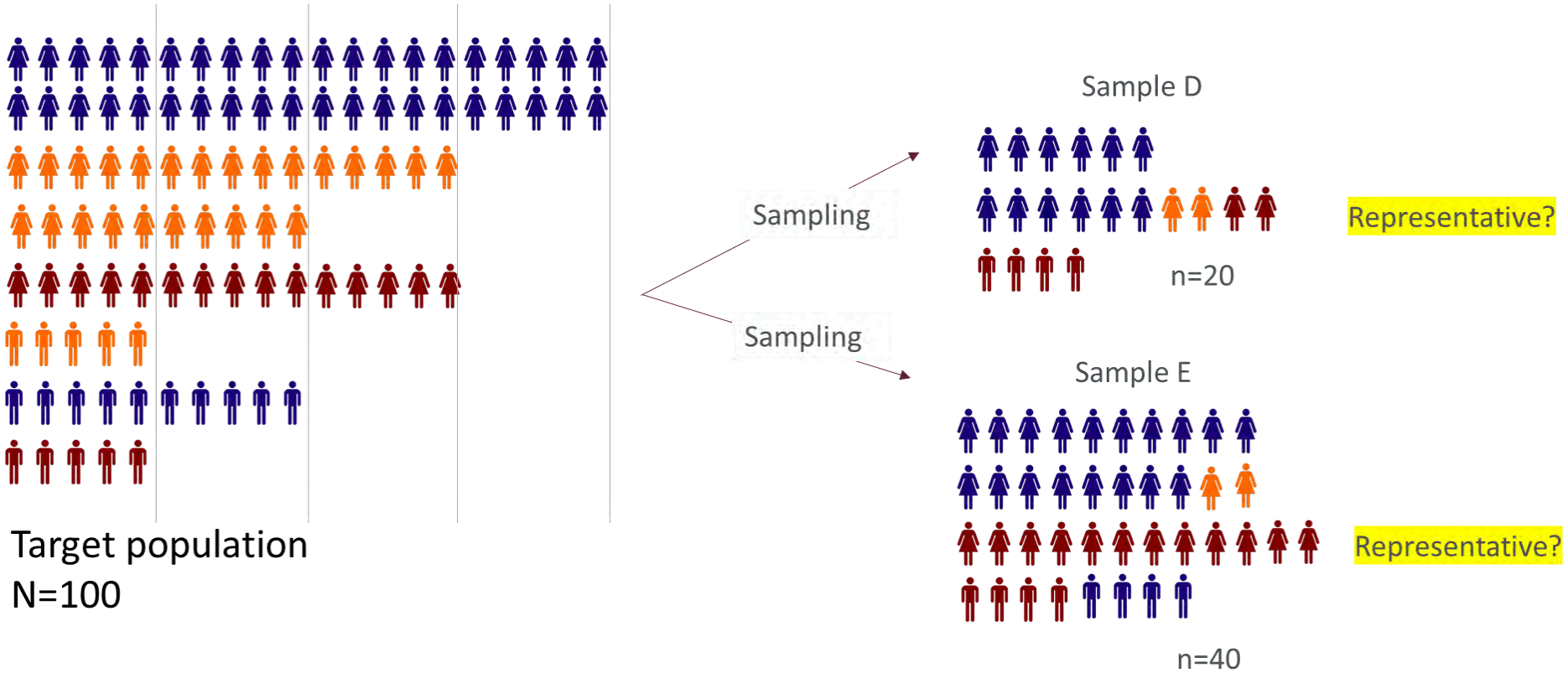


Representativeness



Characteristics of interest: Gender & hair colour

Representativeness



Characteristics of interest: Gender & hair colour

GENERALISABILITY

Sample size

% of population sampled

Statistical concept

Coverage

Sampling strategy

Research question

REPRESENTATIVENESS

Kamper (2020)

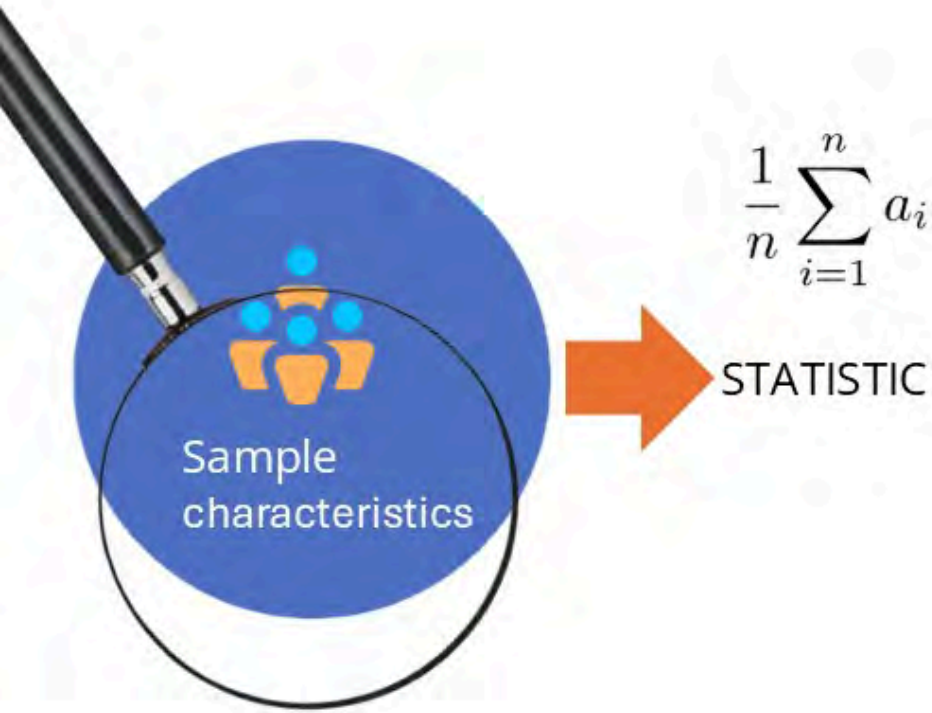
Estimation

What is the average systolic blood pressure in the population?

LibreTexts (2014)

Estimation

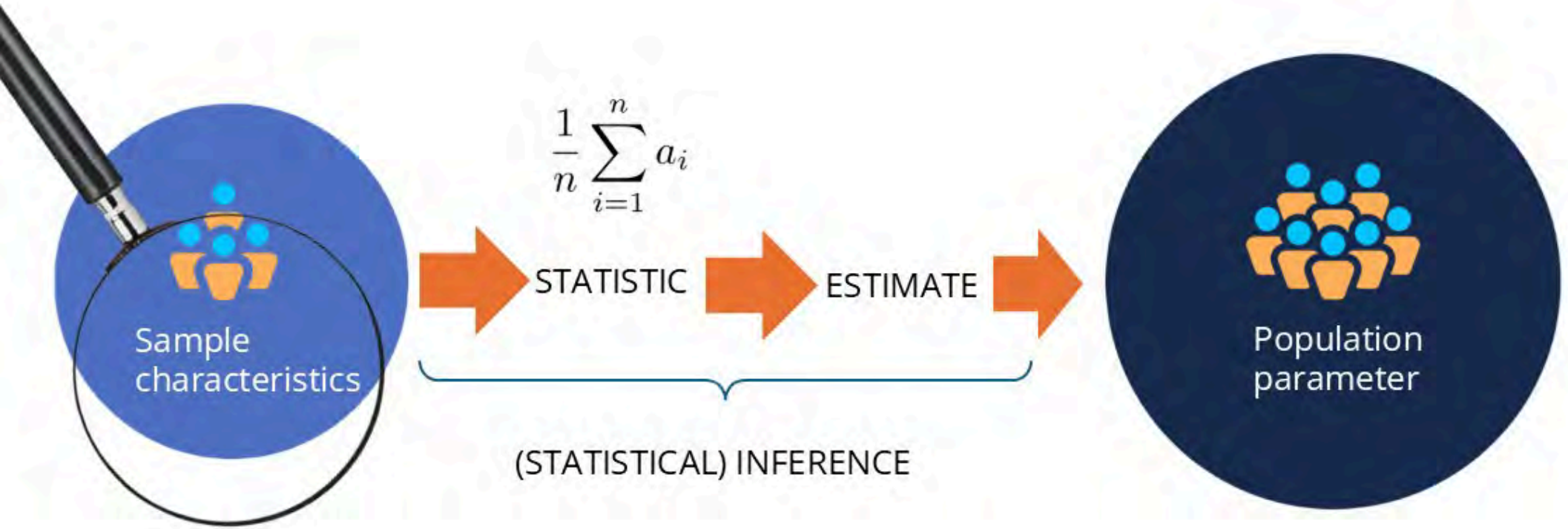
Estimation



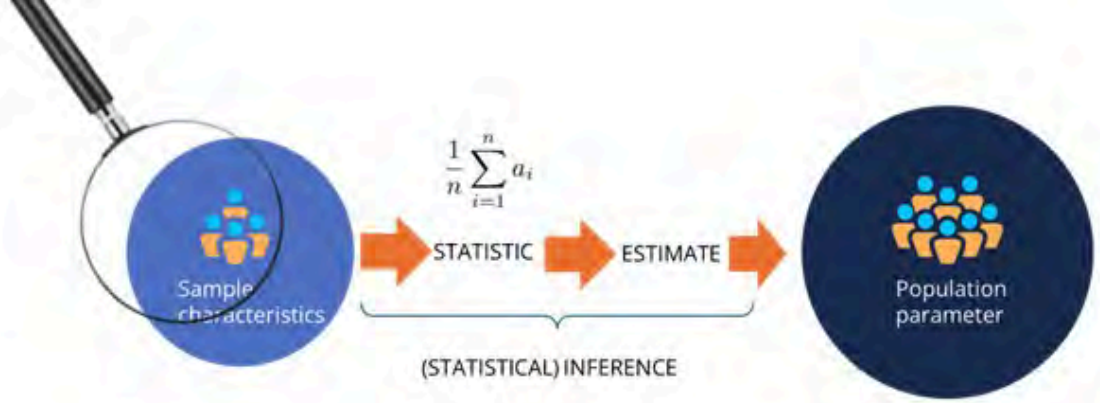
$$\frac{1}{n}\sum_{i=1}^{n} a_i$$

STATISTIC

Estimation


$\frac{1}{n}\sum_{i=1}^{n} a_i$
Sample characteristics
STATISTIC
ESTIMATE
Population parameter
(STATISTICAL) INFERENCE

Estimation



Respondent	SBP [mmHg]
H_100_437_1	133.3
H_101_138_4	85.9
H_104_54_6	87.4
H_106_200_4	81.1
H_106_263_5	77.1
H_15_216_2	64.5
H_18_293_1	162.6
H_2_149_2	47.9
H_2_156_4	85.1
H_2_75_4	116.2
H_24_192_1	120.1
H_24_499_6	94.1
H_27_149_2	139.4
H_29_42_1	95.2
H_32_143_2	86
H_32_199_6	125.9
H_37_16_1	87.4
H_37_344_1	81.1

ESTIMATE

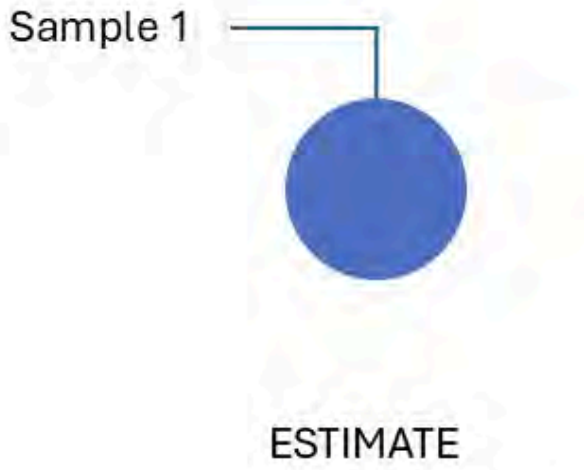
$$\overline{SBP}_{sample} = \frac{\sum_1^n SBP_i}{n} = 123 \text{ mmHg}$$

STATISTIC

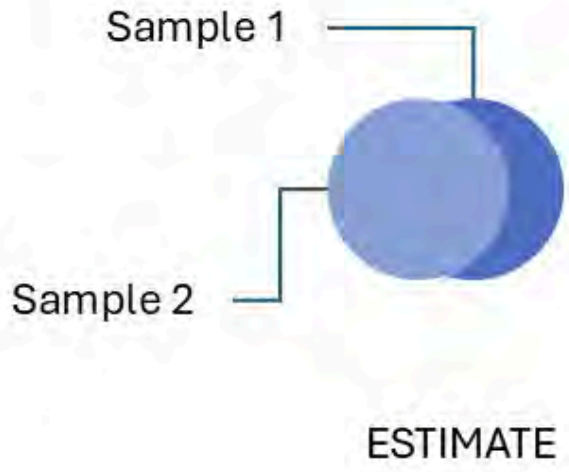
Our best estimate of the population mean systolic blood pressure in our target population is 123 mmHg

Estimation

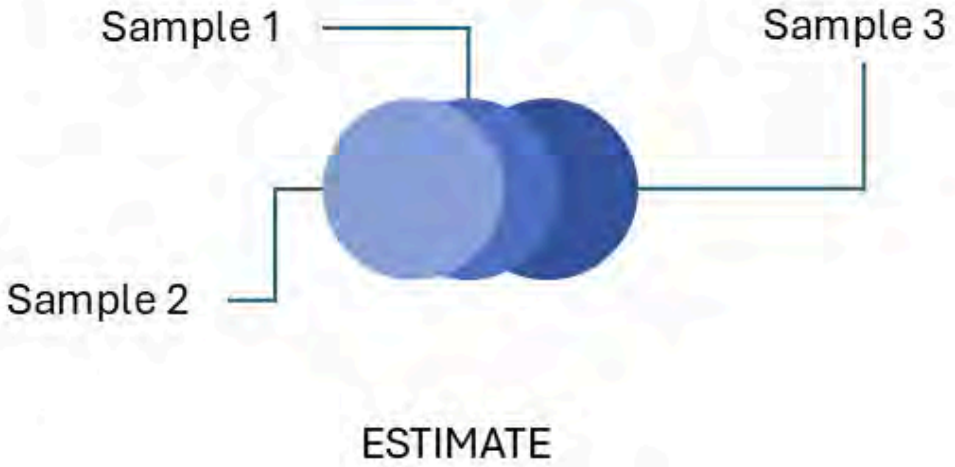
Estimation



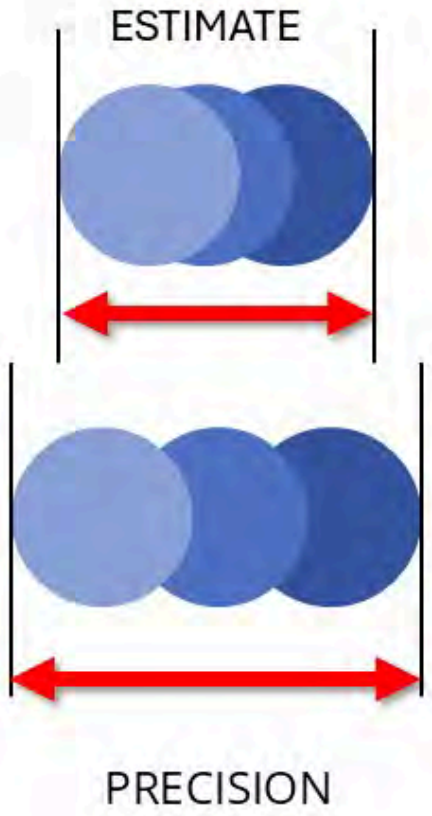
Estimation



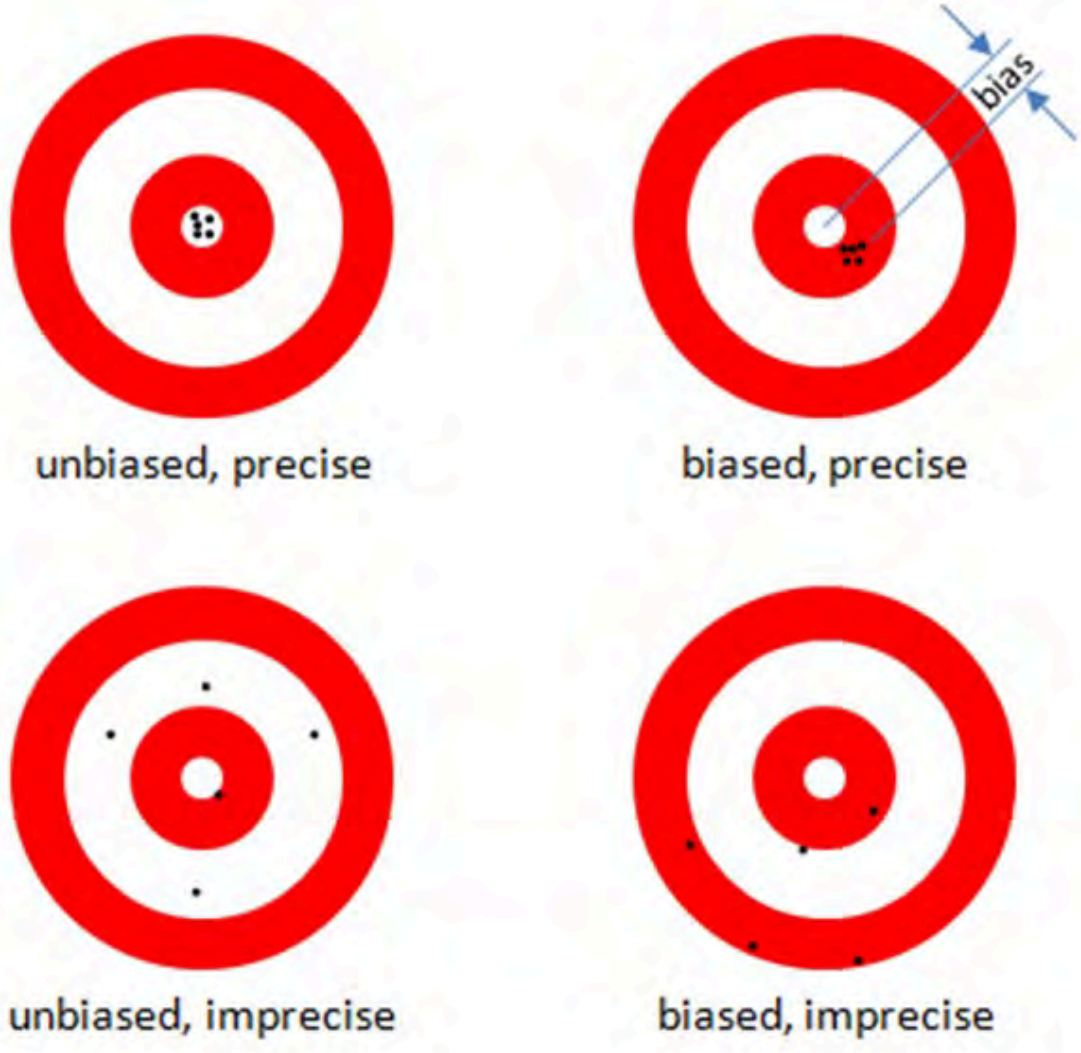
Estimation



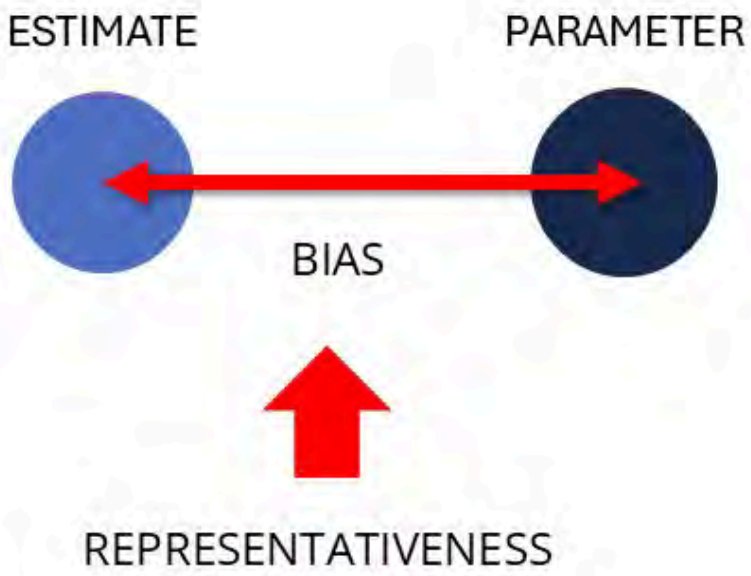
Estimation

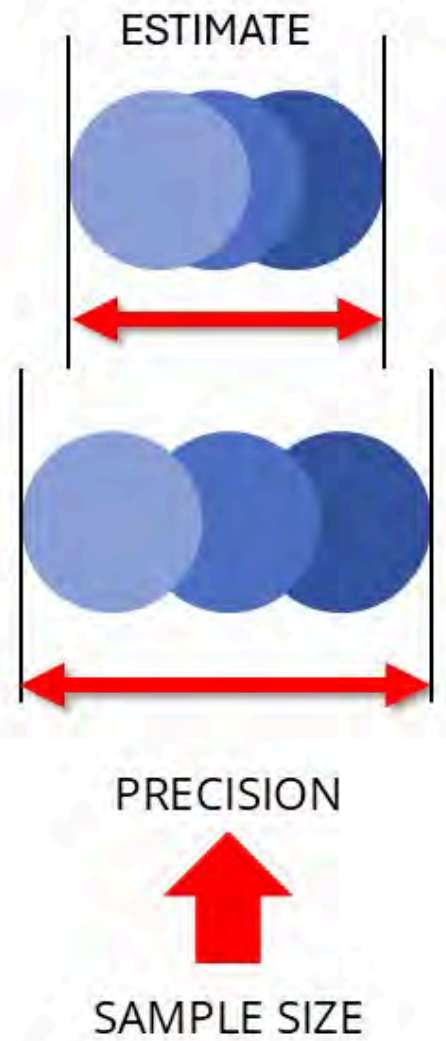


Estimation

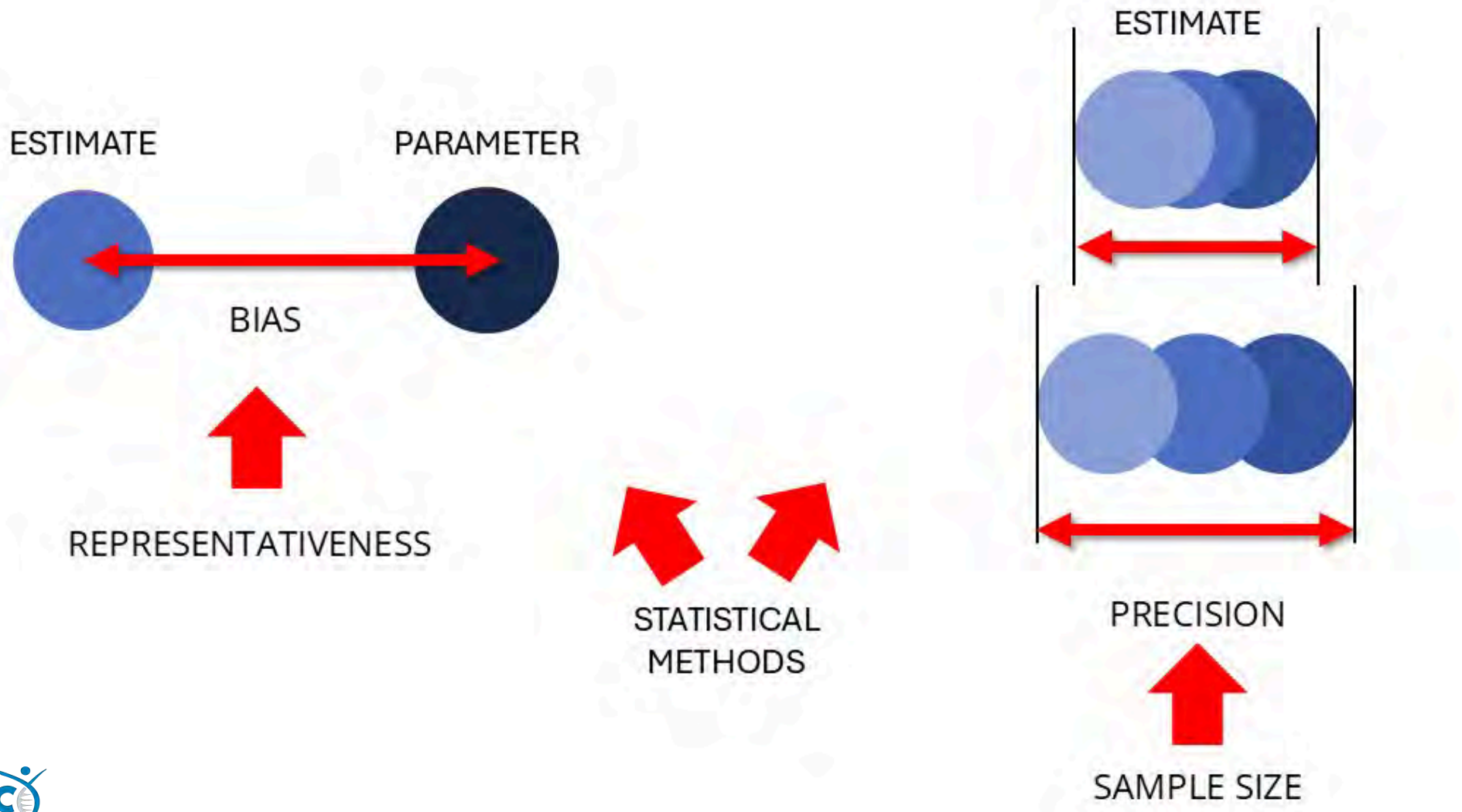

bias
unbiased, precise
biased, precise
unbiased, imprecise
biased, imprecise

Estimation

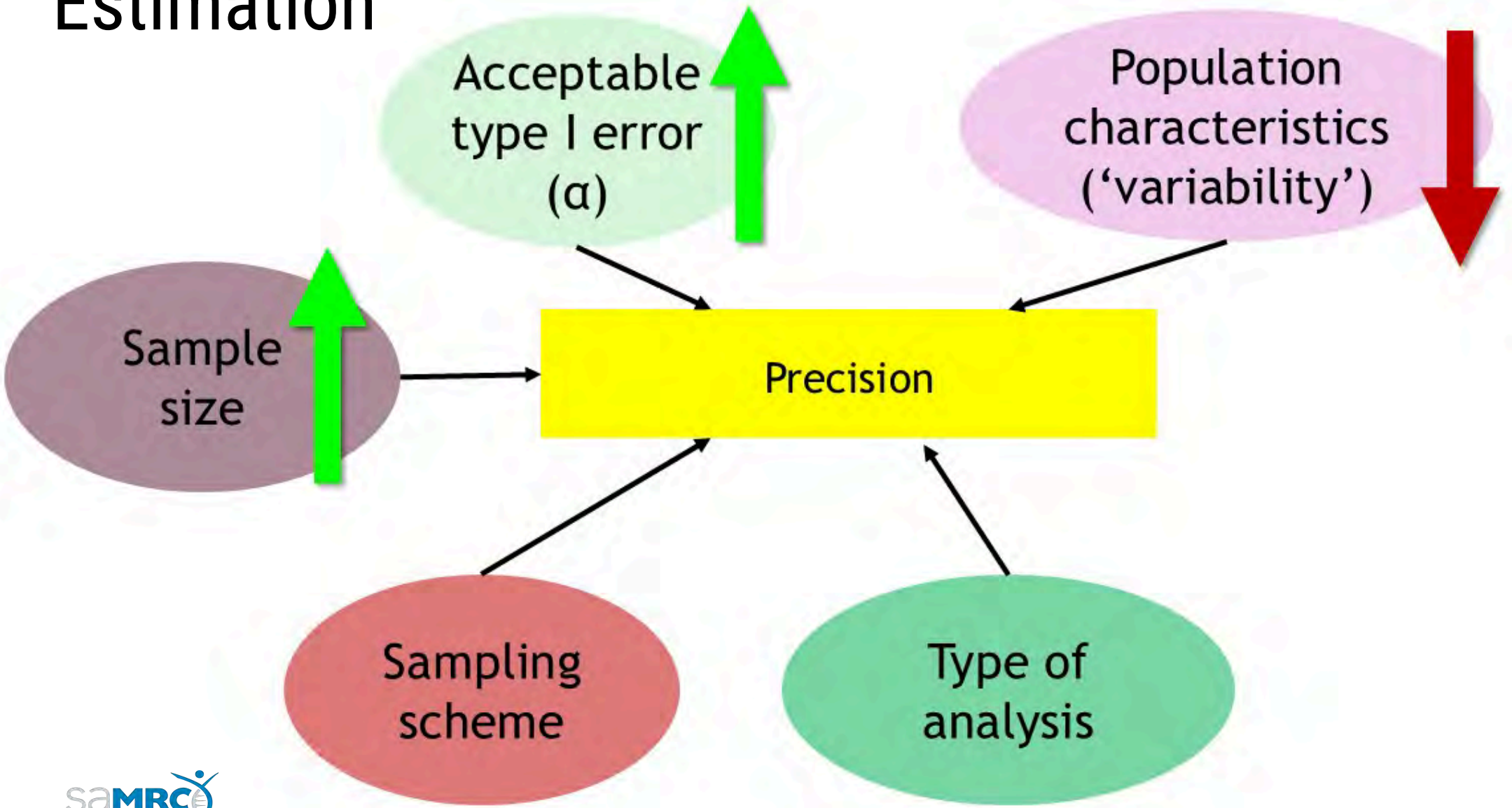

ESTIMATE
PARAMETER
BIAS
REPRESENTATIVENESS


ESTIMATE
PRECISION
SAMPLE SIZE

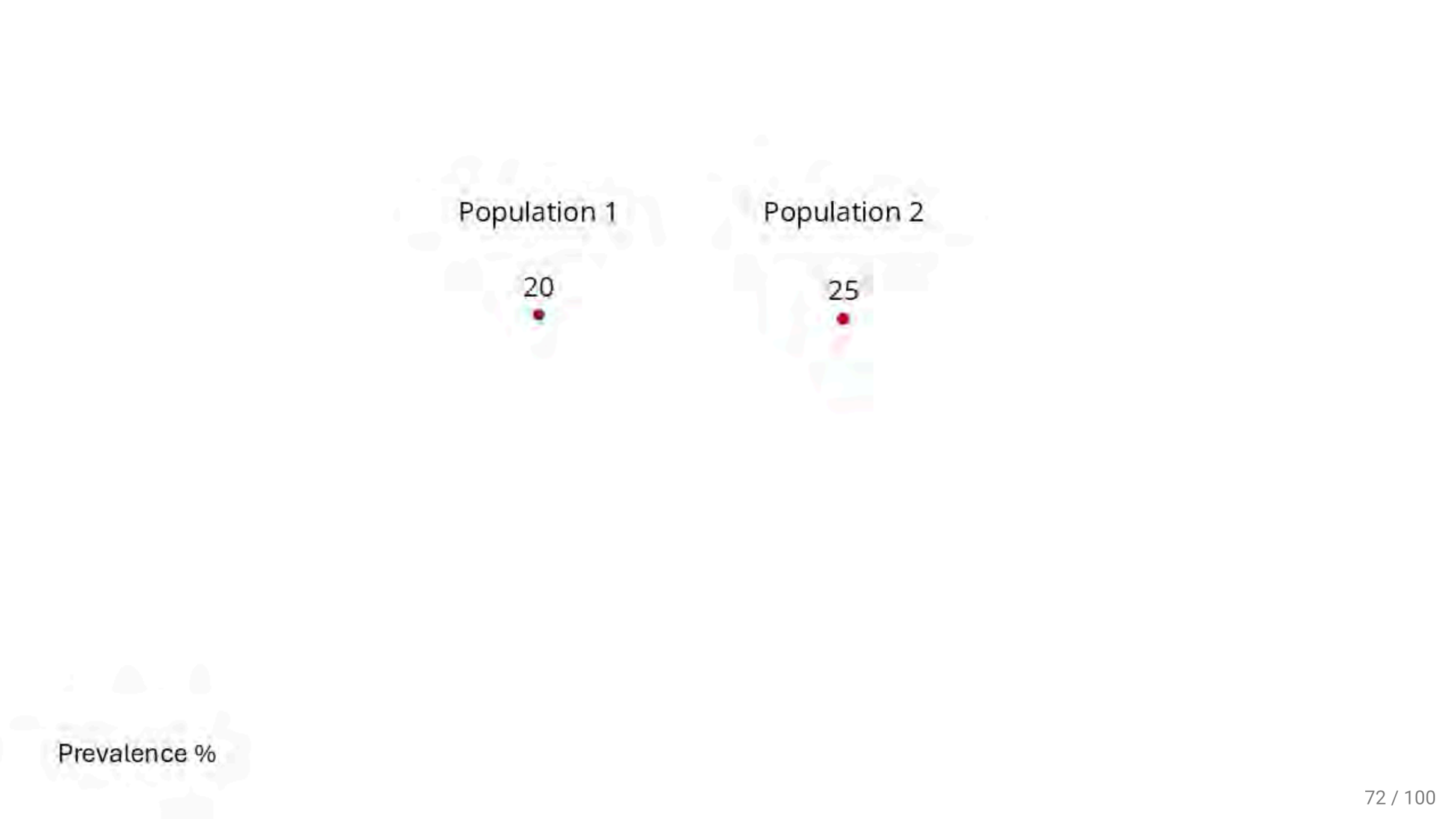
Estimation


ESTIMATE
PARAMETER
BIAS
REPRESENTATIVENESS
STATISTICAL METHODS
ESTIMATE
PRECISION
SAMPLE SIZE

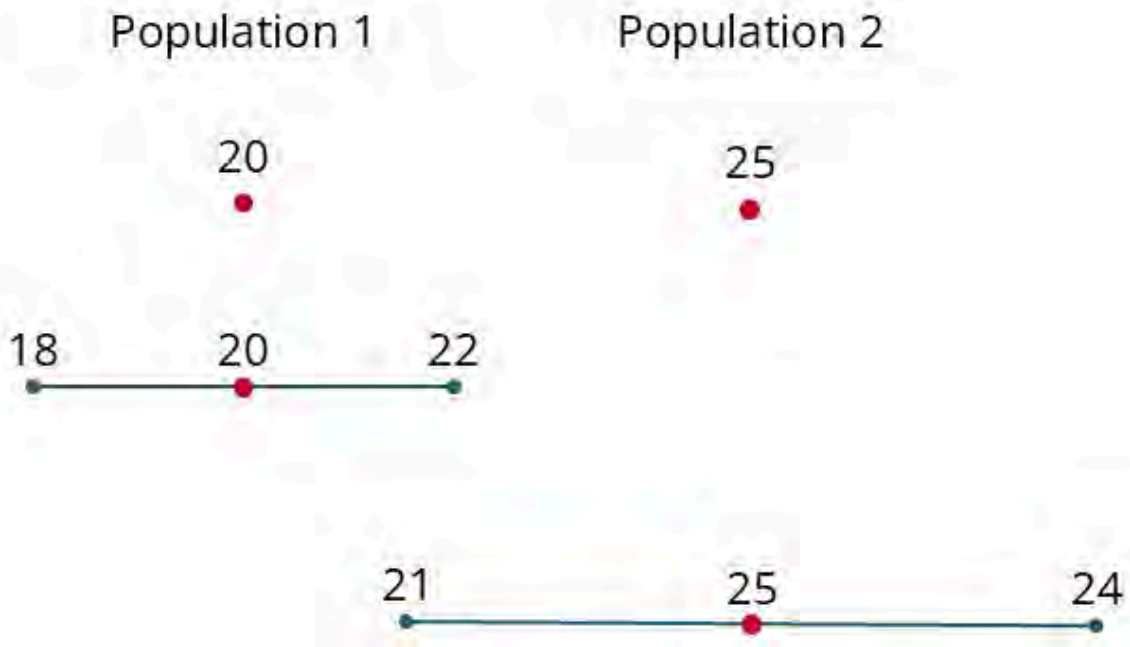
Estimation


Acceptable type I error (α)
Population characteristics ('variability')
Sample size
Precision
Sampling scheme
Type of analysis

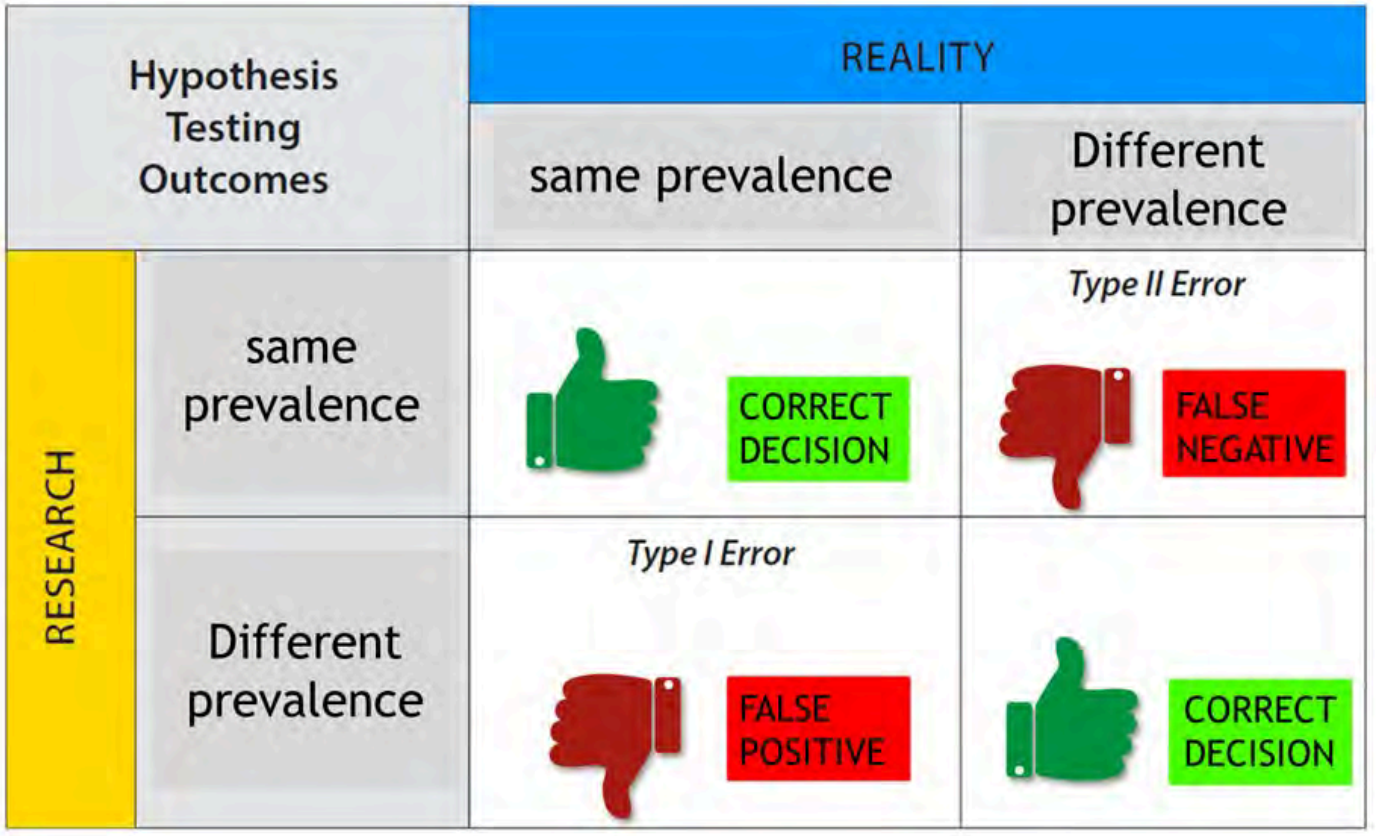
Does the prevalence of diabetes in Population 1 differ from the prevalence in Population 2?


Population 1
Population 2
20
25
Prevalence %

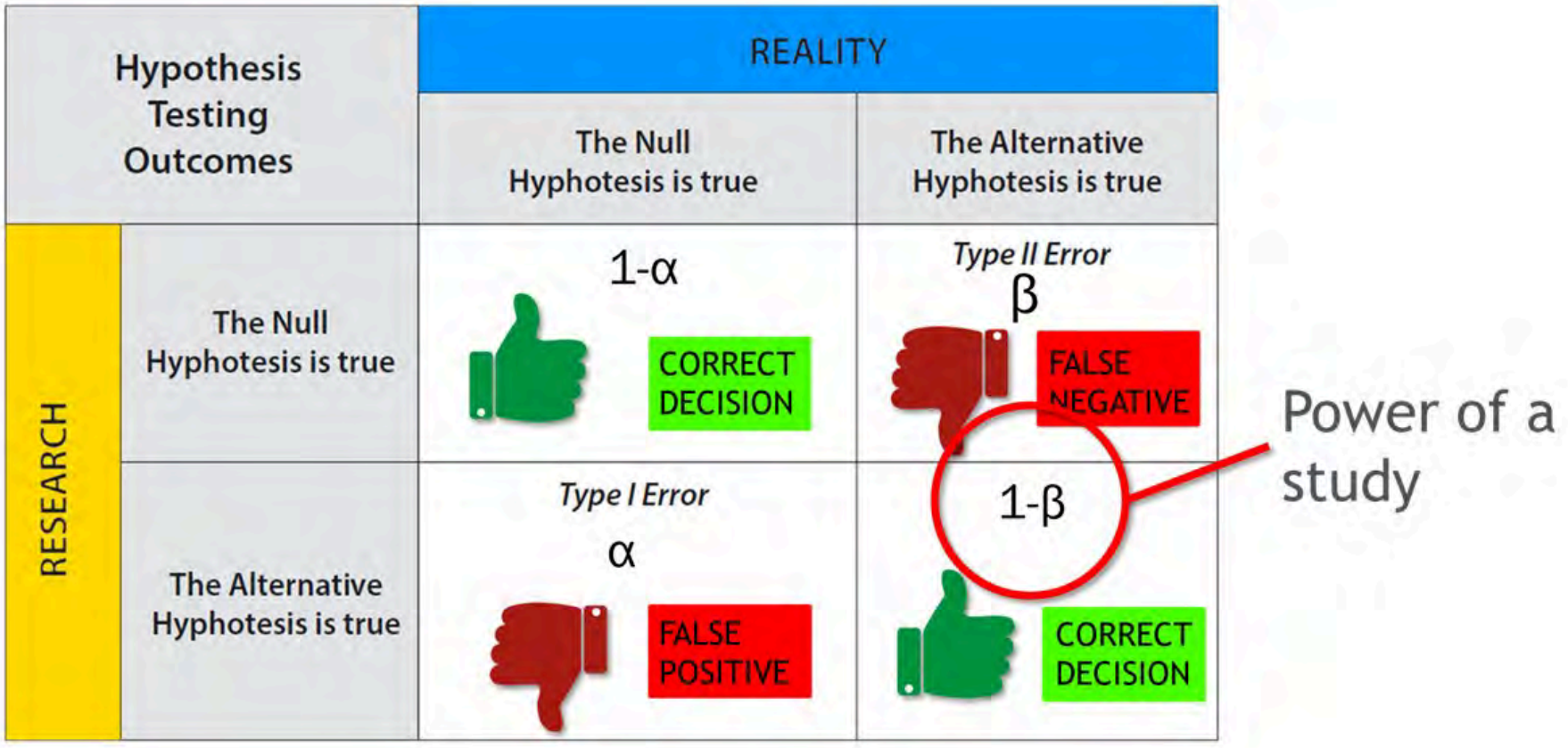
Prevalence %


Population 1
Population 2
20
25
18
20
22
21
25
24
Prevalence %

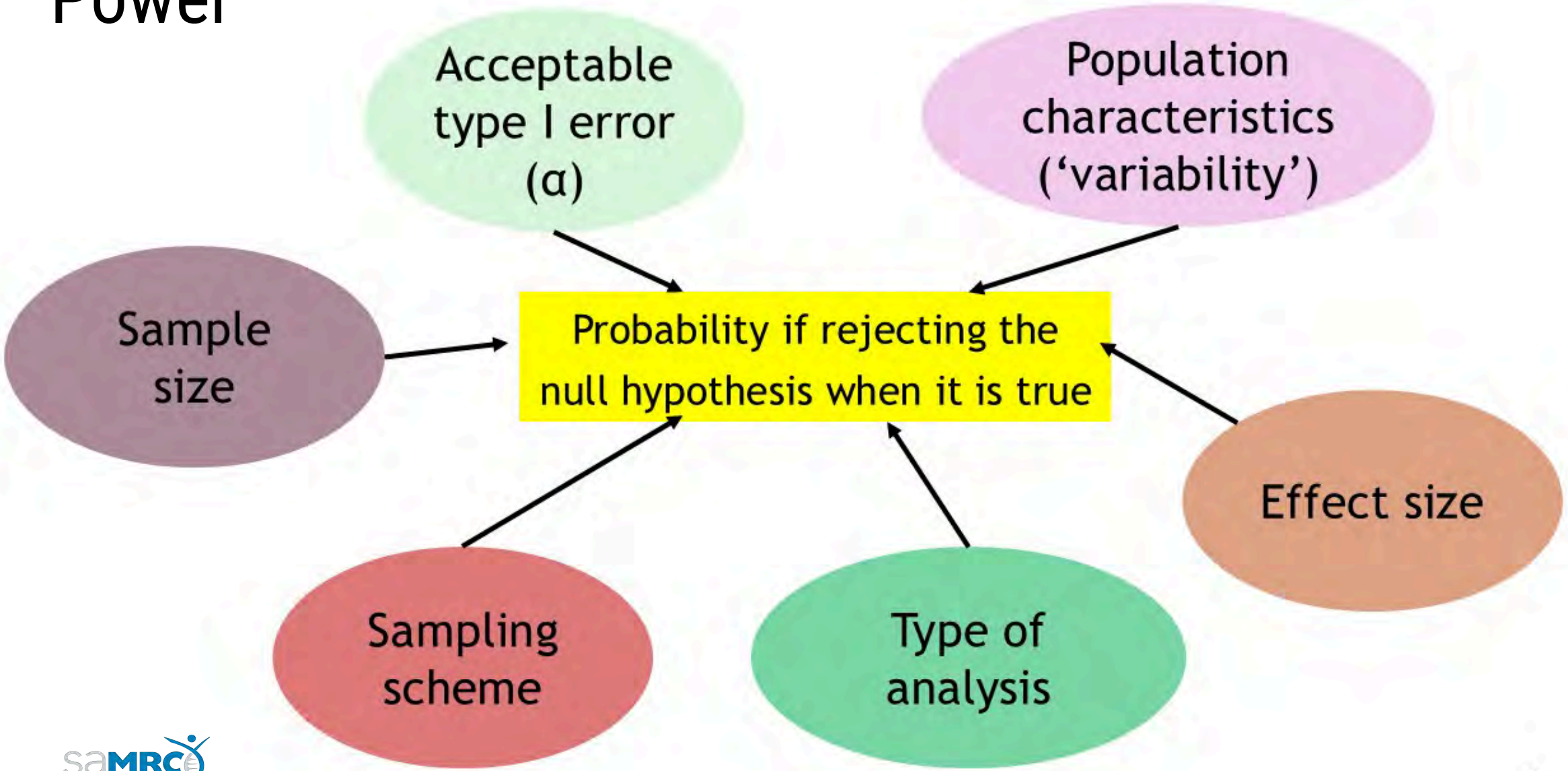
Power of a study



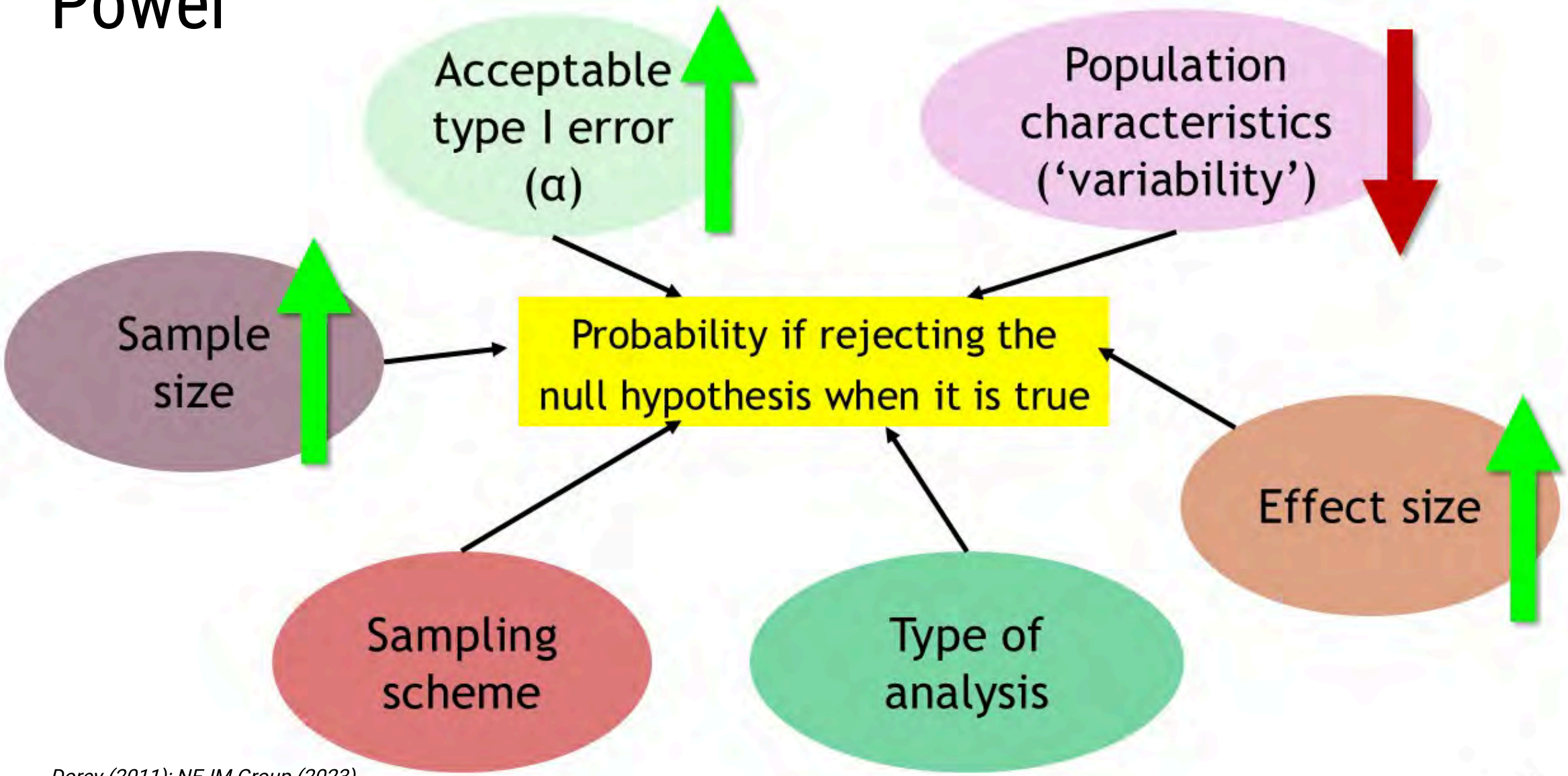
Statistically speaking



Power


Acceptable type I error (α)
Population characteristics ('variability')
Sample size
Probability if rejecting the null hypothesis when it is true
Effect size
Sampling scheme
Type of analysis

Power



Dorey (2011); NEJM Group (2023)

...a needle in a haystack

...a needle in a haystack

...a needle in a haystack

Sample size

...and its relationship with precision and power

Power calculations

Standard studies
Advanced Statistical Training

Standard studies
No Statistical Training

Complex studies
Basic/no Statistical Training

Standard studies
Basic Statistical Training

https://clincalc.com/stats/samplesize.aspx

ASK a statistician

Power calculations: an example

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

N - sample size

P - prevalence or proportion of event

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

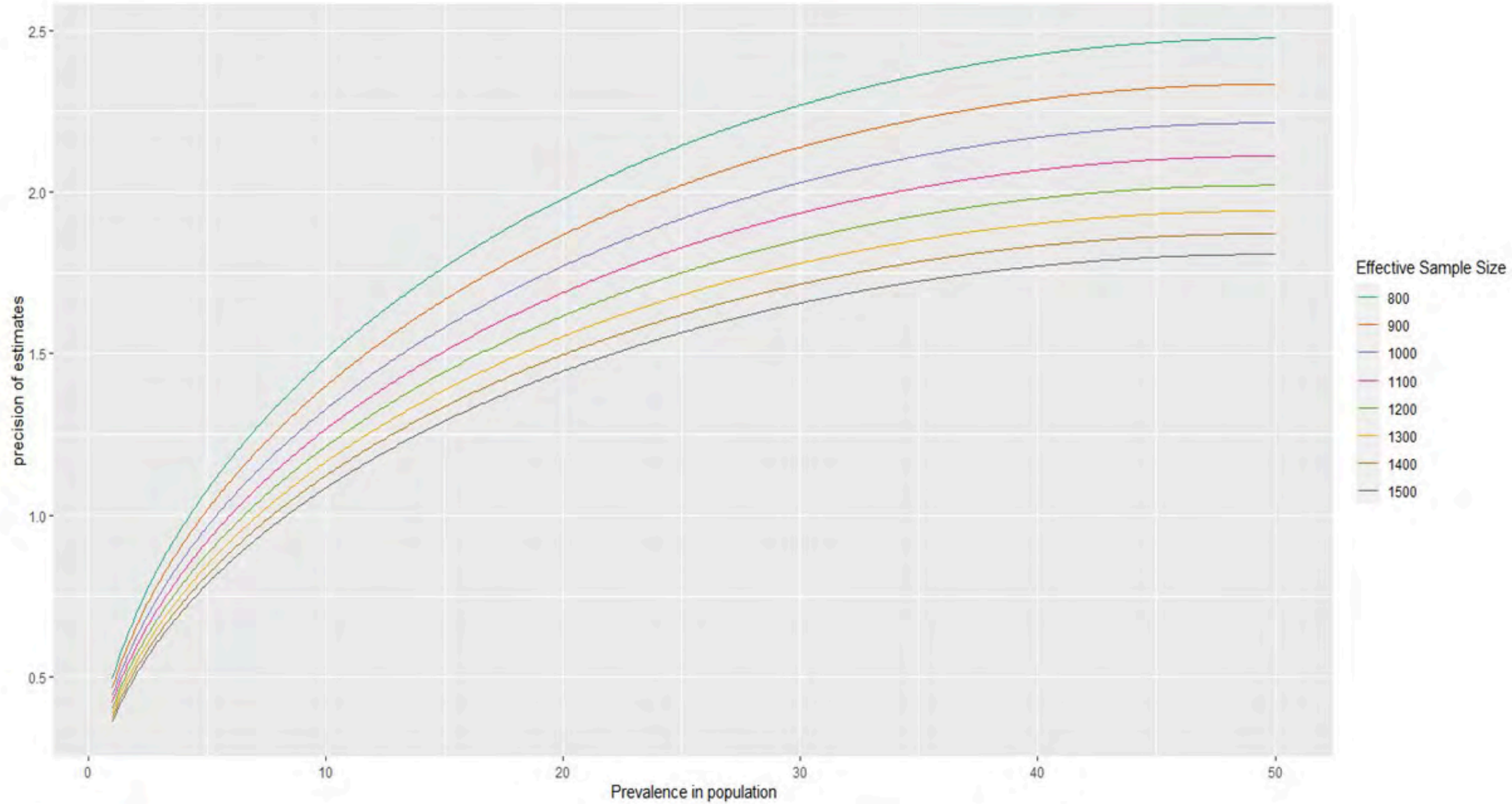
D - design effect reflects the sampling design used in the survey type of study. This is 1 for simple random sampling and higher values (usually 1 to 2) for other designs such as stratified, systematic, cluster random sampling

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

Population characteristics and power

	Margin of error (ME) (for CI 95%)	
Population Size	**10%**	**5%**
100	50	80
500	81	218
1000	88	278
10,000	96	370
100,000	96	383
1.000,000	97	384

(Adapted form Serdar et al. 2021:22)



Serdar, Cihan, Yücel, and Serdar (2021); Kadam and Bhalerao (2010); Naing, Nordin, Abdul Rahman, and Naing (2022)

Population characteristics?

The SurveyLab

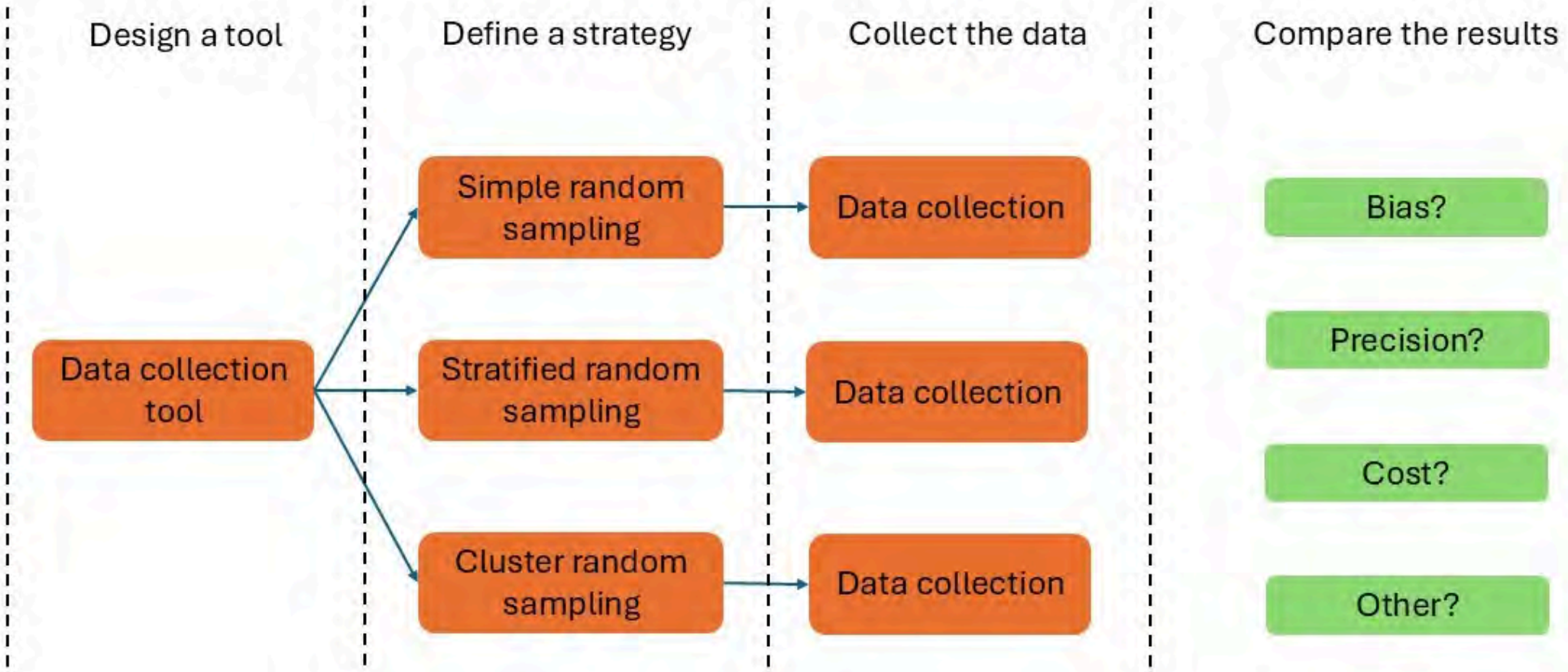
Navigating the virtual environment

https://bit.ly/SurveyLab_workshop

Conducting a survey

Sampling strategies, tools and data collection

What is the prevalence of COPD in our population?


Design a tool
Define a strategy
Collect the data
Compare the results
Data collection tool
Simple random sampling
Stratified random sampling
Cluster random sampling
Data collection
Data collection
Data collection
Bias?
Precision?
Cost?
Other?

Sample size

$$n = \frac{z^2_{\alpha/2} * p(1-p)}{e^2}$$

$$e = precision$$

Point estimate

$$est = p = \frac{N_{yes}}{N_{yes} + N_{no}}$$

95% CI

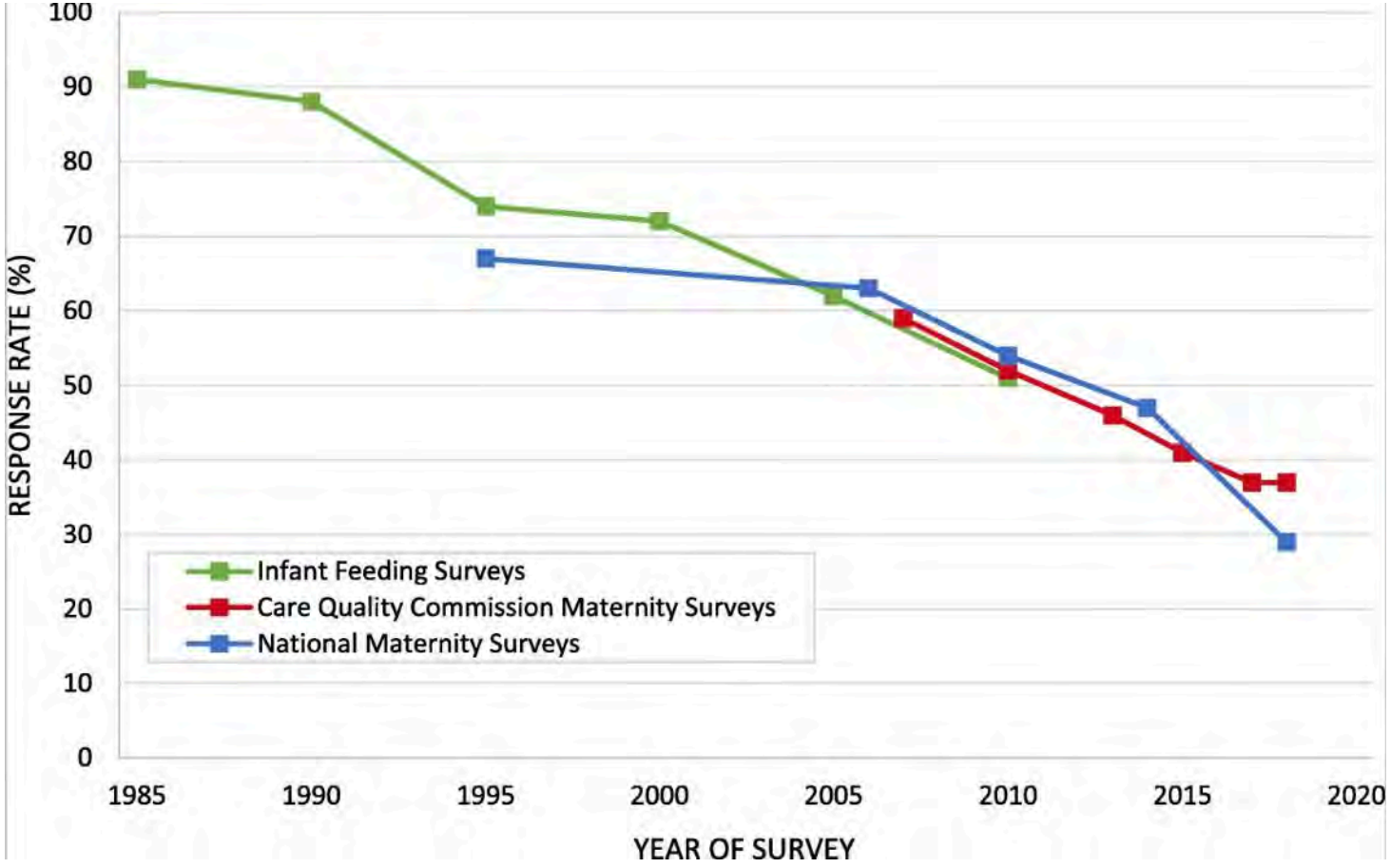
$$ci = p \pm z_{\alpha/2} se$$

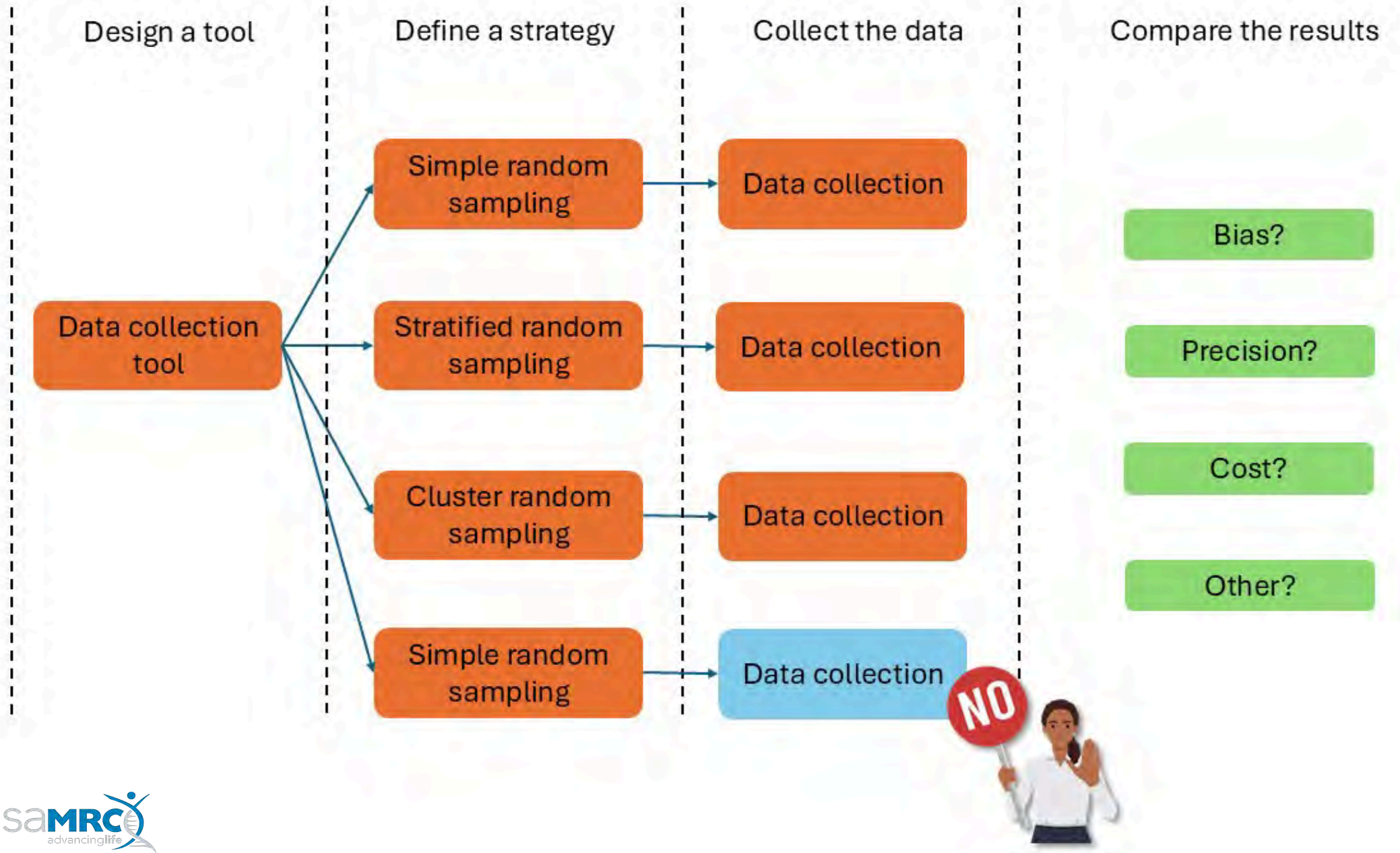
$$se = \sqrt{\frac{p(1-p)}{n}}$$

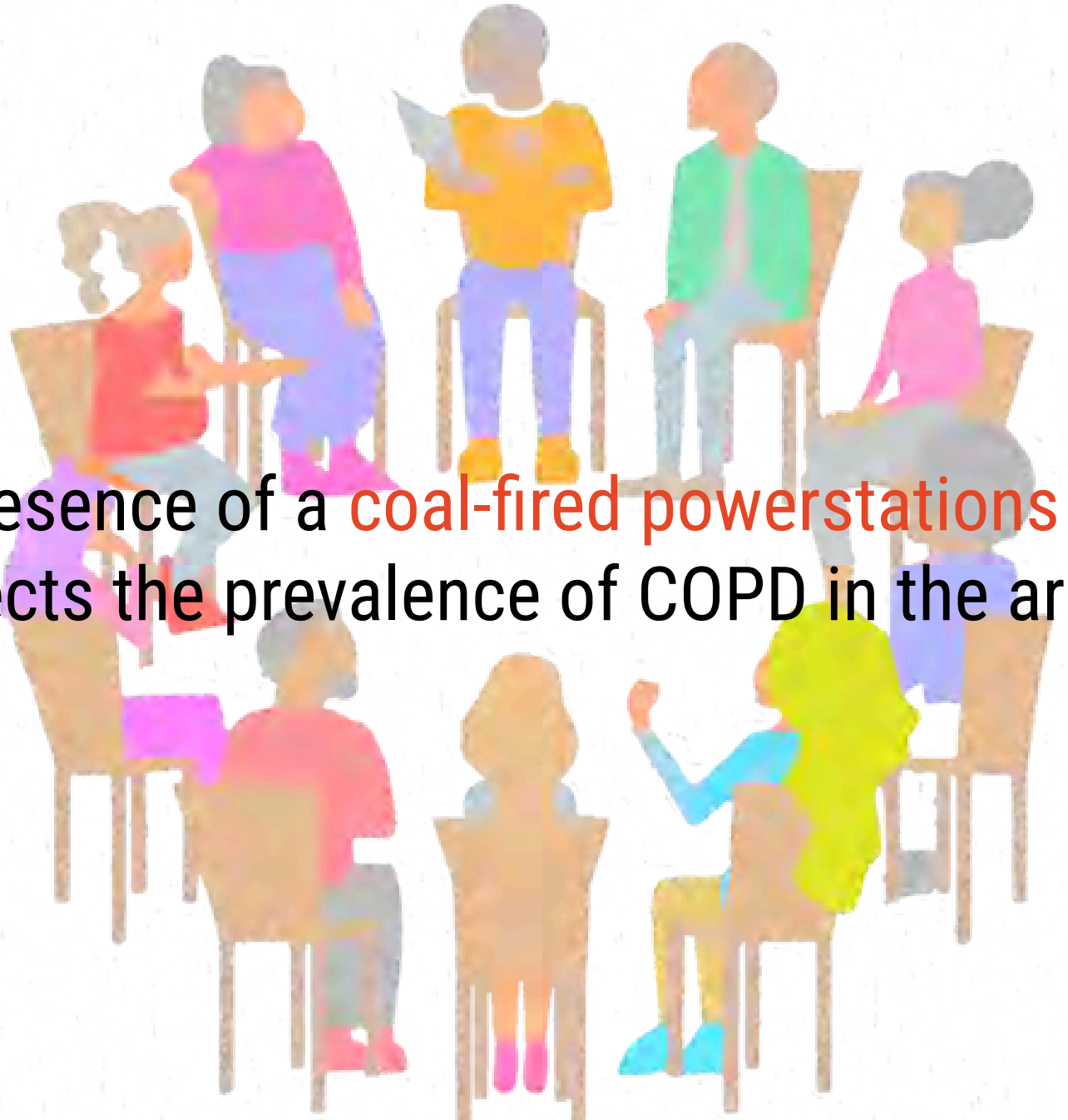
$$z_{\alpha/2} = critical\ z - score\ for\ \alpha/2$$

(in R: *qnorm(0.05/2, mean = 0, sd = 1)*)

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

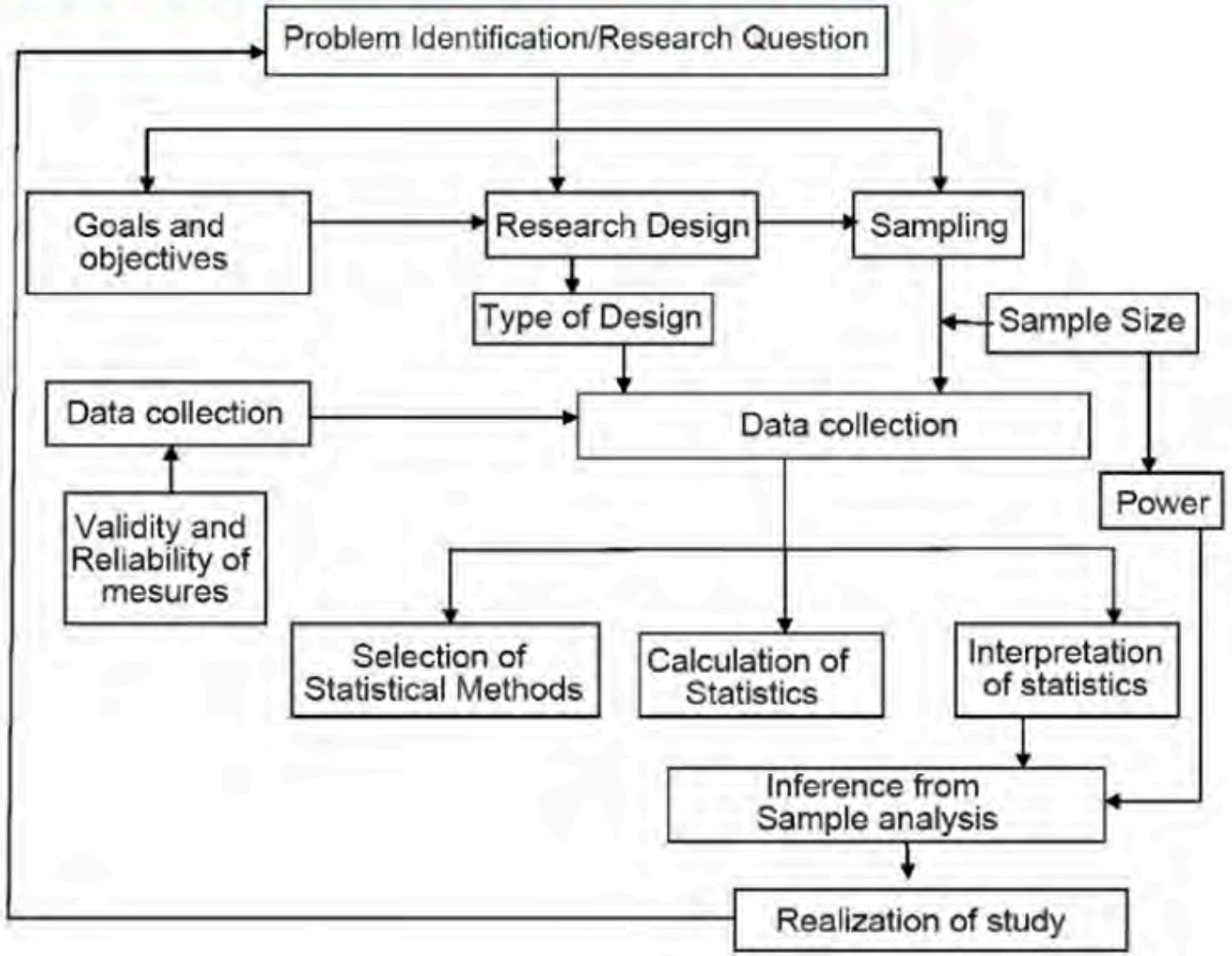



Design a tool
Define a strategy
Collect the data
Compare the results
Data collection tool
Simple random sampling
Stratified random sampling
Cluster random sampling
Simple random sampling
Data collection
Data collection
Data collection
Data collection
Bias?
Precision?
Cost?
Other?
NO



Does the presence of a coal-fired powerstations in the town affects the prevalence of COPD in the area?

Research design framework



Suresh, Thomas, and Suresh (2011b)

Discussion & recap

References

Banerjee, A. and S. Chaudhury (2010). "Statistics without Tears: Populations and Samples". In: *Industrial Psychiatry Journal* 19.1, pp. 60-65. ISSN: 0972-6748. DOI: 10.4103/0972-6748.77642. (Visited on Sept. 22, 2025).

Cornesse, C. (2018). "Representativeness and Response Quality of Survey Data". Mannheim.

Currie, P. K., J. K. Musango, and N. D. May (2017). "Urban Metabolism: A Review with Reference to Cape Town". In: *Cities* 70, pp. 91-110. ISSN: 0264-2751. DOI: 10.1016/j.cities.2017.06.005. (Visited on Sept. 17, 2025).

Dorey, F. J. (2011). "In Brief: Statistics in Brief: Statistical Power: What Is It and When Should It Be Used?" In: *Clinical Orthopaedics and Related Research* 469.2, pp. 619-620. ISSN: 0009-921X. DOI: 10.1007/s11999-010-1435-0. (Visited on Sept. 22, 2025).

Groves, R. M., F. J. Fowler Jr, M. P. Couper, et al. (2009). *Survey Methodology*. John Wiley & Sons. ISBN: 1-118-21134-0.

Groves, R. M. and L. Lyberg (2010). "Total Survey Error: Past, Present, and Future". In: *Public Opinion Quarterly* 74.5, pp. 849-879. ISSN: 0033-362X. DOI: 10.1093/poq/nfq065. (Visited on Jun. 06, 2022).

Kadam, P. and S. Bhalerao (2010). "Sample Size Calculation". In: *International Journal of Ayurveda Research* 1.1, pp. 55-57. ISSN: 0974-7788. DOI: 10.4103/0974-7788.59946. (Visited on Sept. 22, 2025).

Kamper, S. J. (2020). "Generalizability: Linking Evidence to Practice". In: *Journal of Orthopaedic & Sports Physical Therapy* 50.1, pp. 45-46. ISSN: 0190-6011. DOI: 10.2519/jospt.2020.0701. (Visited on Sept. 22, 2025).

Lee, S. (2008). "Population of Inference". In: *Encyclopedia of Survey Research Methods*. Sage Publications, Inc., p. 592.

LibreTexts (2014). *7: Estimation*. https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Introductory_Statistics_(Shafer_and_Zhang)/07 (Visited on Sept. 22, 2025).

Naing, L., R. B. Nordin, H. Abdul Rahman, et al. (2022). "Sample Size Calculation for Prevalence Studies Using Scalex and ScalaR Calculators". In: *BMC Medical Research Methodology* 22.1, p. 209. ISSN: 1471-2288. DOI: 10.1186/s12874-022-01694-7. (Visited on Sept. 04, 2024).

NEJM Group (2023). *How Statistical Power Works \textbar NEJM Evidence*. (Visited on Sept. 22, 2025).

OXFAM (2024). *Understanding Survey Sampling*. https://policy-practice.oxfam.org/resources/understanding-survey-sampling-615904/. (Visited on Sept. 04, 2024).

Raymond, J. and T. E. Darsaut (2025). "Understanding Statistical Populations and Inferences". In: *Neurochirurgie* 71.1, p. 101608. ISSN: 0028-3770. DOI: 10.1016/j.neuchi.2024.101608. (Visited on Sept. 15, 2025).

Rudolph, J. E., Y. Zhong, P. Duggal, et al. (2023). "Defining Representativeness of Study Samples in Medical and Population Health Research". In: *BMJ Medicine* 2.1, p. e000399. ISSN: 2754-0413. DOI: 10.1136/bmjmed-2022-000399. (Visited on Sept. 15, 2025).

Serdar, C. C., M. Cihan, D. Yücel, et al. (2021). "Sample Size, Power and Effect Size Revisited: Simplified and Practical Approaches in Pre-Clinical, Clinical and Laboratory Studies". In: *Biochemia Medica* 31.1, p. 010502. ISSN: 1330-0962. DOI: 10.11613/BM.2021.010502. (Visited on Sept. 22, 2025).

Suresh, K., S. V. Thomas, and G. Suresh (2011b). "Design, Data Analysis and Sampling Techniques for Clinical Research". In: *Annals of Indian Academy of Neurology* 14.4, pp. 287-290. ISSN: 0972-2327.

Tyrer, S. and B. Heyman (2016). "Sampling in Epidemiological Research: Issues, Hazards and Pitfalls". In: *BJPsych Bulletin* 40.2, pp. 57-60. ISSN: 2056-4694, 2056-4708. DOI: 10.1192/pb.bp.114.050203. (Visited on Sept. 04, 2024).

Thank You!