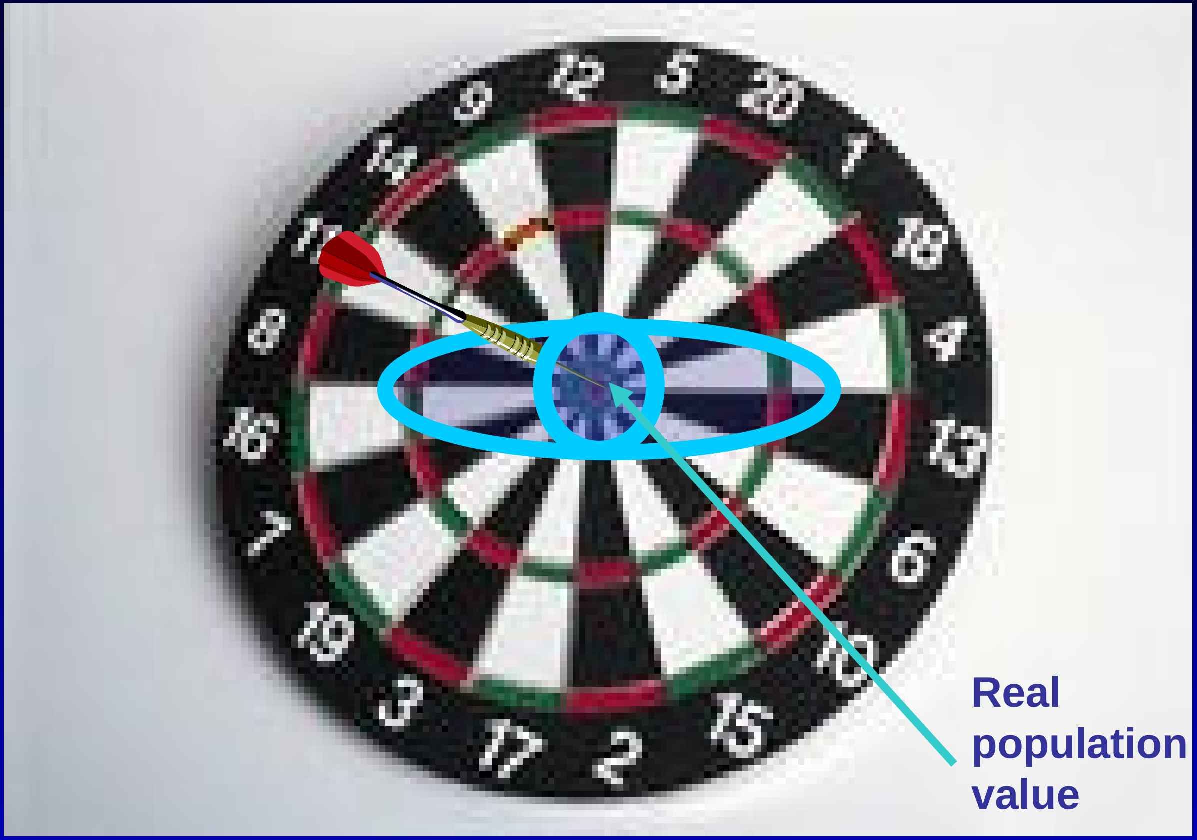
Sample size calculation and development of sampling plan



Overview

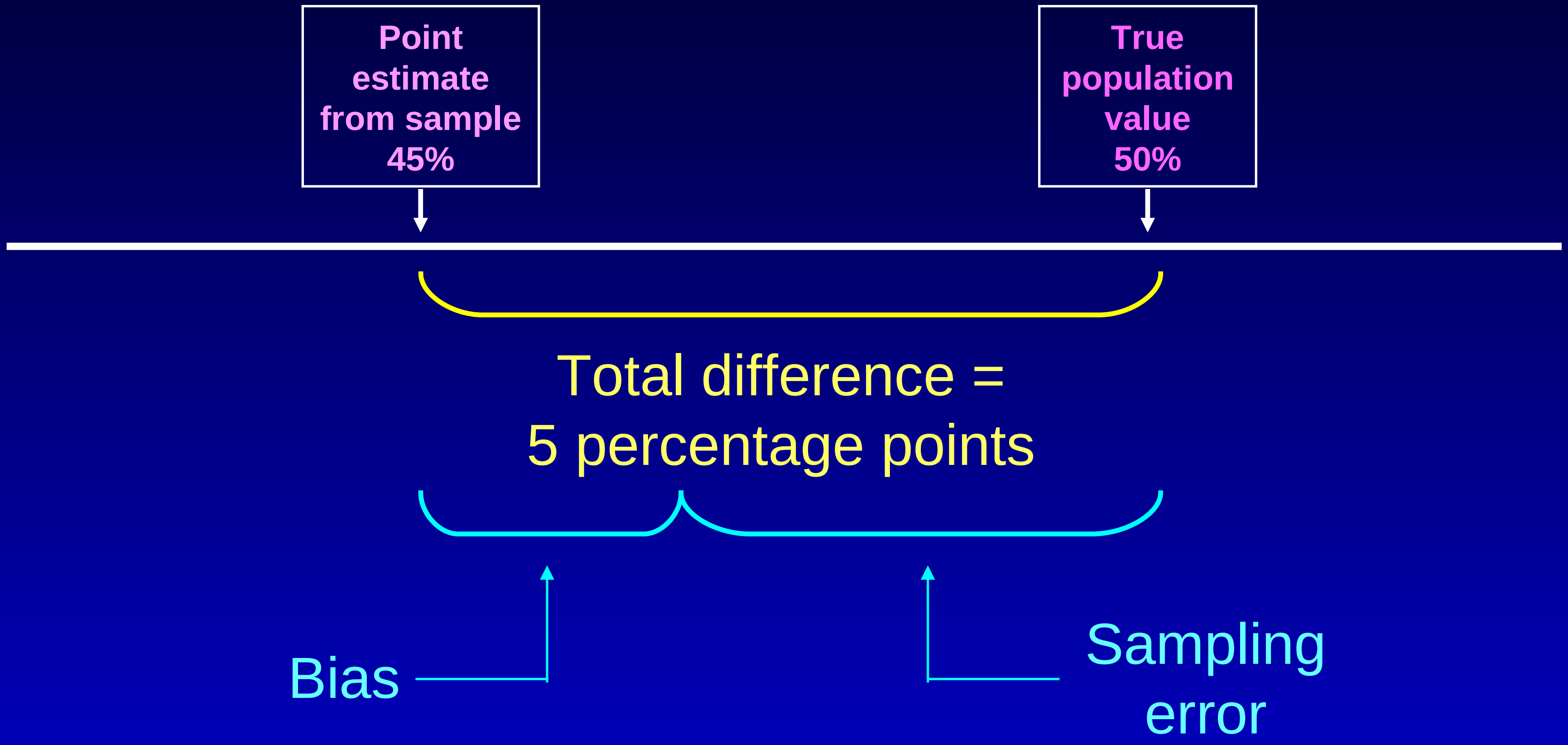
- Bias versus sampling error
- Level of precision
- Calculating sampling size
    - Single survey using random sampling
    - Single survey using two-stage cluster sampling
    - Comparing two surveys
- Drawing the sample
    - First stage
    - Second stage
- Developing a sampling plan

Introduction

Result from survey is never
exactly the same as
the actual value in the population

WHY?

Bias and sampling error



Point estimate from sample 45%

True population value 50%

Total difference = 5 percentage points

Bias

Sampling error

Bias

Results from:

- Enumerator/respondent bias
- Incorrect measurements (anthropometric surveys)
- Selection of non-representative sample
- Likelihood of selection not equal for each sampling unit

Sampling error

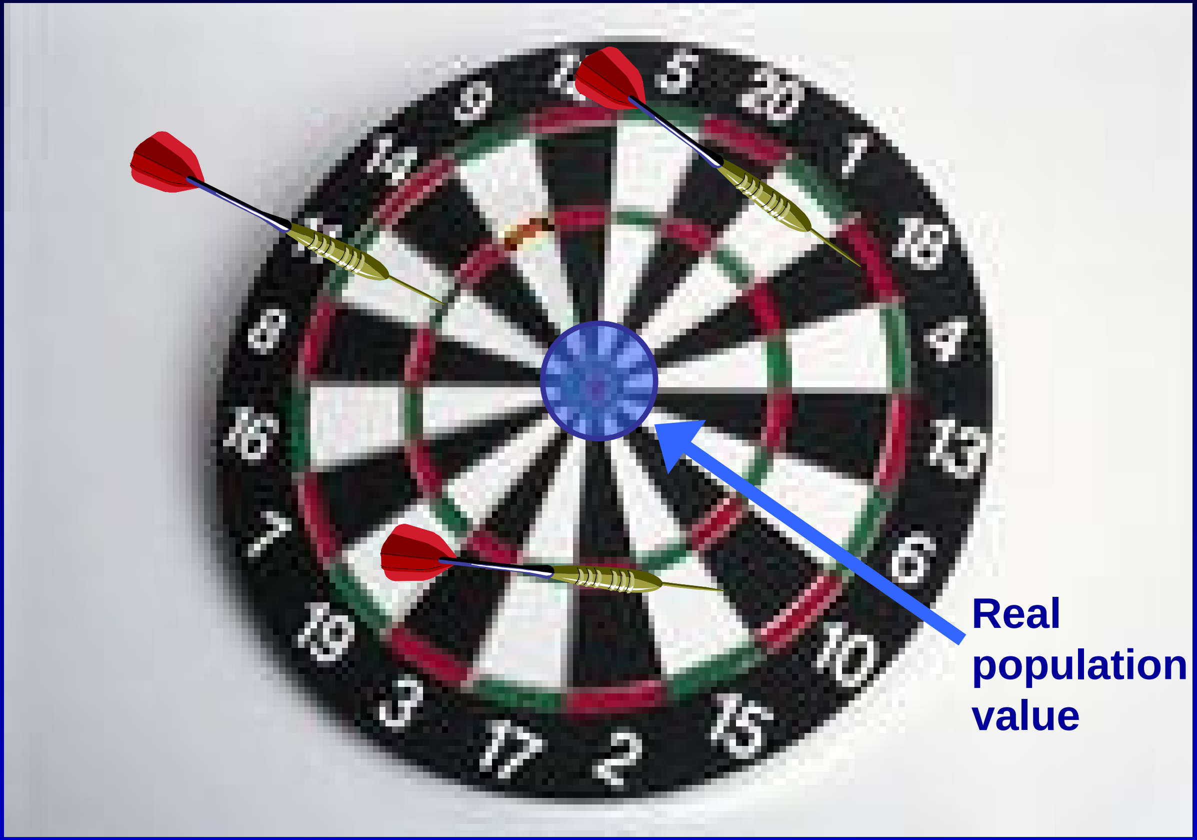
Sampling error

- Difference between survey result and population value due to random selection of sample
- Influenced by:
  - Sample size
  - Sampling method

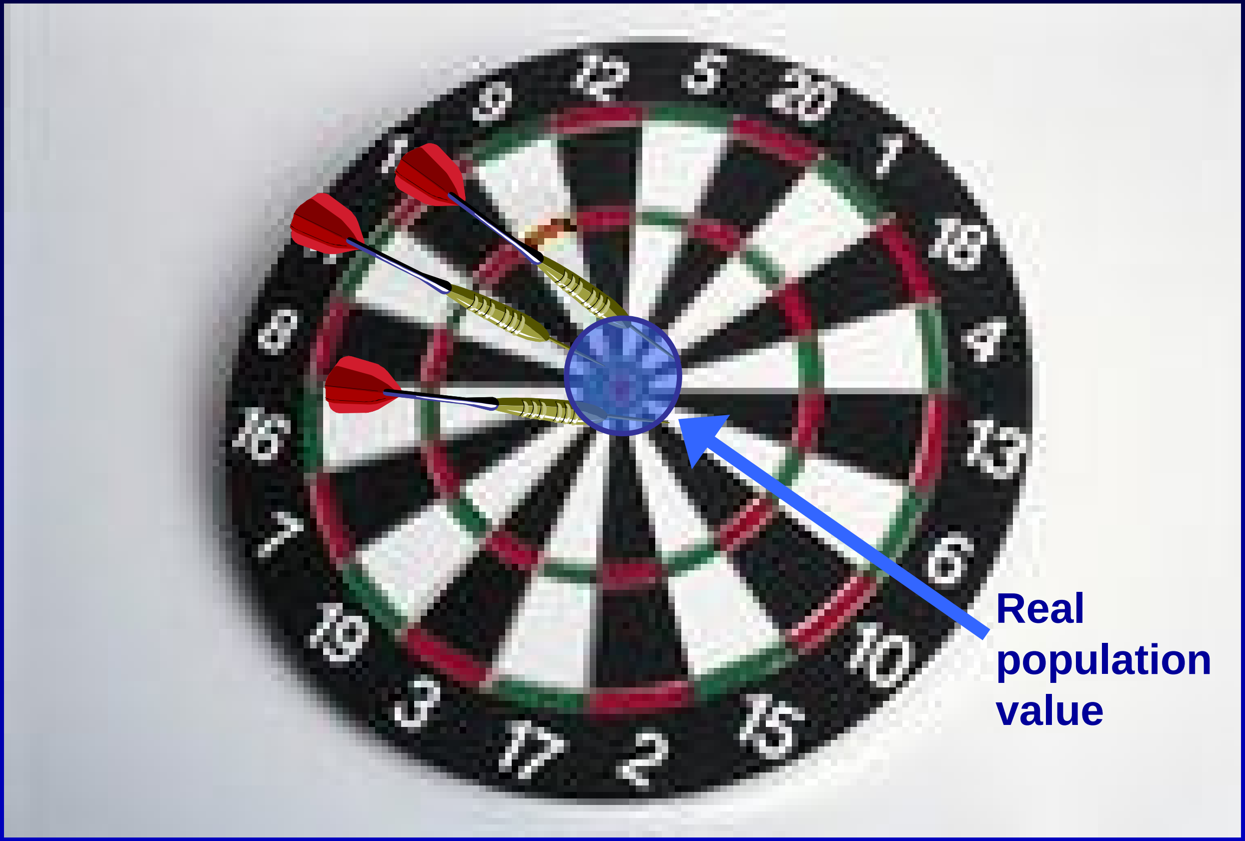
Unlike bias, it can be predicted, calculated, and accounted for

Example 1: Small or large sample? Bias or no bias?

[redacted] bias:



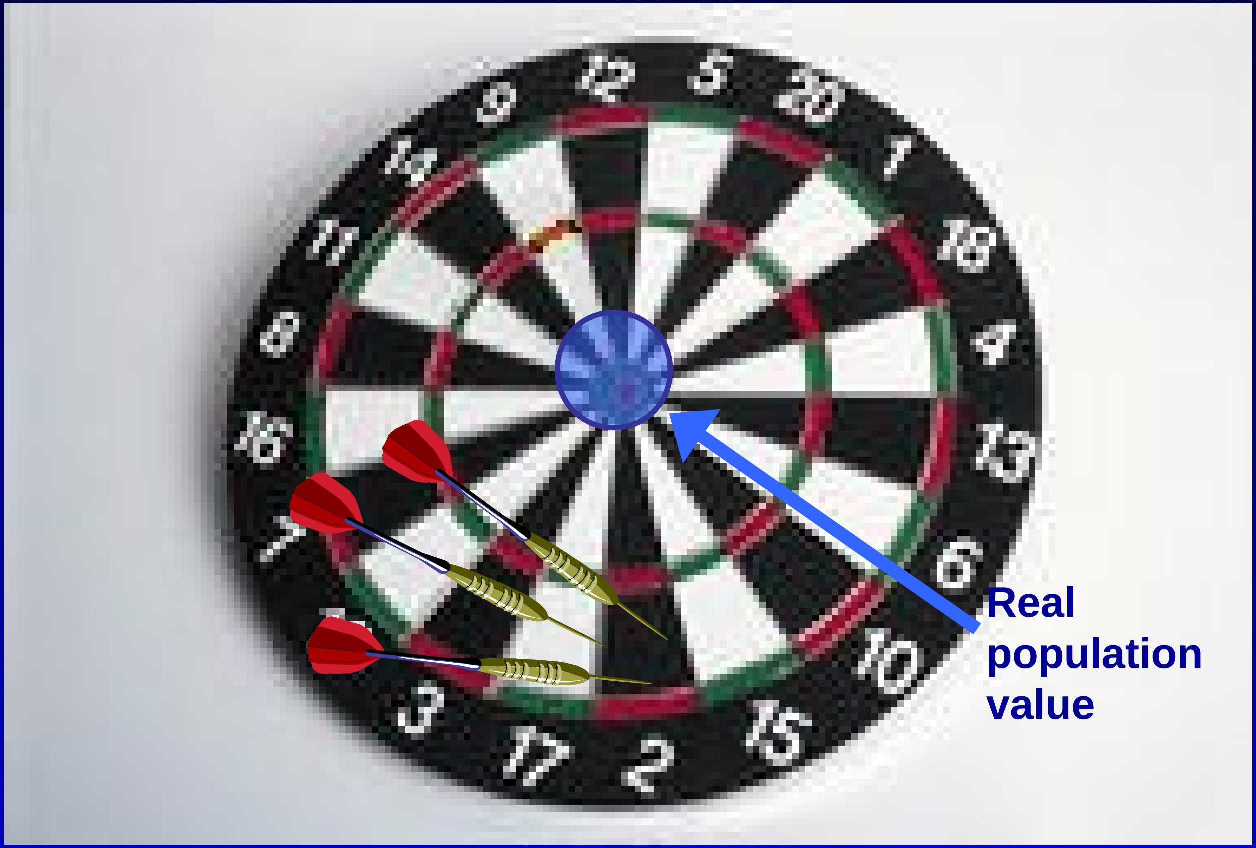
Example 2: Small or large sample? Bias or no bias?



Real population value

Example 3: Small or large sample? Bias or no bias?

[redacted]:



Precision versus bias

- Larger sample size increases precision
  - It does **NOT** guarantee absence of bias
  - Bias may result in very incorrect estimate
- Quality control is more difficult the larger the sample size
- Therefore, you may be better off with smaller sample size, less precision, but much less bias.

Calculating sample size - Introduction

- Sample size calculations can tell you how many sampling units you need to include in the survey to get some required level of precision

FS & nutrition surveys: sampling considerations

If a nutritional survey is conducted as part of CFSVA, additional considerations on sampling are required to reconcile the 2 surveys.

Why?

- Food security analysis has less strict demands on the precision of a single indicator (convergence of evidence) → few % points difference in prevalence of FS is acceptable.
- A difference of a few % points in wasting prevalence can have considerable implications for programmes.

FS & nutrition surveys: sampling considerations (cont.)

The final HH sample size depends on the objective of collecting nutritional info.

a. goal is to study the link between food security & nutrition → smaller sample sizes are sufficient

b. goal is to provide accurate & precise prevalence on nutrition indicators → nutrition sample size has to be adequate (larger)

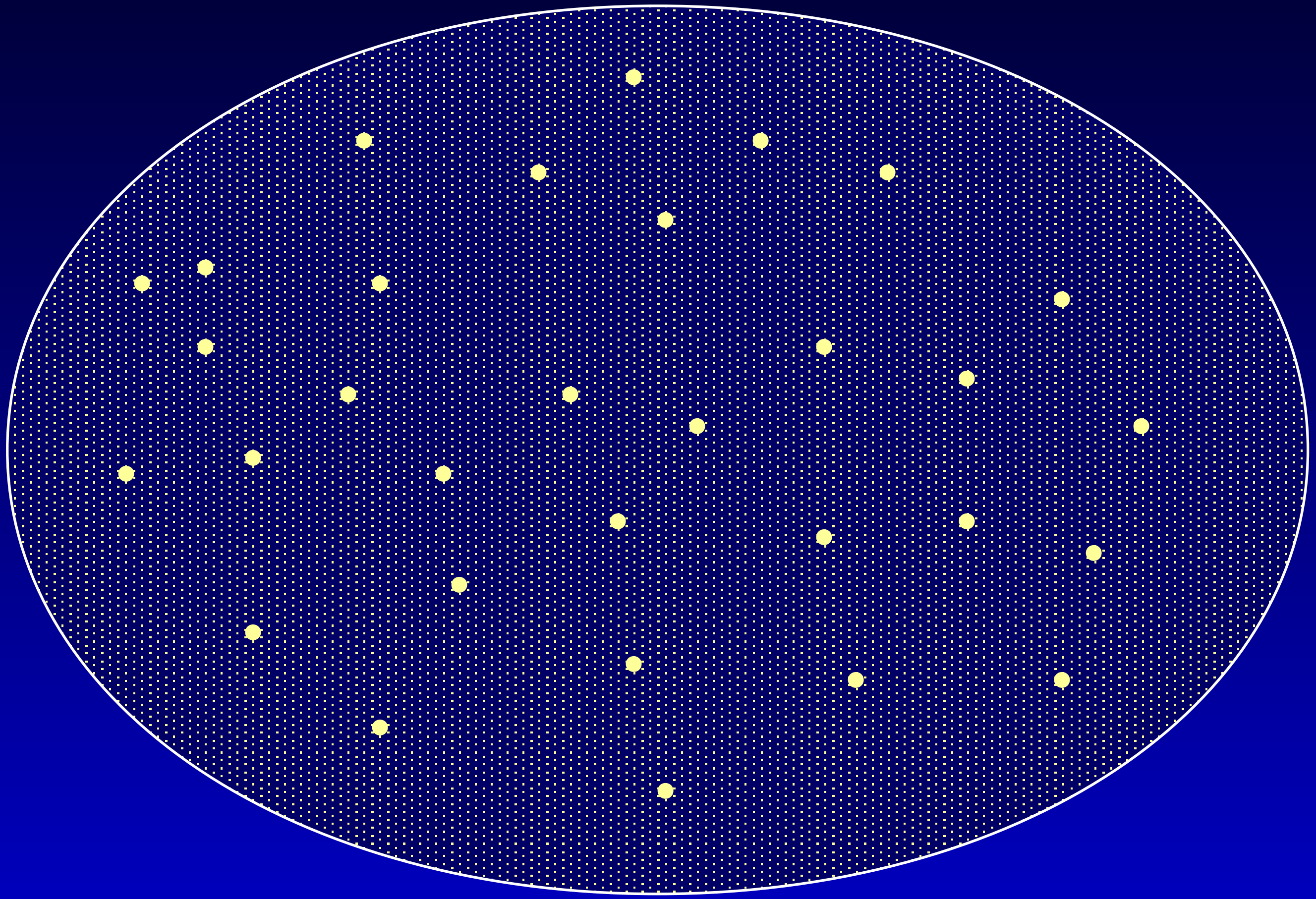
Nutrition survey sample size: 30*30 (?)

Why the 30*30 method was proposed and why is not necessary to compute the nutrition survey sample size?

Proposed to ensure enough precision. It assumes that:

- Prevalence of the core indicator is 50%
- Desired precision is -/+ 5 percentage points
- DEEF = 2
- 15% of the HHs or kids will refuse

Single survey using simple random sampling

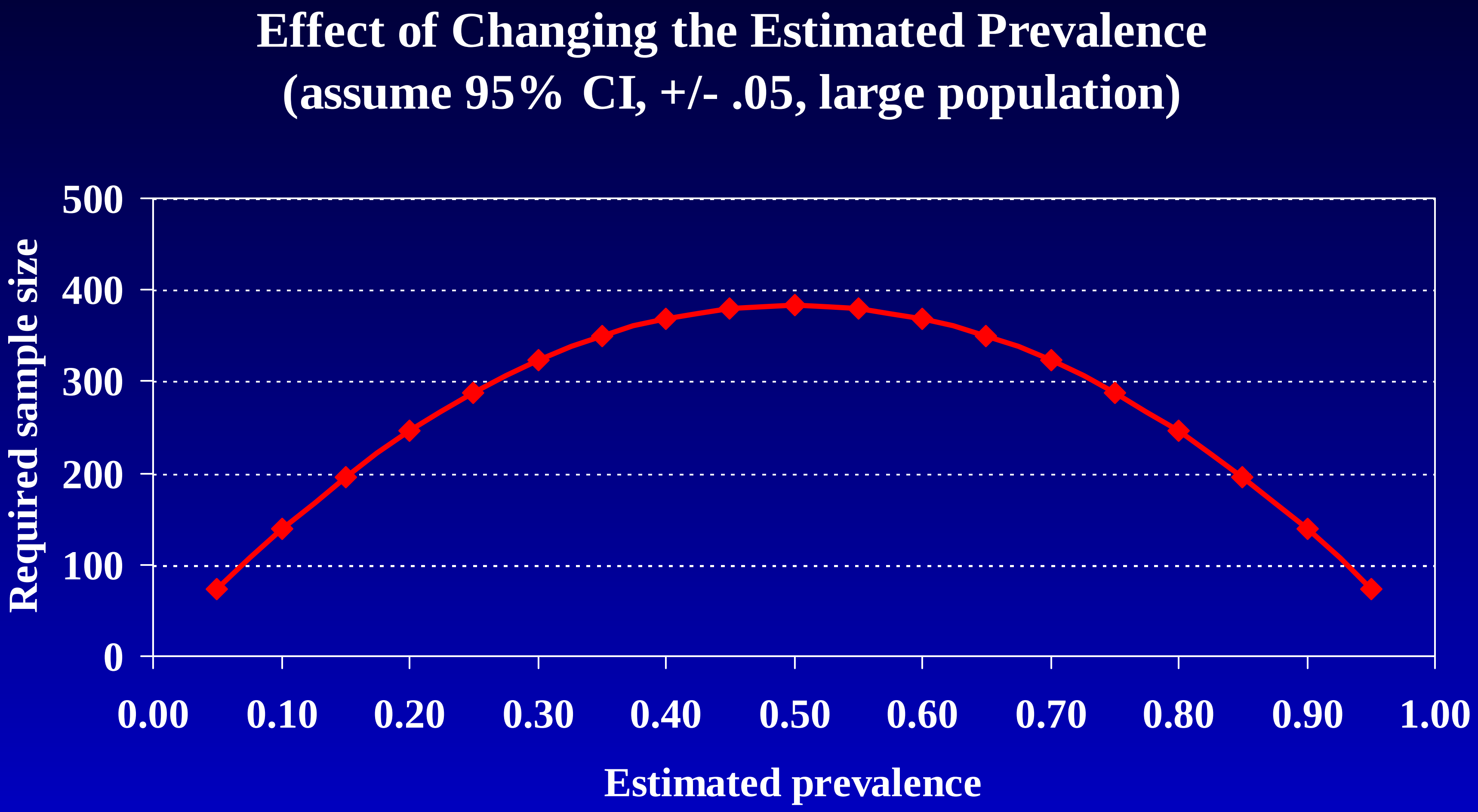


Calculate sample size – single survey using random sample

To estimate sample size, you need to know:

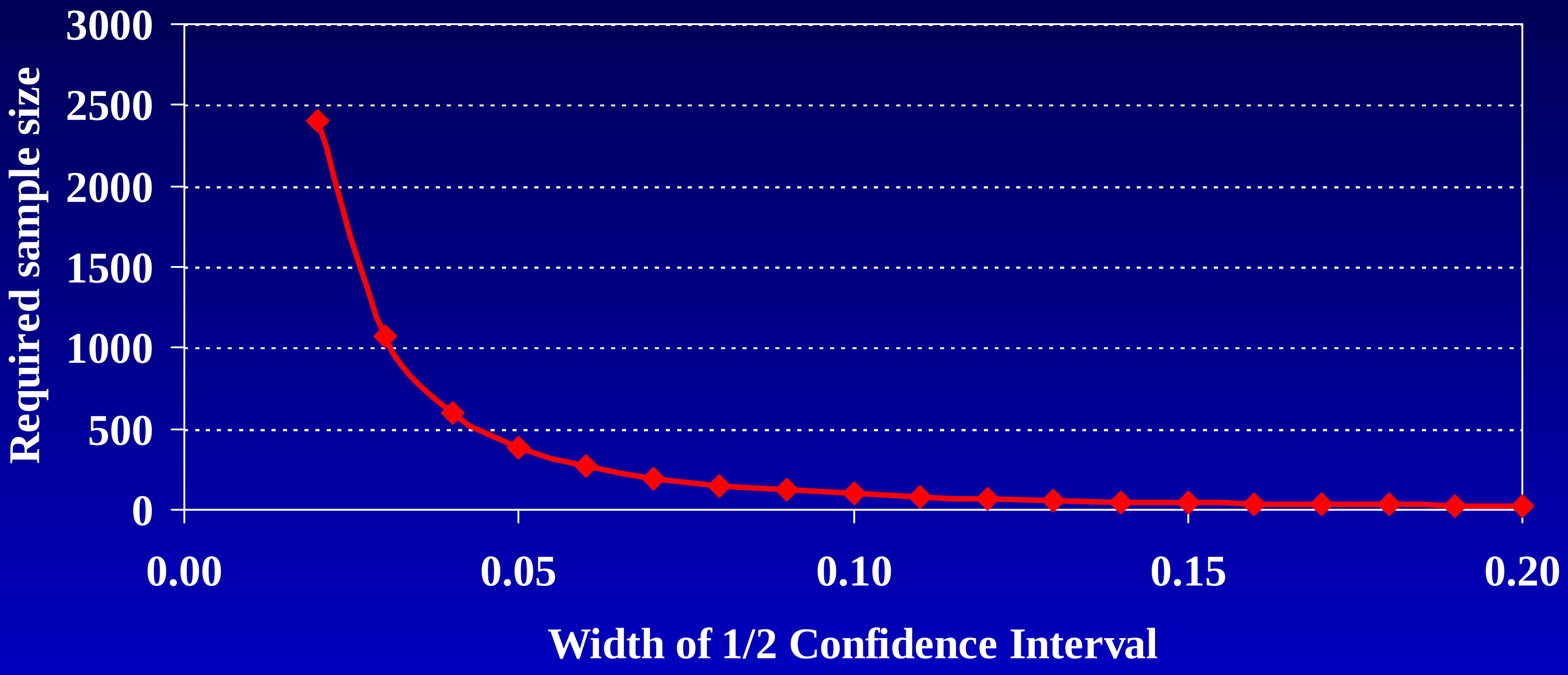
- Estimate of the prevalence of the key indicator (e.g. rate of stunting)
- Precision desired (for example: ± 5%)
- Level of confidence (always use 95%)
- Expected response rate
- *Population*
- *For nutrition surveys:* number of eligible individuals per household

Changing the estimated prevalence



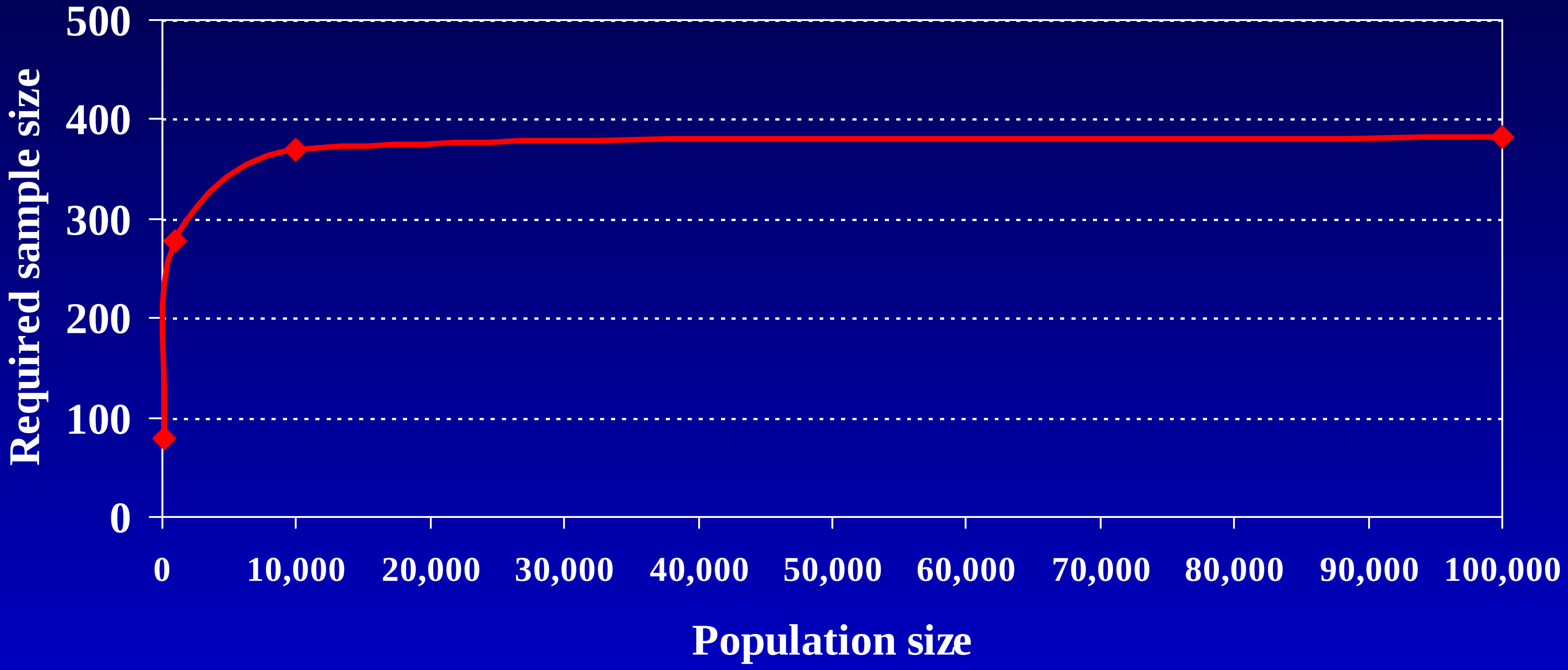
Changing the desired precision

Effect of Changing the Desired Precision
(assume 95% CI, prevalence = 50%, large population)



Changing the population size

Effect of Changing the Population Size
(assume 95% CI, prevalence=.50, +/- .05)



NOTE:

As long as the target population is more than a few thousand people, you do not need to consider it in the sample size.

You do NOT generally need a larger sample size if the population is bigger.

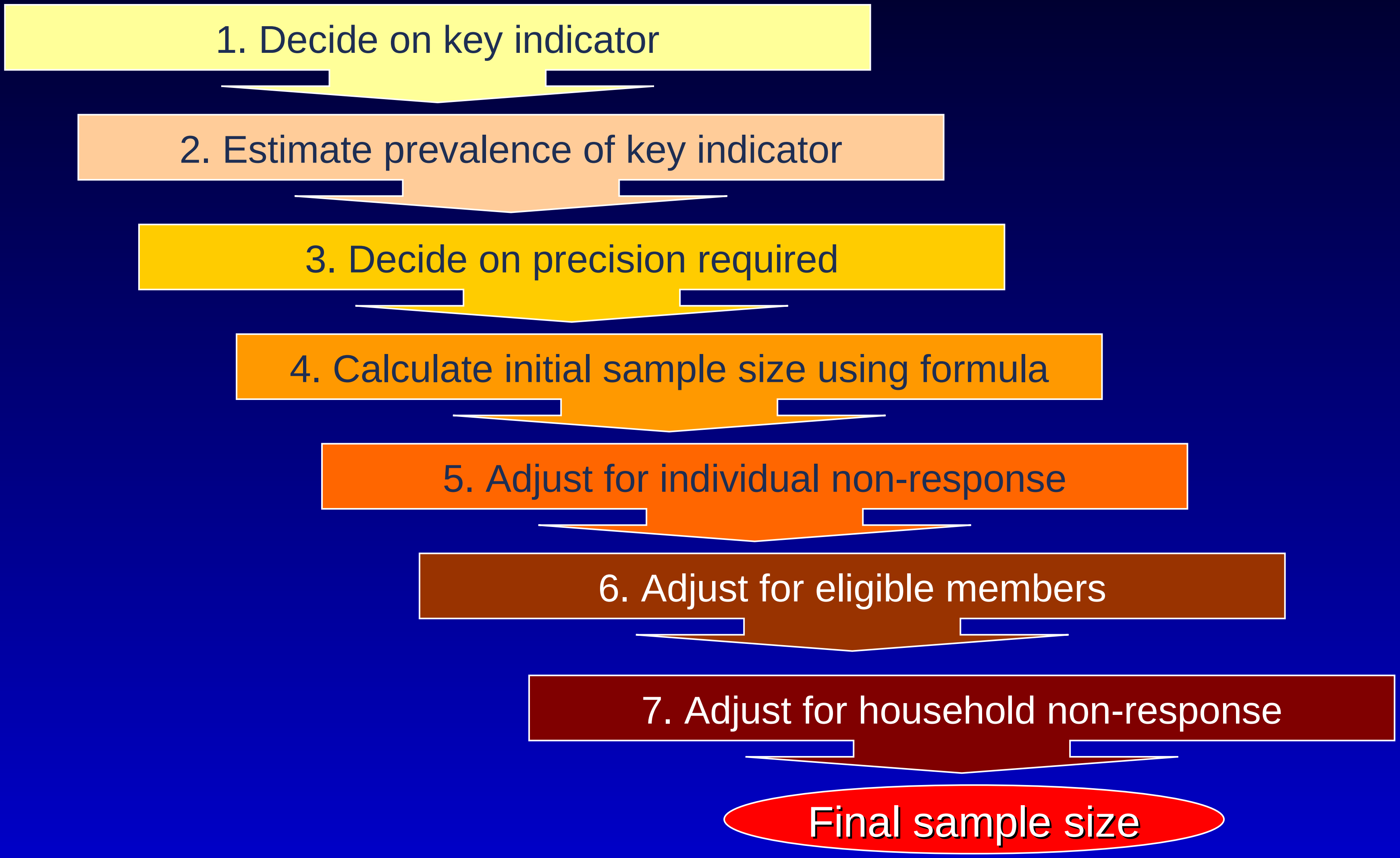
Sample size formula
(single survey using random sampling)

To calculate sample size for estimate of prevalence with 95% confidence limit

$$\boxed{N = \frac{1.96^2 \times (P)(1-P)}{d^2}}$$

1.96 = Z value for 95% confidence limits
P = Estimated prevalence (e.g. 0.3 for 30%)
d = Desired precision (e.g. 0.05 for ± 5%)

Steps for calculating the sample size



1. Decide on key indicator

2. Estimate prevalence of key indicator

3. Decide on precision required

4. Calculate initial sample size using formula

5. Adjust for individual non-response

6. Adjust for eligible members

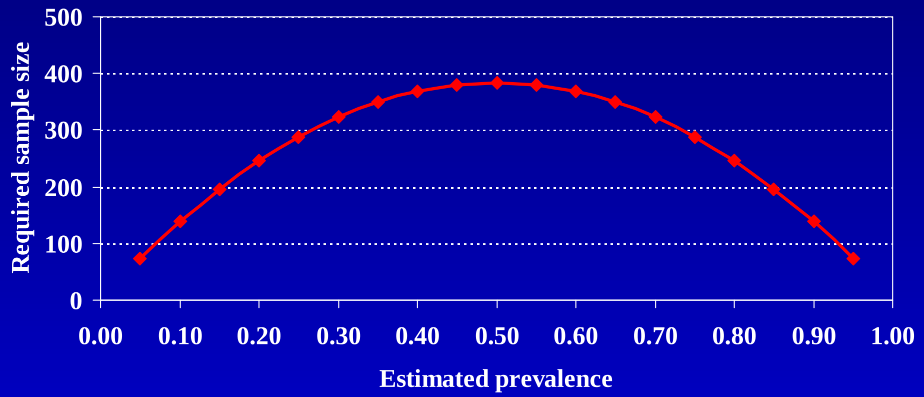
7. Adjust for household non-response

Final sample size

Step 2: Where do get information to make assumption about prevalence?

- Prior surveys
- Qualitative estimates
- “Worst case” scenario: use prevalence of 50%

Effect of Changing the Estimated Prevalence
(assume 95% CI, +/- .05, large population)



Step 3: How do we decide on the needed precision?

- One-time results for advocacy alone does not need much precision (±0.10 good enough)
- Results that you will need to compare against in the future need greater precision (±0.05 if program will have large impact)
- Results you will monitor frequently (e.g. year by year) need even greater precision (±0.02)

Step 4: Calculate required sample size

Survey anemia in children under-5

- Assume prevalence of anemia = 40% (0.40)
- Need precision of +/- 5% (0.05)

$$N = \frac{1.96^2 \times (P)(1-P)}{d^2}$$

1.96 = Z value for 95% confidence limits
P = Estimated prevalence
d = Desired precision

Step 4 (cont.): Calculate required sample size

$$N = \frac{1.96^2 \times (P)(1-P)}{d^2}$$

1.96 = Z value for 95% confidence limits
P = Estimated prevalence
d = Desired precision (for example, 0.05 for ± 5%)

$P = 0.4$

$(1-P) = 0.6$

$d = 0.05$

$$1.96^2 \times \frac{0.4 \times 0.6}{0.05^2} = \mathbf{369}$$

children

Step 5: Adjust sample size for individual non-response

- Some children will refuse or be unavailable
- Therefore, inflate sample size to account for **individual** non-response

Example: If 10% non-response (90% response):

Sample size x 0.9 = new sample size

369 / 0.90 = 410 children

Step 6: Adjust sample size for number of eligible individuals

- If the unit of analysis is the individual within the household there is a need to readjust the number of selected households

Example: If 0.7 children per household:

Sample size (children) / 0.7 = Number of HHs needed

$$\frac{410}{0.7} = \textbf{586 HHs}$$

Step 7: Adjust sample size for household non-response

- Some households will refuse or be unavailable
- Therefore, inflate sample size to account for **household** non-response

Example: If 15% non-response (85% response):

Sample size x 0.85 = new sample size

586 / 0.85 = 689 HHs

Final sample size

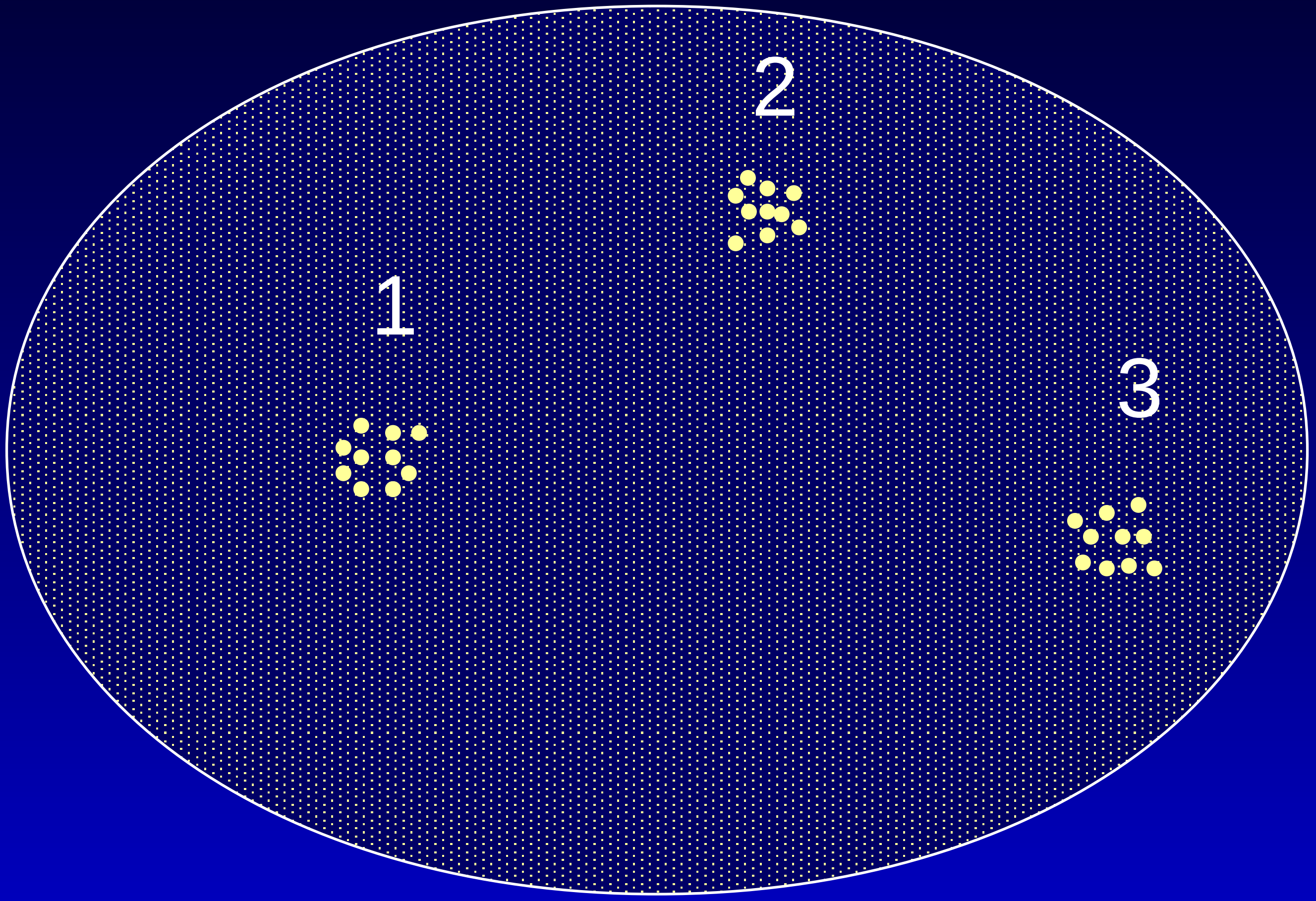
Sample size for more than 1 key indicator

- Decide what are the key indicators
- Calculate sample sizes for each individual key indicator
- Choose largest sample size required

Example:

Key indicator	Children <5	Households
Wasting	300	200
Stunting	600	400
% of HHs with poor food consumption	-	350

Single survey using two-stage cluster sampling



Introduction

- Cluster sampling results in loss of precision compared to simple random sampling.

- When calculating sample size, must increase sample size to obtain the same precision.

- When calculating confidence intervals during data analysis, must take into account cluster sampling.

Calculate sample size – single survey using two-stage cluster sampling

To estimate sample size, you need to know:

- Estimate of the prevalence of the key indicator (e.g. rate of stunting)
- Precision desired (for example: ± 5%)
- Level of confidence (always use 95%)
- Expected response rate
- *Population*
- *For nutrition surveys:* number of eligible individuals per household

IN ADDITION TO THE ABOVE, THE **DESIGN EFFECT**

6. Calculate sample size

Cluster surveys

To calculate sample size for estimate of prevalence with 95% confidence interval taking into account cluster sampling

$$N = \text{DEFF} \times \frac{1.96^2 \times (P)(1-P)}{d^2}$$

DEFF = Design effect

1.96 = Z value for p = 0.05 or 95% confidence intervals

P = Estimated prevalence

d = Desired precision (for example, 0.05 for ± 5%)

Design effect

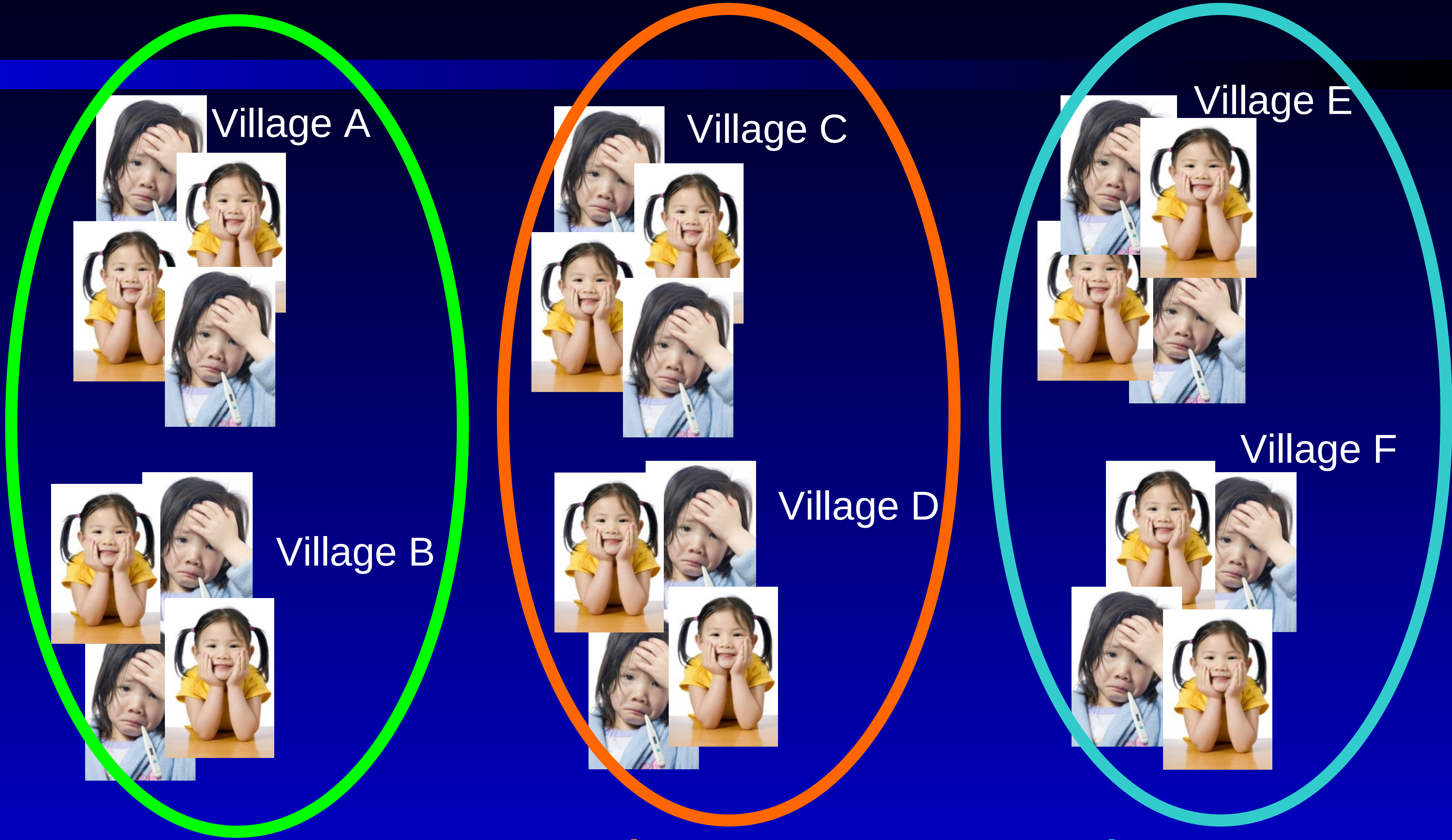
- Design effect increases when
  - Key indicators are highly geographically clustered (e.g. water source, access to health care)
  - When number of clusters are decreased and size of clusters are increased
- To minimize design effect
  - Include more clusters of smaller size
  - Stratify sample into more homogeneous groups

Example: Prevalence of malaria in 6 villages

= child with malaria

= child without malaria

Example 1: Malaria evenly spread throughout population



Village A

Village B

Village C

Village D

Village E

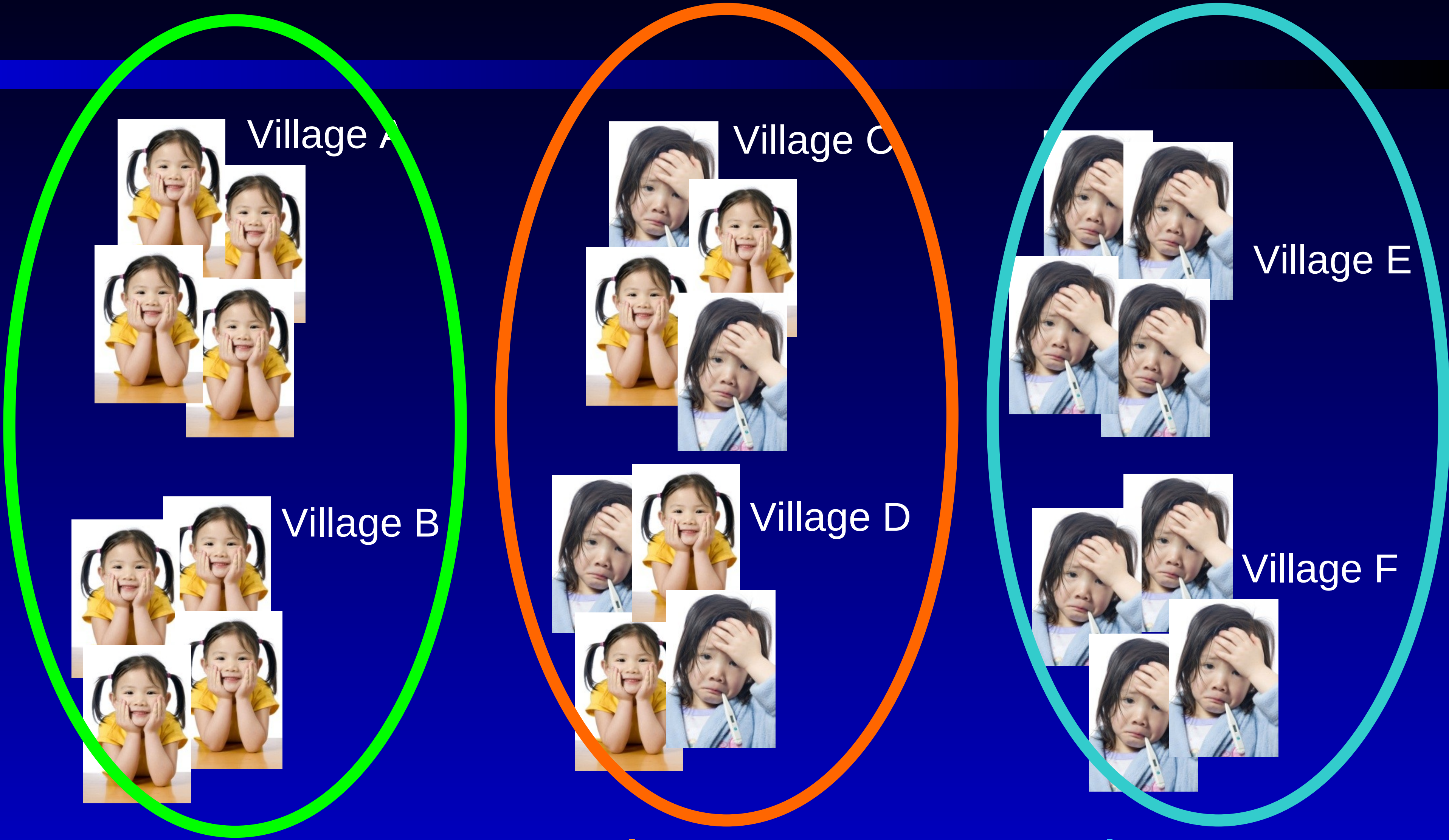
Village F

Prevalence = 50%

Prevalence = 50%

Prevalence = 50%

Example 2: Malaria unevenly spread throughout population (2 clusters)



Village A

Village B

Village C

Village D

Village E

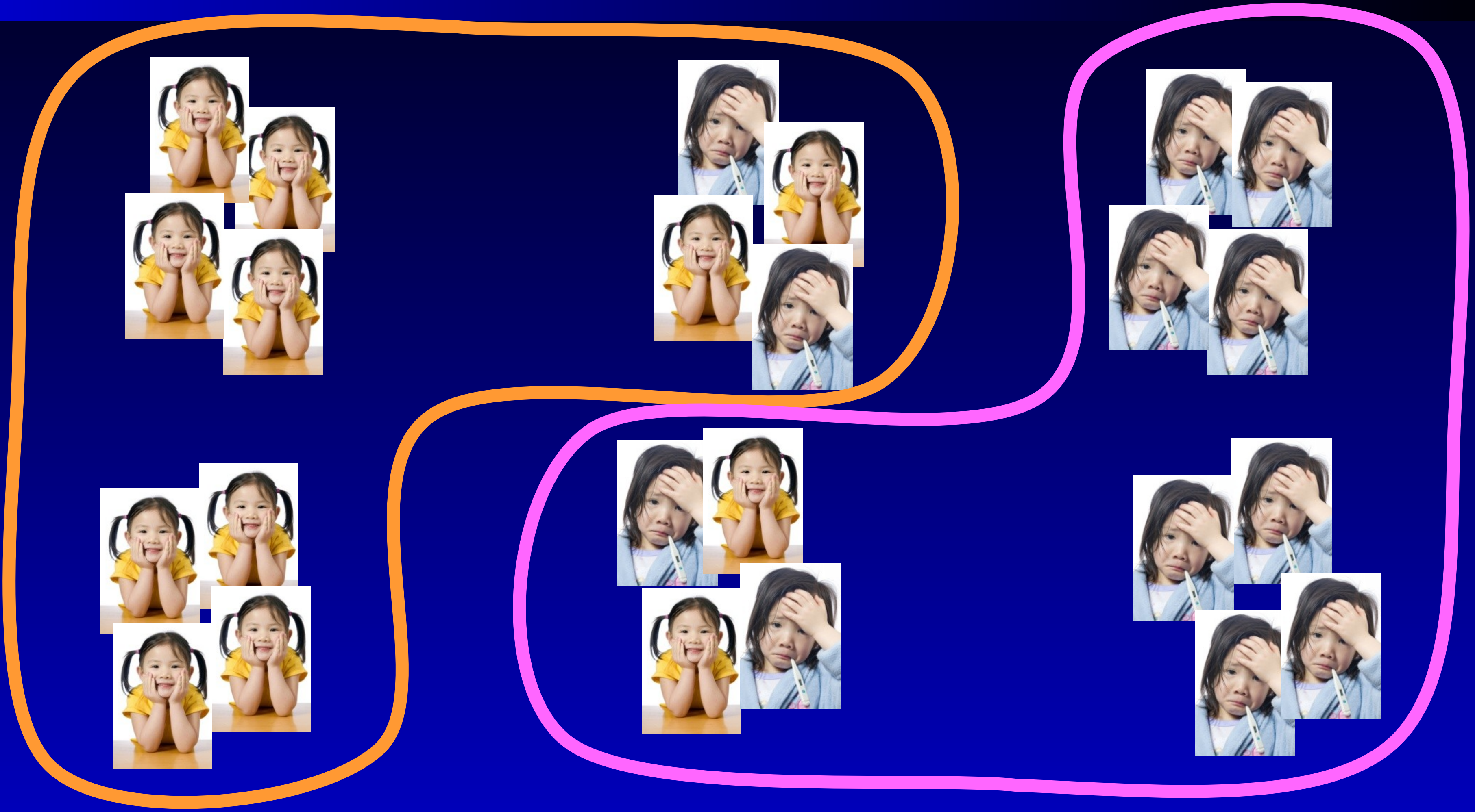
Village F

Prevalence = 0%

Prevalence = 50%

Prevalence = 100%

Example 3: Malaria unevenly spread throughout population (3 clusters)



Prevalence = 17%

Prevalence = 83%

Where do you get design effect to calculate sample size?

- Ideally prior surveys
- We often use ‘2’ as an estimate for a two-stage cluster sampling (comes from immunization coverage in rural Africa)

Example: Design effect for selected key indicators in Mongolia

Key indicator	Design effect assumed
Iodated salt	4.5
Stunting in children	1.5
Nutritional status in mothers	1.3
Anemia in children	2.1
Anemia in mothers	1.9

Why is this so high?

Example: Design effect for selected key indicators in Mongolia

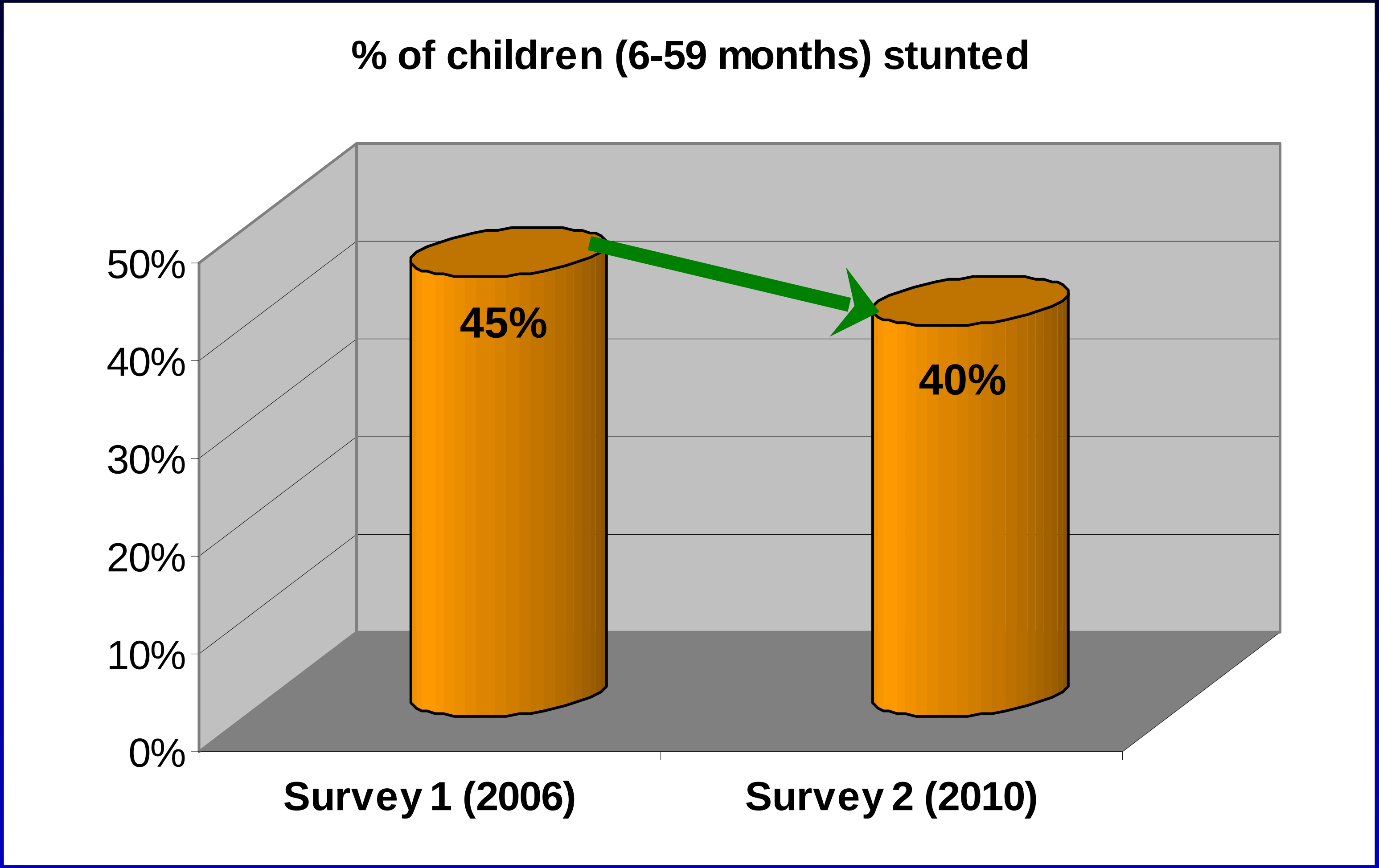
If sample size for simple random sampling = 120

What is sample size for a 2-stage cluster survey if iodated salt is key indicator (DEFF 4.5)?

How do we decide on how many clusters?

- More clusters of smaller size results in smaller design effect
- But more clusters increases cost and time required
- Fewer than **30 clusters** results in high design effect
- But >30 clusters doesn’t decrease design effect much

Comparing two surveys



Comparing two surveys - Introduction

- When comparing 2 surveys
  - Want to be sure any difference is not due only to sampling error
  - Uses different formula
- To calculate sample size, make following assumptions:
  - Estimated prevalence in survey 1
  - Decide difference between 2 surveys you want to be able to detect

Comparing two surveys - Formula

Calculate sample size to compare prevalence estimates from 2 surveys; both surveys with equal sample sizes

$$n = DEFF \times \frac{\left[1.96\sqrt{2\overline{p}\left(1-\overline{p}\right)} + 0.84\sqrt{p_1\left(1-p_1\right) + p_2\left(1-p_2\right)}\right]^2}{(p_1 - p_2)^2}$$

$n =$ sample size for each survey

$DEFF =$ design effect

$1.96 =$ z value for significance level of 0.05

$0.84 =$ z value for power of 0.8

$\overline{p} = \frac{p_1 + p_2}{2}$ (prevalence in combined surveys or)

$p_1 =$ prevalence in survey 1

$p_2 =$ prevalence in survey 2

Comparing two surveys – Formula (cont.)

But no need to memorize these equations:

- Found in most statistics books (e.g. FANTA sampling guidelines)
- Use computer to calculate sample sizes, including EpiInfo, nutri-survey, etc.

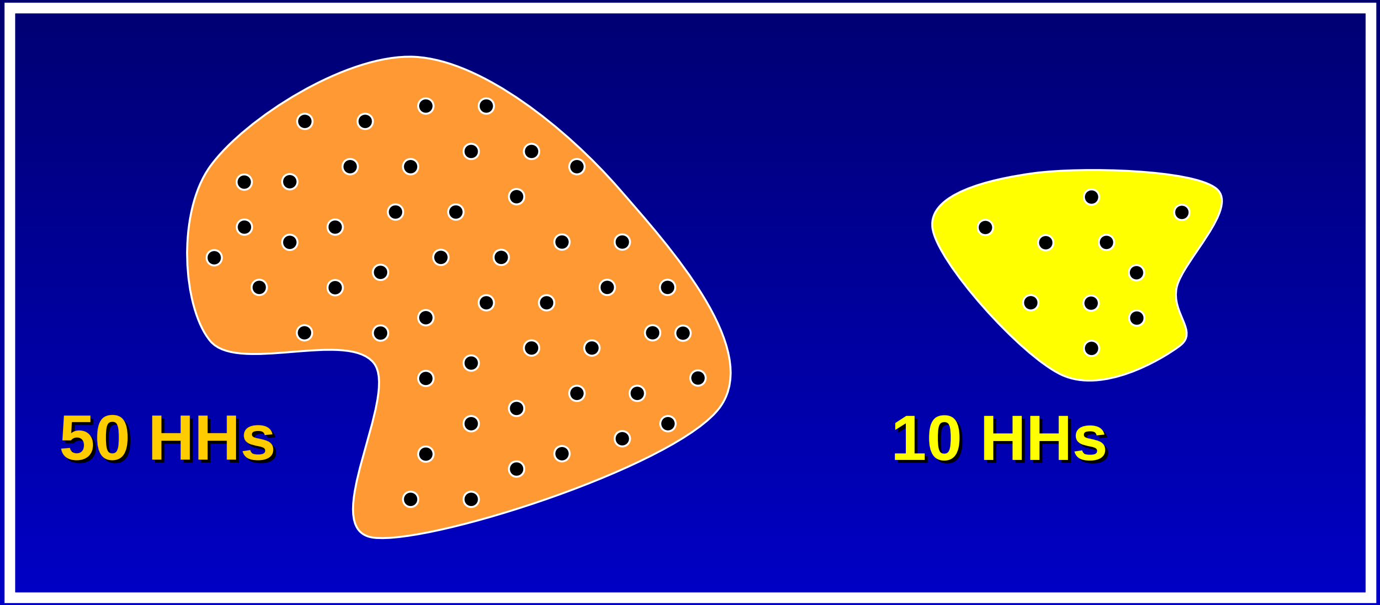
Two-stage cluster survey – Drawing the sample – First stage

First stage – Cluster requirements

- Size
  - Smaller clusters make second stage sampling easier
  - But not too small – each cluster should have the required minimum number of households
- Known boundaries
  - Must be able to tell if individual households belong to a cluster or not
  - Possible clusters: villages, blocks, enumeration areas

What does proportioal to population size (PPS) mean?

- Larger clusters are more likely to be chosen than smaller clusters
- If choose PSUs probability proportional to size (PPS), probability of any single household or person in population being chosen is the same



How to select clusters

Step 1: Make a list of all clusters

Step 2: Add a column with the population of each PSU

Step 3: Add a column with the cumulative populations

Step 4: Calculate the sampling interval as

$$\boxed{\frac{\text{Total population of all clusters}}{\text{Number of clusters in the sample}}}$$

Step 5: Select a random number between 1 and the sampling interval – where this number lies in the cumulative population column is the first cluster

Step 6: Add the sampling interval to the random number – this is the second cluster

Step 7: Continue to add the sampling interval to select clusters

Example: Select 5 out of 17 EAs (PPS)

Enumeration area	No of HHs	Cum no of HHs
A	250	250
[✓ B]	90	340
C	100	440
D	130	570
E	75	645
[✓ F]	490	1135
G	30	1165
H	280	1445
[✓ I]	270	1715
J	40	1755
K	160	1915
L	285	2200
[✓ M]	410	2610
N	200	2810
O	50	2860
[✓ P]	460	3320
Q	25	3345

- Sampling interval: 3345/5=669
- Select random start (randomly one number between 1 and 669; e.g. 300)
- Add sampling interval

1. selected EA	300
2. selected EA	968
3. selected EA	1636
4. selected EA	2304
5. selected EA	2972

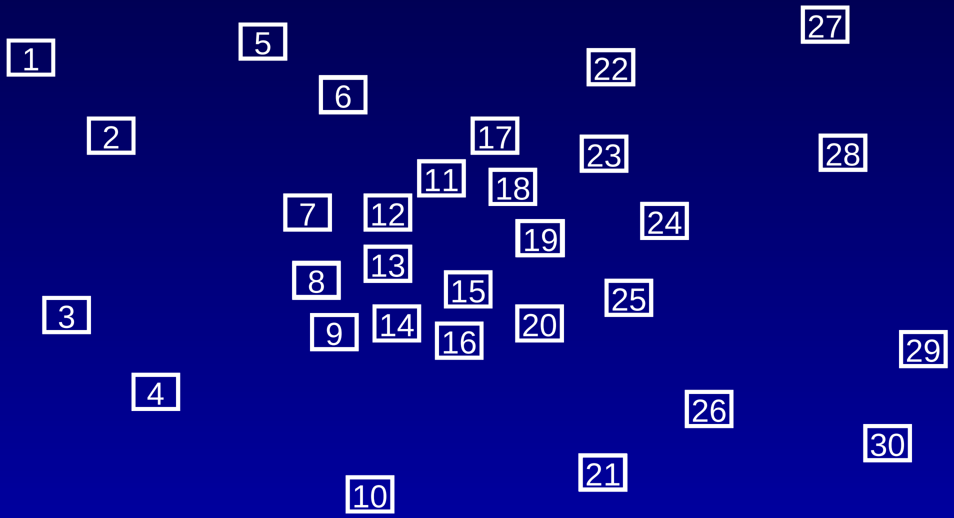
Two-stage cluster survey – Drawing the sample – Second stage

How would you select HHs within a cluster?

- **Option 1:** Simple random sampling
- **Option 2:** Systematic random sampling
- **Option 3:** EPI method (bottle spinning)
- **Option 4:** Spatial sampling
- **Option5:** Segmentation method

Example: Option 1

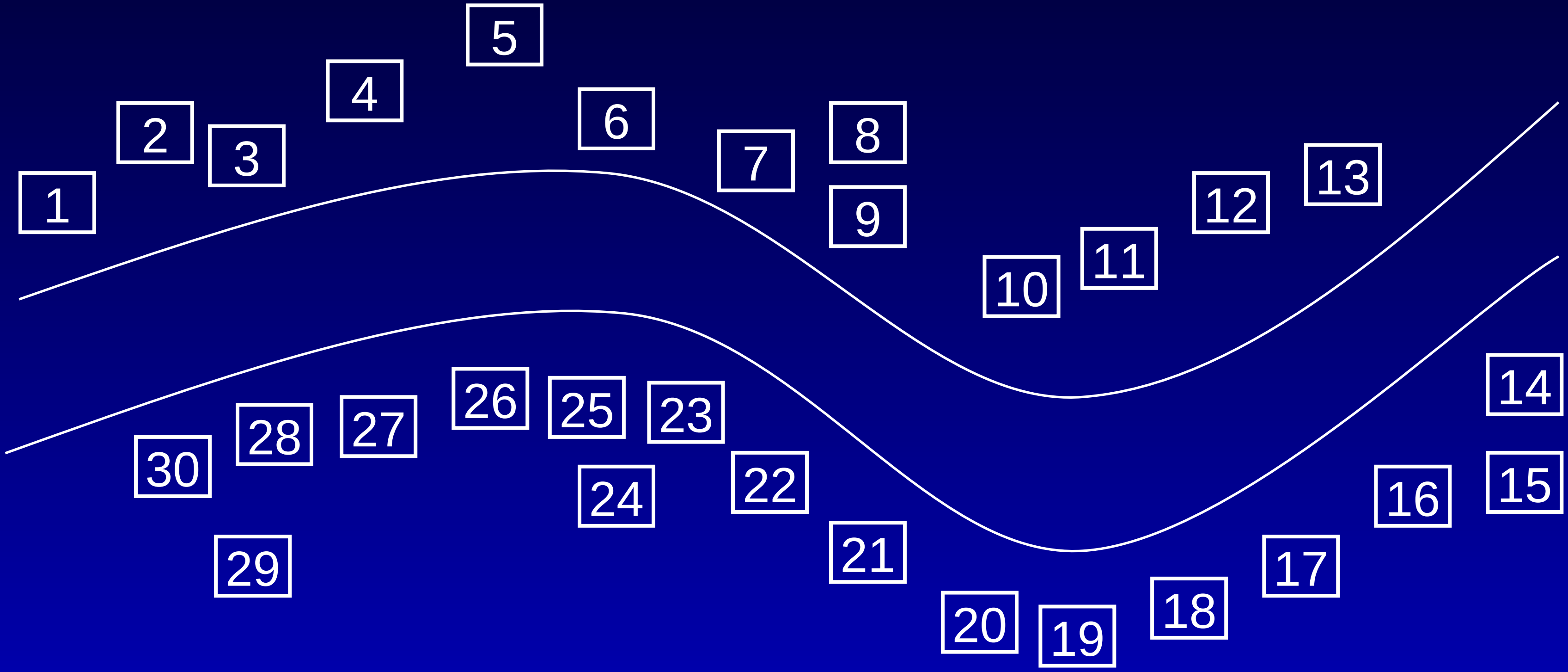
Simple or systematic random sampling of households



	HOUSEHOLD LIST
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
Etc.	

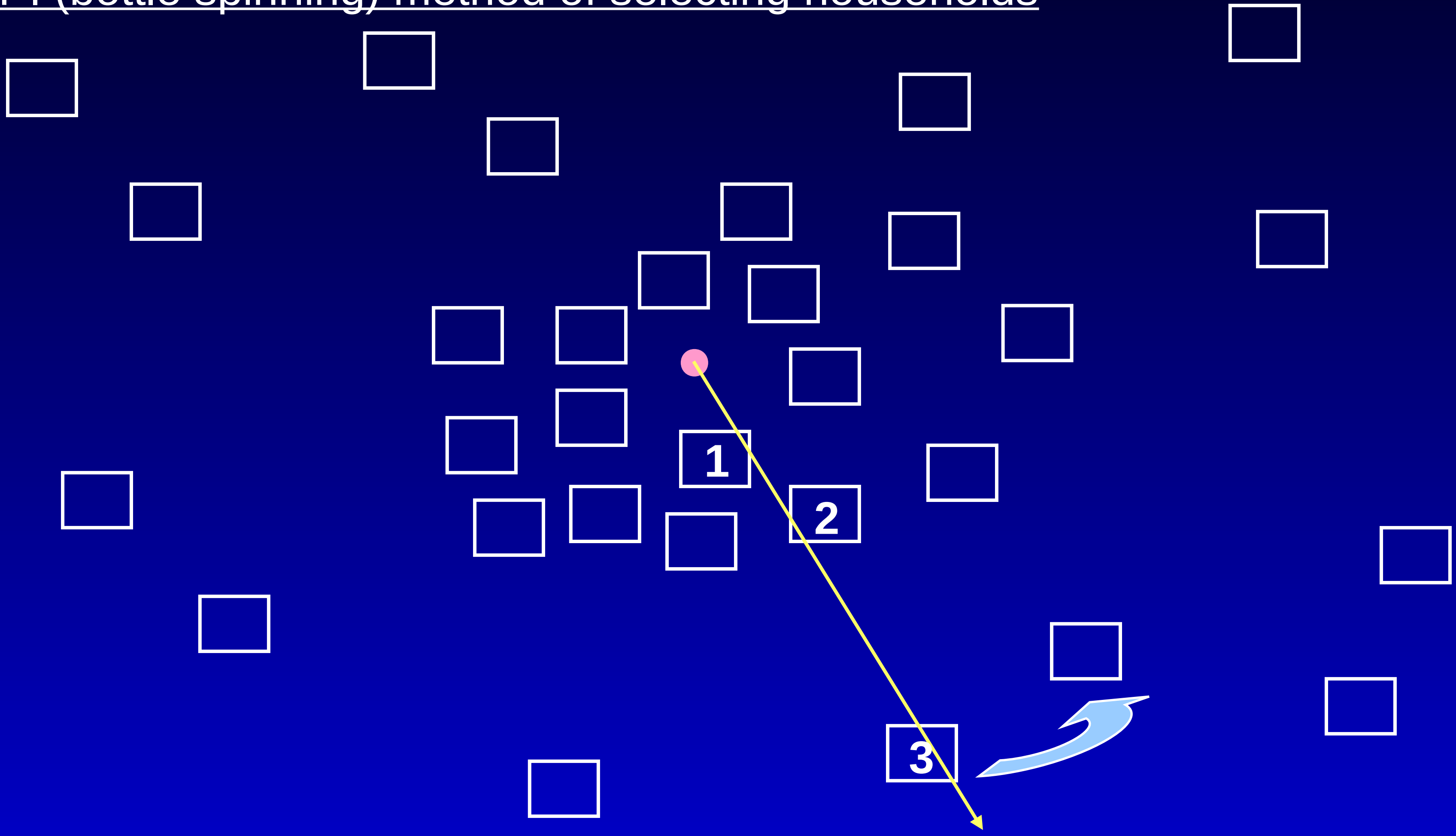
Example: Option 2

Systematic random sampling if HH list is not available



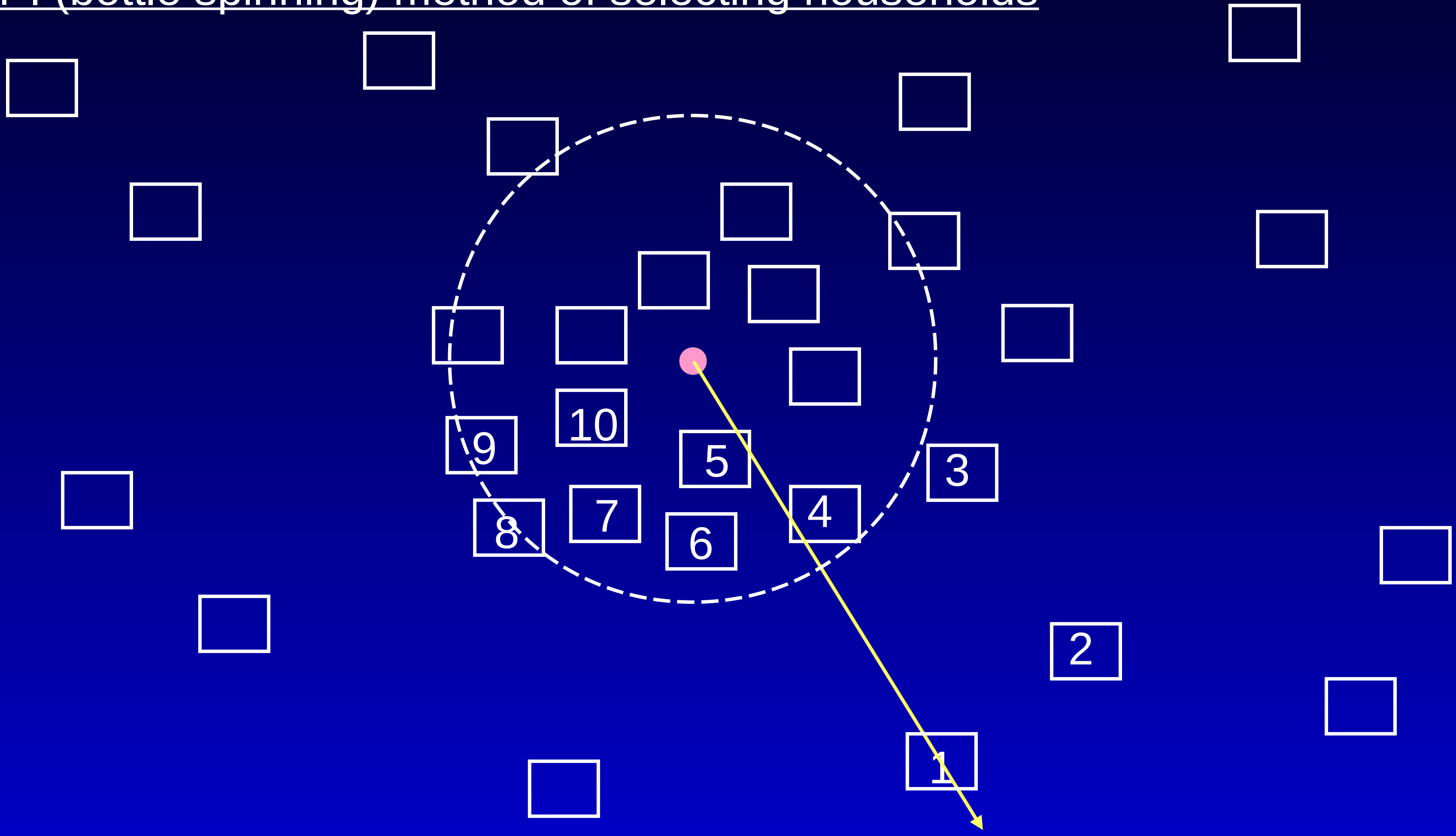
Example: Option 3

EPI (bottle spinning) method of selecting households



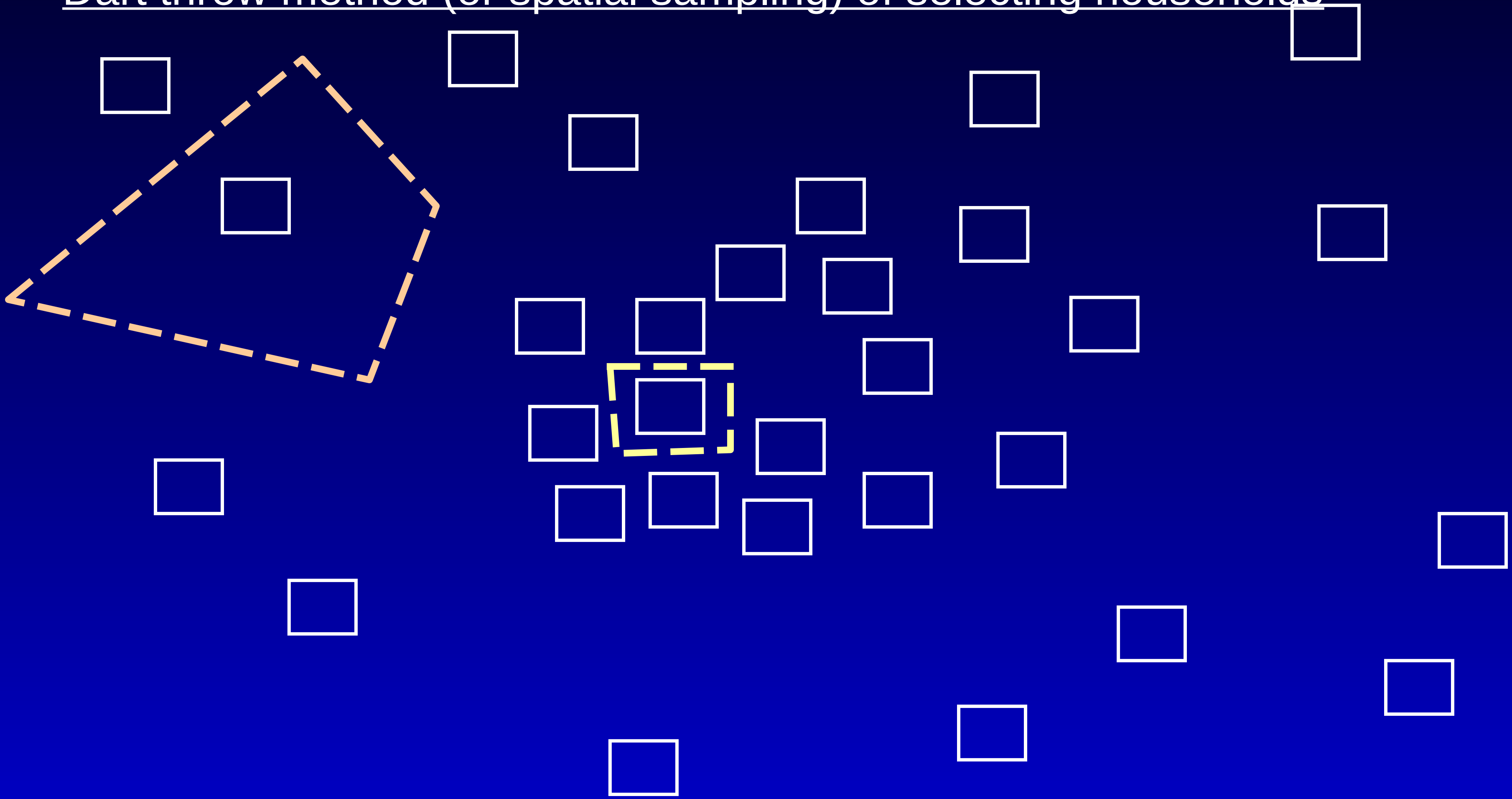
Example: Option 3 (limitations)

EPI (bottle spinning) method of selecting households



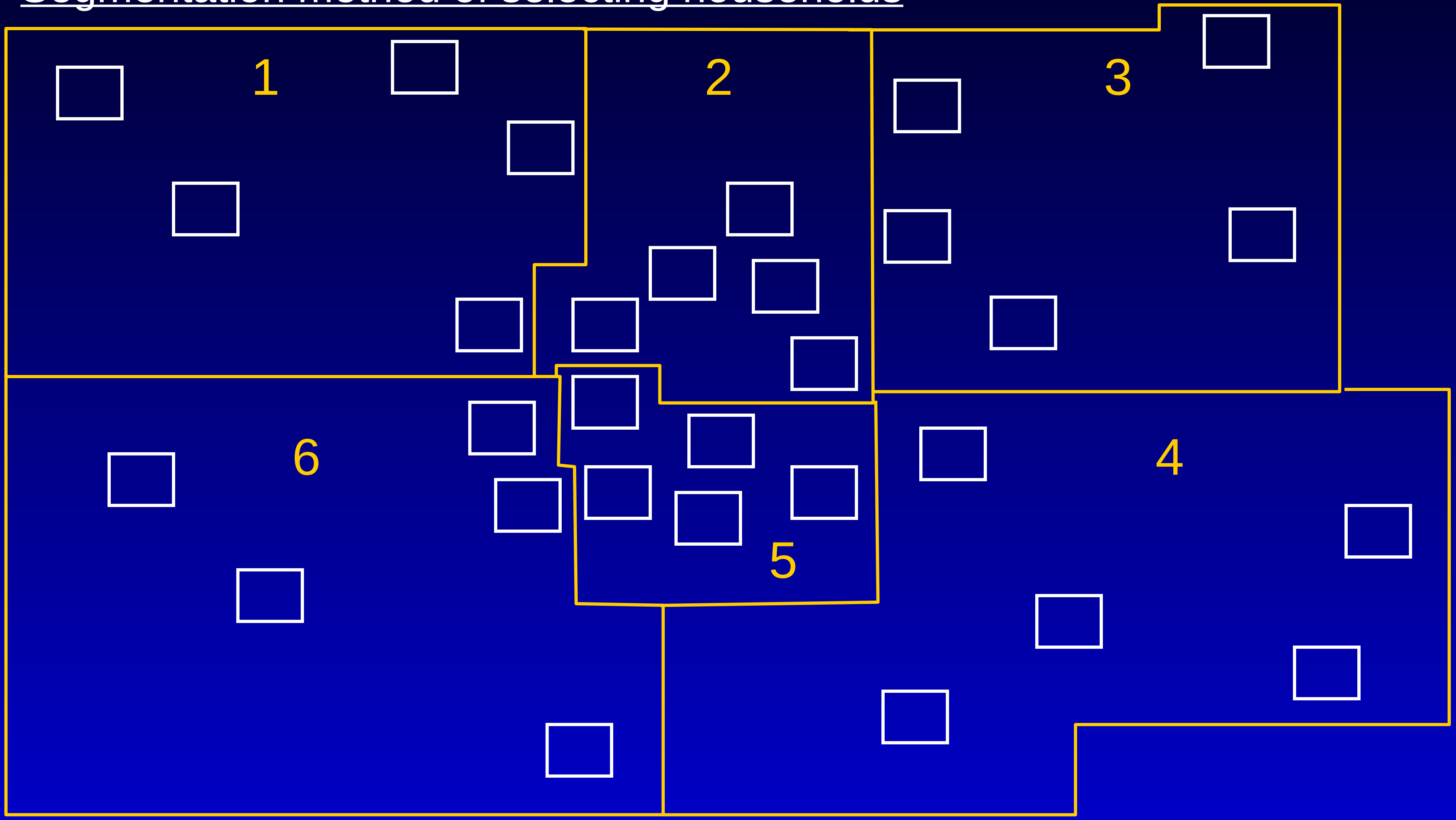
Example: Option 4

Dart throw method (or spatial sampling) of selecting households



Example: Option 5

Segmentation method of selecting households



How to select individuals within HHs?

In a household with more than 1 eligible child, include all in survey or only 1?

- In most cases, include all eligible persons in selected households

<u>Why? (Disadvantages of selecting 1 child)</u>

- Requires additional sampling step
- Produces biased sample
- Requires weighted analysis

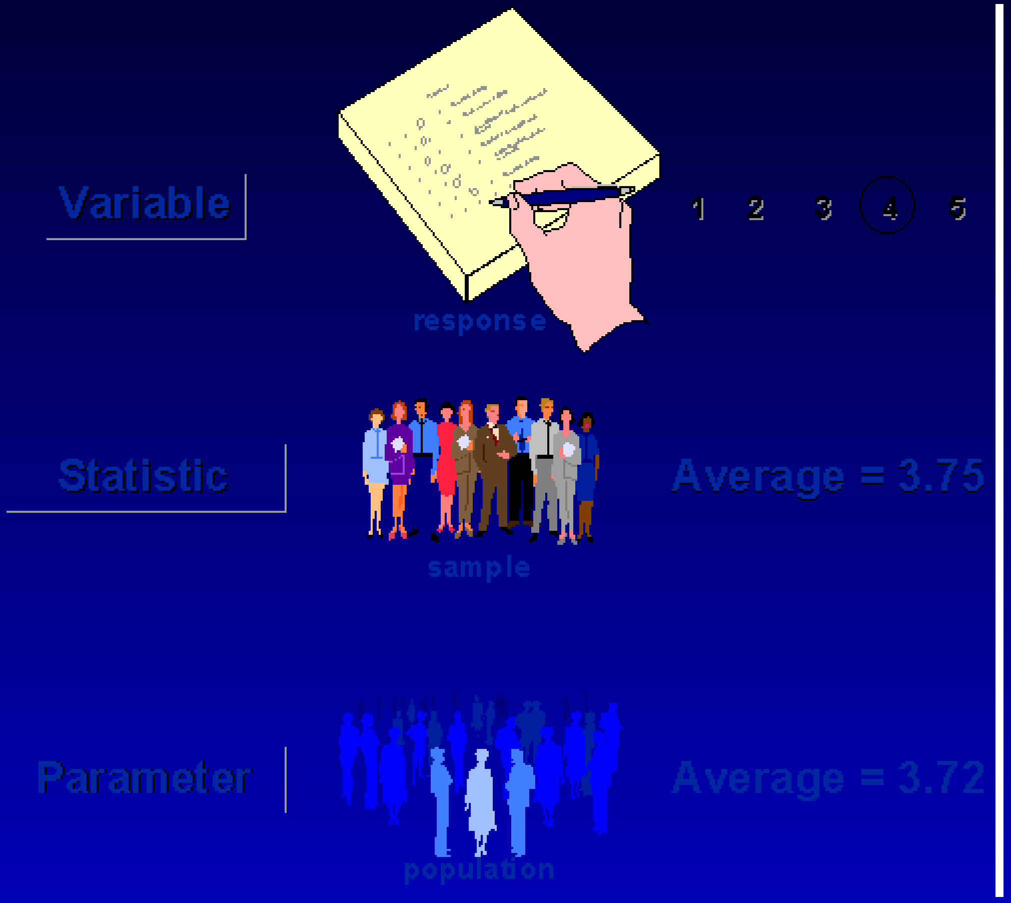
Replacing households (?)

If household unavailable, replace with neighboring household?

Recommendations

- Make an accurate estimate of household non-response when calculating sample size
- Do not replace missing households
- If necessary, select a few more households per cluster to use if household non-response is much greater than predicated

Developing a sampling plan



Variable

response

1 2 3 [✓ 4] 5

Statistic

sample

Average = 3.75

Parameter

population

Average = 3.72

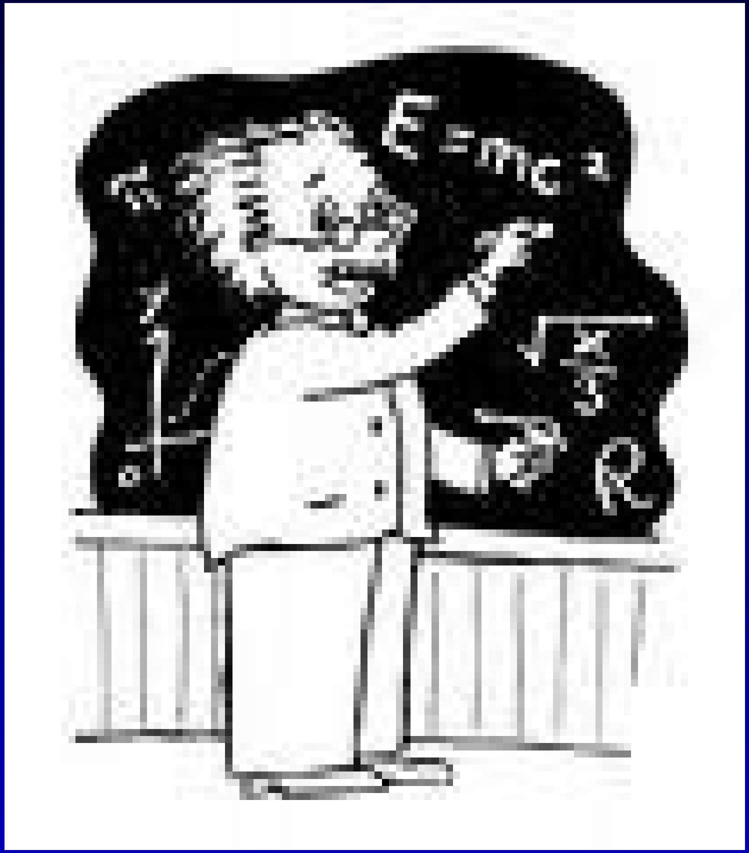
What information should a sampling plan contain

- Objective(s) of assessment/survey
- Sampling methodology
- Sample size calculation (including assumptions)
- Description of how sample is selected (1. stage/ 2. stage)
- Work plan and data collection staff required
- Who is responsible for what?

Resources for sampling

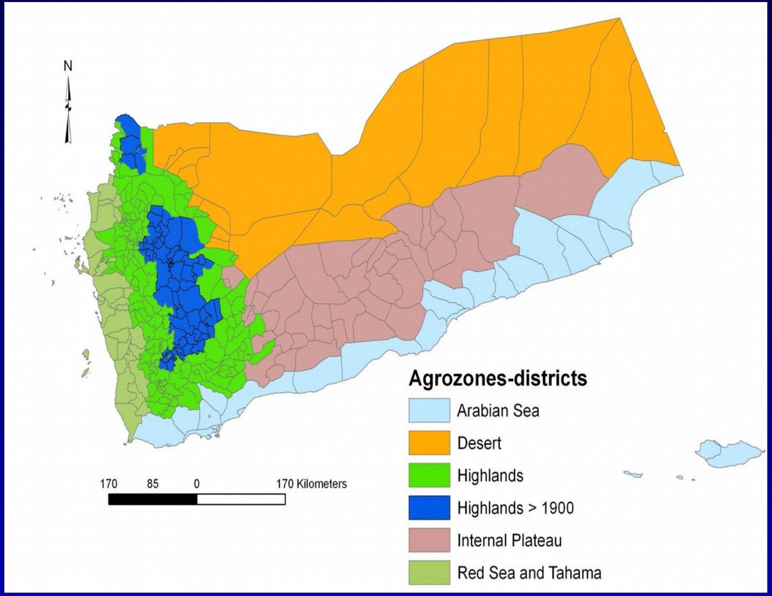
- **EFSA guidelines** **http://www.wfp.org/content/emergency-food-security-assessment-handbook**
- **CFSVA guidelines** http://www.wfp.org/content/comprehensive-food-security-and-vulnerability-analysis-cfsva-guidelines-first-edition
- **CDC/WFP Manual:** Measuring and Interpreting Malnutrition and Mortality http://pgm.wfp.org/index.php/Topics:Nutrition#i._Surveys
- **SMART guidelines** http://www.smartindicators.org/SMART_Methodology_08-07-2006.pdf
- **Software for emergency nutrition assessment:** http://www.nutrisurvey.de/ena_beta/frame.htm
- **OpenEpi sample size calculator:** http://www.sph.emory.edu/~cdckms/Sample%20Size%20Calculation%20for%20a%20proportion%20for%20cluster%20surveys.htm
- **Sampling guidelines for vulnerability analysis:** http://www.wfp.org/content/thematic-guidelines-sampling-guidelines-vulnerability-analysis
- **EFSA: T-square method:** http://www.wfp.org/content/technical-guidance-sheet-no11-using-t-square-sampling-method-estimate-population-size-demographics-a
- **FANTA Sampling Guide:** http://www.fantaproject.org/publications/sampling.shtml

Thank you



Final exercise: Developing sampling plan for case study

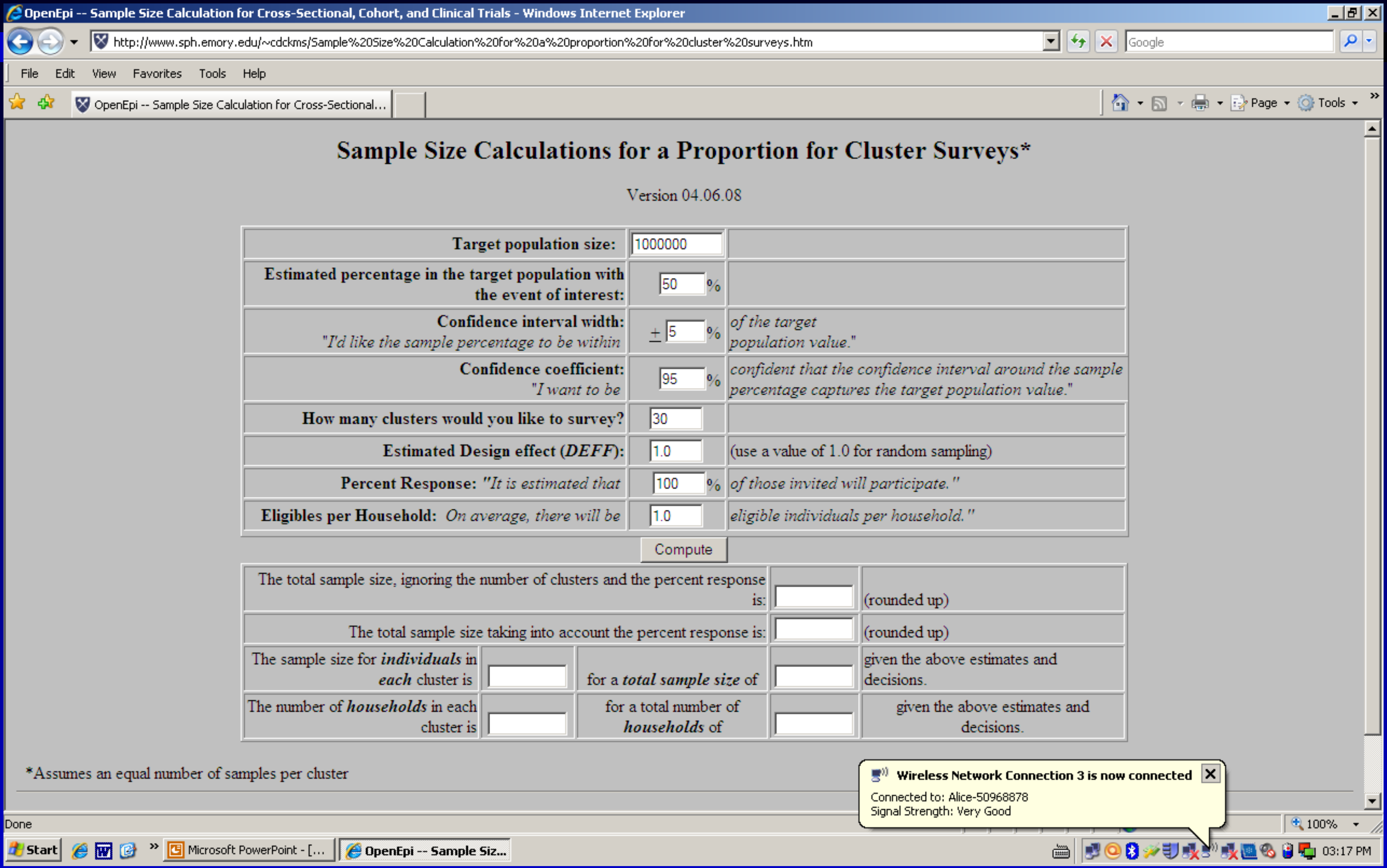
- **Case study:** You are requested to conduct a country-wide survey in Yemen with the objective to assess the level of HH food insecurity and child malnutrition representative at agro-zone level.
- **Key indicators** (prevalence from previous studies):
  - HHs with poor consumption = 11.8%
  - Children (6-59 months) wasted = 12.4%
  - Children (6-59 months) stunted = 53.1%
- **Other factors:**
  - On average, there is 1.2 children (6-59 months) in each HHs
  - It is estimated that 10% of HHs are not available or refuse to respond
  - Survey should take no longer than 4 weeks



Final exercise: Tasks

1) Decide on appropriate sampling method
2) Calculate required sample size taking the key indicators into account
   - Decide on required level of precision
   - Use OpenEpi sample size calculator
3) Estimate how many enumerators will be required if average travel time between clusters is one day and 1 enumerator can complete 5 questionnaires per day
4) Prepare a sampling plan (bullet points)

OpenEpi Sample Size calculator



OpenEpi -- Sample Size Calculation for Cross-Sectional, Cohort, and Clinical Trials - Windows Internet Explorer
http://www.sph.emory.edu/~cdckms/Sample%20Size%20Calculation%20for%20a%20proportion%20for%20cluster%20surveys.htm
Google
File
Edit
View
Favorites
Tools
Help
OpenEpi -- Sample Size Calculation for Cross-Sectional...
Page
Tools
Done
100%
Start
Microsoft PowerPoint - [...
OpenEpi -- Sample Siz...
03:17 PM

Sample Size Calculations for a Proportion for Cluster Surveys*

Version 04.06.08

Target population size:	1000000	
Estimated percentage in the target population with the event of interest:	50 %	
Confidence interval width: *"I'd like the sample percentage to be within*	± 5 %	*of the target population value."*
Confidence coefficient: *"I want to be*	95 %	*confident that the confidence interval around the sample percentage captures the target population value."*
How many clusters would you like to survey?	30	
Estimated Design effect (*DEFF*):	1.0	(use a value of 1.0 for random sampling)
Percent Response: *"It is estimated that*	100 %	*of those invited will participate."*
Eligibles per Household: *On average, there will be*	1.0	*eligible individuals per household."*

Compute

The total sample size, ignoring the number of clusters and the percent response is:				(rounded up)
The total sample size taking into account the percent response is:				(rounded up)
The sample size for ***individuals*** in ***each*** cluster is		for a ***total sample size*** of		given the above estimates and decisions.
The number of ***households*** in each cluster is		for a total number of ***households*** of		given the above estimates and decisions.

*Assumes an equal number of samples per cluster

Wireless Network Connection 3 is now connected
Connected to: Alice-50968878
Signal Strength: Very Good