MRC Centre for Global Infectious Disease Analysis
IMPERIAL

Malaria Molecular Surveillance Study Design Workshop

Module 1: Sampling from a population

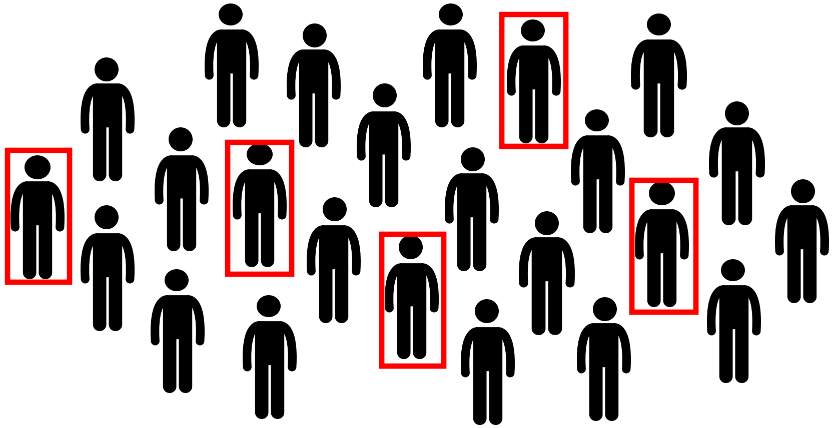
Prevalence estimation

We often want to estimate prevalence of:

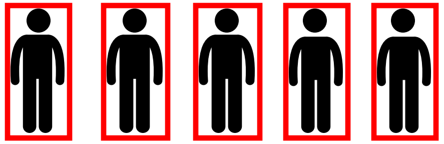
- Drug resistance mutations
- Diagnostic resistance mutations
- etc…

We can’t test everyone, we need to ***sample*** the population

What is a “population”?



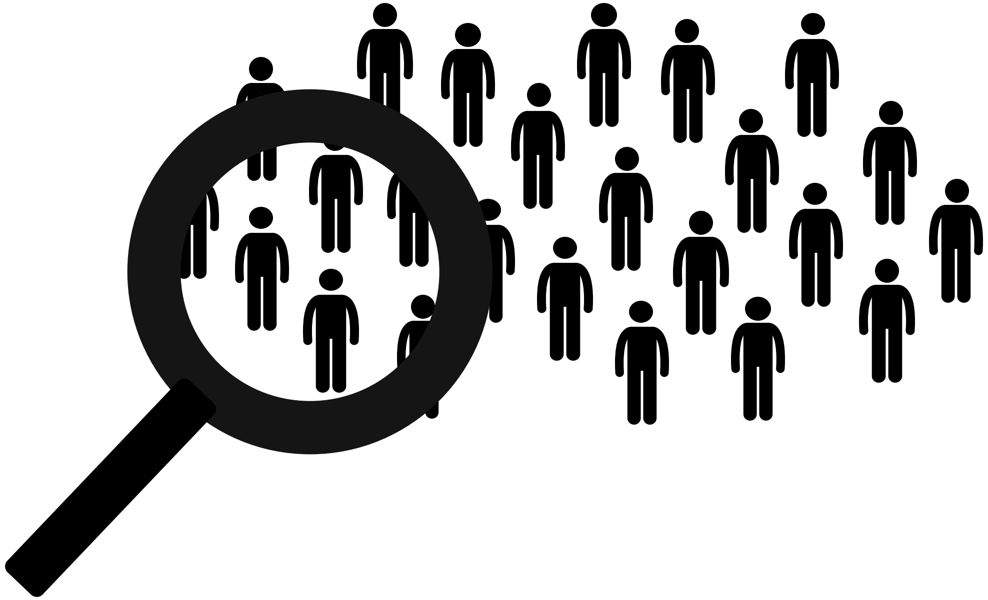
Population
The complete pool of people we are interested in



Sample
A subset of people that we use to draw conclusions about the population

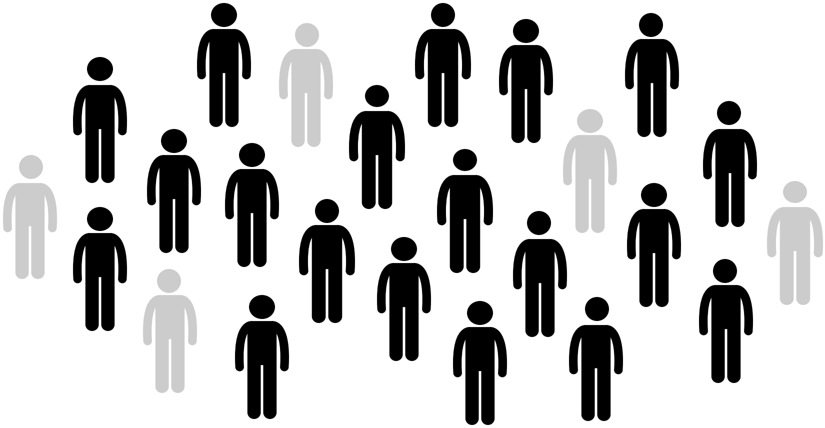
What is the target population?

Target population = sub-group of the population that is the focus of research interest



External validity = how generalizable are our results

What is the accessible population?



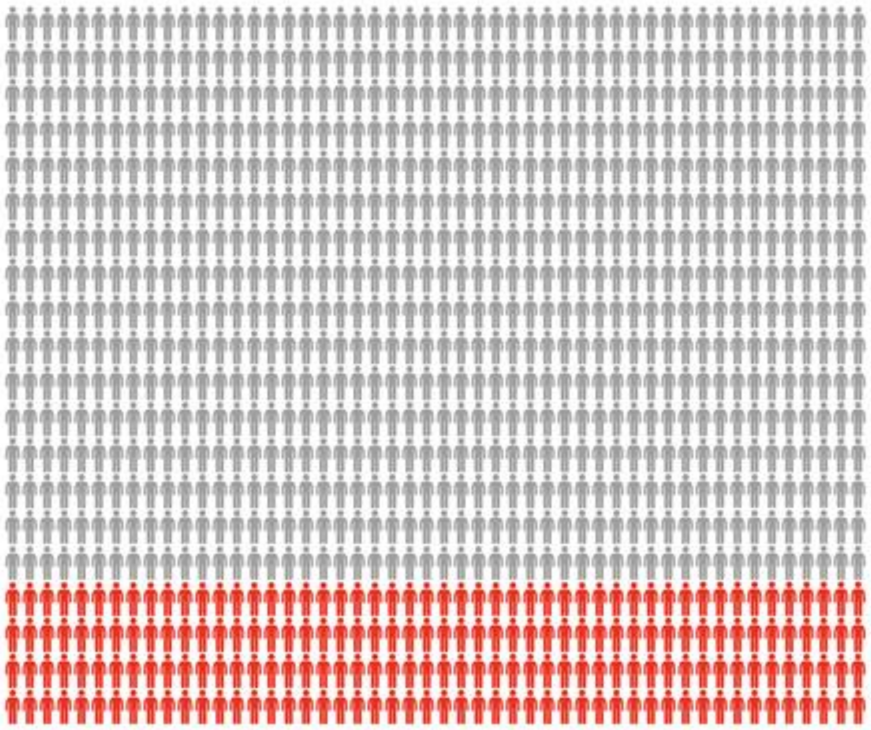
Accessible population

The portion of the target population to which you have reasonable access

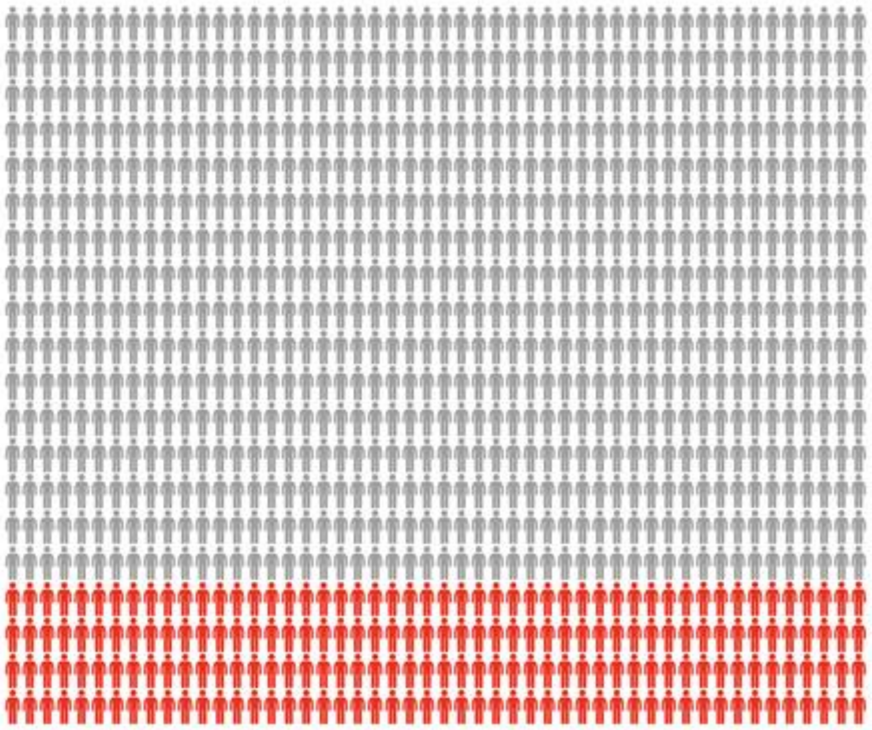
Sampling variation and uncertainty

Population size: 1000

Population prevalence: 20%



Sampling variation and uncertainty



Population size: 1000

Population prevalence: 20%

Sample size: 50

Sample prevalence: 20%

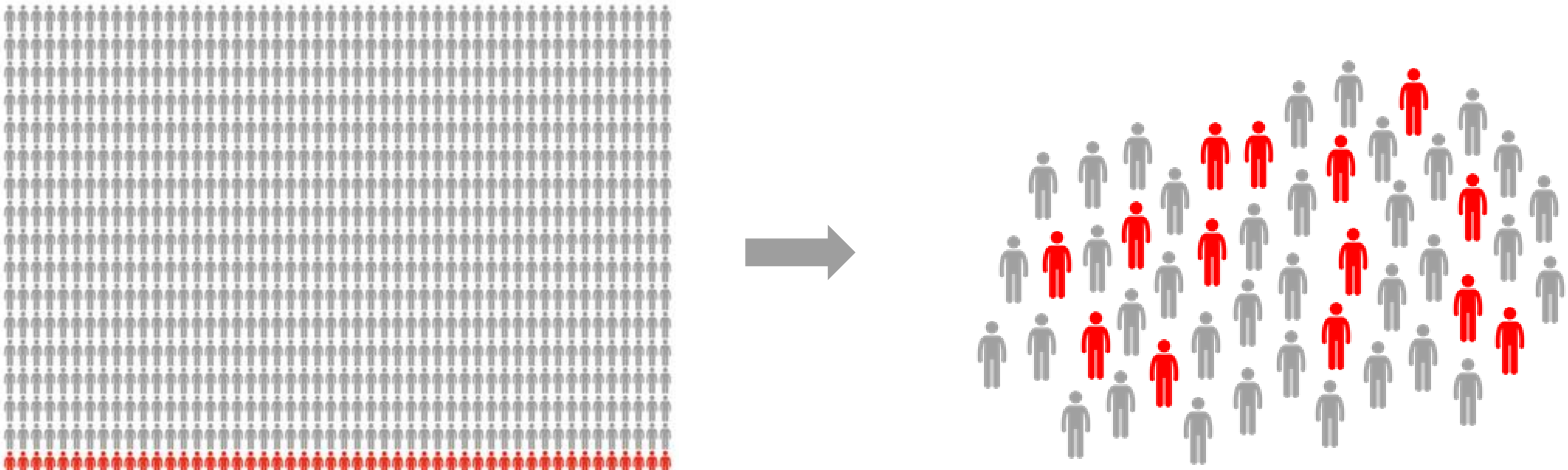
Sampling variation and uncertainty

Population size: 1000

Population prevalence: 20%

Sample size: 50

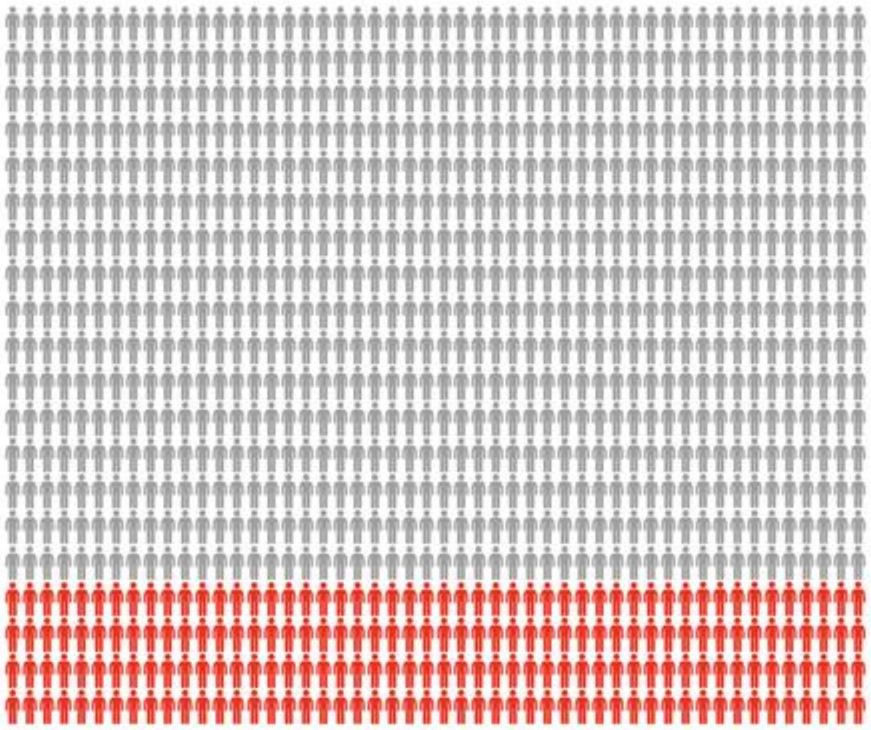
Sample prevalence: 28%



Sampling variation and uncertainty

Population size: 1000

Population prevalence: 20%



Sample size: 50

Sample prevalence: 14%

Example.

Estimating the prevalence of SP drug resistance mutations

Population	All individuals in the region
Target population	Children under 6 years old in the region
Accessible population	Children visiting health facilities for routine vaccination

Hypothetical study results

We conducted our study and found **8 out of 50 children** with the SP mutation

What is the prevalence?

Hypothetical study results

We conducted our study and found **8 out of 50 children** with the SP mutation

What is the prevalence?

Number observed / sample size = **8 / 50**

Hypothetical study results

We conducted our study and found **8 out of 50 children** with the SP mutation

What is the prevalence?

Number observed / sample size = **8 / 50**

This is a point estimate. But, how *confident* are we in this estimate?

The Wald confidence interval

$$CI = \hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Point estimate

$$CI = \boxed{\hat{p}} \pm z_{1-\alpha/2}\sqrt{\frac{\boxed{\hat{p}}(1-\boxed{\hat{p}})}{n}}$$

Point estimate = **8/50 = 0.16**

Critical Z value

$$CI = \hat{p} \pm \boxed{z_{1-\alpha/2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Alpha (α) = 0.05 (95% confidence)

1-(α/2) = **0.975** -> corresponds to ***Z = 1.96***

Critical Z value

$$CI = \hat{p} \pm \boxed{1.96} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Standard error

$$CI = \hat{p} \pm 1.96 \boxed{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Putting it all together

$$CI = 0.16 \pm 1.96 \sqrt{\frac{0.16(1-0.16)}{50}}$$

0.16 ± 0.10

or

16% ± 10%

Summary

- We sample the population to **draw conclusions** about the population as a whole
  - *E.g. what is the prevalence of pfdhps K540E?*

- We can use the **Wald confidence interval** to calculate our 95% Confidence Interval (uncertainty around our point estimate)

Any questions?

Module 1 activity

Format: Interactive R code, accessed through the web

- Define target population
- Differentiate between population and sample
- Calculate the 95% CI
- Explore sampling variability
- Understand effect of sample size on CIs

Workshop materials

https://mrc-ide.github.io/MMS-SD_workshop/