AGRODEP

Household survey data course

Dakar, 8-10 October 2012

Random sampling

IFPRI®

HarvestChoice

BETTER CHOICES, BETTER LIVES

THE WORLD BANK

LSMS

Living Standards Measurement Study

Random Sampling

- Random Sampling (a.k.a. Scientific Sampling) is a selection procedure that gives each element of the population a non-zero, known positive probability of being included in the sample

- Mathematical theory is available to predict the probability distribution of the Sampling Error (*the error caused by observing a sample instead of the whole population*) and Confidence Intervals

- Other sampling procedures (purposive sampling, quota sampling, etc.) cannot do that

Random Sampling techniques

- **Single stage**, equal probability sampling
  - Simple Random Sampling (SRS)
  - Systematic sampling with equal probability
- **Multi-stages sampling**
- **Stratified sampling**

In real life those techniques are usually combined in various ways – most sampling designs are complex

Techniques in Random Sampling

Single stage, equal probability sampling/1

- Random selection of n "units" from a population of N units, so that each unit has an equal probability of selection
  - N (population) → n (sample)
  - Probability of selection (sampling fraction) = f = n/N

Is the most basic form of probability sampling and provides the theoretical basis for more complicated techniques

Techniques in Random Sampling

Single stage, equal probability sampling/2

- Advantage
  - self-weighting (simplifies the calculation of estimates and variances)
- Disadvantages
  - A list of all households in the country is generally not available to select the sample from (in other words, we don't have a good sample frame)
  - Difficult management
  - May entail high transportation costs

Techniques in Random Sampling

Single stage, equal probability sampling/3

1. **Simple Random Sampling**. The investigator mixes up the whole target population before grabbing "n" units. A Simple Random Sample is self-weighted

2. **Systematic Random Sampling.** The N units in the population are ranked 1 to N in some order (e.g., alphabetic). To select a sample of n units, calculate the step *k* (k= N/n) and take a unit at random, from the 1st k units and then take every k^{th} unit.

Sample variance & standard error

- Uncertainty is measured by the standard error ($\hat{e}$).
- Variance of the sample mean of an SRS of 'n' units for a population of size 'N':

$$\hat{e}^2 = Var(\bar{x}) = \left(\frac{N-n}{N-1}\right)\frac{Var(X)}{n} \approx \left(1-\frac{n}{N}\right)\frac{Var(X)}{n}$$

- Measure of sampling error. Depends on 3 factors:
  - (1 - n/N) = Finite Population Correction (fpc)
  - n = sample size
  - $Var(X)$ = Population variance. Unknown, but can be estimated without bias by:

$$\hat{s}_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

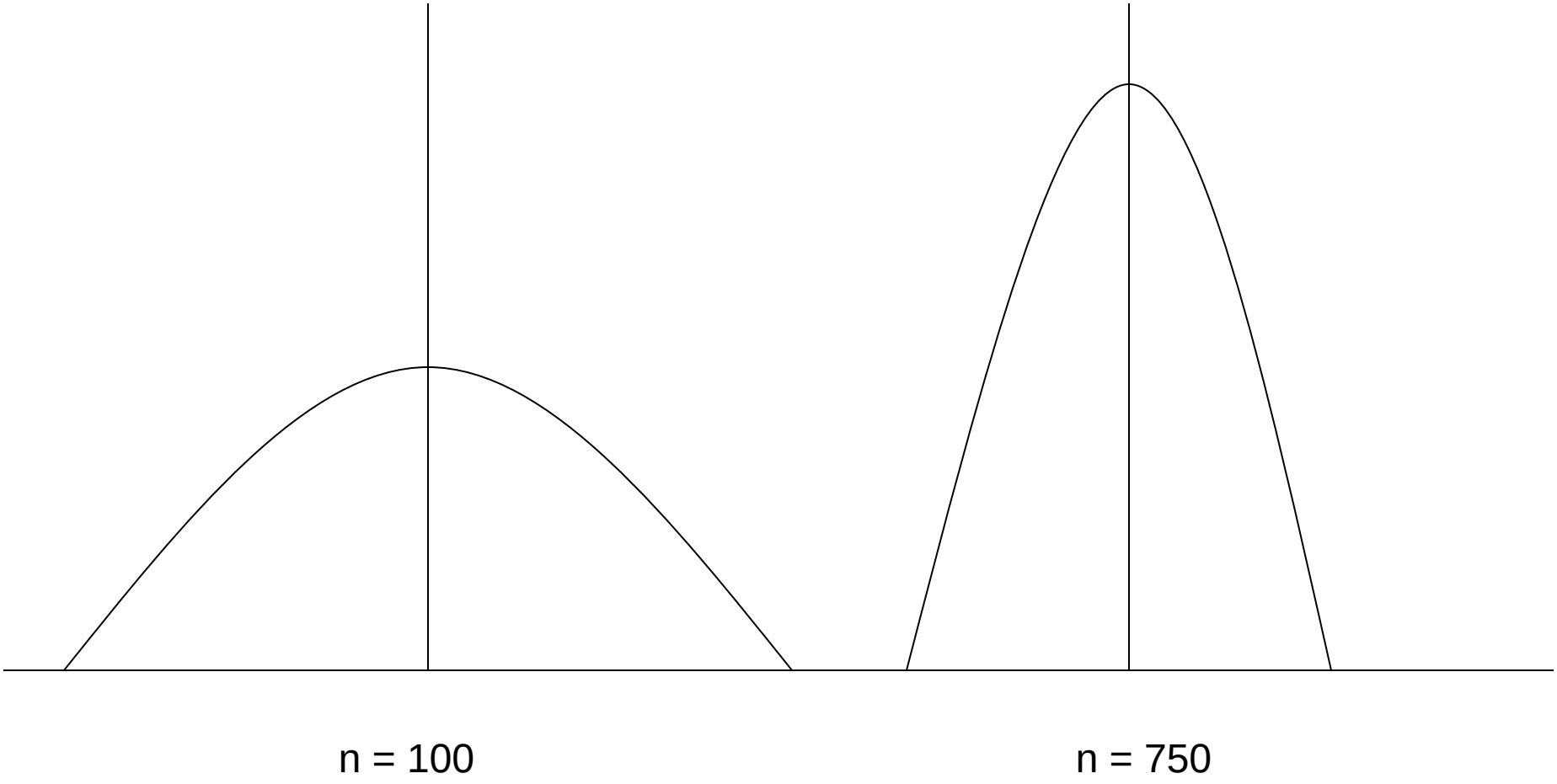
Standard Deviation vs Standard Error

Population	Sample
σ^2 = variance of the population	s^2 = variance of the sample
$\frac{\sigma}{\sqrt{N}}$ = standard deviation around the mean	$\frac{s}{\sqrt{n}}$ = standard error

Difference: The standard deviation is a descriptive statistic. It expresses the degree to which individuals in the population differ from the mean of the population. The standard error is an estimate of how close to the population mean your sample mean is likely to be.

Standard errors decrease with sample size. Standard deviations are left unchanged.

Sample Standard Error



Bigger samples have smaller standard errors around the mean

Sample Variance for Proportions

- A proportion *P* (or prevalence) is equal to the mean of a dummy variable.
- In this case *Var(P) = P(1-P)*, and

$$Var(\hat{p}) \approx \frac{\hat{p}(1-\hat{p})}{n-1}$$

Confidence intervals

- Estimates obtained from random samples can be accompanied by measures of the uncertainty associated with the estimate called confidence intervals.

- It is not sufficient to simply report the sample proportion obtained by a candidate in the sample survey, we also need to give an indication of how accurate the estimate is.

Confidence intervals for proportions

In a sample of 1,000 electors, 280 of them (28 percent) say they will vote Green.

$$e \approx \sqrt{Var(\hat{p})} \approx$$

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{\frac{0.28 \times 0.72}{999}}$$

Standard error is 1.42 percent.

Confidence intervals for averages

$$\bar{x} \pm t_{\alpha} \times \hat{e}(\bar{x})$$

where:

t_{α} = 1.28 for confidence level α = 80%

t_{α} = 1.64 for confidence level α = 90%

t_{α} = 1.96 for confidence level α = 95%

t_{α} = 2.58 for confidence level α = 99%

Sample Size

The required sample size n is determined by

- The variability of the parameter ***Var(X)***

Though this is unknown…

- The maximum margin of error ***E*** we are willing to accept
- How confident we want to be in that the error of our estimation will not exceed that maximum

For each confidence level α there is a coefficient t_α

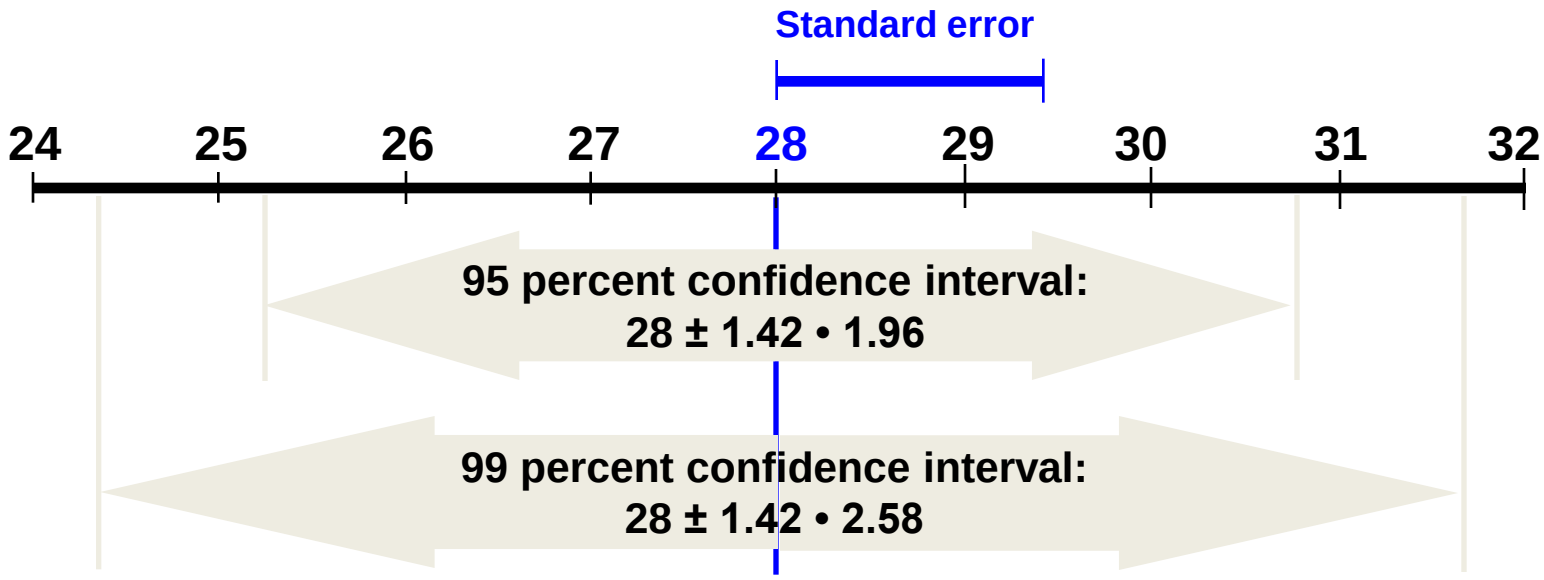
- The size of the population

(not very important)

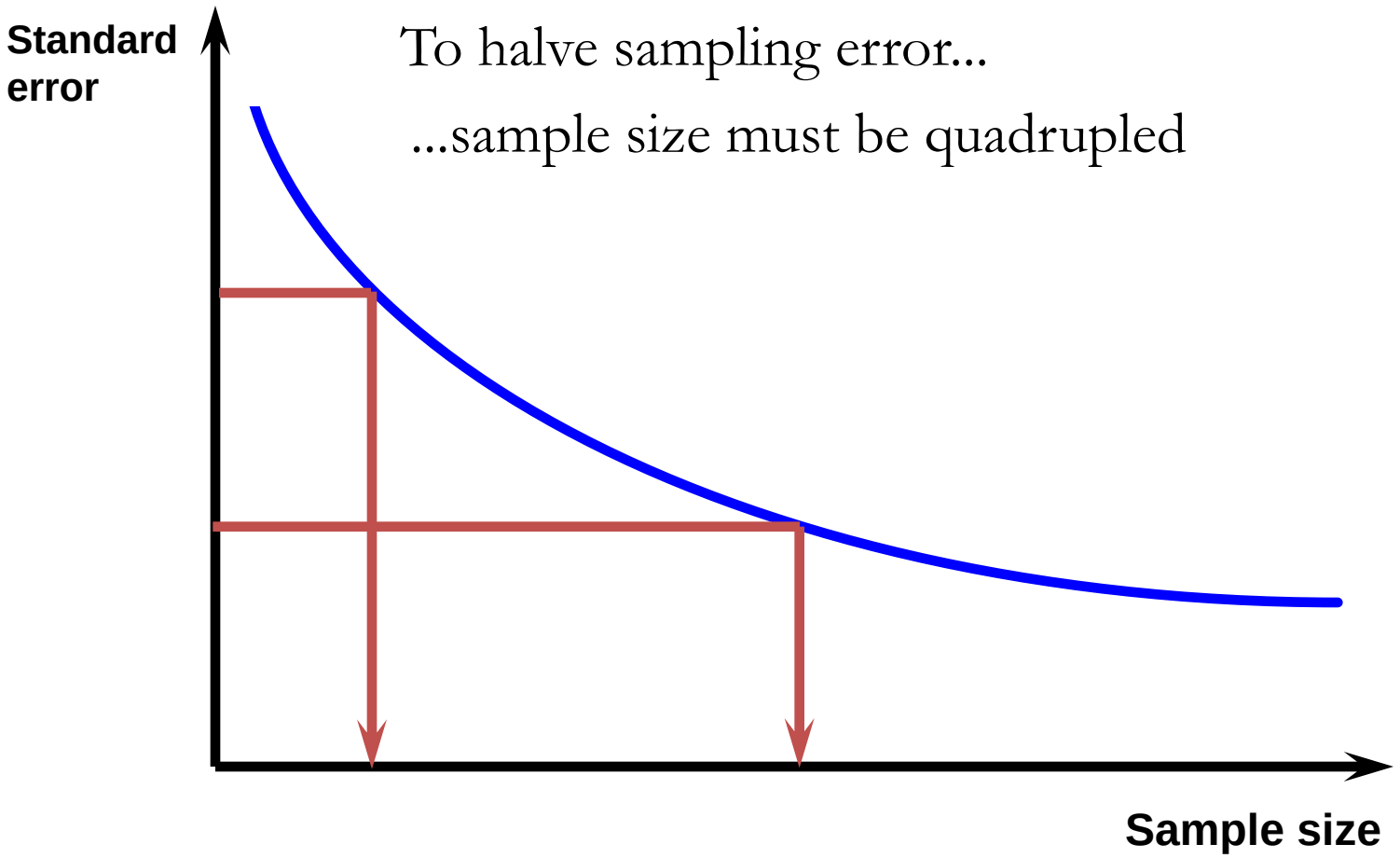
$$n_\infty = \frac{t_\alpha^2 \times Var(X)}{E^2} \qquad n_\infty = \frac{t_\alpha^2 \times P(1-P)}{E^2} \qquad n_N = \frac{n_\infty}{1 + n_\infty / N}$$

Confidence intervals

In a sample of 1,000 households, 280 households (28 percent) say they will vote Green. Standard error is 1.42 percent.



Sampling error and sample size

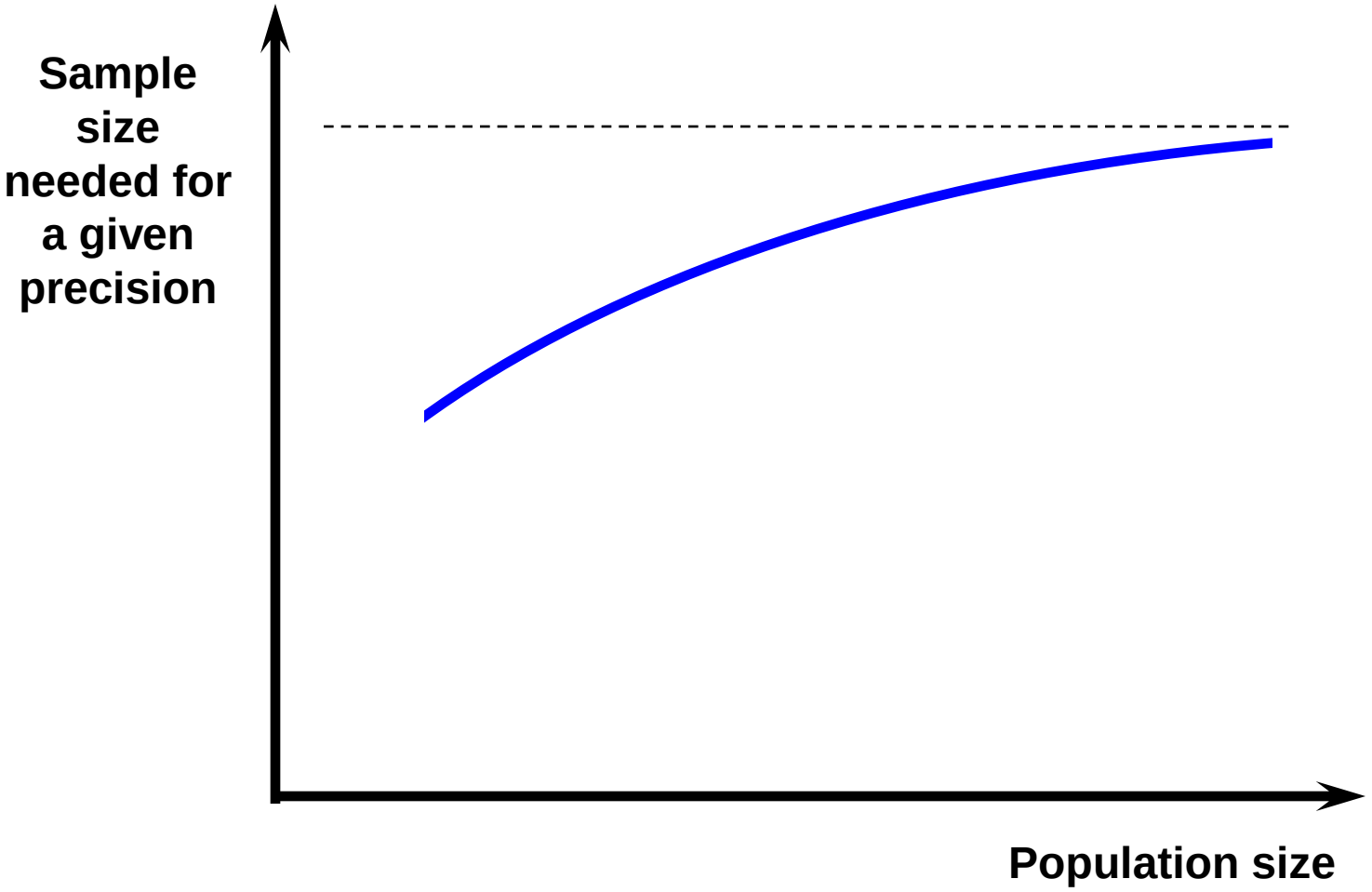


Sample size and population size

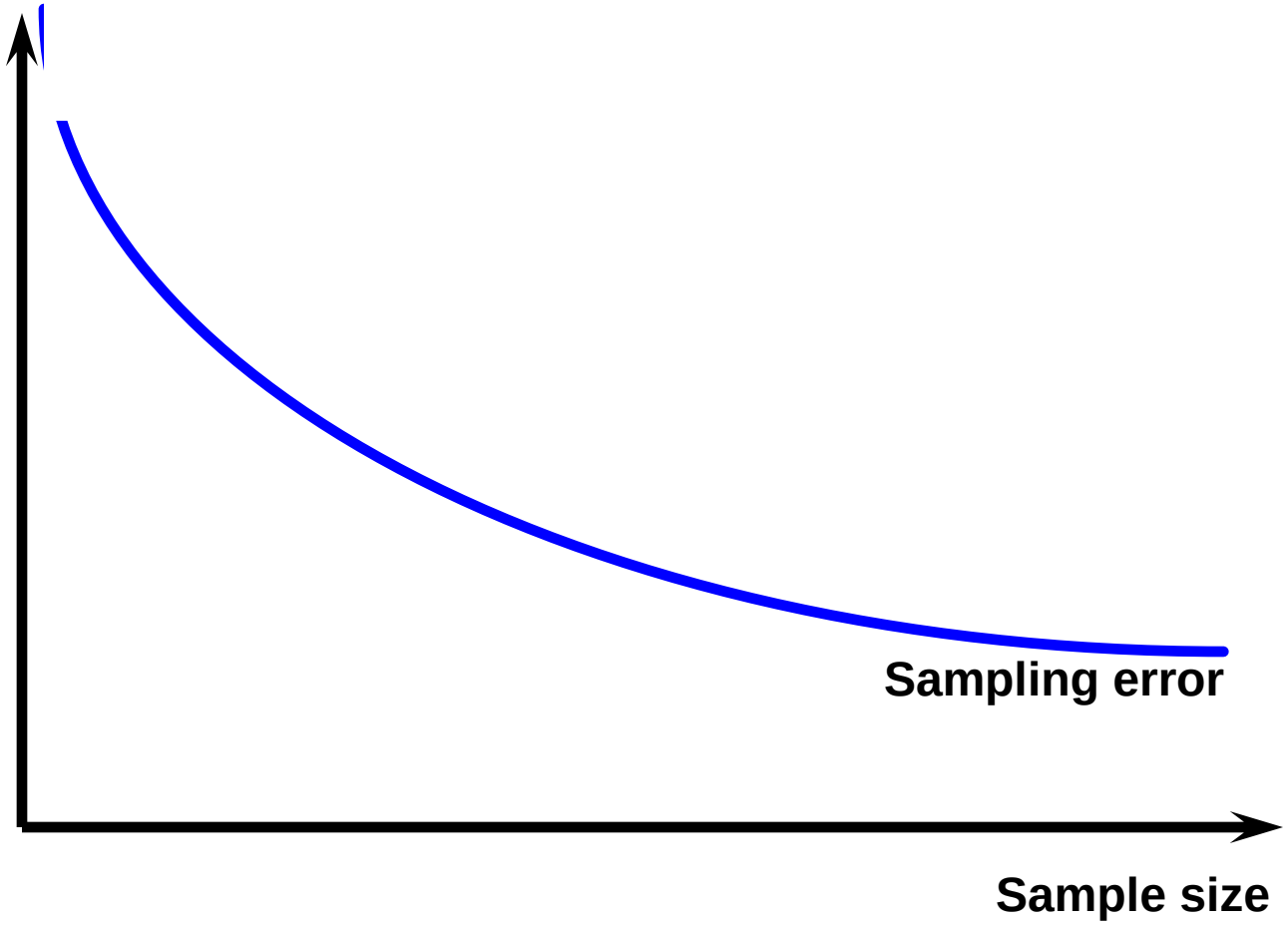
Standard error e when estimating a prevalence P in a sample of size n taken from a population of size N

$$e = \sqrt{1 - \frac{n}{N}} \sqrt{\frac{P(1-P)}{n}}$$

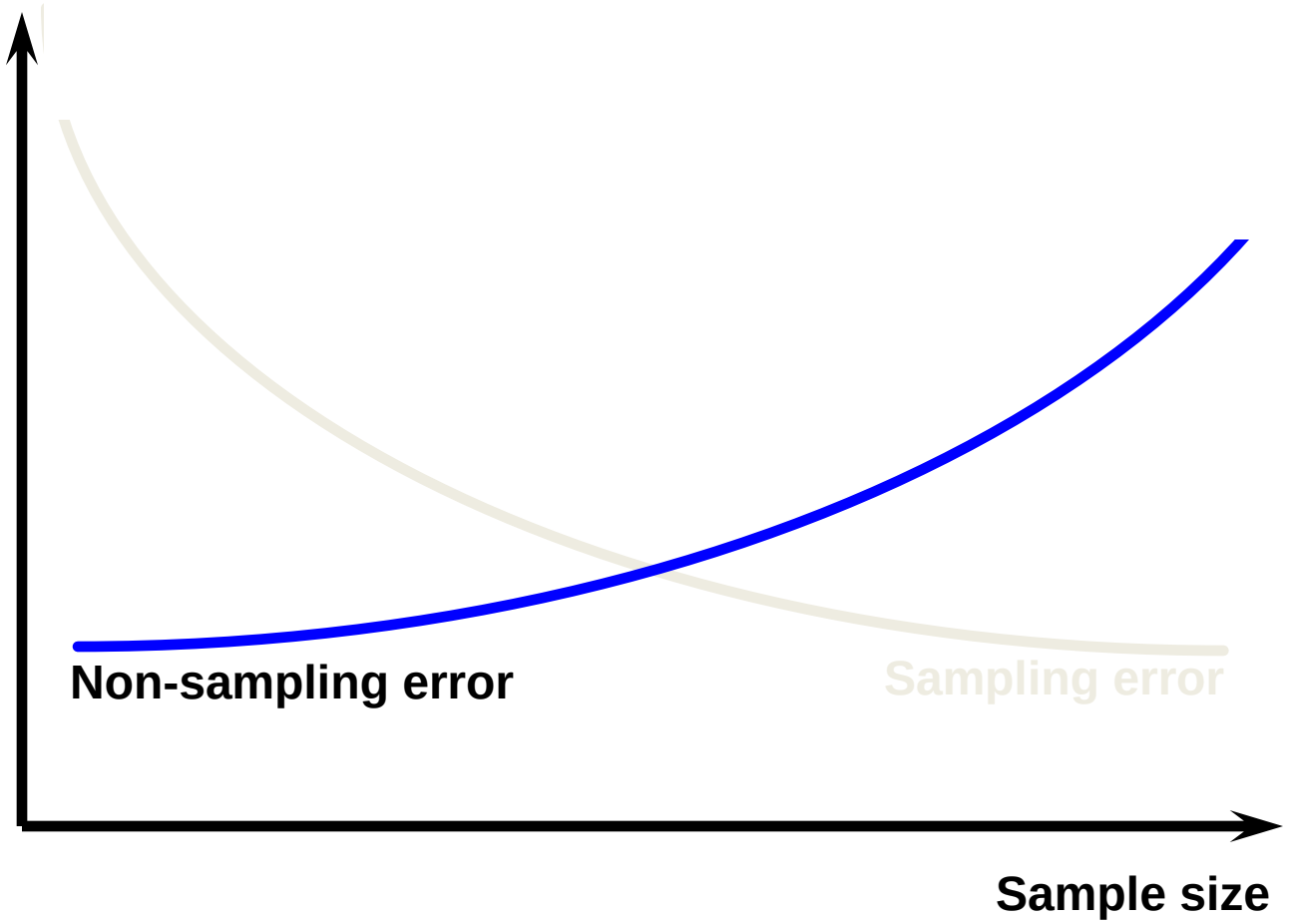
Sample size and population size



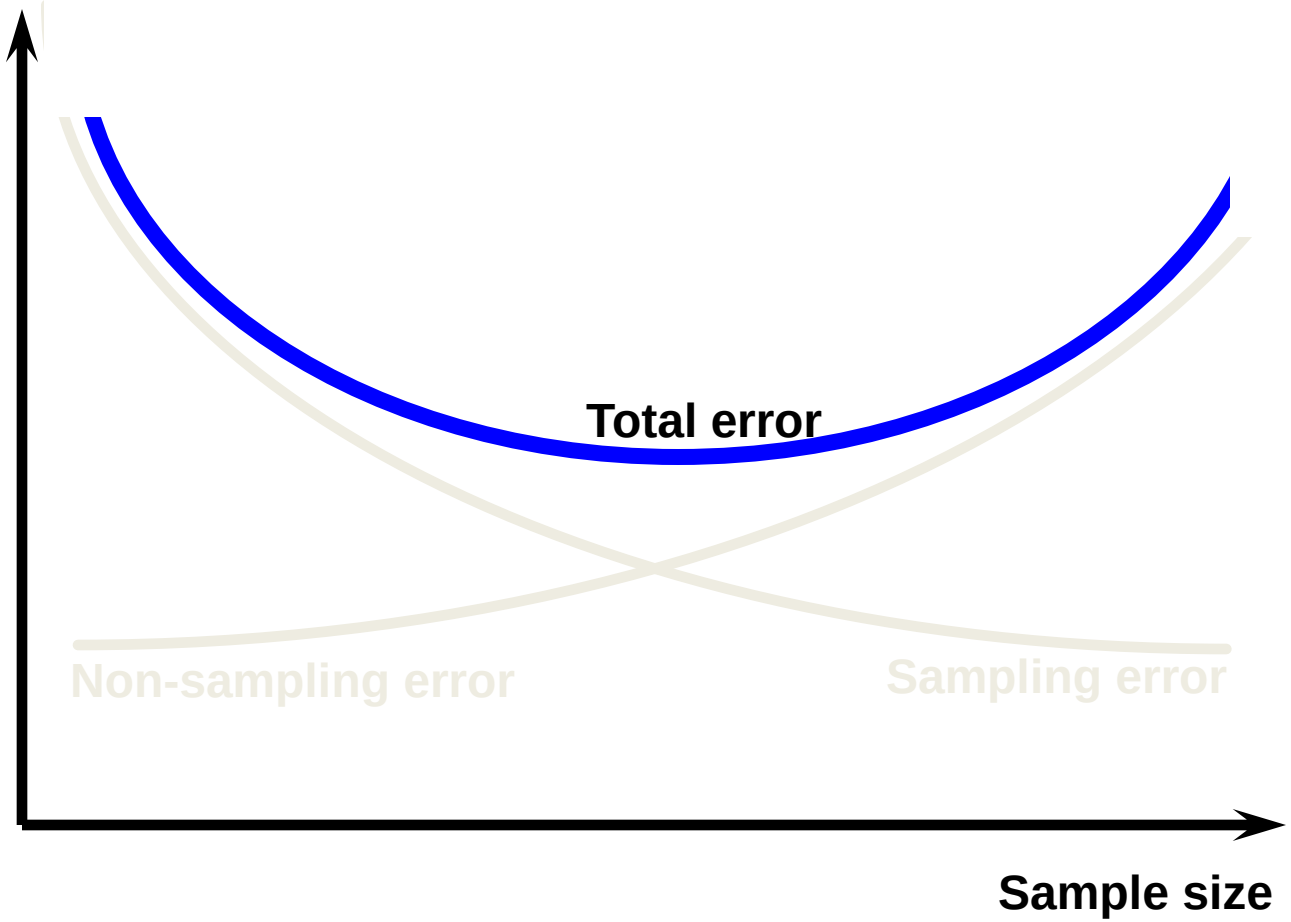
Sampling vs. non-sampling errors



Sampling vs. non-sampling errors



Sampling vs. non-sampling errors



Sample size

- Variability of factor in population (unknown)
- Acceptable margin of error
- Degree of confidence

sample design

Sample size (cont'd)

- As sample size increases, the distribution of the sample estimates becomes more concentrated around the expected value (CLT)
  - Level of precision increases
  - Confidence intervals become more narrow
- Standard error inversely proportional to square root of sample size

Determining sample size needed for specified level of precision

- Need to estimate the standard deviation (S or σ)

- In the case of a proportion, need to determine the approximate value of the proportion P being estimated
  - Maximum sample size needed with P = 0.5 – conservative value

Determining sample size to estimate a proportion P

$$n = \frac{t^2 \times PQ}{E^2}$$

where:

E = acceptable margin of error

Example: Estimate proportion of persons voting for a certain candidate, within ± 3%, with a confidence level of 95%:

Assuming a large population and P=0.5,

$$n = \frac{1.96^2 * 0.5 * 0.5}{0.03^2} = 1{,}067$$

Absolute and relative errors

Formula $e = \sqrt{\frac{p(1-p)}{n}}$ gives the absolute error

But we are often interested in the relative error $\frac{e}{p} = \sqrt{\frac{(1-p)}{pn}}$

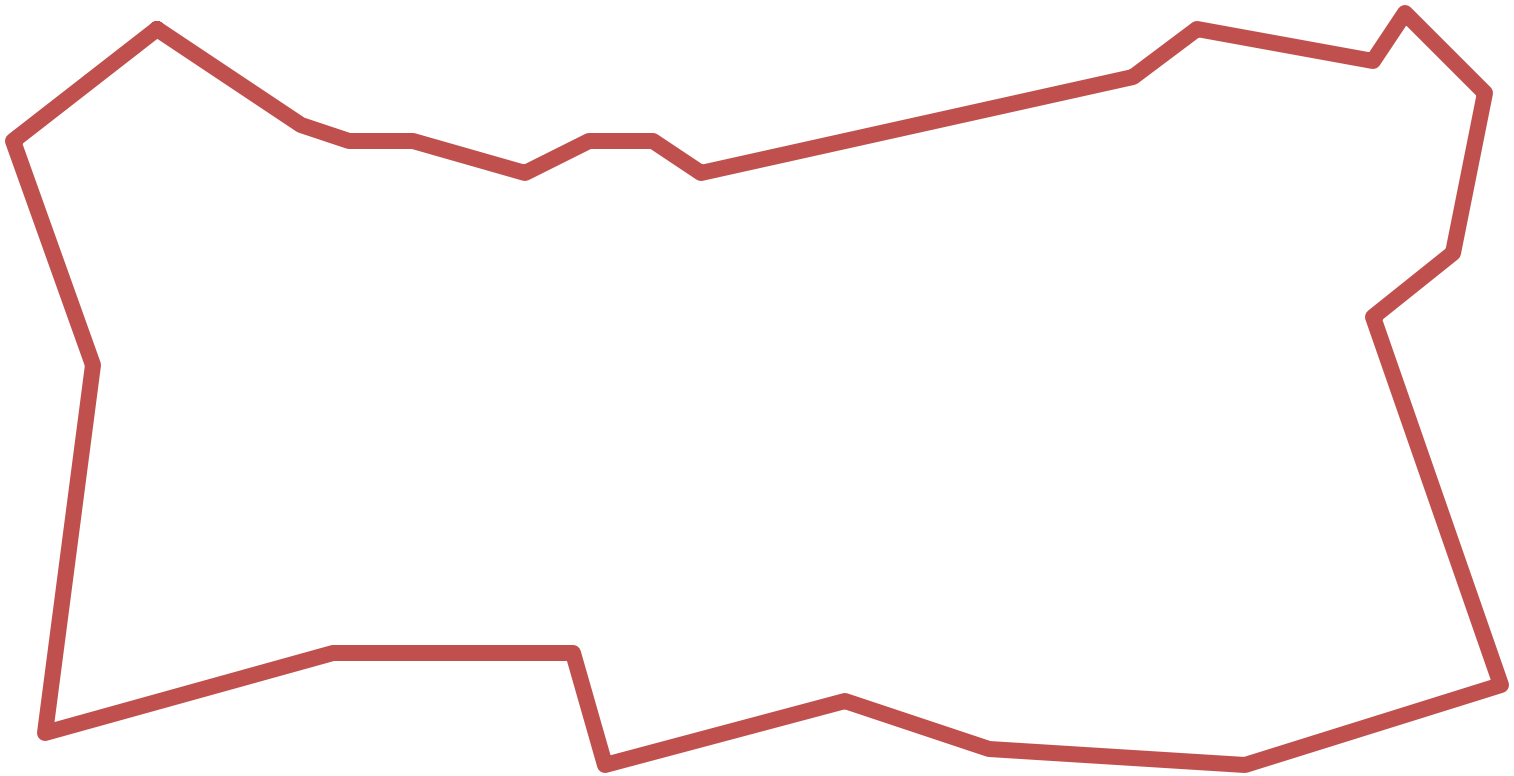
For rare events (small p), the relative error can be large, even with very big samples

This may be the case of some of the MDG's

- Infant / maternal mortality
- HIV/AIDS prevalence
- Extreme poverty

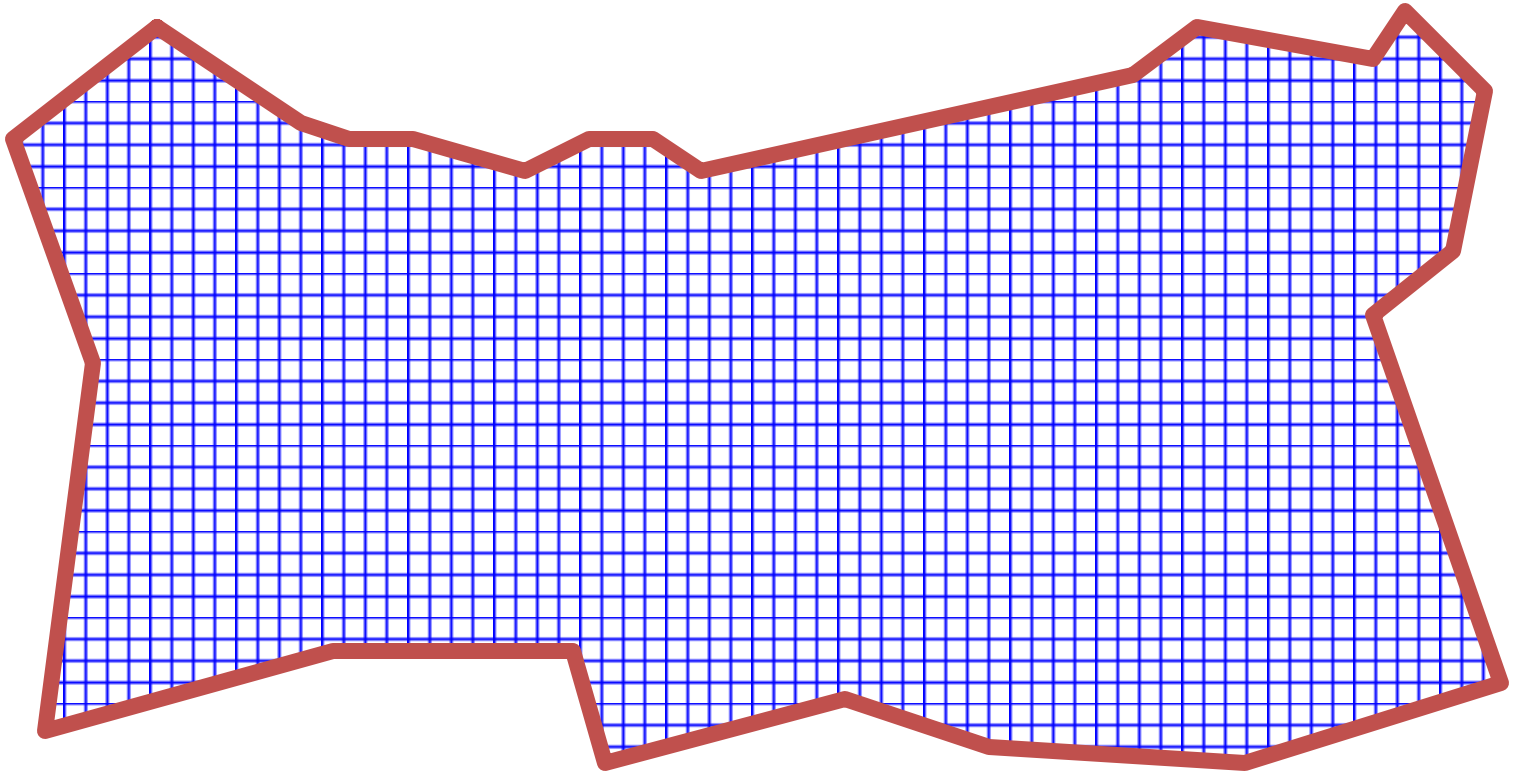
Two-stage sampling

- The country



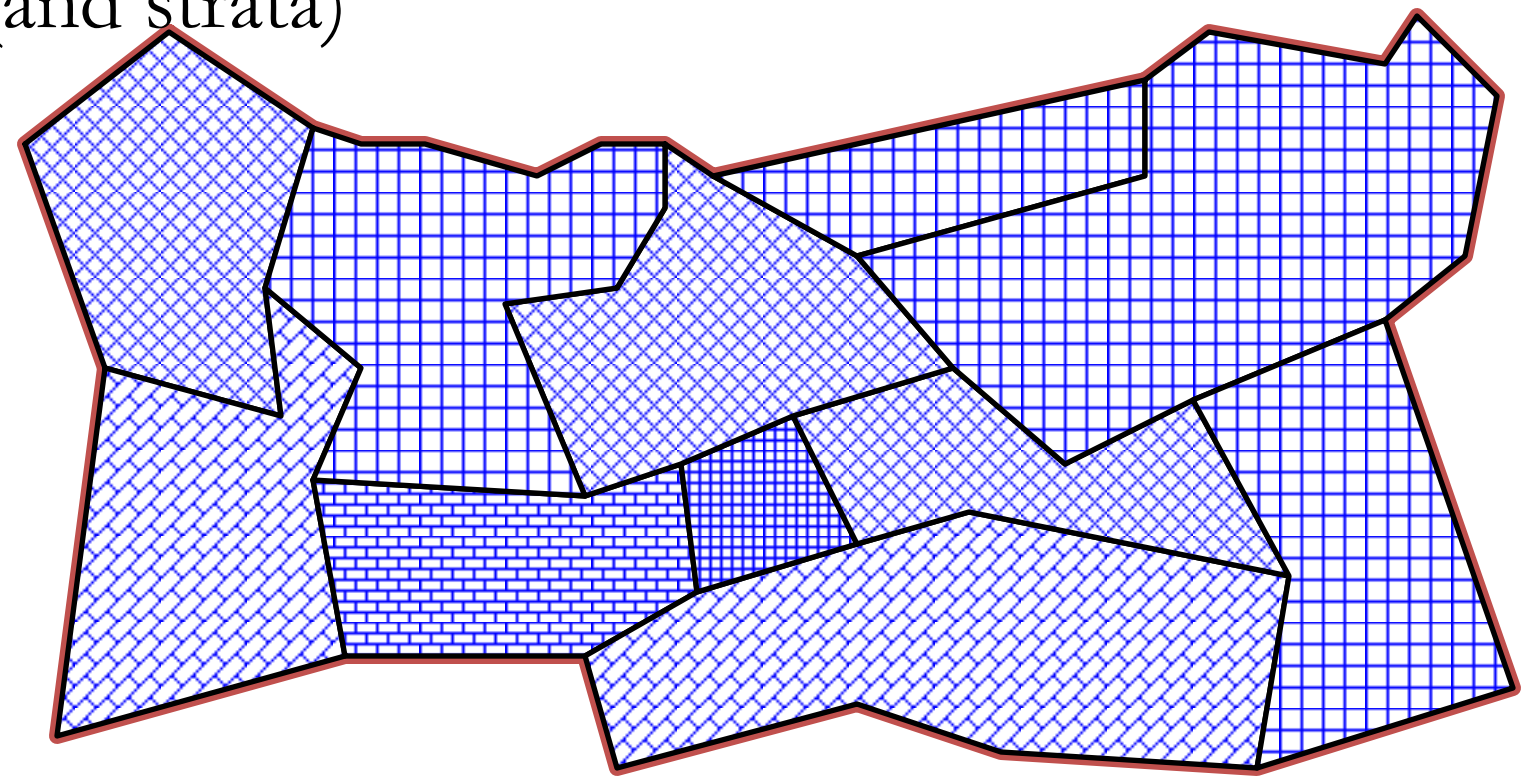
Two-stage sampling

- The country can be divided…



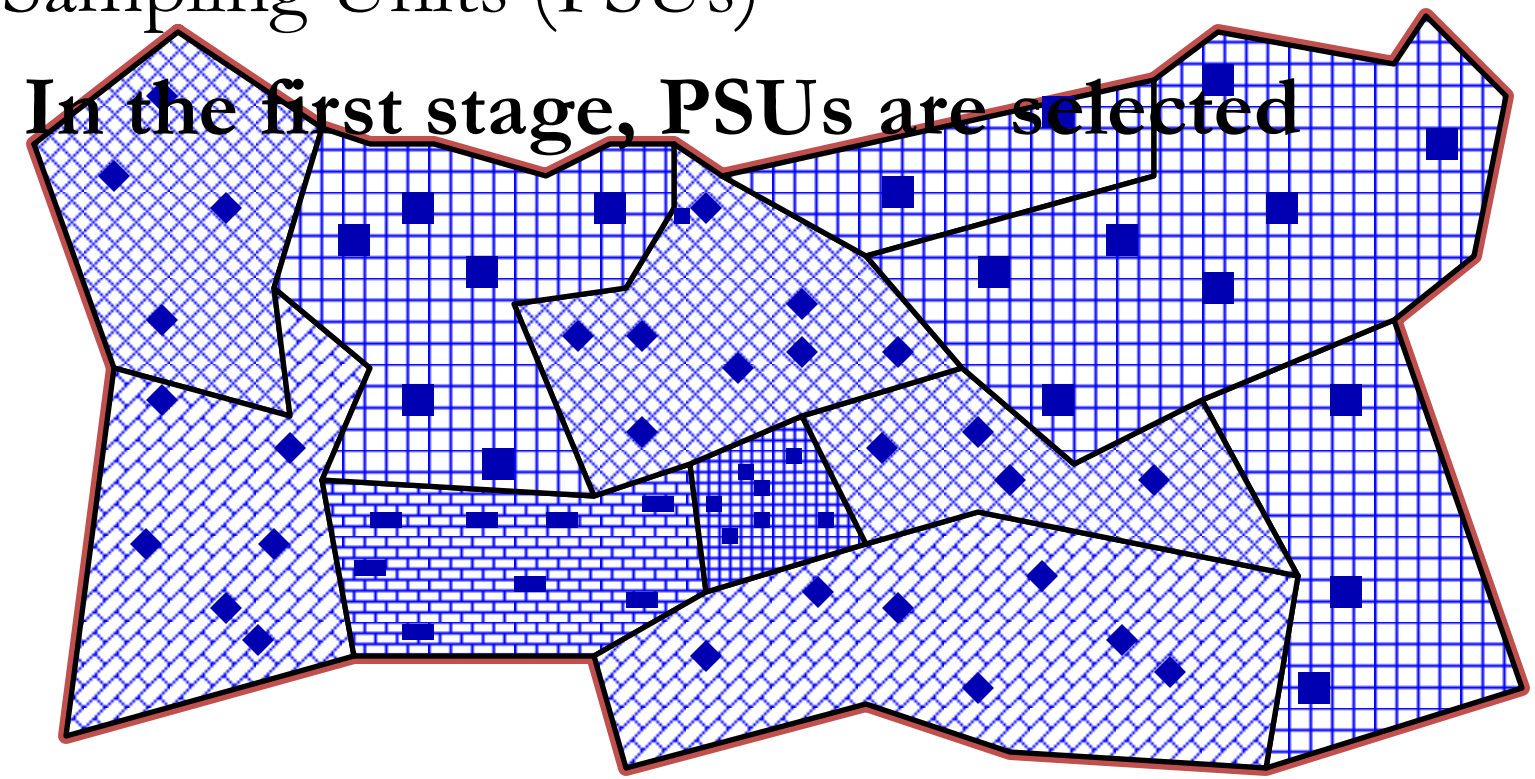
Two-stage sampling

- …into small Primary Sampling Units (PSUs) (and strata)



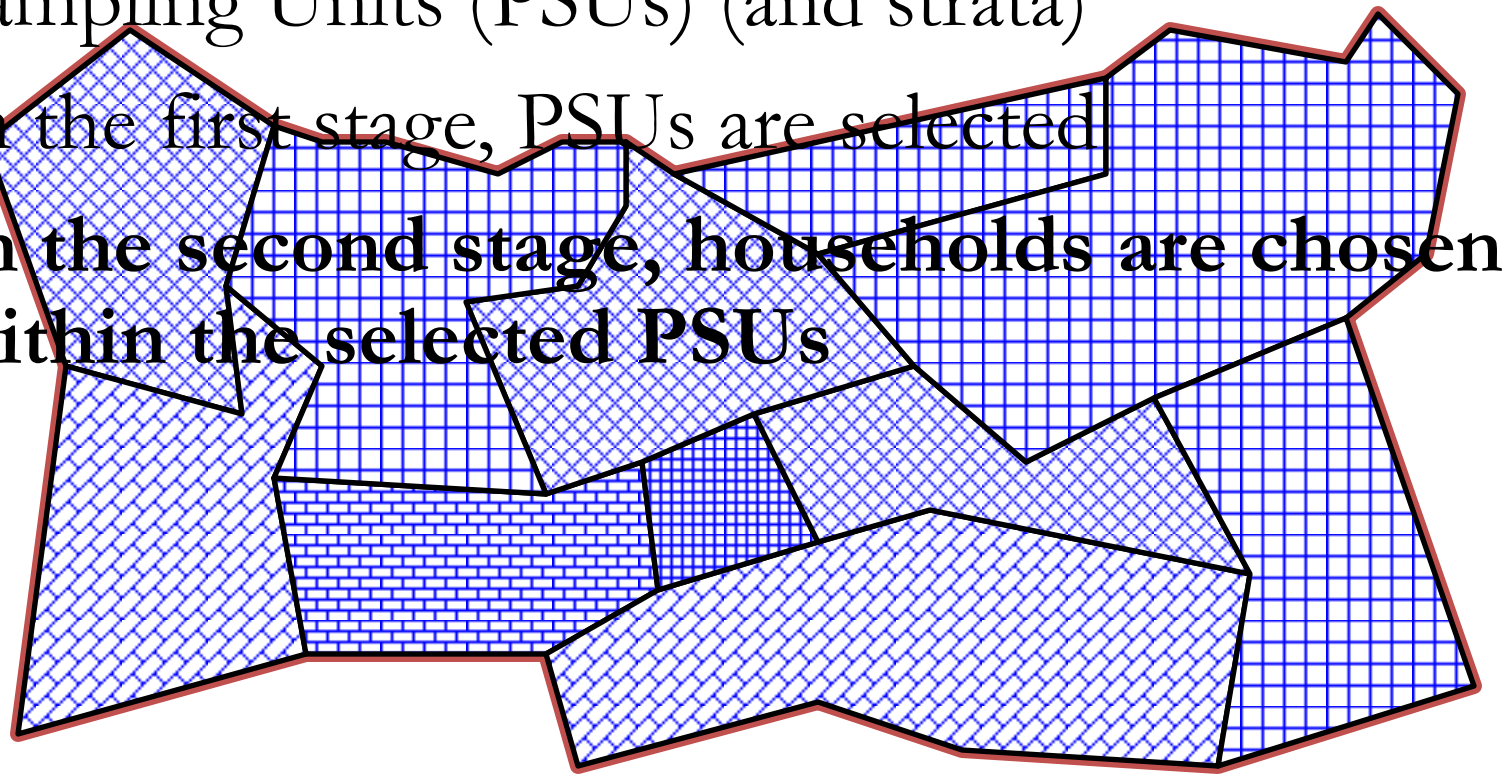
Two-stage sampling

- The country is divided into small Primary Sampling Units (PSUs)
- **In the first stage, PSUs are selected**



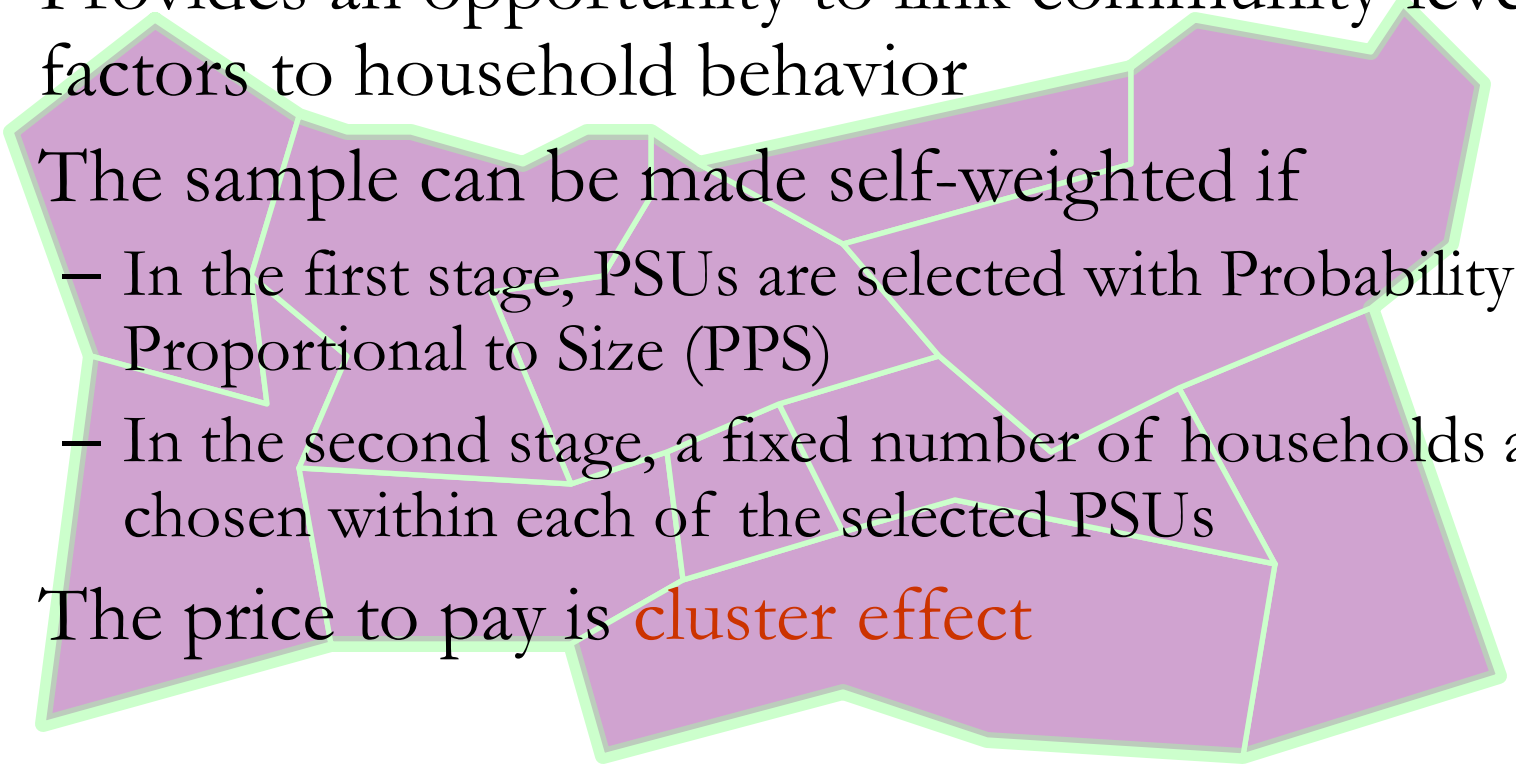
Two-stage sampling

- The country is divided into small Primary Sampling Units (PSUs) (and strata)
- In the first stage, PSUs are selected
- **In the second stage, households are chosen within the selected PSUs**



Two-stage sampling

- Solves the problems of Simple Random Sampling
- Provides an opportunity to link community-level factors to household behavior
- The sample can be made self-weighted if
  - In the first stage, PSUs are selected with Probability Proportional to Size (PPS)
  - In the second stage, a fixed number of households are chosen within each of the selected PSUs
- The price to pay is cluster effect



Techniques in Random Sampling

Two-stage sampling

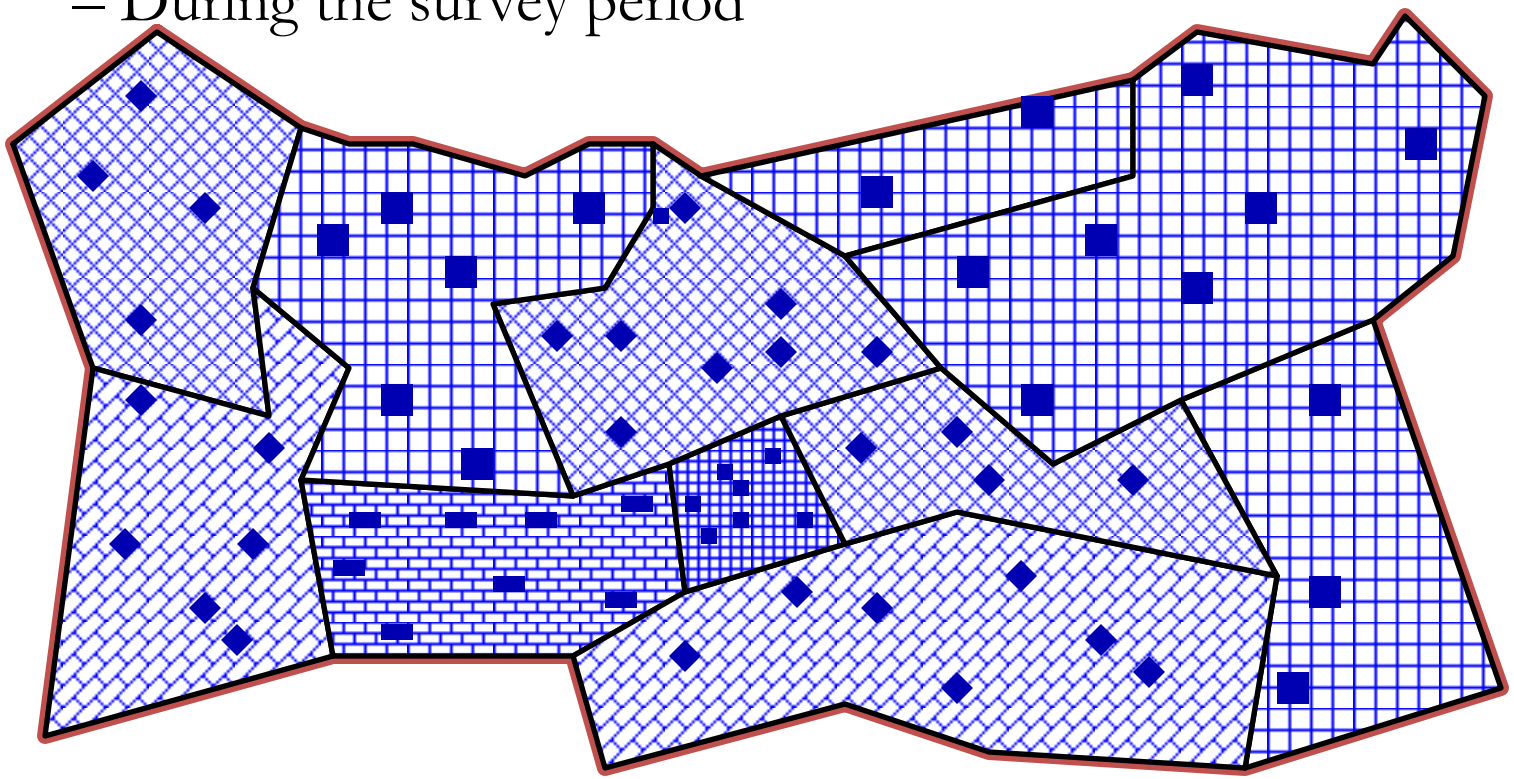
- Units of analysis are divided into groups called Primary Sampling Units (PSUs)
- A sample of PSUs is selected first
- Then a sample of units is chosen in each of the selected PSUs

This technique can be generalized
(multi-stage sampling)

Techniques in Random Sampling

Two-stage sampling

- In planning the survey, selected PSUs should be allocated
  - Among teams
  - During the survey period



First stage sample frame: The list of Census Enumeration Areas

- Exhaustive
- Unambiguous
- Linked with cartography
- Measure of size (for PPS selection) 
- Up to date (?)
- Area Units of adequate size

Second stage sample frame: The household listing operation

What is involved?	Training, organization, supervision
How long does it take?	50-80 households per enumerator/day
How much does it cost?	~15% of the total cost of fieldwork
How much earlier than the survey?	As close as possible
Is it always needed?	Yes (almost)
Dwellings or households?	A dwelling listing is more permanent
Who draws the sample?	Ideally, central staff
Asking extra questions during listing	Can be dangerous. Do it only if there are very good reasons
Can new technologies help?	Yes (GPS)

Techniques in Random Sampling

Stratified sampling

- The population is divided into mutually exclusive subgroups called strata.
- Then a random sample is selected from each stratum.
- Common examples : Urban / Rural, Provinces, Male / Female

Techniques in Random Sampling

Stratified sampling

- Parts of the country may need to be excluded from the sample for security or other reasons (excluded strata)