

Sampling



**General introduction to sampling methods in
epidemiology and some applications to food
microbiology study**

October 2006 - Hanoi

Stéphanie Desvaux, François Roger, Sophie Molia

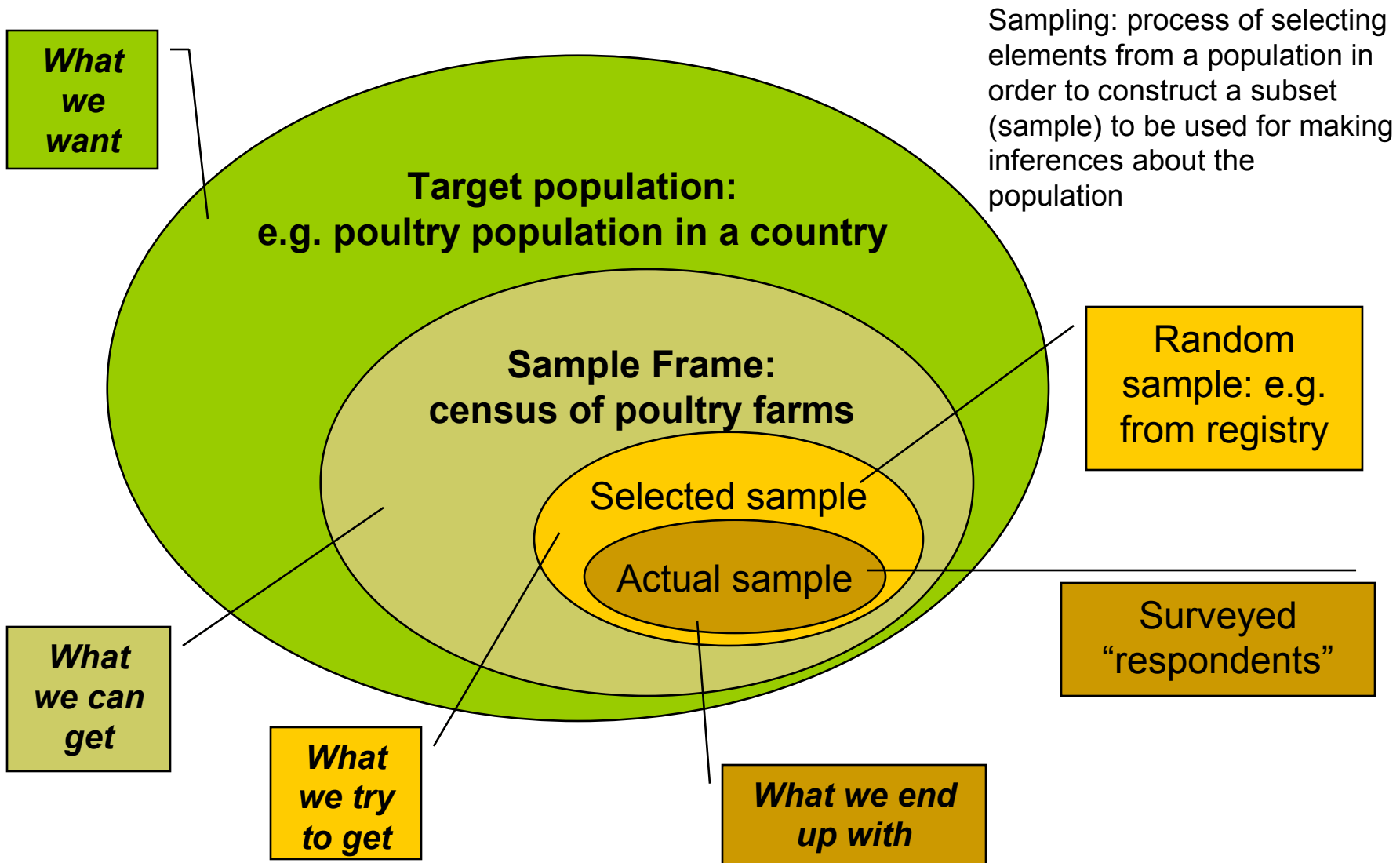
CIRAD Research Unit Epidemiology and Ecology of Animal Diseases

Sampling

- Sampling sensu stricto
 - Process of selecting elements from a population in order to construct a subset (sample) to be used for making inferences about the population
 - Designing a sampling process
 - Sources of errors
 - Types of sampling
 - Calculating a sample size

- Field sampling
 - Logistics
 - Collection
 - Shipment

Sampling principle



Why sample?

- Objectives of sampling
 - **Detect** a disease / Detect the presence of a microorganisms or contaminants (epidemiological surveillance)
 - **Measure** the frequency of a disease present in the country / control the level of a microorganisms or contaminant in a raw material, intermediate or final products (epidemiological surveillance & descriptive epidemiological surveys)
 - **Estimate the association** between potential risk factors and disease (analytical epidemiological surveys)

- Advantages of sampling
 - Information obtained more rapidly, more easily and for a lesser cost than when working with whole population

Sources of errors in sampling

□ **Systematic error = Bias**

Bias occurs when there is a tendency to produce results that differ in a **systematic manner** from the true values

- Lack of bias = good accuracy

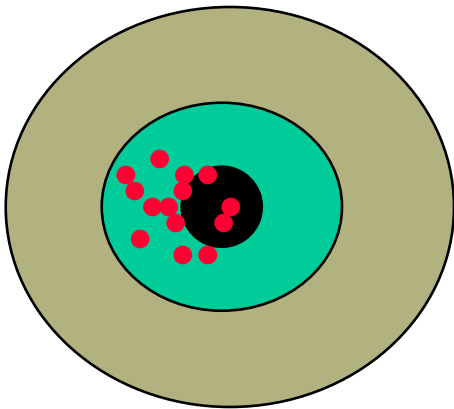
□ **Random error**

Random error = is the divergence, **due to chance alone**, of an observation on a sample from the true population value

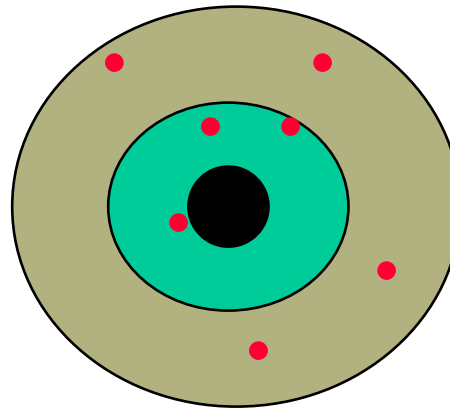
- Lack (or minimized) random error = good precision

Quality of an estimate

**Precision
& validity**

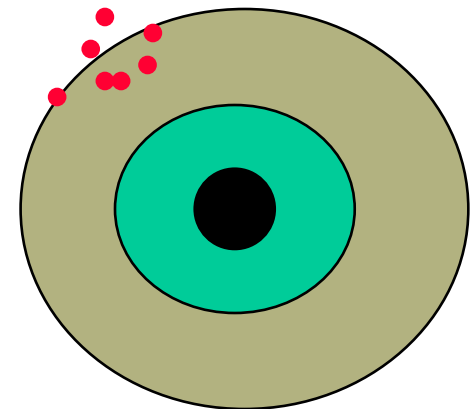


No precision



**Random
error**

**Precision but
no validity**



**BIAS =
Systematic
error**

Sources of bias

- ❑ Bias - or systematic error - may lead to over- or underestimation of the occurrence or of the strength of an association
- ❑ The sources of bias in epidemiology are many and over 30 specific types of bias have been identified. The main biases are:
 - Selection bias
 - Information bias
 - Measurement bias (Uncalibrated laboratory instruments...)



Sources of random error

- The major sources of random error are:
 - individual/biological variation
 - sampling error
 - Measurement error
- Random error
 - can be minimized
 - but can never be completely eliminated
 - **since we can study only a sample of the population**
 - individual variation always occurs and no measurement is perfectly accurate

Sampling error occurs as part of the process of selecting animals within a larger population
- **The best way to reduce sampling error is**
 - to ensure that the sample is **really random**
 - and that it is of **sufficient size**

Main sample qualities

- We want our sample to be
 - **Representative** of the population it was drawn from
 - use random sampling method
 - **Precise**
 - use an adequately large sample size
- Remember!!!
 - Increasing sample size does not compensate for bias due to a non random sample!!!

Designing a sampling process

- Define your objective
 - Estimate prevalence, detect a foreign disease.../ detect presence of dangerous microorganism or control level of normal flora
- Define the unit of concern
 - Village, herd, flock, individual
 - Animal at slaughter house / batch of a product / ...
- Define target population
 - Bovines > 36 months, any poultry holding
 - Daily, monthly production in a plant / products sold on market...
- Obtain sampling frame
 - Not always easy or up to date (village>farm>holding> animals)
- Decide how precise your answer needs to be
 - The more precise, the more \$!
- Determine sampling procedure and size

Sampling methods

- Non probability sampling
 - Eg: convenience sampling: units where easier to obtain info

- Probability sampling = random sampling
 - Simple random sampling
 - Systematic random sampling
 - Stratified random sampling
 - Cluster sampling
 - Multistage sampling

Every element in the population has a known probability of being included in the sample



Simple random sampling

- Each individual is assigned a number, then a sample of these numbers is selected using some random process
 - Flip a coin, throw dice, draw numbers from a hat
 - Table of random numbers
 - Random number generation with Excel
 - Statistical software

- Formal random method
- Each unit has an equal chance of being selected



06318	25019	79125	56761
96593	58161	04253	18420
13614	44281	01807	52716
05585	31616	71810	96974
50424	17376	31496	61292
54533	60100	31988	76790
40558	41321	46058	16234
23047	17672	16026	50248
57302	42154	26360	82660
88418	23878	87780	88651
89309	67546	62072	90626
02338	99984	20600	37661
52795	28307	90374	21686
04828	62488	46999	99007
37018	55650	64280	49617

Table of random numbers

Simple random sampling

□ Advantages

- Simple
- Practicable for small, compact populations

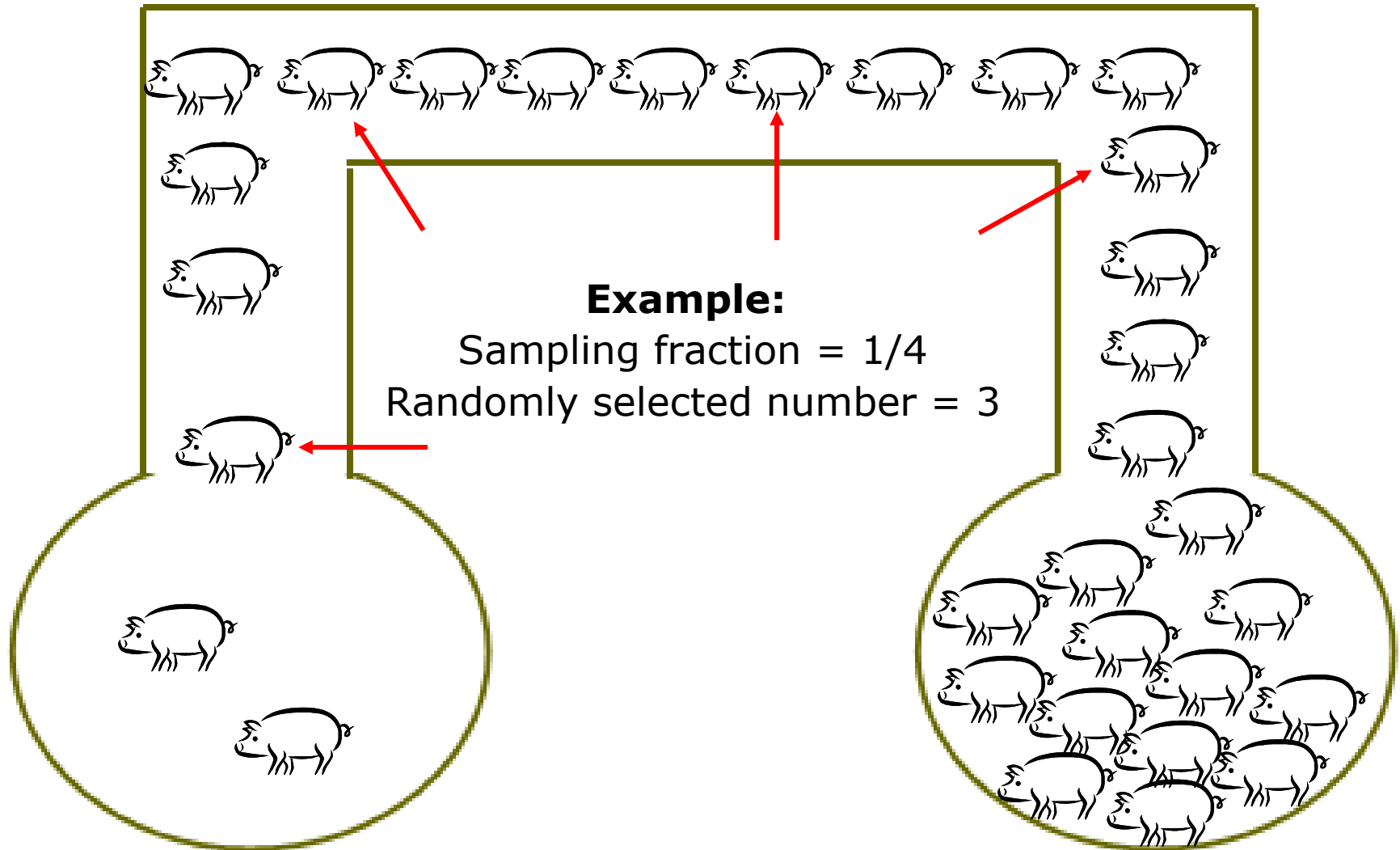
□ Disadvantages

- Can only be used when all animals or sampling unit in population are identified
- Expensive
- Lack of precision when the population is heterogeneous (does not show the differences between ages, sexes, husbandry systems...)

Systematic random sampling

- ❑ You do not have individually identified animals → no sampling frame → simple random sampling not possible
- ❑ Choose a sampling fraction
 - For example you know you have 2000 birds in a farm and you want to sample 40 birds; sampling fraction = $40/2000 = 1/50$
- ❑ Randomly select a number between 1 and 50
 - For example 13
- ❑ Select unit #13, then unit # $13+50=63$, then unit # $63+50 = 113$, then #163, #213... at regular interval until you have 40 units
- ❑ Good approximation to simple random sampling

Systematic random sampling



Systematic random sampling

□ Advantages

- Does not require identification of all animals in the population
- Can be simple to use

□ Disadvantages

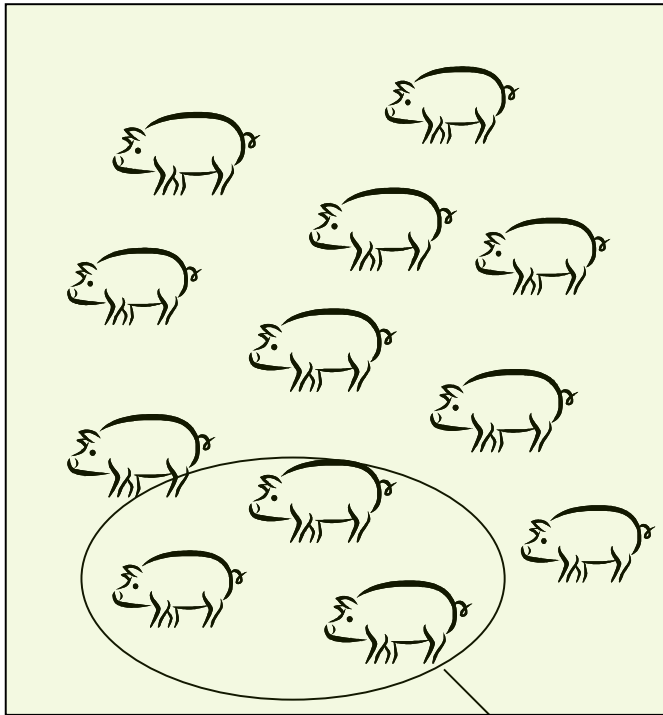
- Beware of cyclic pattern in the order of the units
- Lack of precision when population is heterogeneous

Stratified random sampling

- The sampling frame is divided into strata based on factors which are likely to influence the level of the characteristic being estimated
 - Example: age, breed, sex, husbandry system / type of market
- Then simple or systematic random sampling is performed within each stratum
- The size of the sample within each stratum can vary with number of units of each stratum or the expected prevalence in each stratum

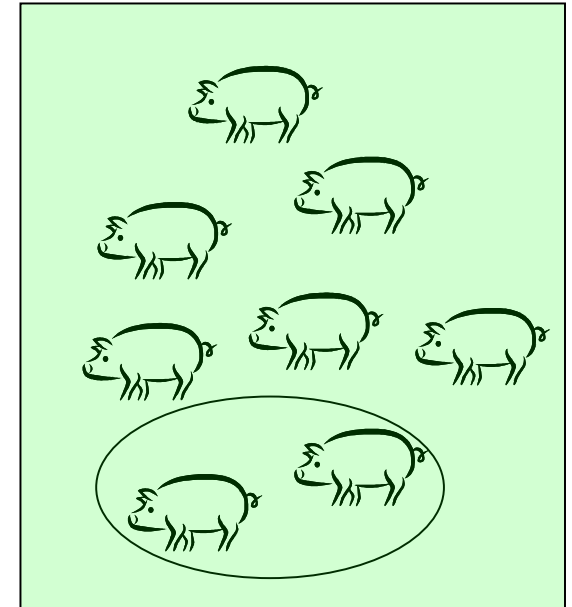
Stratified random sampling

Strata 1



Random
sample
from
strata 1

Strata 2



Random
sample
from
strata 2

Sample

Stratified random sampling

□ Advantages

- Improves precision
- Can ensure samples are large enough to get desired precision for each strata
- May be easier for administrative reasons

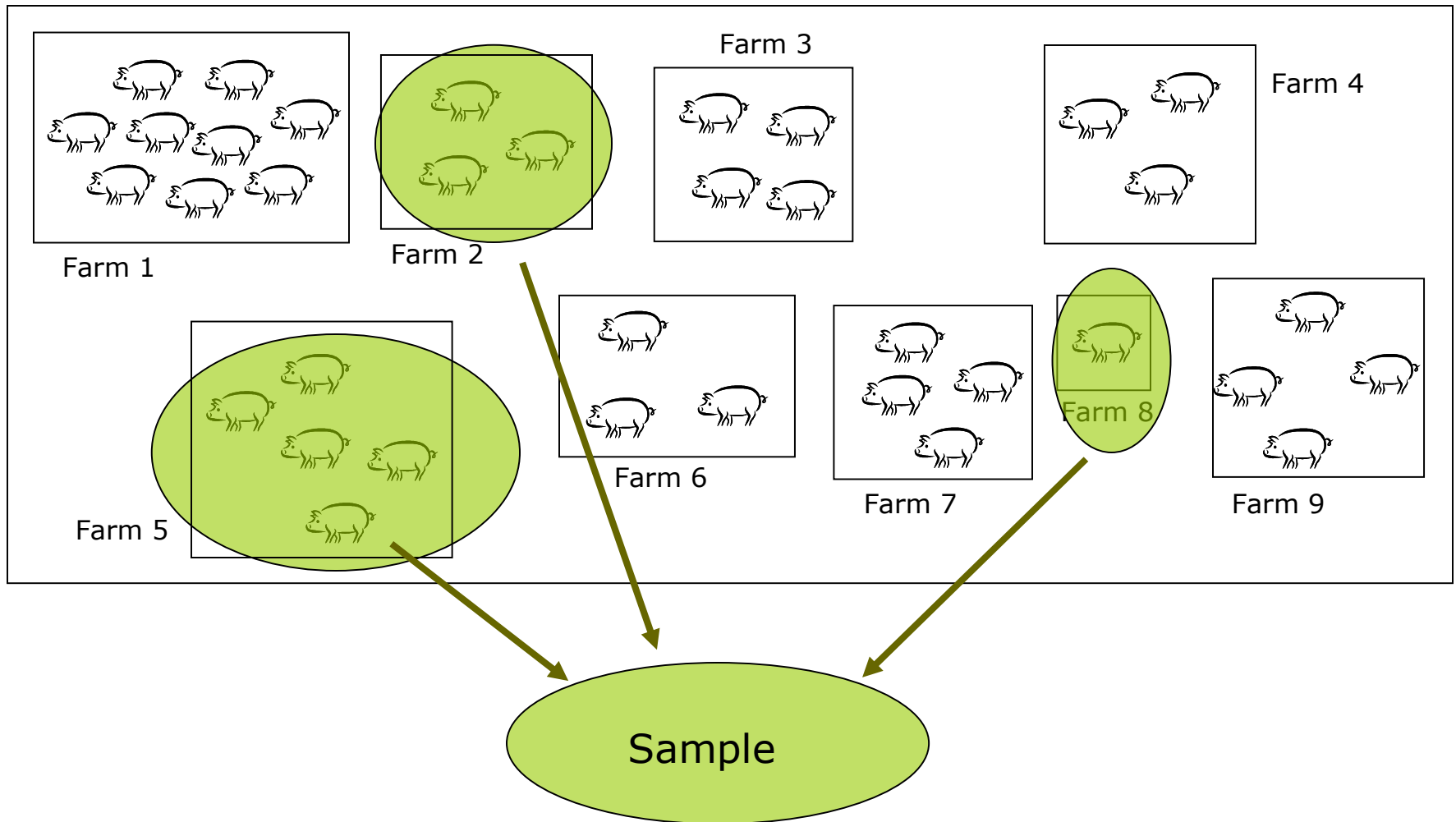
□ Disadvantages

- The status of each sampling unit with regard to the stratifying factor must be known

Cluster sampling

- Clusters or groups of animals are randomly selected
 - Example of clusters: herd, litter, pen...
- All animals within selected clusters are sampled
- Need a list of the clusters but not the individual animals

Cluster sampling



Cluster sampling

□ Advantages

- Reduction of costs and time in field studies
- Does not require a complete list of all units / animals

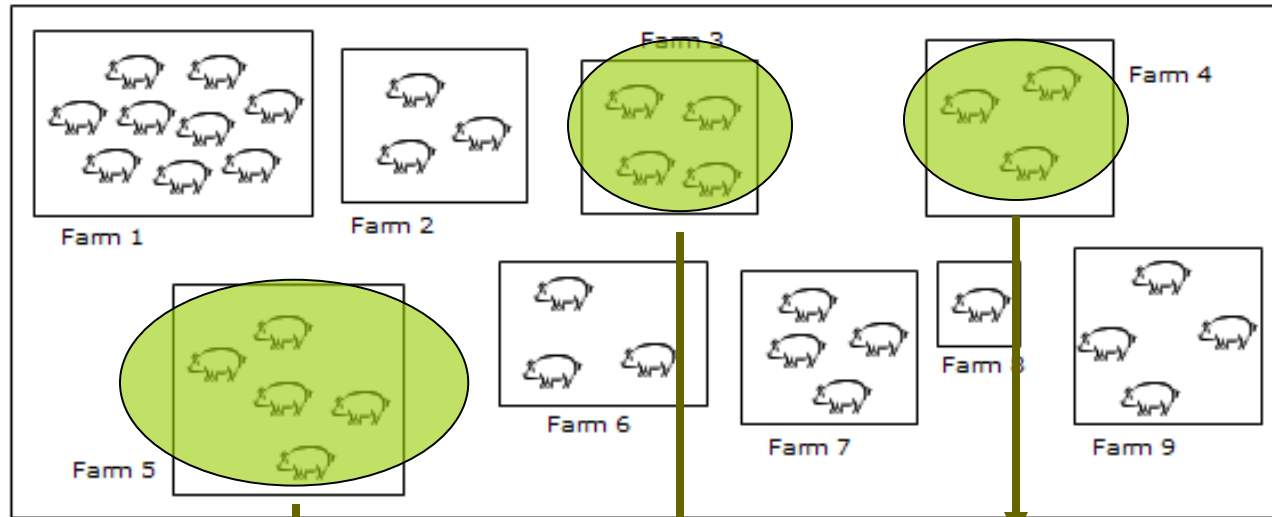
□ Disadvantages

- It can reduce precision because animals within a cluster tend to be more similar to each other than animals in other clusters
- The analysis is more difficult

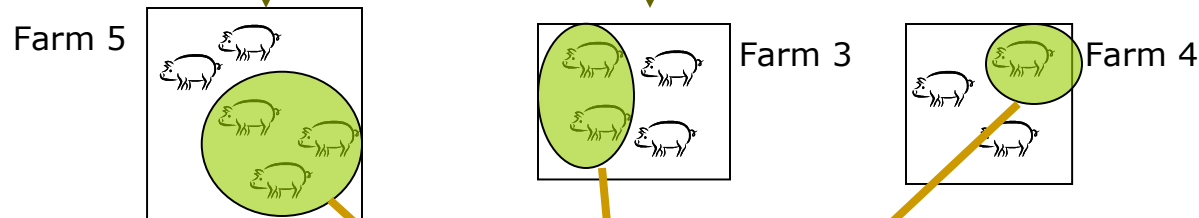
Multistage sampling

- Is a more complex type of cluster sampling
- Clusters or groups of animals are randomly selected, then within these groups a further sample of individuals or groups are selected
- It can extend to three or more stages
- Often recommended to select sample units for each stage with a probability proportional to the number of units they contain

Multistage sampling



1st stage sampling



2nd stage sampling



Multistage sampling

□ Advantages

- Flexible method – the number of primary and secondary units can be varied to account for the differences in costs of sampling
- Does not require a complete list of units / animals

□ Disadvantages

- Can reduce precision
- Requires a more complex analysis

Sampling methods: summary

- Non probability sampling



avoid whenever possible

- Probability sampling = random sampling



**Choose method
depending on objective,
context , means**

Calculating a sample size

- Different procedures depending on objective
 - **Quantitative approach:** determine the frequency of the disease → estimate the prevalence with a predetermined confidence interval
 - **Qualitative approach:** detect if the disease is present → estimate the presence or absence with respect to a confidence threshold
 - **Comparison between means or percentages** → estimate the chance of getting the observed effect or difference if the null hypothesis is true = if there is no difference between the groups

Confidence interval

□ Definition

- The confidence interval for a population prevalence P is a range of values within which we are confident (generally 95% confident = with a risk of error of 5%) that the real prevalence is

□ Calculation

$$P = p \pm 1.96 * \sigma \quad \text{or} \quad P = [p - 1.96 * \sigma ; p + 1.96 * \sigma]$$

- Where p = proportion estimated from the sample
 σ = standard error of the proportion

when $n/N < 10\%$

$$\sigma = \sqrt{\frac{pq}{n}}$$

when $n/N > 10\%$

$$\sigma = \sqrt{\left(1 - \frac{n}{N}\right) \times \frac{pq}{n}}$$

n = sample size

N = population size

Quantitative approach

- What do we need to calculate the sample size necessary to estimate the prevalence with a 95% confidence interval?
 - An estimate of the prevalence we're trying to determine(!)
 - A desired level of precision
 - Relative precision or absolute precision
 - An idea of the total population size N
 - To see if the sample size n that we obtain is $< 10\% N$
 - If not we have to adjust our sample size

Quantitative approach

$$n = \frac{1.96^2 \times \text{Prev} \times (1-\text{Prev})}{(\text{Absolute precision})^2}$$

□ Example:

■ Leptospirosis survey

- Estimate of prevalence: $\text{Prev} = 30\% = 0.3$
- Relative precision = $15\% = 0.15$
Same as absolute precision = $4.5\% = 0.045$
- Farm population size $N = 1000$

$$n = \frac{1.96^2 \times 0.3 \times (0.7)}{(0.045)^2} = 399$$

- Problem: $n > 10\% N \rightarrow$ adjustment

$$n' = \frac{1}{(1/n) + (1/N)} = 285$$

What if I have no idea whatsoever of Prev??

\rightarrow Take Prev = 50%

Qualitative approach

- What do we need to detect a disease in a study population by only examining a part of the population with a chosen risk of error?
 - The maximum level of risk of error accepted (α)
 - 5% accepted risk of error = 95% confidence level
 - The detectable prevalence in the population tested
 - Extrapolated from previous studies or general knowledge
 - The larger the detectable prev, the smaller the sample size
 - The size of the population
 - If “finite population” ($n > 10\% N$), probability that the next animal/farm to be drawn is infected increases as sampling is carried on

Qualitative approach

$$n = \frac{\text{Log } (a)}{\text{Log } (1 - p)}$$

□ Example:

■ Detection of H5N1 in wild birds

- Estimate of prevalence: (Chen et al, 2006: six H5N1 isolated from apparently healthy migratory ducks at Poyang Lake out of 4316 tested) $\text{Prev} = 0.14\% = 0.0014$
- Accepted risk of error = 5%
- Wild bird population size $N = 200\ 000$

$$n = \frac{\text{Log } (0.05)}{\text{Log } (1 - 0.0014)} = 2138$$

- Statistically, if no positive wild bird is found positive, the population is supposed to be considered free of H5N1...

Others ways of calculating sample sizes

- Statistic tables
 - Example for an infinite population

Prev	$\alpha = 5\%$	$\alpha = 1\%$
0.01%	29 956	46 050
0.1%	2995	4603
0.2%	1497	2301
0.5%	598	919
1%	299	459
5%	59	90
10%	29	44
20%	14	21
50%	5	7

Others ways of calculating sample sizes

□ WinEpiscope

- Statistical software
- Can be downloaded for free on

<http://www.clive.ed.ac.uk/cliveCatalogueItem.asp?id=B6BC9009-C10F-4393-A22D-48F436516AC4>



WinEpiscope 2.0

File Tests Samples Analysis Models Windows

Estimate Mean
Estimate Difference between Means
Estimate Percentage
Estimate Difference between Percentages
Detection of Disease
Threshold Value
Unmatched Case-Control
Matched Case-Control
Cohort

Sample Size: Detection of Disease #2

Sample Size Maximum n° positives Level of Confidence

Input of DATA:

Population Size: 200000
Prevalence (%): 0,14
Level of Confidence (%): 95 %

RESULTS:

N° of Diseased Animals: 280
Required sample size: 2127
Sampling fraction (%): 1.06

	N° Diseased	% Diseased	Sample Size	N° Diseased	% Diseased	Sample Size
1	11	0,01	190000,1	11	0,01	47679,6
2	12	0,01	155278,1	12	0,01	44183,2
3	13	0,01	126319,1	13	0,01	41162,1
4	14	0,01	105425,1	14	0,01	38526,0
5	15	0,01	90143,0	15	0,01	36206,0
6	16	0,01	78606,6	16	0,01	34148,7
7	17	0,01	69631,7	17	0,01	32311,9
8	18	0,01	62467,7	18	0,01	30662,2
9	19	0,01	56624,6	19	0,01	29172,4
10	20	0,01	51771,9	20	0,01	27820,3

Calculate Close

In food microbiology

- Classical statistical schema are difficult to apply
- There are sampling standards for control purpose, set up by the international commission:
 - 2 or 3 classes protocol according to the risk for human health