Surveys

- Information on disease and associated events, such as productivity, can be obtained from surveys . These involve counting members of an aggregate of units and measuring their characteristics.
- They may be conducted to estimate either a continuous variable such as weight and milk yield, or discrete events such as diseased animals. An important application of surveys in epidemiology is estimation of the prevalence of clinical disease, infection, or seropositive animals from samples of an animal population.

A survey of prevalence can involve:

- a single sample, either to determine prevalence or to determine whether or not disease is present in a group of animals;
- two samples, to compare prevalence;
- three or more samples.

Sampling: some basic concepts

- The validity of sampling theory is based on the assumption that an aggregate of units can be divided into representative subunits, and that characteristics of the aggregate can be estimated from the subunits.

Some definitions

- The target population is the total population about which information is required. Ideally, this should be the population at risk.
- The study population is the population from which a sample is drawn.
- These two populations should be the same. However, for reasons of practicality, this may not be possible.

Sampling populations

Epidemiologists frequently examine populations to:

• Detect the presence of a disease;

• Demonstrate that a disease is not present within a population; and

• Establish the level of occurrence of a disease within a population.

Types of sampling

There are two main types of sampling:

1. . non-probability sampling in which the choice of the sample is left to the investigator;
2. probability sampling in which the selection of the sample is made using a deliberate, unbiased process, so that each sampling unit in a group has an equal probability of being selected; this is the basis of random sampling.

Probability sampling methods

A - Simple random sampling

- A simple random sample is selected by drawing up a list of all animals or other relevant sampling units (e.g., herds) in the study population, and then selecting the sampling units randomly, as described above.

B - Systematic sampling

- Systematic sampling involves selection of sampling units at equal intervals, the first animal being selected randomly.
- For example, if one animal in every 100 were required, then the first animal would be selected randomly from the first 100. If this were animal 63,then the sample would comprise animals, 63, 163, 263, 363 and so on. Systematic sampling is used frequently in industrial quality control, such as selecting samples of goods on a conveyor belt.

C - Stratified sampling

- A stratified random sample is obtained by dividing the study population into exclusive groups (strata), then randomly sampling units from all of the individual strata.
- For example, the strata may be different ranges of herd or flock size, or different geographical regions.
- Stratification can improve the accuracy of a sample because it overcomes the tendency of a simple random sample to either over-represent or under-represent some sections of the sampling frame.

- Thus, if a simple random sample of animals in all herds in a country were selected, it is possible that no animals in very small herds would be chosen. Stratification, which ensures that each group in the population is represented, overcomes this problem.
- The number of sampling units selected from each stratum can be determined by several methods.
- A common method is proportional allocation, where the number of sampling units selected is proportional to the number in each stratum.

D - Cluster sampling

- Sometimes, strata are defined by geographical locations, such as different countries, shires, parishes and villages, or by other categories such as veterinary practices or periods of time during which samples are selected.
- The strata are then termed clusters. Sampling from all of these clusters can be time-consuming and costly.
- This disadvantage can be overcome by selecting a few clusters, and sampling the animals only in these clusters; for example, animals in a few villages or herds could be sampled. This is cluster sampling.
- Commonly, all animals in each selected cluster are sampled; this is one-stage cluster sampling.
- A sample also may be selected in more than one stage. Thus, a sample of clusters can be selected, followed by sub-sampling of some animals in the clusters (in contrast to all animals in one-stage cluster sampling) This procedure is therefore called two-stage cluster sampling; the clusters are the primary units,and the selected members of the sub-samples are the secondary units.
- If the secondary units are the individual members of the study population, there is no point in going further.

What sample size should be selected?

- The question that should be answered in all sample surveys is 'How many animals should be chosen for the survey?' An answer cannot be given without considering the objectives and circumstances of the investigation.

- The choosing of sample size depends on no statistical and statistical considerations. The former include the availability of manpower and sampling frames. The latter are the **desired precision** of the estimate of prevalence and the **expected prevalence** of the disease.

Precision of the estimate of prevalence

- The ability of an estimator to determine the true population value of a variable (i.e., the estimator's precision) can be expressed in terms of the bound on the error of estimation that can be tolerated.
- The error can be defined either absolutely or relatively. For example, an acceptable absolute error of ±2% of a prevalence of 40% represents an acceptable range of 38-42%. A relative error of ±2% of the same prevalence corresponds to 2% of 40%, that is 40% ±0.8%, representing an acceptable range of 39.2-40.8% .

Expected prevalence of the disease

- It may appear paradoxical to suggest that some idea of disease prevalence is necessary before a survey is undertaken, because the objective of the survey is to determine the prevalence.
- However, a general notion is required; if the prevalence is thought to be close to either 0% or 100%, then the confidence interval for a given sample size will be narrower than if the prevalence were close to 50%, that is, fewer animals will be required in the sample to achieve a stipulated width of confidence interval in the former case.
- Information on prevalence might be obtained from other related surveys. However, frequently this information is not available and so estimates have to be made that may be little more than informed guesses ('guestimates').

Estimation of disease prevalence

Simple random sampling

- The approximate sample size required to estimate prevalence in a large (theoretically 'infinite') population can be determined for a defined precision and level of confidence.
- The limits of the associated interval indicate the specified bounds within which the estimate will lie with the defined level of confidence.
- The relevant formula for a 95% confidence

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

where: n = required sample size;

Pexp = expected prevalence;

d = desired absolute precision

- This is acceptable if the size of the study population is large in relation to the sample. However, as the size of the sample relative to the study population increases, the variance of the estimator of the mean of the study population is decreased and the width of the confidence interval is reduced accordingly.
- Therefore, in relatively small populations, it is possible to select a smaller sample than one from a theoretically infinite population to achieve the same degree of precision.
- The required sample size, is given by the following formula:

$$n = \frac{\mathrm{N}_{\mathrm{x}}\,\mathrm{n}}{\mathrm{N} + \mathrm{n}}$$

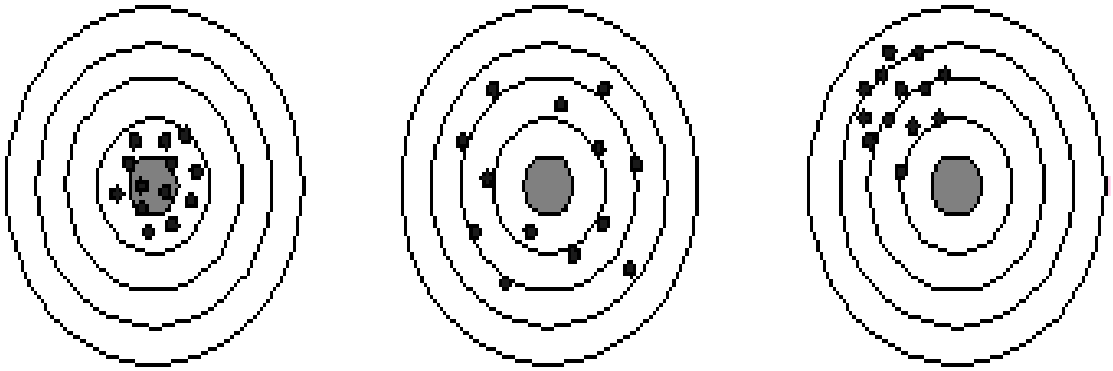
- Where (n) is the sample size, based on an infinite population (obtained from the formula above and (N) is the size of the study population.

Sources of error and how to reduce error

When you derive an estimate from a sample, you want it to be precise and accurate.

- A precise estimate has confidence intervals that are small.
- An accurate has confidence intervals that are centered on the true population value.

- There are two types of error that can exist within a sample estimate: random errors and bias.
- The difference between random error and bias may be explained using the following diagram:



The distribution of bullets fired at the target on the left show little evidence of random error and bias. The distribution of the bullets fired at the centre target show a high degree of random error and a low degree of bias. The distribution of the bullets fired at the target on the right show a low degree of random error and a high degree of bias.

Random error

A. Random error is caused by chance. A random selection of individuals taken to make up a sample will differ slightly from each other. These differences will result in sample estimates that differ slightly from each other and also from the target population. Random error is the inherent error that arises from using a sample to make a measurement of a population. The influence of random error may be reduced by:

1. Increasing the size of the sample taken. Using the central limit theorem it can be demonstrated that a fourfold increase in sample size will result in a halving of the confidence interval.
2. Modifying the sample selection procedure to ensure that only the target group is sampled. For example, you may be interested in the performance of only one particular breed of dairy cow. You can design the study to ensure that you

sample animals only from farms that contain this breed of cow. Stratified sampling is a technique that reduces sample variance by dividing the population into individual strata. Each stratum contains individuals that are similar, and so the variance within strata is less than the variation between strata. You would typically obtain samples from individual strata that have less variation than similar-sized samples obtained from the whole (unstratified) population.

3. Using an appropriate scale of measurement. Ratio estimators may result in areduction in confidence intervals in some situations. Suppose, for example, that you wish to determine whether farmed lambs have reached the correct weight for sale. You could take a sample of lambs and estimate the average weight of the sample and from that an associated confidence interval. If the weight of lambs in the population is quite variable and you do not select a large sample it is likely that the associated confidence interval will be wide (and will include the target value). An alternative is to dichotomously classify each lambs weight within the sample with respect to the target weight (i.e. describe it as either above or below target weight). You can then calculate an estimate of the proportion of lambs that have obtained target weight (along with associated confidence intervals). You are more likely to produce narrow confidence intervals for this ratio estimate and arethus able to make a more confident decision regarding the sale of the lambs.

B - Bias

- Bias is caused by systematic error, a systematic error being one that is inherent to the technique being used that results in a predictable and repeatable error for each observation.
- Bias may present itself in two ways:

1. Non-observational errors are due to inappropriate sample selection. These errors may arise from failure to include an important group of individuals within the sampling frame (resulting in their exclusion from selection), or as a result of missing data. In some situations data may be missing from a particular group of individuals within the sample.

2. Observational errors are due to inappropriate measurements. These may be attributable to false responses (i.e. participants make untrue statements) or to measurement errors.