

Exercise - sampling

Sampling units

Carefully read the following examples and then choose your preferred sampling unit. Explain your choice.

1. You wish to estimate the economic losses arising from lameness in sheep in a region. You decide to conduct a survey to estimate the prevalence of lameness in sheep in the region.
 2. You wish to conduct a survey to determine the incidence of foot-and-mouth disease (FMD) outbreaks in pigs from an intensively farmed endemic region of a country during the last year.
 3. You believe that poor stockyards and overcrowding of cattle in abattoirs before slaughter is contributing to carcass bruising. You plan to conduct a study to investigate this proposed risk factor.
-
1. Lameness is a disorder that affects individuals, so the sampling unit is the sheep. You need to obtain a sample of sheep that is representative of the total sheep population within the region and measure the prevalence of lameness levels within the sample.
 2. The key word in this question is 'outbreak.' An outbreak occurs within a group of individuals. Thus our unit of interest is the intensive pig farm. Because FMD is highly contagious, spread within an infected piggery is likely to be rapid if preventive measures are not implemented. Therefore the positive diagnosis of FMD within one or more pigs on an intensive pig farm would represent an outbreak on that farm. An estimate of the incidence of outbreaks of FMD in intensive piggeries within the region in the last year would be obtained from a random sample of intensive piggeries within the region.
 3. Your hypothesis is that poor facilities and inappropriate animal handling practices at individual abattoirs contribute to the bruising of carcasses. The unit of interest is therefore the abattoir. You need to obtain a random sample of abattoirs within the region or country. The abattoir facilities can be measured and an assessment made of the animal handling practices of the abattoir. This information can be compared to the data on carcass bruising for each abattoir.

Subpopulations

It is frequent in animal production systems to divide animals into separate groups. Dairy farms (for example) manage lactating and non-lactating stock as separate groups. These divisions can make the collection of a representative sample from a population difficult. Give three examples from livestock enterprises where free mixing of animals is prevented. List all of the sub-groups that may be present in each. How would you obtain a representative sample from each enterprise?

Extensive beef cattle farms: cows and unweaned calves, weaned calves, yearlings and heifers, bulls.

Intensive pig production: farrowing sows, dry sows, weaners, growers, finishers, boars.

Extensive sheep production: wethers, rams, weaners.

The samples required will depend upon the purpose of sampling. If you wish to estimate the prevalence of a particular gene within the population, then every animal on the farm should have an equal probability of selection into your sample. A sampling frame could be made using the individual animal identification. Individuals can then be selected at random from the sampling frame. In this way, animals have an equal chance of selection into the sample irrespective of their actual physical position on the farm.

Alternatively, you might wish to selectively sample from subgroups on the farm. The incidence of abortion on farms is only obtained by examining a sample obtained from the population of pregnant females. A sampling frame of pregnant females would be used to select the sample. In the extensive beef farm example above, the pregnant female population would be contained within the main herd of cows, but would also include the population of yearling heifers. The yearling heifers are likely to be carried separately from the main herd of cows.

Stratified sampling

Suppose you wish to determine the prevalence of disease within the pig population of a region. Previous surveys have indicated that 70% of the region's pigs are located in very large, intensive specialised pig farms, 20% of pigs are found within smaller farming units (frequently as a secondary industry on large dairy farms), and 10% of pigs are kept singly within small plots around towns (by people whose major occupation is not farming). With proportional stratification, a sample would be selected at random from within each stratum such that the aggregated sample would consist of 70%

pigs obtained from the large intensive farms, 20% pigs obtained from the smaller pig farms, and 10% pigs obtained from small plots near towns.

1. Explain why it is important for each stratum of pigs to be represented in this sample for the prevalence survey.
2. Assume that the disease that you are investigating is leptospirosis. Combine your knowledge of leptospirosis with the description of the farming systems. Is the epidemiology of leptospirosis likely to vary between the different strata?
 1. To produce an accurate estimate of the population prevalence you must ensure that all groups of pigs within the population are represented in the sample. You have three distinct groups (strata) of pigs within the population: intensively farmed pigs, small-unit pigs and individually housed pigs. It is likely that many diseases of pigs are not evenly distributed among the different strata types. Thus, in order to represent the whole pig population accurately, you must ensure that your sample contains representatives from each stratum. Also, because the strata are of different sizes, you must ensure that the composition of your sample is similar to the composition of the pig population. For example, if a stratum contained 65% of the total population of animals, then 65% of animals within your sample should originate from this stratum.
 2. Leptospirosis is a contagious disease caused by various leptospiral bacteria. Transmission is mostly via skin or mucous membrane contact with contaminated urine or other infected pigs or rodents. Vaccination programs are effective at limiting spread of disease and reducing maintenance of the organism within the population. It is likely that there will be differences between the pig farm strata in: (a) level of pig-to-pig contact, (b) the amount of contact between pigs and rodents, and (c) the extent of vaccination coverage of the population. These epidemiological differences between the strata are likely to result in a corresponding variation in the incidence and prevalence of leptospirosis.

Cluster sampling

Suppose that you wish to conduct a survey to determine if overfishing of any species of ocean fish is occurring within a country. There is no registration requirement to fish in this country, and so a sampling frame of individuals who fish cannot be drawn. However, all ocean-going fishing boats must be based at deep harbours (in other words, the boats are clustered around deep harbours). A study of a map of the country indicates that there are 30 deep harbours capable of supporting ocean-fishing vessels. A random selection of harbours is made. All the fishing boats moored within each selected harbour are identified and listed. A random selection of boats is made and the catch from each boat is examined on a designated day. The quantity of each fish species present on each boat is estimated. This is used to provide an estimate of the total fishing pressure for each species of interest.

1. How many sampling stages were employed in this example?
2. Describe your approach to the following scenario. When the first selected harbour is examined and the list of ocean-going fishing vessels is drawn for sampling, you notice that there are two distinct types of boats present. One boat is smaller and designed for daytime fishing only (it has no sleeping cabins). The other type of boat is much larger and designed for longer and more extensive fishing trips (it has sleeping cabins for the crew). When you examine the catch from each type of boat, you notice that the catch from the smaller boat consists predominantly of fish species that live close to shore. The larger fishing boat's catch is dominated by larger migratory species of fish. You wish to obtain accurate statistics on all the fish species that you have identified on both types of boat. In light of this new information, what modifications would you make to your sampling procedure?
 1. This was a two-stage sampling strategy. We first randomly selected the harbours (stage 1) and then randomly selected the boats within each harbour (stage 2).
 2. The problem is that we have two different strata of fishing boats: boats that are used close to shore and boats that are used in deeper waters. The different types of boat catch different species of fish. In order to obtain accurate estimates of fishing pressure for each species of fish within the study you need to draw up separate sampling frames of fishing boats for each boat stratum. You can then randomly select a stratified sample of boats from each harbour based upon the relative size of each boat stratum. Alternatively, you could base the boat stratum sampling proportion on the proportion of the total harbour fish catch made by each boat size stratum. Either way, we can now produce accurate estimates of the total fishing pressure for each fish species of interest for each harbour and for the country.

Bias 1

Examine the following three examples and decide whether the disease estimate obtained from the sample will be an accurate representation of the disease level in the population. You may assume that the diagnostic tests employed

are perfect --- in other words, that they have 100% sensitivity and 100% specificity. If you consider an estimate to be inaccurate, outline how you would modify the study design to improve the result.

1. There are 1000 sheep in a mob. You take a sample of 100 sheep at random and find that 5 are diseased. You conclude that the disease prevalence in the mob is approximately 5%.
2. You notice that many of the free-roaming village chickens in a region are sick. You decide to identify the main disease that is affecting these chickens and to estimate the prevalence within the population. You visit some chicken farms in the area, sample 500 birds from farm sheds and diagnose coccidiosis in 100 birds. You conclude that coccidiosis is the main disease of village chickens and that the disease is present at a prevalence of approximately 20%.
3. You wish to determine the prevalence of tuberculosis in a native species of wild animal within a certain area. This animal is difficult to catch, so you decide to collect and examine all dead specimens found within the study area. You collect 150 specimens and diagnose tuberculosis in 23 individuals. You conclude that the prevalence of tuberculosis in the wild population is 15%.
 1. This sampling process uses individual sheep as the sampling unit. Assuming sampling has been completely random, the prevalence estimate obtained should be accurate (i.e. without bias). In practice, a systemic random sampling approach would be used on sheep farms.
 2. You have sampled from the wrong population in this example. You observed disease within the free-roaming chicken population of a village, but measured disease from a sample of housed birds. It is very likely that the diseases and exposures of free-roaming village chickens are different from those of housed farmed chickens. You cannot accurately infer that free-roaming chickens have an identical disease distribution to housed farmed chickens. Thus your results are subject to sampling bias. Your results may, however, provide good insight into the disease distribution of housed farmed birds, although this depends on your method of sampling the chicken farms. If you sampled the chicken farms randomly within the region, then your results are likely to be unbiased. If you sampled purposively (i.e. selected the closest farms or the biggest farms) then your results will be biased.
 3. You have inferred that the disease distribution within the population of dead animals is the same as the disease distribution within the population of live animals. This is almost certainly incorrect: we would all agree that being dead is less healthy than being alive! It is quite likely that the prevalence of tuberculosis lesions in dead animals is greater than the prevalence of tuberculosis lesions in live animals. Your population of interest is the population of live animals within the region. In order to determine the prevalence of tuberculosis within this population you must sample from that population. This is likely to be very difficult because you must trap a sample of animals that is representative of the population. Often traps are adept at catching certain age groups of animals (such as adults or immature animals) or rely upon individual behaviour characteristics (such as food-gathering individuals).

Bias 2

Identify and classify the bias present in each of the following studies:

1. A survey is planned to determine the prevalence of disease within the horse population of a developed country. A list of all racehorse trainers is compiled and a random selection of trainers is made. All horses present in the stables of the selected trainers are examined.
2. A study has been conducted to identify the major cause of lameness diagnosis made by clinical veterinarians upon examination of cattle presenting with gait abnormalities within a region. A random selection of clinic records of veterinary visits to cattle with gait abnormalities is drawn from the veterinary practices within the region. When the clinic case records data are examined, it is noticed that 10% of selected cows and 80% of selected bulls did not have the foot lifted for veterinary examination and thus no foot measurement was recorded.
3. A milk company reports that residues of a banned parasiticide are regularly being found in samples taken from milk tankers. You decide to try to identify the farmers who continue to use this banned product. A questionnaire is sent to all of the company's farmers (i.e. a census). When you examine the responses from each farmer you notice that all farmers record that they are not currently using the banned parasiticide.
4. You have collected serum samples from animals as part of a study designed to measure the seroprevalence of a disease within a country. Serum samples are sent at random to one of two diagnostic laboratories for analysis. You notice that the seroprevalence reported from laboratory A is 10%, while laboratory B reports a seroprevalence of 3%. You send each laboratory 100 test serum samples taken from known positive animals. Laboratory A classifies 98 of the 100 positive test samples correctly, whereas laboratory B correctly classifies 30 of the 100 positive test samples.

1. This is an example of a non-observational error. It has arisen because a group of the population has been excluded from the sampling frame. Not all horses in the country are racehorses. Many horses are kept simply for pleasure and these have been excluded from the sampling frame.
2. Again, this is a non-observational error. This time, the problem has arisen because data were not obtained from an important group within the population: the bulls. Relatively more bulls than cows with gait abnormalities did not have a foot examination or foot measurement recorded. While this is physically understandable (bulls are much heavier and more difficult to handle than cows), it is not sound sampling technique. There is now a significant bias in this study. Ideally, all the necessary measurements should be taken from sampled animals irrespective of the difficulties associated with collection of that data.
3. This is an example of an observational error. You have factory evidence suggesting that dairy farmers are using the banned chemical; however, all farmers deny currently using the banned parasiticide. The farmers who are using the banned parasiticide have made incorrect statements. Many may have deliberately answered incorrectly in order to avoid possible repercussions.
4. This is also an observational error; however, this time the error is due to a measurement error occurring within laboratory B. The diagnostic test used in laboratory B has an unacceptable low sensitivity and as such is failing to diagnose the presence of disease in a significant proportion of diseased individuals.

Estimating the prevalence of disease

It is decided to do a survey to estimate the prevalence of disease X in a population of cattle. Three experts are asked for their opinions about the expected prevalence and they reply: 75%, 50% and 25%. Assuming that there are 1 million head of cattle in the study area, a desired absolute precision of 5% and a desired confidence level of 95%.

1. Calculate the needed sample size according to the three expert opinions.
2. When prevalence is unknown and you have absolutely no idea about its expected value, what prevalence estimate should you use for the sample size calculation?


$z = 1.96$
 $P = 0.75$
 Absolute error = 0.05

$$n = [z^2 \times (1 - P) \times P] \div \epsilon^2$$

$$n = [1.96^2 \times (1 - 0.75) \times 0.75] \div (0.05 \times 0.05)$$

$$n = 288$$

When the expected prevalence is 75%, 288 samples are required.

library(epiR) 
 epi.samplesize(N = 1E06, sd = 0.75, epsilon = 0.05, method = "proportion",
 conf.level = 0.95)

When the expected prevalence is 25%, 50%, and 75% the exact formula specifies 288, 384, and 288 samples, respectively. When prevalence is unknown, an expected prevalence of 50% provides the largest estimated sample size.

Intramammary antibiotics

A survey is to be conducted in a state where there are 2500 veterinary practices. The purpose of the survey is to estimate the average retail price of a single dose of intramammary antibiotic. An estimate is needed that is within 10% of the true value of the average retail price in the state. Data collected earlier indicates an average price of \$7.00 with a standard deviation of \$1.40. How many practices should be included in the survey to be 95% confident that the surveyed value will be within 10% of the average retail price?

$z = 1.96$
 $SD = 1.40$
 Absolute error = $0.10 \times 7.00 = 0.70$

$$n = [z^2 \times SD^2] \div \epsilon^2$$

$$n = [1.96 \times 1.96 \times 1.40 \times 1.40] \div (0.07 \times 0.07)$$
$$n = 15$$

```
library(epiR)
epi.simplesize(N = 2500, sd = 1.40, epsilon = 0.07, method = "mean", conf.level = 0.95)
```

A sample of 15 practices are required to meet the conditions of the survey.

Calf body weights

We want to estimate the mean bodyweight of calves on a large dairy farm. On examination, we reckon that the minimum weight in the mob is 90 kg and the maximum weight 150 kg. We would like to be 99% certain that our estimate is within 5 kg of the true mean. How many calves should we weigh?

For a variable that is normally distributed the standard deviation is the difference between the variable minimum and maximum divided by 6. In this example we've estimated minimum and maximum body weight to be 90 kg and 150 kg, respectively. This gives an estimated standard deviation of $(150 - 90) \div 6 = 10$ kg.

$$z = 2.58$$
$$SD = 10$$
$$\text{Absolute error} = 5$$

$$n = [z^2 \times SD^2] \div \epsilon^2$$
$$n = [2.58 \times 2.58 \times 10 \times 10] \div (5 \times 5)$$
$$n = 27$$

```
library(epiR)
epi.simplesize(N = 1E06, sd = 10, epsilon = 5, method = "mean", conf.level = 0.99)
```

We need to weigh at least 27 calves to be 99% certain that the sample mean is within 5 kg of the population mean.