Surveillance System Assessment using epiR

Mark Stevenson and Evan Sergeant

12 Nov 2026

Contents

1 Definitions 1
2 An approach for thinking about surveillance design and assessment 2
3 Representative sampling 3
4 Risk-based sampling 7
5 Analysis of passive surveillance data 13
6 Miscellaneous functions 14
References 15

Surveillance is defined as the on-going systematic collection, collation and interpretation of accurate information about a defined population with respect to disease and/or infection, closely integrated with timely dissemination of that information to those responsible for control and prevention measures (Thacker and Berkelman 1988; World Organisation for Animal Health 2022).

Surveillance is a tool for monitoring changes in health related events in a defined population with specific goals relating to: (1) the detection of disease incursions, both new and emerging; (2) the assessment of progress in terms of control or eradication of selected diseases and pathogens; (3) demonstration of disease freedom for trading partners; and (4) identification of hazards or risk factors for disease outbreaks.

This vignette provides instruction on the way R and `epiR` (specifically, the surveillance functions within `epiR`) can be used for: (1) the design of disease surveillance programs; and (2) the design of surveillance programs to provide a quantitative basis for claims for disease freedom.

1 Definitions

Design prevalence. The design prevalence — also known as the minimum detectable prevalence, maximum acceptable or permissible prevalence, and minimum expected prevalence, P^*, (P star) — is a fixed value for prevalence used for testing the hypothesis that disease is present in a population of interest. The null hypothesis is that disease is present in the population at a prevalence equal to or greater than the design prevalence. If a sufficient number of samples are collected and all return a negative result we can reject the null hypothesis and accept the alternative hypothesis, concluding that the prevalence of disease in the population is less than the specified design prevalence.

A design prevalence is not related to any actual prevalence of disease in the population under study. It is not subject to uncertainty or variability and therefore doesn't need to be described using a distribution.

Cluster-level design prevalence refers to a design prevalence assigned at the cluster (e.g., village, herd or household) level.

Unit-level design prevalence refers to a design prevalence assigned at the individual unit (e.g., cow, sheep, bird) level. The unit-level prevalence of disease can be applied either within clusters (e.g., herds, flocks, villages) or across broader, unclustered populations (e.g., human populations or wildlife).

Surveillance system. A surveillance system (SS) is a set of procedures to collect, collate and interpret information about a defined population with respect to disease.

Surveillance system components. Surveillance system components (SSCs) are the individual activities (e.g., on-farm testing, abattoir surveillance, disease hotlines) that make up a surveillance system.

Surveillance units. Surveillance units (SUs) are the individual units that get assessed within each surveillance system component. For the surveillance system components listed (on-farm testing, abattoir surveillance, disease hotlines) the corresponding surveillance units would be individual animals, carcasses and individual phone reports, respectively.

Surveillance unit sensitivity. Surveillance unit sensitivity (SUSe) is defined as the average probability that a surveillance system unit (selected from those processed) will return a positive surveillance outcome, given that disease is actually present in the population at a level equal to or greater than the specified design prevalence.

Surveillance component sensitivity. Surveillance component sensitivity (SCSe) is the average probability that a surveillance system component will return a positive surveillance outcome, given disease is present in the population at a level equal to or greater than the specified design prevalence.

Surveillance system sensitivity. Surveillance system sensitivity (SSSe) is defined as the average probability that a surveillance system (as a whole) will return a positive surveillance outcome, given disease is present in the population at a level equal to or greater than the specified design prevalence.

2 An approach for thinking about surveillance design and assessment

The first thing to consider when you're designing a surveillance program is to consider the sampling method that will be used. If the aim of a surveillance program is to detect a specific pathogen the sampling method will usually be either **representative** or **risk-based**. Other options include the situation where you might observe every individual in a population (a **census**) or where you might take no active steps to collect surveillance data but instead rely on stakeholders to report their observations on a voluntary basis (**passive surveillance**).

Once you have the method of sampling defined think about how the surveillance system will be designed and how you might assess the surveillance system once it has been designed and implemented.

In terms of design, firstly specify the sampling method to be used and then determine how many surveillance system units will be sampled (usually, to achieve a pre-specified surveillance system sensitivity).

Once samples have been collected and tested or if you are making an assessment of an existing surveillance system, you then might want to answer the question: If the disease of interest is actually present in the population what is the chance that the surveillance system will actually detect it? This question can be expressed in three ways:

1. What is the surveillance system sensitivity (SSSe)? or
2. What is the probability that the prevalence of disease is less than the specified design prevalence? or
3. What is the surveillance system's negative predictive value?

The remainder of this vignette follows the structure outlined above. For each sampling method (representative, risk-based, census and passive) we provide notes and examples on the use of `epiR` for sample size estimation, estimation of surveillance system sensitivity and estimation of the probability of disease freedom. While 'estimation of the probability of disease freedom' is the name assigned to the last group of analyses a more correct label would be 'estimation of the probability that the prevalence of disease is less than a specified design prevalence' (i.e., the negative predictive value of the surveillance system). Be aware that we can only truly demonstrate disease freedom if every member of the population at risk is assessed using a test with perfect diagnostic sensitivity and perfect diagnostic specificity.

3 Representative sampling

Sample size estimation

The sample size functions for surveillance representative sampling in `epiR` fall into two classes: sampling to achieve a defined probability of disease freedom and sampling to achieve a defined surveillance system sensitivity.

The surveillance system sensitivity sample size functions include those for simple random sampling and two stage sampling. Two stage sampling is the preferred (indeed, the only practical approach) when a population is organised in clusters (e.g., cows within herds, households within villages). With two stage sampling clusters (herds, villages) are sampled first and then, from within each selected cluster, individual surveillance units (animals, individuals) are sampled. In this context we talk about herds and villages as **primary sampling units** (PSUs) and animals and individuals as **secondary sampling units** (SSUs).

Table 1: Table 1: Functions in epiR to estimate sample size using representative population sampling data.

Sampling	Outcome	Detail	Function
Representative	Pr DF [a]	Imperf Se, perf Sp	rsu.sspfree.rs
Representative	SSe	Imperf Se, perf Sp	rsu.sssep.rs
Two stage representative	SSe	Imperf Se, perf Sp	rsu.sssep.rs2st
Representative	SSe	Imperf Se, perf Sp, known N	rsu.sssep.rsfreecalc
Pooled representative	SSe	Imperf Se, perf Sp	rsu.sssep.rspool

[a] Pr DF: Probability of disease freedom.

EXAMPLE 1

A cross-sectional study is to be carried out to confirm the absence of brucellosis in dairy herds using a bulk milk tank test assuming a design prevalence of 5%. Assume the total number of dairy herds in your study area is unknown and large and the bulk milk tank test to be used has a diagnostic sensitivity of 0.90 and specificity of 1.00. How many herds need to be sampled to achieve a surveillance system sensitivity (`se.p`) of 95%? That is, what is the probability that disease will be detected if it is present in the population at the designated design prevalence?

```
library(epiR)
rsu.sssep.rs(N = NA, pstar = 0.05, se.p = 0.95, se.u = 0.90)
#> [1] 66
```

A total of 66 herds need to be sampled and tested.

This question can be asked in another way. If our prior estimate of the probability that the population of herds is free of disease is 0.50 and we believe that there's a 1% chance of disease being introduced into the population during the next time period, how many herds need to be sampled to be 95% confident that disease is absent (i.e., less than the design prevalence) if all tests are negative?

```
rsu.sspfree.rs(N = NA, prior = 0.50, p.intro = 0.01, pstar = 0.05, pfree = 0.95, se.u = 0.90)
#> $n
#> [1] 65
#>
#> $se.p
#> [1] 0.9484106
#>
#> $adj.prior
```

```
#> [1] 0.495
```

A total of 65 herds need to be sampled (similar to the value calculated above). Note that function `rsu.sssep.rs` returns the sample size to achieve a desired surveillance system sensitivity ('what's the probability that disease will be detected?'). Function `rsu.sspfree.rs` returns the sample size to achieve a desired (posterior) probability of disease freedom.

Now assume that it is known that there are 500 dairy herds in your study area. Revise your sample size estimate to achieve the desired surveillance system sensitivity in light of this new information.

```
rsu.sssep.rs(N = 500, pstar = 0.05, se.p = 0.95, se.u = 0.90)
#> [1] 63
```

A total of 63 herds need to be sampled and tested.

The sample size calculations presented so far assume the use of a test with perfect specificity (that is, if a sample returns a positive result we can be 100% certain that the herd is positive and disease is actually present in the population).

Consider the situation where a test with imperfect specificity is used. Imperfect specificity presents problems for disease freedom surveys. If a positive test result is returned, how sure can we be that it is a true positive as opposed to a false positive? The `rsu.ss.rsfreecalc` function returns the required sample size to confirm the absence of disease using a test with imperfect diagnostic sensitivity and specificity based on the methodology implemented in the standalone software 'Freecalc' (Cameron and Baldock 1998).

EXAMPLE 2

We'll continue with the brucellosis example introduced above. Imagine the test we're using has a diagnostic sensitivity of 0.90 (as before) and specificity of 0.98. How many herds need to be sampled to be 95% certain that the prevalence of brucellosis in dairy herds is less than the design prevalence if less than a specified number of tests return a positive result?

```
rsu.sssep.rsfreecalc(N = 5000, pstar = 0.05, mse.p = 0.95,
   msp.p = 0.95, se.u = 0.90, sp.u = 0.98, method = "hypergeometric",
   max.ss = 32000)$summary
#>     n    N c pstar         p1      se.p      sp.p
#> 1 221 5000 8  0.05 0.04990684 0.9500932 0.9648964
```

A population sensitivity of 95% is achieved with a total sample size of 221 herds, assuming a cut-point of 8 or more positive herds are required to return a positive survey result.

Note the substantial increase in sample size when diagnostic specificity is imperfect (221 herds when specificity is 0.98 compared with 65 when specificity of the test was perfect). The relatively low design prevalence in combination with imperfect imperfect specificity means that false positives are likely to be a problem in this population so the number tested needs to be (substantially) increased. Increase the design prevalence to 0.10 to assess its effect on estimated sample size.

```
rsu.sssep.rsfreecalc(N = 5000, pstar = 0.10, mse.p = 0.95,
   msp.p = 0.95, se.u = 0.90, sp.u = 0.98, method = "hypergeometric",
   max.ss = 32000)$summary
#>    n    N c pstar         p1      se.p      sp.p
#> 1 82 5000 4   0.1 0.04947366 0.9505263 0.9754309
```

The required sample size decreases to 82 and the cut-point to 4 positives due to: (1) the expected reduction in the number of false positives; and (2) the greater difference between true and false positive rates in the first example compared with the second.

EXAMPLE 3

Now consider the situation where individual surveillance units (e.g., animals) are aggregated within groups

called 'clusters' (e.g., herds). With this type of system **two-stage cluster sampling** is a commonly used approach for disease surveillance studies.

With two stage cluster sampling herds (clusters, PSUs) are sampled first and then surveillance units are sampled from each sampled cluster, as secondary sampling units (SSUs). This means that we have two sample sizes to calculate: the number of PSUs and the number of SSUs from within each PSU.

For this example we assume that there are 20,000 at risk herds in our population and we do not know the number of animals present in each herd. This disease is not very common among herds but if a herd is positive the prevalence is expected to be relatively high, so we set the herd-level design prevalence to 0.005 and the within-herd design prevalence to 0.05. The test we will use at the surveillance unit level has a diagnostic sensitivity of 0.90 and a diagnostic specificity of 1.00. The target sensitivity of disease detection at the population level is 0.95.

How many herds (PSUs) need to be sampled if you want to be 95% certain of detecting at least one infected herd if the between-herd prevalence of disease is greater than or equal to 0.005?

```
rsu.sssep.rs(N = 20000, pstar = 0.005, se.p = 0.95, se.u = 0.90)
#> [1] 656
```

We need to sample a total of 656 herds.

How many animals (SSUs) need to be sampled from each herd (PSU) if we want to be 95% certain of detecting at least one infected animal if the within-herd prevalence of disease is greater than or equal to 0.05?

```
rsu.sssep.rs(N = NA, pstar = 0.05, se.p = 0.95, se.u = 0.90)
#> [1] 66
```

Within each selected herd we need to sample at least 66 animals.

We can calculate the required number of herds (PSUs) to sample and the required number of animals to sample from within each herd (SSUs) in a single step using the function `rsu.sssep.rs2stage`:

```
rsu.sssep.rs2st(H = 20000, N = NA, pstar.c = 0.005, pstar.u = 0.05, se.p = 0.95,
   se.c = 0.95, se.u = 0.90)
#> $clusters
#>       H nsample
#> 1 20000     622
#>
#> $units
#>    N nsample
#> 1 NA      66
```

Estimation of surveillance system sensitivity and specificity

Table 2: Table 2: Functions in epiR to estimate surveillance system sensitivity (SSSe) using representative population sampling data.

Sampling	Outcome	Detail	Function
Representative	SSSe	Imperf Se, perf Sp	rsu.sep.rs
Two stage representative	SSSe	Imperf Se, perf Sp	rsu.sep.rs2st
Representative	SSSe	Imperf Se, perf Sp, mult comp	rsu.sep.rsmult
Representative	SSSe	Imperf Se, imperf Sp	rsu.sep.rsfreecalc
Pooled representative	SSSe	Imperf Se, perf Sp	rsu.sep.rspool
Representative	SSSe	Imperf Se, perf Sp	rsu.sep.rsvarse

Table 2: Table 2: Functions in epiR to estimate surveillance system sensitivity (SSSe) using representative population sampling data.

Sampling	Outcome	Detail	Function
Representative	SSSp	Imperf Sp	rsu.spp.rs

EXAMPLE 4

Three hundred samples are to be tested from a population of animals to confirm the absence of disease. The total size of the population is unknown. Assuming a design prevalence of 0.01 and a test with diagnostic sensitivity of 0.95 will be used what is the surveillance system sensitivity? That is, what is the probability that disease will be detected if it is present in the population at or above the specified design prevalence?

```
rsu.sep.rs(N = NA, n = 300, pstar = 0.01, se.u = 0.95)
#> [1] 0.9429384
```

The probability that this surveillance strategy will detect disease if it is present in the population at or above the specified design prevalence (the surveillance system sensitivity) is 0.943.

EXAMPLE 5

Thirty animals from five herds ranging in size from 80 to 100 head are to be sampled to confirm the absence of a disease. Assuming a design prevalence of 0.01 and a test with diagnostic sensitivity of 0.95 will be used, what is the sensitivity of disease detection for each herd?

```
N <- seq(from = 80, to = 100, by = 5)
n <- rep(30, times = length(N))

herd.sep <- rsu.sep.rs(N = N, n = n, pstar = 0.01, se.u = 0.95)
sort(round(herd.sep, digits = 2))
#> [1] 0.28 0.30 0.32 0.34 0.36
```

The sensitivity of disease detection for each herd ranges from 0.28 to 0.36.

EXAMPLE 6

Assume 73 samples were tested at two different labs, using different tests. Laboratory 1 tested 50 samples with the standard test which has a diagnostic sensitivity of 0.80. Laboratory 2 tested the remaining 23 samples with a different test which has a diagnostic sensitivity of 0.70. What is the surveillance system sensitivity of disease detection if we set the design prevalence to 0.05?

```
# Diagnostic test sensitivities and the number of samples tested at each laboratory:
se.t1 <- 0.80; se.t2 <- 0.70
n.lab1 <- 50; n.lab2 <- 23

# Create a vector of test sensitivities for each sample:
se.all <- c(rep(se.t1, times = n.lab1), rep(se.t2, times = n.lab2))
rsu.sep.rsvarse(N = n.lab1 + n.lab2, pstar = 0.05, se.u = se.all)
#> [1] 0.9971275
```

If the design prevalence is 0.05 the estimated surveillance system sensitivity is 0.997.

Estimation of the probability of disease freedom

Table 3: Table 3: Functions in epiR to estimate the probability of disease freedom using representative population sampling data.

Sampling	Outcome	Detail	Function
Representative	Pr DF [a]	Imperf Se, perf Sp	rsu.pfree.rs
Representative	Equ Pr DF	Imperf Se, perf Sp	rsu.pfree.equ

[a] Pr DF: Probability of disease freedom.

EXAMPLE 7

You are the epidemiologist for a land-locked country in central Asia. You have developed a surveillance program for a given disease which has an estimated surveillance system sensitivity (SSSe) of 0.65. The disease of interest is carried by live animals and you know that the frequency of illegal importation of animals into your country (and therefore the likelihood of disease incursion) is greater during the warmer months of the year (June to August).

Plot the probability of disease freedom assuming surveillance testing is carried out each month. Include on your plot the probability of disease incursion to show how it changes during the year. Previous surveillance work indicates that the probability that your country is actually free of disease is 0.50.

```
library(ggplot2); library(lubridate); library(scales)

# Define a vector disease incursion probabilities (January to December):
p.intro <- c(0.01,0.01,0.01,0.02,0.04,0.10,0.10,0.10,0.08,0.06,0.04,0.02)

rval.df <- rsu.pfree.rs(se.p = rep(0.65, times = 12), p.intro = p.intro,
   prior = 0.50, by.time = TRUE)

# Re-format rval.df ready for for ggplot2:
gdat.df <- data.frame(mnum = rep(1:12, times = 2),
   mchar = rep(seq(as.Date("2020/1/1"), by = "month", length.out = 12), times = 2),
   class = c(rep("Disease introduction", times = length(p.intro)),
             rep("Disease freedom", times = length(p.intro))),
   prob = c(rval.df$PIntro, rval.df$PFree))

# Plot the results:
ggplot(data = gdat.df, aes(x = mchar, y = prob, group = class, col = class)) +
  theme_bw() +
  geom_point() +
  geom_line() +
  scale_colour_manual(values = c("red", "dark blue")) +
  scale_x_date(breaks = date_breaks("1 month"), labels = date_format("%b"),
     name = "Month") +
  scale_y_continuous(limits = c(0,1), name = "Probability") +
  geom_hline(aes(yintercept = 0.95), linetype = "dashed", col = "blue") +
  guides(col = guide_legend(title = "")) +
  theme(legend.position = "inside", legend.position.inside = c(0.8,0.5))
```

4 Risk-based sampling

With risk-based sampling we modify the intensity of sampling effort across the population of interest according to risk (as opposed to representative sampling where the probability that an individual unit is sampled is uniform across the population of interest). When our objective is to detect the presence of

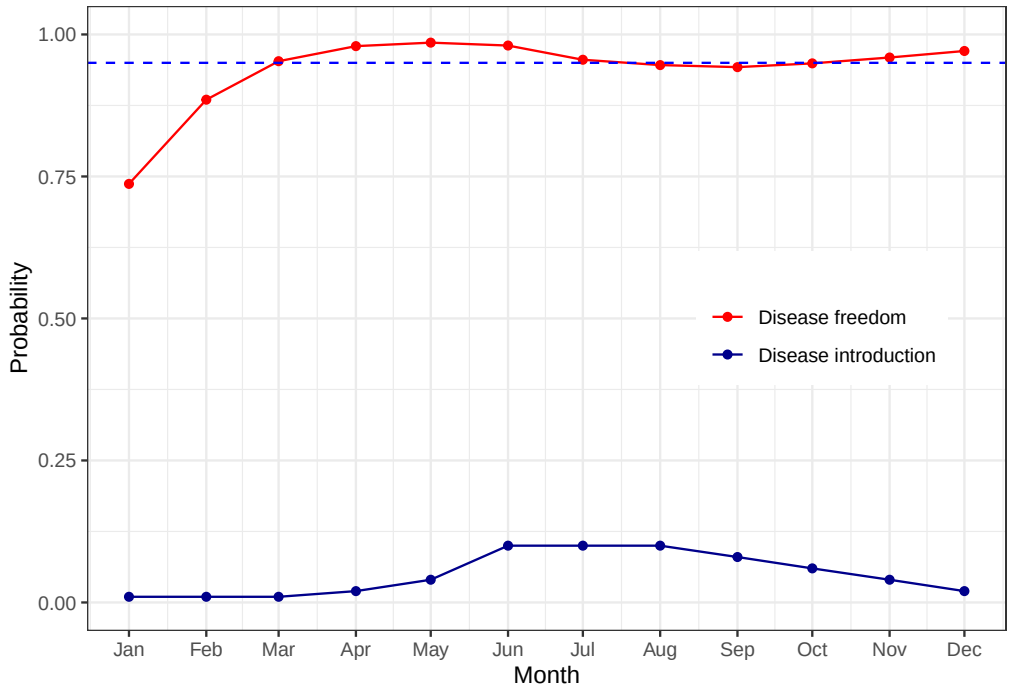


Figure 1: Line plot showing the probability of disease freedom (red) and the probability of disease introduction (blue) as a function of calendar date.

disease risk-based sampling makes intuitive sense: We concentrate our search effort on those sections of the population where we believe we're more likely to detect disease (i.e., where the risk of disease is high).

How many samples do I need?

The sample size functions all relate to sampling to achieve a defined surveillance system sensitivity.

Table 4: Table 4: Functions in epiR to estimate sample size using risk based sampling data.

Sampling	Outcome	Detail	Function
Risk-based	SSSe	Single Se for RGs, perf Sp [a]	rsu.sssep.rbsrg
Risk-based	SSSe	Multiple Se within RGs, perf Sp	rsu.sssep.rbmrg
Risk-based	SSSe	Two stage sampling, 1 risk factor	rsu.sssep.rb2st1rf
Risk-based	SSSe	Two stage sampling, 2 risk factors	rsu.sssep.rb2st2rf

[a] RGs: Risk groups.

EXAMPLE 8

You are working with a disease of cattle where the prevalence of disease is believed to vary according to herd type. The risk of disease is 5 times greater in dairy herds and 3 times greater in mixed herds compared with the reference category, beef herds. The proportion of dairy, mixed and beef herds in the population of interest is 0.10, 0.10 and 0.80, respectively. Assume you intend to distribute your sampling effort 0.4, 0.4 and 0.2 across dairy, mixed and beef herds, respectively.

Within each of the three risk groups a single test with a diagnostic sensitivity of 0.95 will be used. How many herds need to be sampled if you want to achieve 95% system sensitivity for a prevalence of disease in the population of greater than or equal to 1%?

```
# Generate a matrix listing the proportions of samples for each test in each
# risk group (the number of rows equal the number of risk groups, the number
# of columns equal the number of tests):
m <- rbind(1,1,1); m
#>      [,1]
#> [1,]    1
#> [2,]    1
#> [3,]    1

rsu.sssep.rbmrg(pstar = 0.01, rr = c(5,3,1), ppr = c(0.1,0.1,0.8),
   spr = c(0.4,0.4,0.2), spr.rg = m, se.p = 0.95, se.u = 0.95)
#> $n
#>       se.u 0.95 total
#> rr 5         59    59
#> rr 3         59    59
#> rr 1         29    29
#> total       147   147
#>
#> $epi
#> [1] 0.03125 0.01875 0.00625
#>
#> $mean.se
#> [1] 0.95 0.95 0.95

# The design prevalence (pstar) is 0.01. The relative risks of disease (rr) for
# dairy, mixed and beef herds are 5, 3, and 1, respectively. The proportions of
# dairy, mixed and beef herds in the population (ppr) are 0.1, 0.1 and 0.8,
# respectively. You intend to allocate 0.4, 0.4 and 0.2 of your sampling effort
# (spr) to dairy, mixed and beef herds, respectively. All herds in each of the
# three risk groups receive the same test so object m is a one column, three
# row matrix comprised of 1s.
```

A total of 147 herds need to be sampled: 59 dairy, 59 mixed and 29 beef herds.

Now assume that one of two tests will be used for each herd. The first test has a diagnostic sensitivity of 0.92. The second test has a diagnostic sensitivity of 0.80. The proportion of dairy, mixed and beef herds receiving the first test is 0.80, 0.50 and 0.70, respectively (which means that 0.20, 0.50 and 0.30 receive the second test, respectively). Recalculate the required sample size.

```
# Generate a matrix listing the proportions of samples for each test in each
# risk group (the number of rows equal the number of risk groups, the number
# of columns equal the number of tests):
m <- rbind(c(0.8,0.2), c(0.5,0.5), c(0.7,0.3)); m
#>      [,1] [,2]
#> [1,]  0.8  0.2
#> [2,]  0.5  0.5
#> [3,]  0.7  0.3

rsu.sssep.rbmrg(pstar = 0.01, rr = c(5,3,1), ppr = c(0.1,0.1,0.8),
   spr = c(0.4,0.4,0.2), spr.rg = m, se.p = 0.95, se.u = c(0.92,0.80))
#> $n
#>       se.u 0.92 se.u 0.8 total
#> rr 5         52       12    64
#> rr 3         32       32    64
#> rr 1         22        9    31
```

```
#> total       106       53   159
#>
#> $epi
#> [1] 0.03125 0.01875 0.00625
#>
#> $mean.se
#> [1] 0.896 0.860 0.884

# The design prevalence (pstar) is 0.01. The relative risks of disease (rr) for
# dairy, mixed and beef herds are 5, 3, and 1, respectively. The proportions of
# dairy, mixed and beef herds in the population (ppr) are 0.1, 0.1 and 0.8,
# respectively. You intend to allocate 0.4, 0.4 and 0.2 of your sampling effort
# (spr) to dairy, mixed and beef herds, respectively. The proportion of dairy,
# mixed and beef herds receiving the first test is 0.80, 0.50 and 0.70. The
# proportion of dairy, mixed and beef herds receiving the second test is
# (1 - 0.80), (1 - 0.50) and (1 - 0.70) i.e., 0.20, 0.50 and 0.30, respectively.
# Object m is a three row, two column matrix listing these probabilities.
# The rows of object m must sum to one.
```

A total of 159 herds need to be sampled: 64 dairy, 64 mixed and 31 beef herds.

EXAMPLE 9

A cross-sectional study is to be carried out to confirm the absence of disease using risk based sampling. Assume a population level design prevalence of 0.01 and there are 'high', 'medium' and 'low' risk areas where the risk of disease in the high risk area compared with the low risk area is 5 and the risk of disease in the medium risk area compared with the low risk area is 3. The proportions of the population at risk in the high, medium and low risk area are 0.10, 0.10 and 0.80, respectively.

Half of your samples will be taken from individuals in the high risk area, 0.30 from the medium risk area and 0.20 from the low risk area. You intend to use a test with diagnostic sensitivity of 0.90 and you'd like to take sufficient samples to return a population sensitivity of 0.95. How many units need to be sampled to meet the requirements of the study?

```
rsu.sssep.rbsrg(pstar = 0.01, rr = c(5,3,1), ppr = c(0.10,0.10,0.80),
   spr = c(0.50,0.30,0.20), se.p = 0.95, se.u = 0.90)
#> $total
#> [1] 148
#>
#> $n
#> [1] 74 44 30
#>
#> $epinf
#> [1] 0.03125 0.01875 0.00625
#>
#> $adj.risk
#> [1] 3.125 1.875 0.625
```

A total of 147 units needs to be sampled to meet the requirements of the study: 74 from the high risk area, 45 from the medium risk area and 28 from the low risk area.

EXAMPLE 10

A cross-sectional study is to be carried out to confirm the absence of disease using risk based sampling. Assume a design prevalence of 0.02 at the cluster (herd) level and a design prevalence of 0.10 at the surveillance unit (individual animal) level. Clusters are categorised as being either high, medium or low risk with the probability of disease for clusters in the high and medium risk area 5 and 3 times the probability of disease

in the low risk area. The proportions of clusters in the high, medium and low risk area are 0.10, 0.20 and 0.70, respectively. The proportion of samples from the high, medium and low risk area will be 0.40, 0.40 and 0.20, respectively.

Surveillance units (individual animals) are categorised as being either high or low risk with the probability of disease for units in the high risk group 4 times the probability of disease in the low risk group. The proportions of units in the high and low risk groups are 0.10 and 0.90, respectively. All of your samples will be taken from units in the high risk group.

You intend to use a test with diagnostic sensitivity of 0.95 and you'd like to take sufficient samples to be 95% certain that you've detected disease at the population level, 95% certain that you've detected disease at the cluster level and 95% at the surveillance unit level. How many clusters and how many units need to be sampled to meet the requirements of the study?

```
rsu.sssep.rb2st2rf(rr.c = c(5,3,1), ppr.c = c(0.10,0.20,0.70),
   spr.c = c(0.40,0.40,0.20),
   pstar.c = 0.02,
   rr.u = c(4,1), ppr.u = c(0.1, 0.9), spr.u = c(1,0),
   pstar.u = 0.10,
   se.p = 0.95, se.c = 0.95, se.u = 0.95)
#> $clusters
#> $clusters$total
#> [1] 83
#>
#> $clusters$n
#> [1] 33 33 17
#>
#> $clusters$epinf
#> [1] 0.05555556 0.03333333 0.01111111
#>
#> $clusters$adj.risk
#> [1] 2.7777778 1.6666667 0.5555556
#>
#>
#> $units
#> $units$total
#> [1] 9
#>
#> $units$n
#> [1] 9 0
#>
#> $units$epinf
#> [1] 0.30769231 0.07692308
#>
#> $units$adj.risk
#> [1] 3.0769231 0.7692308
```

A total of 82 clusters needs to be sampled: 33 from the high risk area, 33 from the medium risk area and 16 from the low risk area. A total of 9 units should be sampled from each cluster.

Surveillance system sensitivity

Table 5: Table 5: Functions in epiR to estimate surveillance system sensitivity using risk based sampling data.

Sampling	Outcome	Detail	Function
Risk-based	SSSe	Varying Se, perf Sp	rsu.sep.rb
Risk-based	SSSe	Varying Se, perf Sp, 1 risk factor	rsu.sep.rb1rf
Risk-based	SSSe	Varying Se, perf Sp, 2 risk factors	rsu.sep.rb2rf

EXAMPLE 11

You have been asked to provide an assessment of a surveillance program for *Actinobacillus hyopneumoniae* in pigs. It is known that there are high risk and low risk areas for *A. hypopneumoniae* in your country with the estimated probability of disease in the high risk area thought to be around 3.5 times that of the probability of disease in the low risk area. It is known that 10% of the 1784 pig herds in the study area are in the high risk area and 90% are in the low risk area.

The risk of *A. hypopneumoniae* is dependent on age, with adult pigs around five times more likely to be *A. hypopneumoniae* positive compared with younger (grower) pigs. Pigs from 20 herds have been sampled: 5 from the low-risk area and 15 from the high-risk area. All of the tested pigs were adults: no grower pigs were tested. The ELISA for *A. hypopneumoniae* in pigs has a diagnostic sensitivity of 0.95.

What is the surveillance system sensitivity if we assume a design prevalence of 1 per 100 at the cluster (herd) level and 5 per 100 at the surveillance system unit (pig) level?

```
# There are 1784 herds in the study area:
H <- 1784

# Twenty of the 1784 herds are sampled. Generate 20 herds of varying size:
set.seed(1234)
hsize <- rlnorm(n = 20, meanlog = log(10), sdlog = log(8))
hsize <- round(hsize + 20, digits = 0)

# Generate a matrix listing the number of growers and finishers in each of the
# 20 sampled herds. Assume that anywhere between 80% and 95% of the pigs in
# each herd are growers:
set.seed(1234)
pctg <- runif(n = 20, min = 0.80, max = 0.95)
ngrow <- round(pctg * hsize, digits = 0)
nfini <- hsize - ngrow
N <- cbind(ngrow, nfini)

# Generate a matrix listing the number of grower and finisher pigs sampled from
# each herd. Fifteen pigs from each herd are sampled. If there's less than
# 15 pigs we sample the entire herd:
nsgrow <- rep(0, times = 20)
nsfini <- ifelse(nfini <= 15, nfini, 15)
n <- cbind(nsgrow, nsfini)

# The herd-level design prevalence is 0.01 and the individual pig-level design
# prevalence is 0.05:
pstar.c <- 0.01
pstar.u <- 0.05

# For herds in the high-risk area the probability being A. hyopneumoniae
```

```
# positive is 3.5 times that of herds in the low-risk area. Ninety percent of
# herds are in the low risk area and 10% are in the high risk area:
rr.c <- c(3.5,1)
ppr.c <- c(0.1,0.9)

# We've sampled 15 herds from the high risk area and 5 herds from the low risk
# area. Above, for vector rr.c, the relative risk for the high risk group is
# listed first so the vector rg follows this order:
rg <- c(rep(1, times = 15), rep(2, times = 5))

# The probability being A. hyopneumoniae positive for finishers is 5 times that
# of growers. For the matrices N and n growers are listed first then finishers.
# Vector rr.u follows the same order:
rr.u <- c(1,5)

# The diagnostic sensitivity of the A. hyopneumoniae ELISA is 0.95:
se.u <- 0.95

tmp.rb2st <- rsu.sep.rb2st(H = H, N = N, n = n,
   pstar.c = pstar.c, pstar.u = pstar.u,
   rg = rg, rr.c = rr.c, rr.u = rr.u,
   ppr.c = ppr.c, ppr.u = NA,
   se.u = se.u)
tmp.rb2st$se.p
#> [1] 0.3171584
```

The estimated surveillance system sensitivity of this program is 0.32.

Repeat these analyses assuming we don't know the total number of pig herds in the population and we have only an estimate of the proportions of growers and finishers in each herd.

```
# Generate a matrix listing the proportion of growers and finishers in each of
# the 20 sampled herds:

ppr.u <- cbind(rep(0.9, times = 20), rep(0.1, times = 20))

# Set H (the number of clusters) and N (the number of surveillance units within
# each cluster) to NA:
tmp.rb2st <- rsu.sep.rb2st(H = NA, N = NA, n = n,
   pstar.c = pstar.c, pstar.u = pstar.u,
   rg = rg, rr.c = rr.c, rr.u = rr.u,
   ppr.c = ppr.c, ppr.u = ppr.u,
   se.u = se.u)
tmp.rb2st$se.p
#> [1] 0.2078001
```

The estimated surveillance system sensitivity is 0.21.

5 Analysis of passive surveillance data

Estimation of surveillance system sensitivity and specificity

EXAMPLE 12

There are four 'steps' in a (passive) disease detection process for disease X in your country: (1) an infected animal shows clinical signs of disease; (2) a herd manager observes clinical signs in a disease animal and

calls a veterinarian; (3) a veterinarian responds appropriately to the disease investigation request (taking, for example, appropriate samples for laboratory investigation); and (4) the laboratory conducts appropriate tests on the submitted samples and interprets the results of those tests correctly. The probabilities for each step in the disease detection pathway (in order) are thought to be 0.10, 0.20, 0.90 and 0.99, respectively.

Assuming the probability that a unit actually has disease if it is submitted for testing is 0.98, the sensitivity of the diagnostic test used at the unit level is 0.90, the population is comprised of 1000 clusters (herds), five animals from each cluster (herd) investigated for disease are tested and the cluster-level design prevalence is 0.01, what is the sensitivity of disease detection at the cluster (herd) and population level?

```
rsu.sep.pass(step.p = c(0.10,0.20,0.90,0.99), pstar.c = 0.01,
   p.inf.u = 0.98, N = 1000, n = 5, se.u = 0.90)
#> $se.p
#> [1] 0.164565
#>
#> $se.c
#> [1] 0.01781959
```

The sensitivity of disease detection at the cluster (herd) level is 0.018. The sensitivity of disease detection at the population level is 0.16.

6 Miscellaneous functions

Adjusted relative risks

EXAMPLE 13

For a given disease of interest you believe that there is a 'high risk' and 'low risk' area in your country. The risk of disease in the high-risk area compared with the low-risk area is 5. A recent census shows that 10% of the population are resident in the high-risk area and 90% are resident in the low-risk area. Calculate the adjusted relative risks for each area.

```
rsu.adjrisk(rr = c(5,1), ppr = c(0.10,0.90))
#> [1] 3.5714286 0.7142857
```

The adjusted relative risks for the high and low risk areas are 3.6 and 0.7, respectively.

Design prevalence back calculation

EXAMPLE 14

The population size in a provincial area in your country is 193,000. In a given two-week period a total of 7764 individuals have been tested for COVID-19 using an approved PCR which is believed to have a diagnostic sensitivity of 0.85. All of the individuals tested have returned a negative result. What is the maximum prevalence required to provide system sensitivity of 0.95 if COVID-19 is actually present in this population (i.e., what is the back-calculated design prevalence)? Express your result as the number of COVID-19 cases per 100,000 head of population.

```
rsu.pstar(N = 193000, n = 7764, se.p = 0.95, se.u = 0.85) * 100000
#> [1] 44.61341
```

If the 7764 individuals have all returned a negative test result (using a test with 85% sensitivity) we can be 95% confident that COVID-19, if it is present, is present at a prevalence of 44 cases per 100,000 or less. What is the probability that the prevalence of COVID-19 in this population is less than or equal to 10 cases per 100,000?

```
rsu.sep(N = 193000, n = 7764, pstar = 10 / 100000, se.u = 0.85)
#> [1] 0.4890517
```

If all of the 7764 individuals returned a negative test we can only be 48% confident that the prevalence of

COVID-19 in the province is less than 10 per 100,000. How many need to be tested to be 95% confident that the prevalence of COVID-19 is less than or equal to 10 cases per 100,000? Return to the sample size functions covered earlier:

```
rsu.sssep.rs(N = 193000, pstar = 10 / 100000, se.p = 0.95, se.u = 0.85)
#> [1] 31586
```

To be 95% confident that the prevalence of COVID-19 is less than or equal to 10 cases per 100,000 a total of 31,586 individuals need to be tested.

References

Cameron, AR, and FC Baldock. 1998. "A new probability formula for surveys to substantiate freedom from disease." *Preventive Veterinary Medicine* 34: 1–17.

Thacker, SB, and RL Berkelman. 1988. "Public health surveillance in the United States." *Epidemiological Reviews* 10: 164–90.

World Organisation for Animal Health. 2022. "Terrestrial Animal Health Code." In. Paris: World Organisation for Animal Health.