Sample Size Estimation using epiR

Mark Stevenson and Evan Sergeant

12 Nov 2026

Contents

1 Prevalence estimation 1
2 Prospective cohort studies 2
3 Case-control studies 2
4 Non-inferiority trials 2
5 Estimation of a population parameter, one-stage cluster sampling 3
6 Estimation of a population parameter, two-stage cluster sampling 3
References 5

A review of sample size calculations in (veterinary) epidemiological research is provided by Stevenson (2021).

1 Prevalence estimation

The expected seroprevalence of brucellosis in a population of cattle is thought to be in the order of 15%. How many cattle need to be sampled and tested to be 95% certain that our seroprevalence estimate is within 20% of the true population value. That is, from 15 - (0.20 × 0.15) to 15 + (0.20 × 0.15 = 0.03) i.e., from 12% to 18%. Assume the test you will use has perfect sensitivity and specificity. The size of the population of cattle at risk is unknown so we set `N = NA`:

```
library(epiR)
epi.sssimpleestb(N = NA, Py = 0.15, epsilon = 0.20, error = "relative",
   se = 1, sp = 1, nfractional = FALSE, conf.level = 0.95)
#> [1] 545
```

A total of 545 cows need to be sampled to meet the requirements of the study. Let's now imagine that we have a reasonable estimate of the size of the cattle population at risk, 4000. Re-run `epi.sssimpleestb`, specifying `N`:

```
epi.sssimpleestb(N = 4000, Py = 0.15, epsilon = 0.20, error = "relative",
   se = 1, sp = 1, nfractional = FALSE, conf.level = 0.95)
#> [1] 480
```

If the size of the population at risk is 4000, 480 cows need to be sampled to meet the requirements of the study. When a value for `N` is provided in `epi.sssimpleestb` the function automatically applies a finite correction factor. In this example if we assumed an infinite (i.e., very large) population size the required sample size was 545. When the population size was (only) 4000, the required sample size reduced to 480.

2 Prospective cohort studies

A prospective cohort study of dry food diets and feline lower urinary tract disease (FLUTD) in mature male cats is planned. A sample of cats will be selected at random from the population of cats in a given area and owners who agree to participate in the study will be asked to complete a questionnaire at the time of enrollment. Cats enrolled into the study will be followed for at least 5 years to identify incident cases of FLUTD. The investigators would like to be 0.80 certain of being able to detect when the risk ratio of FLUTD is 1.4 for cats habitually fed a dry food diet, using a 0.05 significance test. Previous evidence indicates that the incidence risk of FLUTD in cats not on a dry food (i.e., 'other') diet is around 50 per 1000. Assuming equal numbers of cats on dry food and other diets are sampled, how many cats should be sampled to meet the requirements of the study?

```
epi.sscohortt(irexp1 = 70/1000, irexp0 = 50/1000, FT = 5, n = NA,
   power = 0.80, r = 1, design = 1, sided.test = 2, nfractional = FALSE,
   conf.level = 0.95)$n.total
#> [1] 2080
```

A total of 2080 subjects are required (1040 exposed and 1040 unexposed).

It's important to remember that you can use the `epi.sscohortt` (and other sample size functions in `epiR`) to return the study power if a value for `n` is provided. Continuing the example above, imagine that only 1500 cats were enrolled into the study. What is the expected study power? Here we set `n = 1500` and `power = NA` in `epi.sscohortt`:

```
epi.sscohortt(irexp1 = 70/1000, irexp0 = 50/1000, FT = 5, n = 1500,
   power = NA, r = 1, design = 1, sided.test = 2, nfractional = FALSE,
   conf.level = 0.95)$power
#> [1] 0.6629312
```

If only 1500 cats are enrolled into the study the expected study power is 0.66.

3 Case-control studies

A case-control study of the relationship between white pigmentation around the eyes and ocular squamous cell carcinoma in Hereford cattle is planned. A sample of cattle with newly diagnosed squamous cell carcinoma will be compared for white pigmentation around the eyes with a set of controls. Assuming an equal number of cases and controls, how many study subjects are required to detect an odds ratio of 2.0 with 0.80 power using a two-sided 0.05 test? Previous surveys have shown that around 0.30 of Hereford cattle without squamous cell carcinoma have white pigmentation around the eyes.

```
epi.sscc(N = NA, OR = 2.0, p1 = NA, p0 = 0.30, n = NA, power = 0.80,
   r = 1, phi.coef = 0, design = 1, sided.test = 2, conf.level = 0.95,
   method = "unmatched", nfractional = FALSE, fleiss = FALSE)$n.total
#> [1] 282
```

If the true odds for squamous cell carcinoma in exposed subjects relative to unexposed subjects is 2.0, we'll need to enrol 141 cases and 141 controls (282 cattle in total) to reject the null hypothesis that the odds ratio equals one with probability (power) 0.80. The Type I error probability associated with the test of this null hypothesis is 0.05.

4 Non-inferiority trials

Suppose a pharmaceutical company would like to conduct a clinical trial to compare the efficacy of two antimicrobial agents when administered orally to patients with skin infections. Assume the true mean cure rate of the treatment is 0.85 and the true mean cure rate of the control is 0.65. We consider a difference of less than 0.10 in cure rate to be of no clinical importance (i.e., delta = 0.10). Assuming a one-sided test size of 5% and a power of 80%, how many subjects should be included in the trial?

```
epi.ssninfb(treat = 0.85, control = 0.65, delta = 0.10, n = NA,
   r = 1, power = 0.80, nfractional = FALSE, alpha = 0.05)$n.total
#> [1] 50
```

A total of 50 subjects need to be enrolled in the trial, 25 in the treatment group and 25 in the control group.

5 Estimation of a population parameter, one-stage cluster sampling

An aid project has distributed cook stoves in a single province in a resource-poor country. At the end of three years, the donors would like to know what proportion of households are still using their donated stove. A cross-sectional study is planned where villages in a province will be sampled and then all households (approximately 75 per village) will be visited to determine if their donated stove is still in use. A pilot study of the prevalence of stove usage in five villages showed that 0.46 of householders were still using their stove and the intracluster correlation coefficient (ICC) for stove use within villages is in the order of 0.20. If the donor wanted to be 95% confident that the survey estimate of stove usage was within 10% of the true population value, how many villages (primary sampling units) need to be sampled?

```
epi.ssclus1estb(N.psu = NA, b = 75, Py = 0.46, epsilon = 0.10,
   error = "relative", rho = 0.20, se = 1, sp = 1, nfractional = FALSE,
   conf.level = 0.95)$n.psu
#> [1] 96
```

A total of 96 villages need to be sampled to meet the requirements of the study.

Continuing the example above, we are now told that the number of households per village varies. The average number of households per village is 75 with a 0.025 quartile of 40 households and a 0.975 quartile of 180. Assuming the number of households per village follows a normal distribution the expected standard deviation of the number of households per village is in the order of (180 - 40) $\div$ 4 = 35. How many villages (primary sampling units) need to be sampled?

```
epi.ssclus1estb(b = c(75,35), Py = 0.46, epsilon = 0.10,
   error = "relative", rho = 0.20, se = 1, sp = 1, nfractional = FALSE,
   conf.level = 0.95)$n.psu
#> [1] 115
```

A total of 115 villages need to be sampled to meet the requirements of the study.

6 Estimation of a population parameter, two-stage cluster sampling

This example is adapted from Bennett et al. (1991). We intend to conduct a cross-sectional study to determine the prevalence of disease X in a given country. This is a hierarchical data set with individuals (secondary sampling units) clustered within villages (primary sampling units). The expected prevalence of disease X is thought to be around 20%. Previous studies report an intracluster correlation coefficient for disease X to be 0.02. Suppose that we want to be 95% certain that our estimate of the prevalence of disease is within 5% of the true population value and that we'll sample 20 individuals per village. How many villages should be sampled to meet the requirements of the study? Assume that you don't know the total number of villages in the country, so set `N = NA` in `epi.simpleestb`:

```
# From first principles:
n.crude <- epi.sssimpleestb(N = NA, Py = 0.20, epsilon = 0.05 / 0.20,
   error = "relative", se = 1, sp = 1, nfractional = FALSE, conf.level = 0.95)
n.crude
#> [1] 246

# A total of 246 individuals (SSUs) need to be enrolled into the study.

  # Calculate the design effect:
```

```
rho <- 0.02; b <- 20
D <- rho * (b - 1) + 1; D
#> [1] 1.38

# The design effect is 1.38 so our crude sample size estimate needs to be
# increased by a factor of 1.38.

n.adj <- ceiling(n.crude * D)
n.adj
#> [1] 340

# After accounting for lack of independence in the data a total of 340
# individuals (SSUs) need to be enrolled into the study.

# How many villages (PSUs) are required?
ceiling(n.adj / b)
#> [1] 17

# A total of 17 villages need to be sampled to meet the requirements of
# the study.

# Do all of the above using epi.ssclus2estb:
epi.ssclus2estb(N.psu = NA, b = 20, Py = 0.20, epsilon = 0.05 / 0.20,
   error = "relative", rho = 0.02, se = 1, sp = 1, nfractional = FALSE,
   conf.level = 0.95)
#> Warning: The calculated number of primary sampling units (n.psu) is 17. At
#> least 25 primary sampling units are recommended for two-stage cluster
#> sampling designs.
#> $n.psu
#> [1] 17
#>
#> $n.ssu
#> [1] 340
#>
#> $DEF
#> [1] 1.38
#>
#> $rho
#> [1] 0.02
```

A total of 17 villages need to be sampled to meet the requirements of this study. From each sampled village 20 individuals are sampled which means that at total of $17 \times 20 = 340$ individuals are required. Function `epi.ssclus2estb` returns a warning message, reminding you that at least 25 primary sampling units are recommended for two-stage cluster sampling designs.

Continuing the brucellosis prevalence example (above) being seropositive to brucellosis is likely to cluster within herds. Otte and Gumm (1997) cite the intracluster correlation coefficient for Brucella abortus in cattle to be in the order of 0.09. Adjust your sample size of 545 cows to account for lack of independence in the data, i.e., clustering at the herd level. Assume that $b = 10$ animals will be sampled per herd:

```
n.crude <- epi.sssimpleestb(N = NA, Py = 0.15, epsilon = 0.20,
   error = "relative", se = 1, sp = 1, nfractional = FALSE, conf.level = 0.95)
n.crude
#> [1] 545
```

```
rho <- 0.09; b <- 10
D <- rho * (b - 1) + 1; D
#> [1] 1.81

n.adj <- ceiling(n.crude * D)
n.adj
#> [1] 987

# Similar to the example above, we can do all of these calculations
# using epi.ssclus2estb:
epi.ssclus2estb(N.psu = NA, b = 10, Py = 0.15, epsilon = 0.20,
   error = "relative", rho = 0.09, se = 1, sp = 1, nfractional = FALSE,
   conf.level = 0.95)
#> $n.psu
#> [1] 99
#>
#> $n.ssu
#> [1] 986
#>
#> $DEF
#> [1] 1.81
#>
#> $rho
#> [1] 0.09
```

After accounting for clustering at the herd level we estimate that a total of $(545 \times 1.81) = 986$ cattle need to be sampled to meet the requirements of the study. If 10 cows are sampled per herd this means that a total of $(986 \div 10) = 99$ herds are required.

References

Bennett, S, T Woods, W Liyanage, and D Smith. 1991. "A simplified general method for cluster-sample surveys of health in developing countries." *World Health Statistics Quarterly* 44: 98–106.

Otte, JM, and ID Gumm. 1997. "Intra-cluster correlation coefficients of 20 infections calculated from the results of cluster-sample surveys." *Preventive Veterinary Medicine* 31: 147–50.

Stevenson, MA. 2021. "Sample size estimation in veterinary epidemiological research." Journal Article. *Frontiers in Veterinary Science* 7: 539573.