# **Module 6: Using sampling theory to manage disease**

Contributors: Scott Wells, Amy Kinsley, Sandra Godden

Testing a sample of animals within a herd (or kennel or other group) is an example of sampling animals within a population.  Other sampling examples involve selection of a subset of individuals from even larger populations (e.g., group of herds, state, region, country).

In epidemiological evaluations, we often want to measure a **health outcome (**such as disease or infection) in some individuals within a **population**(whose units can be animals but may be herds or regions) to evaluate if this outcome is associated with some characteristic (“**exposure**”) of the individuals that form the population.

In a **census**, every animal in the population is evaluated. Therefore, if data from the whole census is analyzed, the exact level of the health outcome (such as disease) can be obtained. However, obtaining a census is usually too costly to be performed, and measurements are instead performed on a **sample**instead of on a census.

**A sample is a subset of individuals that is (ideally) representative of the population from which it has originated**, so that measurements performed on the sample can be extrapolated (with a certain uncertainty) to the whole population. There are multiple ways of obtaining samples, which each have advantages and disadvantages.

## **Types of populations**

To start, let’s define some terms used in sampling.  In all sample-based evaluations, you should be able to identify the following:

* **External population**: This is the population to which it might be possible to extrapolate the results from the study (depending on the external validity of the study, which will be defined later). This extrapolation is often subjective and depends on the person interpreting the study results.
* **Target population**: This is the immediate population to which the results from the study can be extrapolated (this population is directly linked to the study population).
* **Study population**: This is the population from which the sample has actually originated (so it is formed by all individuals potentially selected to participate in the study, although only a subset of them will be actually part of it).
* **Sample**: the subset of individuals in which the outcome (and sometimes other possible exposures) will be finally measured.

### **Case 6.1. Vaccinating against Lyme disease**

You are trying to decide whether to create educational materials to promote vaccination against Lyme disease in dogs in your veterinary practice area, and would like to estimate the prevalence of dogs that test positive.  Recognizing that you don’t have a list of all dogs in your county and assuming that a history of current vaccination is protective against infection, you ask an interested high school student to select a random sample of 200 of your canine client records and then review these records to assess the proportion of dogs in these records that have tested positive for disease.

Which population is which?

* External population – Dogs in your local area.
* Target population – Dogs in your local area with veterinary care.  If dogs in your area do not have veterinary care, they would not be represented by a sampling system using animal hospital records.
* Study population – All identifiable dogs in your hospital case record system.  This is the population from which the sample (random in this case) is selected.
* Sample population – The dogs identified in the randomly selected case records in the hospital record system with available test results.

In addition, you should recognize your **sampling unit** (the level at which you are measuring the outcome, and the exposure if doing so).  Common sampling unit levels used in veterinary include:

* Animals (diseased/non-diseased)
* Herds/farms/flocks/kennels (herd positive/negative, even if the disease is then measured in the animals in the herd in order to establish if the sampling unit is positive/negative)
* Quarters in cattle udder in a study studying mastitis
* Region in a study comparing factors associated with presence of disease at the region/county level.

## **Types of sampling strategies**

Once you know your populations (target and study) of interest and the level at which you want to measure the outcome, it is time to get the sample. There are different ways of doing so, and they can be classified based on the system to draw the sample:

### **Non-probability sampling**:

The probability of selection of an individual within the (study) population is not known, and there is no explicit method to determine the probability of selection of a given individual.  Samples collected using non-probability sampling are more prone to systematic bias than probability-based sampling.

**Systematic bias** can result from many sources including bias in selection of study subjects. In sampling, selection bias occurs if study subjects are not randomly selected, or are selected randomly from a non-representative subset of the population (this will be discussed in more depth in a future module).

Examples of non-probability sampling:

* Convenience sampling
	+ Selection of an easy-to-sample population like cattle located close to the veterinary clinic, when trying to represent all cattle in a region.
	+ Selection of the first 200 clients with dogs in the animal hospital database (organized alphabetically by last name) for evaluation to estimate Lyme disease prevalence.
	+ Identify people from the database that you personally know (or are your friends) for evaluation to estimate Lyme disease prevalence.
	+ Identify dogs volunteered for evaluation in your study by their clients.  Note: Isn’t sampling the subset of dogs volunteered by their owners for study (due to bringing them to vet clinics and paying for testing) an example of non-probability (and biased) sampling relative to the target population of all dogs in the area?
* Haphazard sampling – selection of samples without an obvious sampling plan, often a mish-mash of available samples from a population, when trying to represent all cattle in a region.

### **Probability sampling**:

All individuals included in the (study) population have a known non-zero probability of being included in the sample. This type of sampling provides the potential to minimize systematic bias (if done correctly). Note: There does not need to be equal probability of inclusion for each individual in the population!

#### **Simple random sampling**:

All individuals in the population have the same chance of being selected. This sampling typically requires having a sampling frame (a list of all the sampling units in the study population).

* + Example:  After downloading the list of clients into a spreadsheet, use a random numbers generator (within a computerized spreadsheet like Excel) to randomly select 200 clients with dogs for evaluation.

#### **Systematic random sampling**:

 Selection of sampling units occurs at a predefined equal interval, the sampling interval (size of the population/sample size).

* + Example:  From your list of 2,000 clients with dogs, select every 10th client record for evaluation.  Or select every 10th client record for client visits to the clinic during a specified time period.

#### **Stratified random sampling**:

The sampling frame is divided into two or more strata (groups) and a random sample is collected from within each stratum.

* + Example:  Categorize the clients by group of dog (Herding, Sporting, Non-Sporting, Working, Hounds, Terriers, and Toy breeds) to determine the proportion of dogs in each group, then randomly select the number of clients in each group (strata) that corresponds to the proportion of dogs in each group.  E.g., Herding dogs = 10 clients, Sporting dogs = 30 clients, etc. for total of 200 clients.
	+ Example:  Randomly select 28 clients from each dog group, for a total of 200 selected.
	+ ***Proportional*stratified random sampling:** The number sampled within each stratum is proportional to the weight of the stratum in the population.
	+ ***Equal*stratified random sampling:** The same number of individuals per stratum is sampled regardless of the size of each stratum.

####  **Cluster sampling**:

The sampling frame is divided into clusters (primary sampling units) and a random selection of those clusters is performed.

* + 1. Example:  Categorize clients with dogs by zip codes (n=5 zip codes), randomly select one zip code, and then review records from all clients in that zip code.
		2. Example:  Categorize clients with dogs by zip codes (n=5), randomly select one zip code, and then randomly select clients from that zip code for review of records.

#### **Multi-stage sampling:**

The sampling process involves different sampling methods at different levels within study (state, county, herd, animal, etc).

* + ***One-stage*cluster sampling:**  All units within the selected clusters are sampled.
	+ ***Two-stage (multi-stage)*cluster sampling:** A random sample within each of the selected clusters is selected.

## **Types of sampling-associated errors**

There are 2 different types of sampling-associated errors to consider when sampling:

#### **Random sampling error**

Since not all the individuals in a population are measured when a sample is taken, the measurement obtained will be subjected to variability (the result will be different depending on the individuals that were actually sampled). This is called the **random sampling error**, and is related to the sample size (a larger sample size leads to a smaller random error).

#### **Systematic error (bias)**

In addition, if the sample selected is not representative of the underlying population, results from the study may be affected by the **systematic error (or bias)**, more on this in a future session).

### **Case 6.2. Estimate average age of cats**

You would like to estimate the average age of the cats in your practice area in order to focus your future veterinary educational programs to current and future clients. The **external population** is the population of cats in your practice area.  The **target population** is the population of cats in your practice area with owners that use veterinary care.

**Option 1:**

From a list of your favorite clients (those who pay their bills), you identify 20 cats and look up the ages of their cats.

The **study population** is all of the cats owned by your clients.

The **sample population** is the 20 cats you identified.

The average age of this population of cats is 6.2 years, as these are cats owned by clients that have used your veterinary services over an extended period of time. This is an example of **non-probability based sampling**, which can lead to systematic bias or error in estimating the average age of cats in your practice area.

**Option 2:**

At a monthly meeting among veterinarians in your practice area, you convince many of the veterinary clinics to join you to conduct a study to evaluate the age distribution of cats that they see in their clinics. From each clinic, a **systematic sampling system** is used to identify the age of every 10th cat in their clinical record system for a 2-month period of time.

The **study population** is all of the client-owned cats in the participating veterinary clinics

The **sample population** is the cats identified using systematic sampling.

Among cats identified from the participating veterinary clinics, the average age was 1.6 years. This is an example of **probability-based sampling**, which should have much reduced systematic error (bias), and therefore provides a much less-biased estimate of the age of cats in your practice area. There still is the possibility of random sampling error, based on the ages of the actual cats identified.

## **How many samples to take from the population?**

This depends on the reason for the sampling, specifically if the sample is collected to **estimate the prevalence of an outcome** (like a disease) in the population, or instead, if the sample is taken to **detect the presence of an outcome** (like a disease) in the population.  We will consider each of these questions in the context of one very common application – Herd-level testing of animals.

### **Herd level testing of animals**

The next step is to apply sampling strategy to useful population-level situations, like addressing the following questions in herds:

1. *Is this herd infected?*  **Partial-herd testing to detect disease in the population.**
2. *What proportion of the animals in this herd are affected*?  **Partial-herd testing to estimate the true prevalence of disease in the population.**
3. Effect of sensitivity and specificity of an imperfect screening test on “herd-level” sensitivity and “herd-level” specificity

**Note**: *“Herd” = herd, flock, gaggle, cattery, kennel, or any other population where animals are managed in a ‘group’.*

**Background on herd-level testing strategies** (from Cannon and Roe, 1982)

Veterinarians are frequently asked to certify the freedom of animal groups or populations from certain diseases.  Here the herd (not the individual animal) is the unit of concern.  Ramifications or actions will be applied to the whole herd, depending on the classification status of the herd after testing.  What testing needs to be undertaken to be able to give such a certification?

#### **Strategy 1. Whole herd testing**

 In some cases, those wanting the certification (e.g., Minnesota Board of Animal Health, BAH) will prescribe tests that must be conducted on all animals in the group (e.g., “whole herd” testing).  In this case, no sampling considerations arise – all animals must be tested.  An example includes testing of all adult cattle in a herd suspected of exposure to bovine tuberculosis (*Mycobacterium bovis*) as part of an eradication program.  If the herd is found to be positive (one or more infected animals are confirmed as infected), then the entire herd is subject to quarantine and may be slaughtered.

#### **Strategy 2. Partial herd testing**

**2a. Partial herd testing to detect disease.**

In other cases, the veterinarian may be asked to certify the disease status of the herd of origin of animals moving interstate or overseas. In many of these cases, the veterinarian may be asked to certify that the herd is free of a certain disease.  In the absence of records of whole herd tests for specified diseases, this requires testing at least a subsample of the animals in the herd.  Typically, the sample size tested should be of sufficient size to give 95% confidence that the disease is not present at a prevalence that would be expected in an infected herd.

**2b.  Partial herd testing to estimate true prevalence.**

Another use for partial herd testing is to estimate the true prevalence of disease in ‘known positive’ herds (e.g., to monitor the prevalence over time to document success of a control program).

 Let’s evaluate each of these partial herd testing situations, using sampling concepts.

#####  **2a*. Is this herd infected?* Partial herd testing to detect the presence of disease**

**Goal of program:**We want to test a subset of animals within the flock to detect whether or not the flock is infected (i.e., if one or more animals in the flock is determined to be infected, the whole flock will be classified as infected).

 You have learned that sampling a subset of animals from a herd will produce an estimate of the infection status of the disease (herd tested positive or herd tested negative).  However, occasionally, some errors in herd classification (as infected or not infected) did occur, primarily due to false-negative (imperfect test sensitivity) or false-positive (imperfect test specificity) results.  The validity of the results will also be influenced by the number of animals tested (the size of the sample).

**When sampling to detect disease, the question is:**  How many animals in this population (kennel, cattery, herd, flock) must I test to be confident of detecting disease “X”, if that disease truly is present?  The size of the sample to be tested to determine whether a disease is present in a particular population depends on:

1. The size of the population in question.
2. If the disease is present, the likely prevalence (design prevalence): The higher the prevalence, the smaller the sample size that is required to detect at least one infected animal.
3. The reliability (confidence) required of the conclusions. By not testing all animals, we will not be able to make an absolute statement (e.g., we are not able to be 100% certain the disease is not present in this group) based on our findings.  However, the larger the sample, the greater the confidence that can be placed on the results.  The sample size can be chosen to give the confidence level desired.  For example, suppose that the sample size has been chosen to provide 95% confidence of detecting antibodies if they are present in at least 10% of the animals in the herd.  This means that, on average, 5 out of 100 herds with 10% prevalence would not be detected as infected by our survey.
4. The sensitivity of the test being used.

###### **Sample size to detect disease (pathogen)**

The sample size to detect the presence of disease, assuming simple random sampling from a population of *infinite* size (>1,000) =

n = (log α) / [ log (1 – D / N)]

 Where:

n          =          required sample size
α          =          1 – confidence level (usually α = 0.05)

D         =          estimated minimum number of diseased animals in the group
                        (population size \* minimum expected prevalence)

N         =          population size

This formula should be adjusted for use in populations of finite size (<1,000 animals in population)

n = {1 – (α)1/D} \* {N-  D-1}

 2

Where:

n = required sample size
α = 1 – confidence level (usually α = 0.05)

D = estimated minimum number of diseased animals in the group
 (population size \* minimum expected prevalence)

N  = population size

*(Note: these formulas assume the test used is 100% sensitive and specific; no false negatives or false positives will occur)*

###### **Case 6.3. Sample size for freedom from disease** (pg. 49 of Dohoo, Martin, Stryhn, 2003)

 Assume you want to document the absence of *Mycoplasma* from a 200-sow herd and that, based on your experience and the literature, a minimum of 20% of sows would have seroconverted if *Mycoplasma* were present in the herd.

If you test 13 sows and get all negative test results, you can state that you are 95% confident that the prevalence of Mycoplasma in the herd is < 20%.  As you don’t believe that the disease would exist at a prevalence < 20%, you are confident that it is not present.  Note: This assumes the test is 100% sensitive and specific.

###### **Case 6.4 - Salmonella in dairy herd**

 A client dairy producer calls you due to an outbreak of clinical diarrhea with mortality in preweaned calves and cows that have recently calved.  From initial testing of tissue samples from cattle that died, the diagnostic laboratory identified Salmonella Dublin.  You decide to collect separate fecal samples from individual cattle by age group to identify the infected groups of animals in order to provide herd management recommendations to the affected groups.  The fecal samples will be submitted to the diagnostic laboratory for bacterial culture to detect the presence of fecal shedding of Salmonella.

1. How many animals would you need to sample from each group of 100 cattle (each group of different ages) in this dairy herd to detect 5% prevalence of infection with 95% confidence in the adult cow herd?

Instead of calculating sample sizes using the formula, the Table below provides the sample size required to detect various minimum levels of infection in different sized herds or flocks with 95% confidence.  Note: this table assumes you are using a perfect test (100% sensitive, 100% specific).

**Table 1. Sample size to detect pathogen (presence vs absence, 95% confidence) using simple random sampling and assuming perfect test performance**

|  | Population size |  |
| --- | --- | --- |
| Prevalence | 50 | 100 | 500 | 1000 | Infinite |
| 0.1% | 50 | 100 | 500 | 950 | 2995 |
| 1% | 50 | 96 | 225 | 258 | 299 |
| 2% | 48 | 78 | 129 | 138 | 149 |
| 5% | 35 | 45 | 56 | 57 | 59 |
| 10% | 22 | 25 | 28 | 29 | 29 |
| 20% | 12 | 13 | 14 | 14 | 14 |

Answer:

1. How many animals would you need to sample from each group of 100 cattle (each group of different ages) in this dairy herd to detect 5% prevalence of infection with 95% confidence in the adult cow herd?

n = (log α) / [ log (1 – D / N)] = 45 animals to be sampled from each group (or Table above).

1. Because the producer’s testing budget is too constrained to allow sampling to detect 5% prevalence with 95% confidence, what sample size is required to detect 10% prevalence with 95% confidence for each age group of 100 animals?

Answer:

1. Because the producer’s testing budget is too constrained to allow sampling to detect 5% prevalence with 95% confidence, what sample size is required to detect 10% prevalence with 95% confidence for each age group of 100 animals?

n = 25 animals per group

Will this level of sampling fit the producer’s budget? Or will you need to adjust the prevalence detected to a higher level?

##### **2b. What proportion of animals in this herd are infected?**

**Using partial herd testing to estimate the true prevalence of disease.**

You previously observed that sampling a subset of animals from a herd will produce an estimate of the ‘apparent prevalence’ of disease within the herd.  However, the apparent prevalence estimate varied considerably among the different sampling events.  And you already learned, if the sensitivity and specificity of the test are known, the true prevalence of disease [P(D+)] in a population can be estimated (using the Rogen-Gladden estimator to estimate true prevalence):

**Estimated true prevalence (TP)** = (Apparent prevalence + Sp – 1)
                                                                        (Se + Sp – 1)

As an example, if the apparent prevalence = 15%, Se = 36.3%, and Sp = 87.6%, then our estimate of true prevalence is 10.9%.

Estimated TP = (15% + 87.6% - 1) / (36.3% + 87.6% - 1) = 10.9%

###### **Sample size to estimate the true prevalence of disease**

Similar to the situation of sampling animals to detect disease with a certain level of confidence (e.g., 95%), the required sample size to **estimate within-herd prevalence** with a certain level of confidence can also be calculated.  Assuming simple random sampling from a population of infinite size (typically >1,000), the sample size (n) required is:

 n = (1.962 \* p \* (1-p)) / d2

n = sample size

p = estimated prevalence

d = desired precision

 You can adjust this sample size formula for a population of finite size (like a herd or flock when the sample size > 10% of population size)

 Adjusted n = 1 / [(1/n) + (1/N)]

 N = population size

 **Table 2. Sample size to estimate prevalence (95% confidence level) using simple random sampling assuming infinite population and perfect test performance**

|  | Desired precision |
| --- | --- |
| Prevalence | 10 | 5 | 1 |
| 1% | 4 | 15 | 380 |
| 5% | 18 | 73 | 1825 |
| 10% | 35 | 138 | 3457 |
| 20% | 61 | 246 | 6147 |
| 40% | 92 | 369 | 9220 |
| 50% | 96 | 384 | 9604 |
| 60% | 92 | 369 | 9220 |
| 80% | 61 | 246 | 6147 |

###### **Case 6.5. Estimation of prevalence fecal shedding of Salmonella**

You are considering the value of using a new vaccine against Salmonella in dairy herds, and one question you would like to address is how best to estimate the prevalence of Salmonella in your client herds.

How many cows would you need to sample to estimate the prevalence of cows shedding Salmonella in feces from a 1000 cow dairy herd?  Based on reports from the scientific literature, assume a 5% estimated prevalence of Salmonella fecal shedding as the expected (design) prevalence.  Identify the sample size needed for 95% confidence of estimating with desired precision of 5%, assuming perfect tests.

ANSWER:

n = 3.84\*(0.05)\*(0.95) / 0.0025 = 73 cows

###### **Case 6.6. Estimating sample size needed in dog kennel**

A client dog kennel owner asks you about what testing would be needed to estimate the prevalence of *Brucella canis* in her kennel.  She has been reading about the economic costs and hidden nature of this disease, and wondering if she should develop a kennel control program for control and prevention.  How many dogs would you need to sample to **estimate a within-herd prevalence** of *B. canis* in a 300 dog kennel with 1% design prevalence and precision +/- 5%?

1. Estimated sample size assuming infinite population size?
2. Sample size with finite population correction?
3. What additional assumptions are needed for these estimate calculations?

ANSWER:

1. Estimated sample size assuming infinite population size?
* n = (1.962 \* p \* (1-p)) / d2= (3.84\* 0.01 \* 0.99) / 0.0025 = 15.2
1. Sample size with finite population correction?
* Adjusted n = 1 / [(1/n) + (1/N)] = 1 / [(1/15) + (1/ 300)] = 14.5
1. What additional assumptions are needed for these estimate calculations?
* Assume 95% confidence level with simple random sampling and perfect test results (100% Se and 100% Sp).

## **Effect of sensitivity and specificity of an imperfect screening test on “herd-level” sensitivity and “herd-level” specificity**

**Definitions**:

**Individual animal level**

***Individual animal test sensitivity***: = Proportion of truly diseased animals that test positive in an individual animal testing program.

***Individual animal test specificity***: = Proportion of truly non-diseased animals that test negative in an individual animal testing program.

**Herd-level**

***Herd level sensitivity***: = Proportion of truly infected herds that test positive in a herd testing program.

***Herd level specificity***: = Proportion of truly disease-free herds that test negative in a herd testing program.

Note: In a herd-testing program, the **herd** (not the individual animal) is the **unit of concern**.  In this situation, the whole herd is considered to be ‘test-positive’ if one or more animals within the herd test positive. For example, if 16 sheep in a flock are tested for scrapie, and 1 or more sheep test positive, then the herd would be declared ‘Scrapie-positive’ with this herd-testing program.  Conversely, if all 16 sheep tested negative, then the herd would be declared ‘Scrapie-free’ with this herd-testing program.

Table 1 above (**Sample size to detect pathogen** (presence vs absence, 95% confidence) using simple random sampling and assuming perfect test performance) is a tool to estimate the sample size necessary to be 95% confident of detecting a disease in a given herd, given assumptions about the herd size and the likely prevalence of disease (if the disease is present).  This table assumes that the test used is perfect (100% sensitive, 100% specific).

However, most of the screening tests available for use are neither 100% sensitive nor 100% specific. “Herd specificity” and “Herd sensitivity” will be influenced by the individual animal level test sensitivity and specificity.  Once the herd-level sensitivity and herd-level sensitivity are known, the evaluation of the predictive values of positive and negative herd results follows the same methods (using a 2x2 table) as were previously described for individual animal-level testing.

Instead, one formula to estimate herd sensitivity and herd specificity of a herd testing program, when using an imperfect test, is given below.  A reference to a more complicated formula can be found in Dohoo et al., 2003. pg. 112.

Herd-level sensitivity = 1 – (1 – Apparent Prevalence)Number tested

Herd-level specificity = (Individual Animal Specificity)Number tested