UKRI MRC Centre for Global Infectious Disease Analysis IMPERIAL

Malaria Molecular Surveillance Study Design Workshop

Module 4: Dealing with over-dispersion in multi-cluster studies

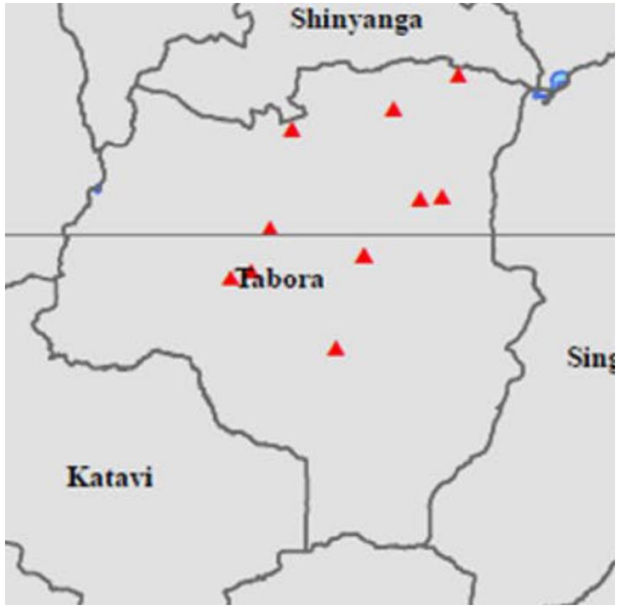
What is a multi-cluster study?

In a multi-cluster study…

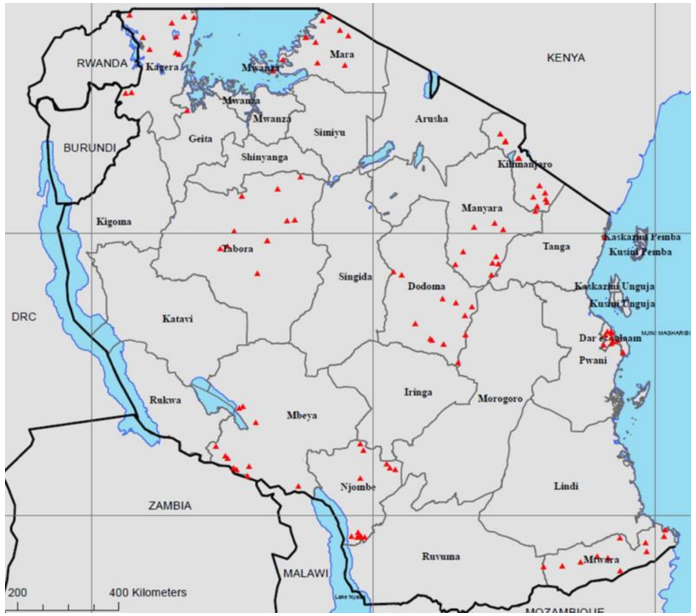
– We conduct the study at several sites (clusters)

– **We aim to draw conclusions at a higher level than the site**

Regional level



Country level



Advantages of multi-cluster studies

We can combine information across sites

- Regional-level estimates aim to draw conclusion about the wider population
- Interventions are often delivered at regional level

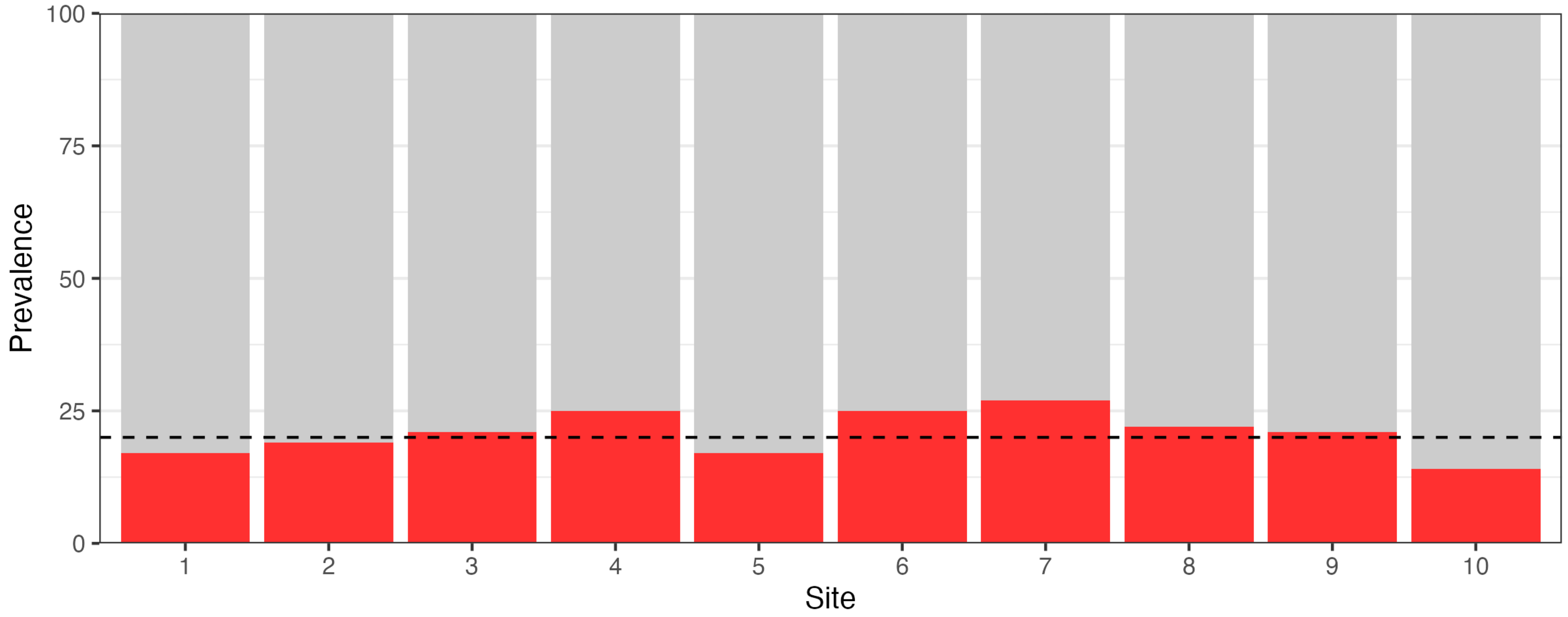
Or we can explore differences between sites

- Are there geographic trends?
- What is the geographic scale of the threat?
- Can we identify cluster-level covariates?

Over-dispersion

- Prevalence study over 10 sites
- 100 samples per site
- global prevalence of 20%

This is what the data spread looks like when samples are perfectly **independent**

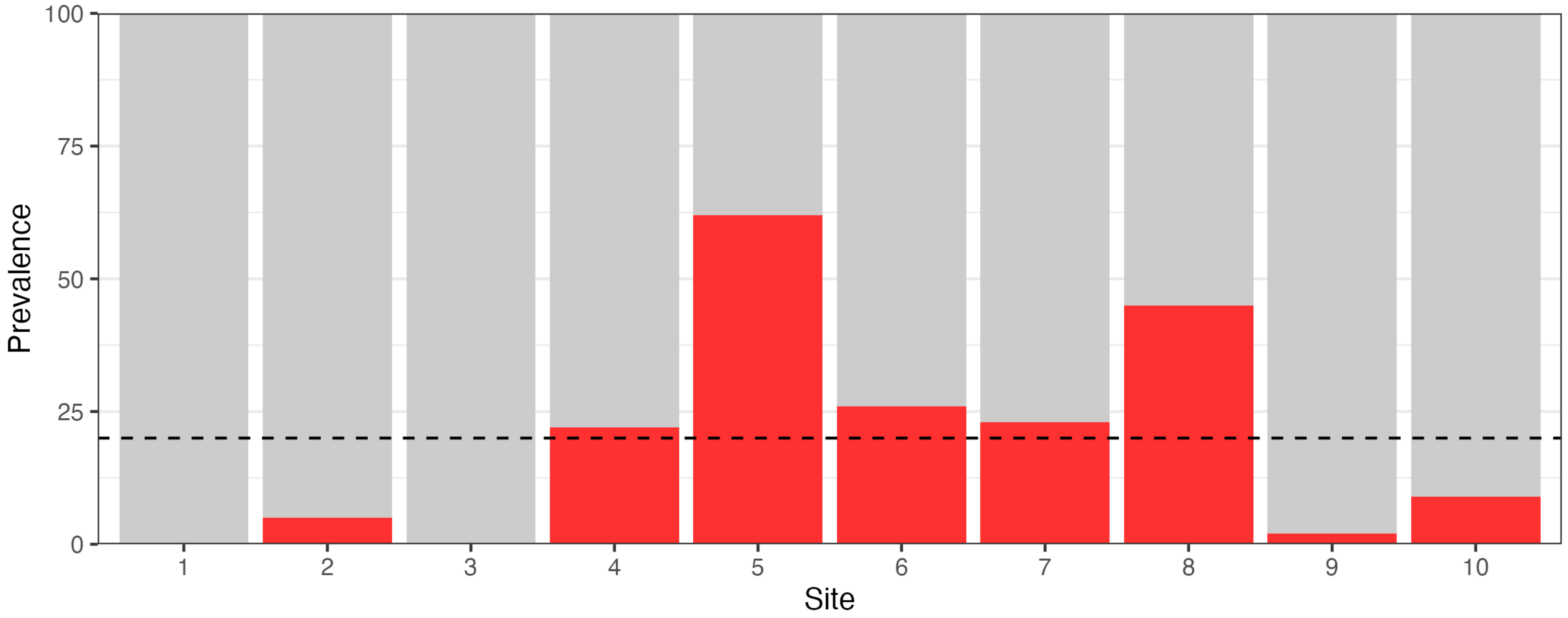


$\hat{p} \approx 20\%$

Over-dispersion

- Prevalence study over 10 sites
- 100 samples per site
- global prevalence of 20%

This is what data really look like!

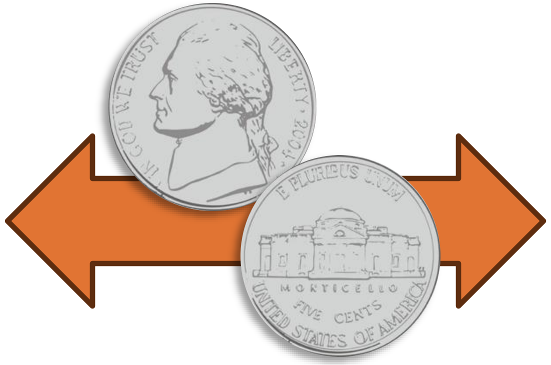


$\hat{p} \approx 20\%$

What causes over-dispersion

Overdispersion

Sites are more **different** than we would expect on average



Intra-cluster correlation

People within sites are more **similar** than we would expect on average

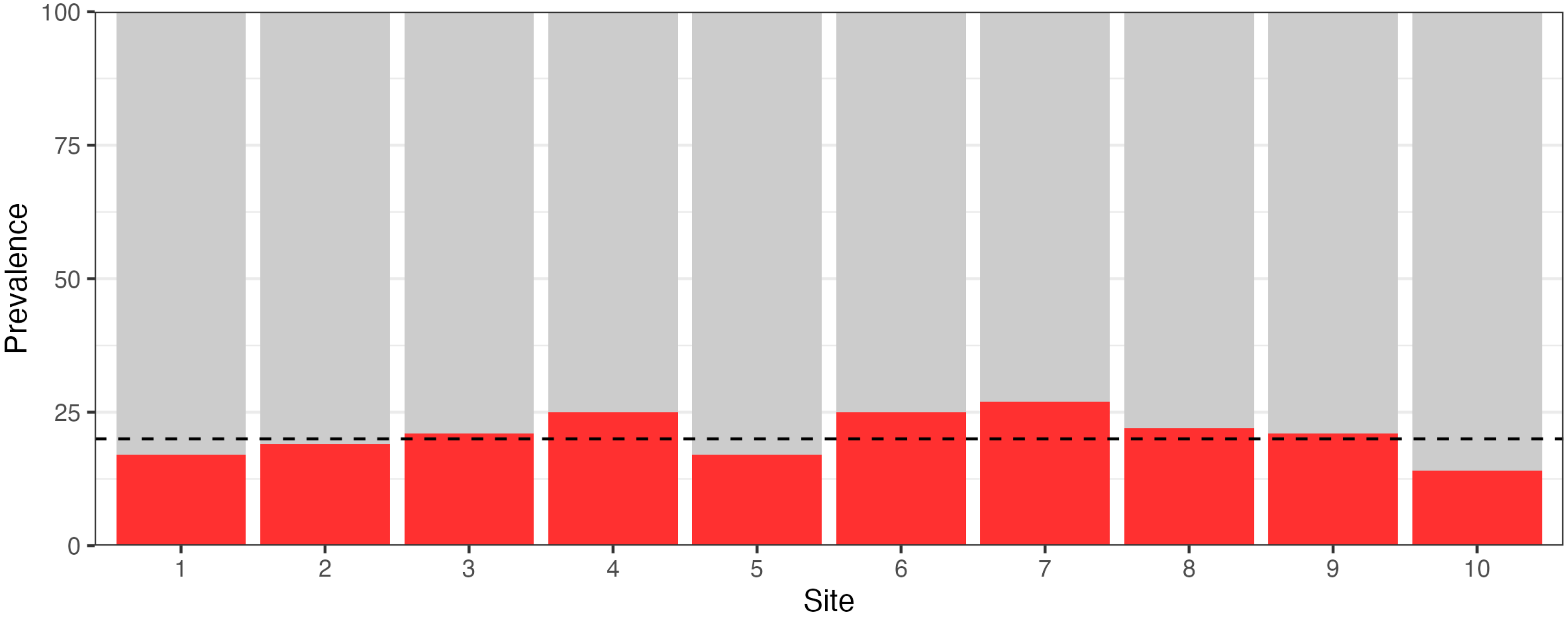
What causes intra-cluster correlation

- – Similar behaviours/customs
- – Similar occupations
- – Shared vector reservoirs
- – Genetic similarities
- – Similar access to healthcare
- – Local transmission and outbreaks

Detecting over-dispersion

There is a level of variability between sites that we expect:

$$\hat{p} \pm \sqrt{\frac{\hat{p}(1 - \hat{p})}{n_i}}$$

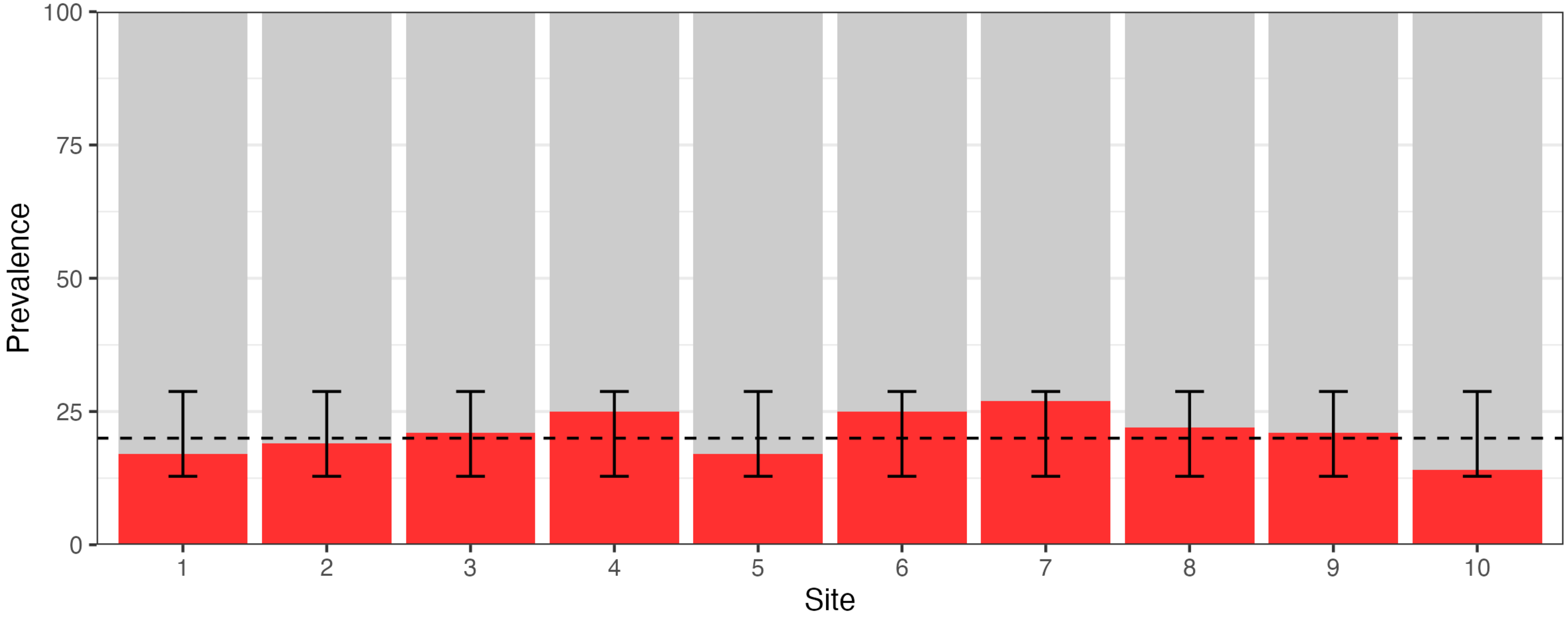


Detecting over-dispersion

There is a level of variability between sites that we expect:

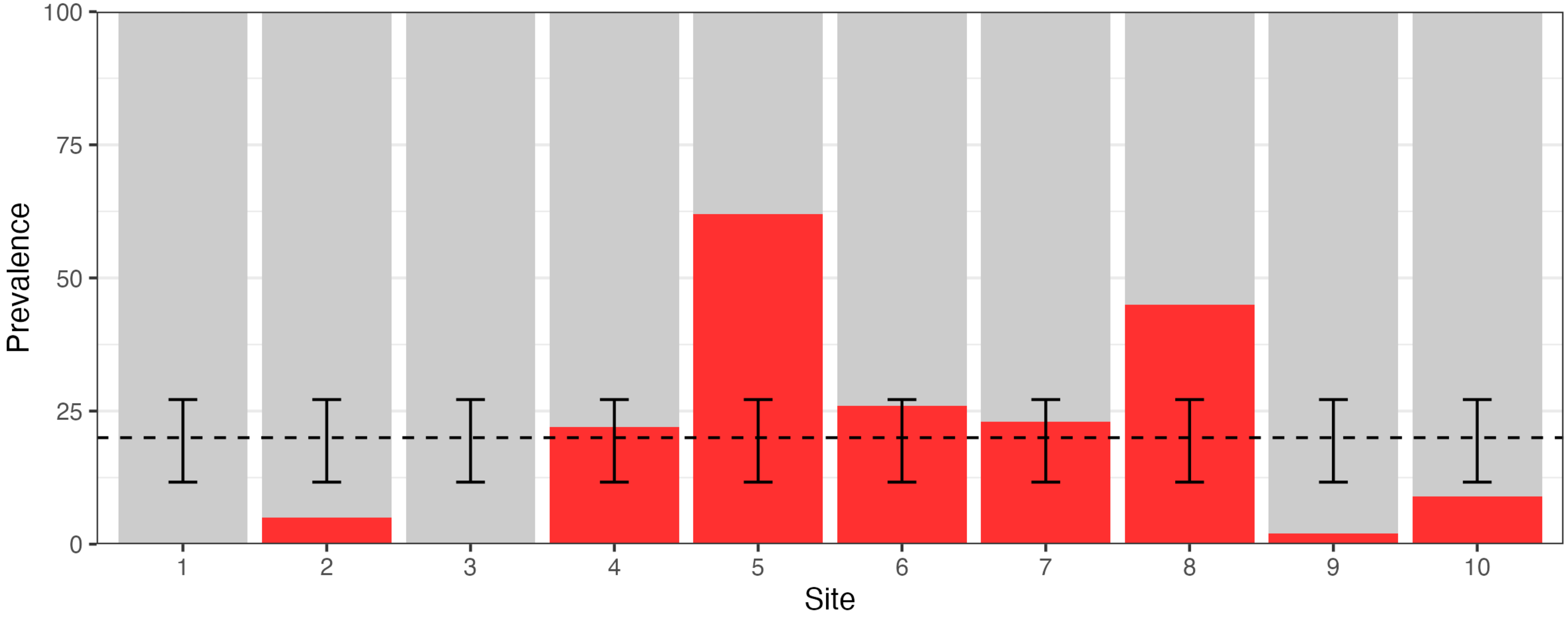
$$\hat{p} \pm \sqrt{\frac{\hat{p}(1 - \hat{p})}{n_i}}$$

95% of sites within this range



Detecting over-dispersion

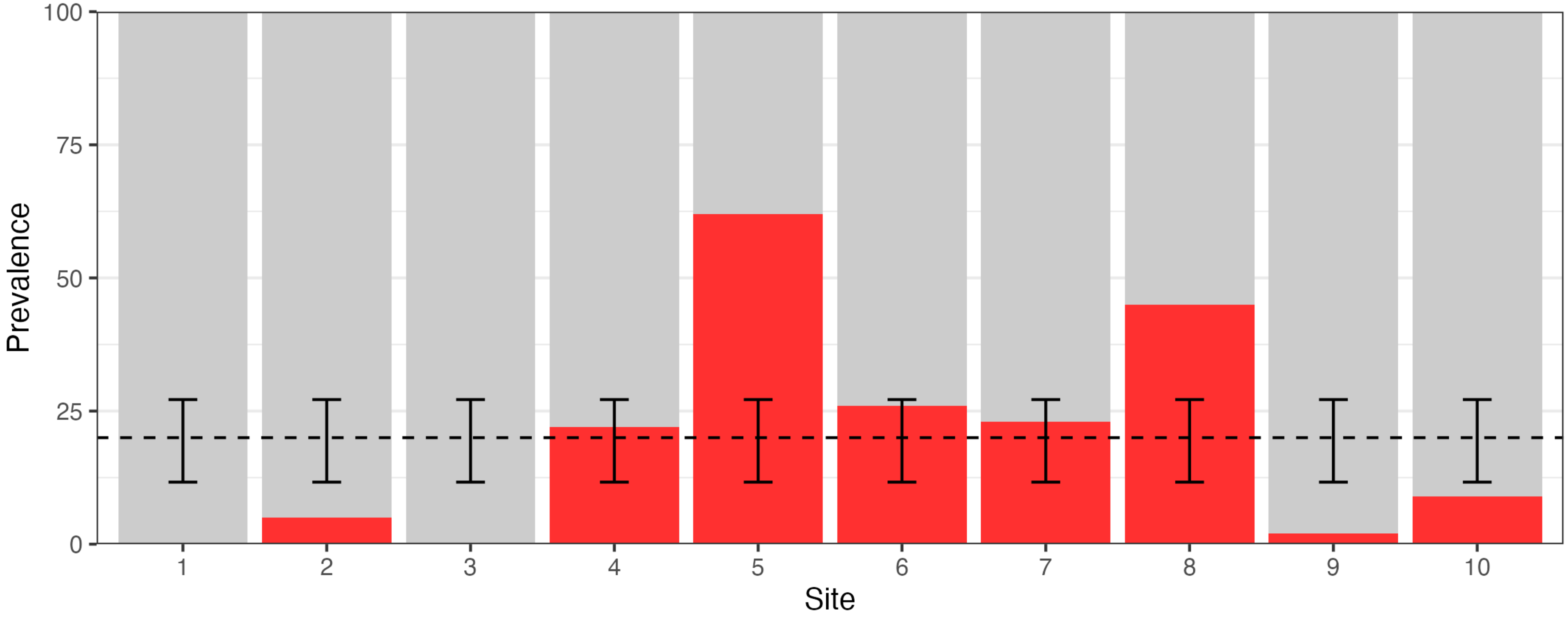
If more than about 10% of sites are outside the range, there is likely some **systematic** overdispersion



Detecting over-dispersion

If more than about 10% of sites are outside the range, there is likely some **systematic** overdispersion

70% outside the range!



Quantifying over-dispersion

1. Design effect

2. Effective sample size

3. Intra-cluster correlation coefficient

The Design Effect

$$D_{\text{eff}} = \frac{\text{What is the variance of my data?}}{\text{What variance would I expect if samples were independent?}}$$

The Design Effect

$$D_{\text{eff}} = \frac{\text{What is the variance of my data?}}{\text{What variance would I expect if samples were independent?}}$$

The design effect is a measure statistical **inefficiency**. A value of $D_{\text{eff}} = 1$ is gold standard (although D_{eff} can be less than 1).

The Design Effect

$$D_{\mathrm{eff}} = \frac{\mathrm{Var}_{\mathrm{obs}}}{\mathrm{Var}_{\mathrm{SRS}}}$$

The Design Effect

$$D_{\text{eff}} = \frac{\text{Var}_{\text{obs}}}{\text{Var}_{\text{SRS}}} = \frac{s^2}{\frac{1}{c}\sum_{i=1}^{c}\frac{\hat{p}(1-\hat{p})}{n_i}}$$

s^2 = sample variance

$\hat{p}$ = global prevalence estimate

c = number of clusters

n_i = sample size in i^{th} cluster

The Design Effect – worked example

Site	Sample size	Prevalence
1	60	0.00
2	80	0.05
3	70	0.00
4	100	0.22
5	40	0.62
6	60	0.26
7	50	0.23
8	90	0.09

See the Excel file

Overdispersion_example.xlsx

to work through steps

(available on course website)

The Design Effect – worked example

Site	Sample size	Prevalence
1	60	0.00
2	80	0.05
3	70	0.00
4	100	0.22
5	40	0.62
6	60	0.26
7	50	0.23
8	90	0.09

$$\mathrm{Var}_{\mathrm{obs}} = 0.0420$$

$$\mathrm{Var}_{\mathrm{SRS}} = 0.0024$$

$$D_{\mathrm{eff}} = 17.73$$

The effective sample size

That’s great…but what does a value $D_{\text{eff}} = 17.73$ really mean?

The effective sample size

That's great…but what does a value $D_{\text{eff}} = 17.73$ really mean?

$$N_{\text{eff}} = \frac{N}{D_{\text{eff}}}$$

The effective sample size

That's great…but what does a value $D_{\text{eff}} = 17.73$ really mean?

$$N_{\text{eff}} = \frac{N}{D_{\text{eff}}}$$

N_{eff} is the number of completely independent samples you would need to achieve the same level of precision as your more complex study design

The effective sample size

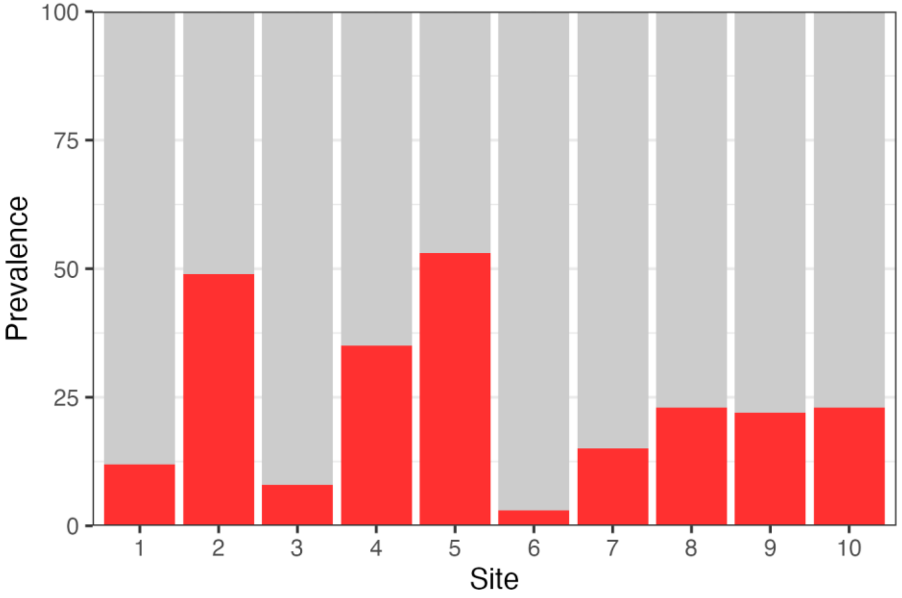
Back to Overdispersion_example.xlsx

The effective sample size

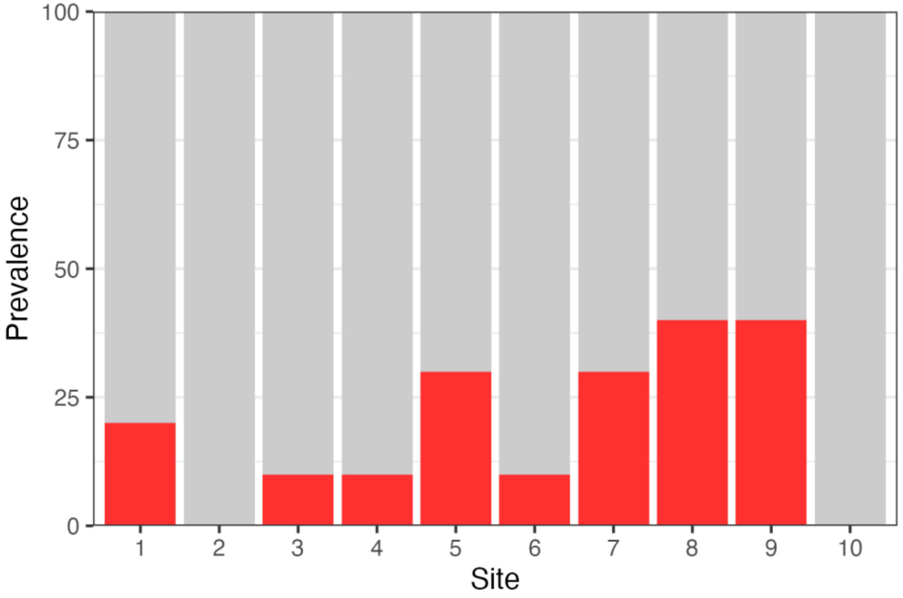
One of these was generated with $N = 100$
the other with $N = 1000$ but $N_{\text{eff}} = 100$

Which one is which?

?



?

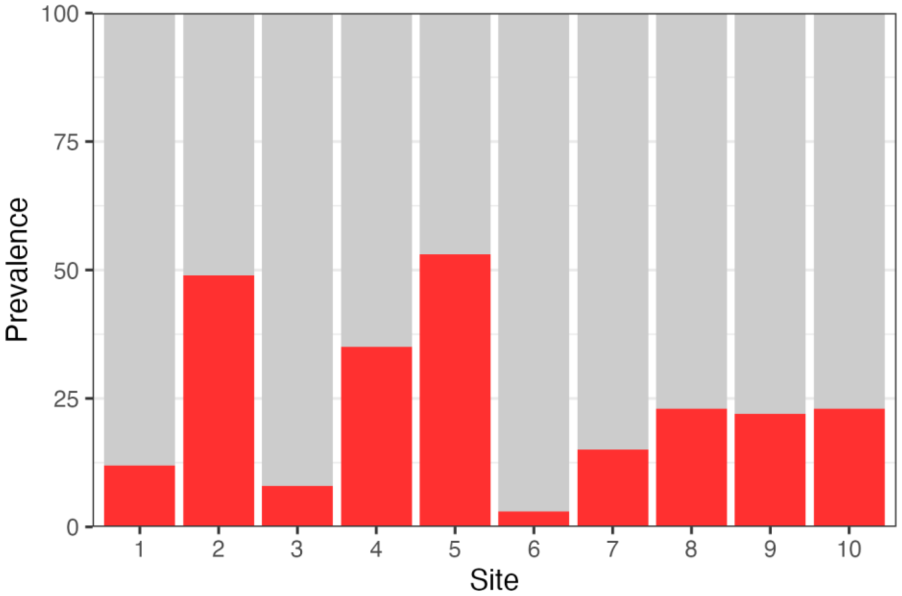


The effective sample size

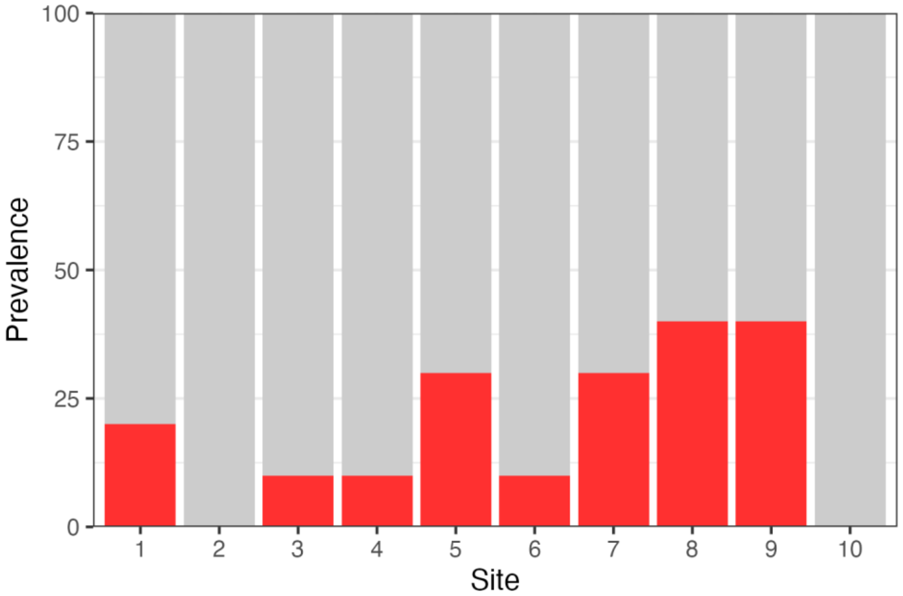
One of these was generated with $N = 100$
the other with $N = 1000$ but $N_{\text{eff}} = 100$

Which one is which?

$N = 1000$



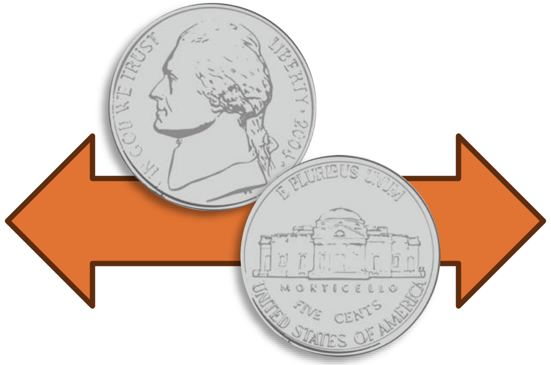
$N = 100$



The intra-cluster correlation coefficient

Overdispersion

Sites are more **different** than we would expect on average



Intra-cluster correlation

People within sites are more **similar** than we would expect on average

The intra-cluster correlation coefficient

The ICC (r) is a value between 0 and 1 that represents the correlation between individuals in the same site.

We can write the design effect in terms of the ICC:

$$D_{\text{eff}} = 1 + (\bar{n} - 1)r$$

$\bar{n}$ = average cluster size

r = ICC

We can write the ICC in terms of the design effect:

$$r = \frac{D_{\text{eff}} - 1}{\bar{n} - 1}$$

The intra-cluster correlation coefficient

Back to Overdispersion_example.xlsx

Why is the ICC useful?

Scenario

Your study population has a true ICC of $r = 0.053$ in terms of *mdr1* N86Y prevalence. Assume we do not know the true ICC.

A pilot study is run over 10 clusters using a sample size of $N = 200$, divided into $n = 20$ per cluster. Over-dispersion is measured, and we find a design effect of $D_{\text{eff}} = 2.0$.

A follow-up study is now planned with a much larger sample size of $N = 10{,}000$, divided into $n = 1000$ per cluster. When designing the study, we assume the same design effect as the pilot, meaning we expect an effective sample size of $N_{\text{eff}} = 5{,}000$, which is still very large.

When the results come in, we measure the design effect on the new data. We find it has increased to $D_{\text{eff}} = 53.9$! We now only have an effective sample size of just $N_{\text{eff}} = 185$!

Why is the ICC useful?

Scenario

Your study population has a true ICC of $r = 0.053$ in terms of *mdr1* N86Y prevalence. Assume we do not know the true ICC.

A pilot study is run over 10 clusters using a sample size of $N = 200$, divided into $n = 20$ per cluster. Over-dispersion is measured, and we find a design effect of $D_{\text{eff}} = 2.0$.

A follow-up study is now planned with a much larger sample size of $N = 10{,}000$, divided into $n = 1000$ per cluster. When designing the study, we assume the same design effect as the pilot, meaning we expect an effective sample size of $N_{\text{eff}} = 5{,}000$, which is still very large.

When the results come in, we measure the design effect on the new data. We find it has increased to $D_{\text{eff}} = 53.9$! We now only have an effective sample size of just $N_{\text{eff}} = 185$!

What happened here!?

Why is the ICC useful?

$$D_{\text{eff}} = 1 + (\bar{n} - 1)r$$

Why is the ICC useful?

$$D_{\text{eff}} = 1 + (\bar{n} - 1)r$$

Pilot study

$\bar{n} = 20$

$r = 0.053$ $\longrightarrow$ $D_{\text{eff}} = 2.0$

Why is the ICC useful?

$$D_{\text{eff}} = 1 + (\bar{n} - 1)r$$

Pilot study

$\bar{n} = 20$

$r = 0.053$

→ $D_{\text{eff}} = 2.0$

Follow-up study

$\bar{n} = 1000$

$r = 0.053$

→ $D_{\text{eff}} = 53.9$

Why is the ICC useful?

$$D_{\text{eff}} = 1 + (\bar{n} - 1)r$$

Pilot study

$\bar{n} = 20$

$r = 0.053$ → $D_{\text{eff}} = 2.0$

Follow-up study

$\bar{n} = 1000$

$r = 0.053$ → $D_{\text{eff}} = 53.9$

- The design effect is not an **intrinsic** measure of the population.
- It relates to **our study** of the population. It **depends on sample size** as well as intrinsic factors.
- D_{eff} is hard to compare objectively between studies, while r is easy.

Recap

1. Design effect
    - A simple measure of statistical inefficiency

2. Effective sample size
    - An intuitive way of measuring efficiency

3. Intra-cluster correlation coefficient
    - Facilitates comparison between studies

How can we design multi-cluster studies?

New versions of formulae (precision, power, sample size etc.) that take over-dispersion into account:

Generalization of Wald interval:

$$\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{N} D_{\text{eff}}}$$

In the design stage, this means we will have to **assume** a value of the design effect, or the ICC

Module 4 activity

Format: Interactive R code, accessed through the web

- Work with the NMCP of Tanzania to analyse data from a multi-site *pfhrp2/3* deletion prevalence study
- Detect and quantify over-dispersion in the data
- Plan a new study that accounts for over-dispersion

Workshop materials

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