FFD: Package to substantiate freedom from disease in R using two-stage sampling

Ian Kopacka
Austrian Agency for Health and Food Safety (AGES)

Abstract

In practise, when conducting surveys to substantiate freedom from disease in large populations two-stage sampling strategies are often used in order to account for herd-level clustering of diseases. Using a modified hypergeometric formula the optimal sample size can elegantly be computed, while incorporating imperfect diagnostic tests and finite populations; see Cameron and Baldock (1998a,b).

In the package `FFD`, tools for calculating optimal sample sizes (on animal and herd level) using sampling strategies "individual sampling" or "limited sampling" (see Ziller, Selhorst, Teuffert, Kramer, and Schlüter (2002)) are implemented. Further, cost optimal sampling strategies, while maintaining constant α-levels can be computed using `FFD`. The package furthermore includes tools for evaluating the a-posteriori significance ($= 1 - \alpha$) corresponding to a specific sample of herds.

Keywords: R, freedom from disease, sample size calculation, individual sampling, limited sampling, a-posteriori alpha-error.

1. Introduction

To meet with standards of trading partners or international organizations, it is often required to prove the absence of certain diseases in certain animal populations using surveys to substantiate freedom from disease. In many cases these surveys are designed in two stages: first the number of herds that needs to be tested is determined, secondly the number of animals that needs to be tested is fixed for each herd. This two-stage sampling accounts for the tendency of most diseases to cluster on a herd-level, i.e., the characteristics of the spread of a disease within a herd might differ from the ability of a disease to spread from one herd to another. E.g., there might be a low percentage of infected herds in a population, while a herd that is infected might show a rather high prevalence. The use of two-stage sampling, however, also has practical advantages. In many cases it is not possible to establish sampling plans purely on animal-level, as this would require a registry of all the animals in the population. Often, such a registry only exists for the herds/holdings in an area containing only the number of animals per holding (but not a unique identifier for those animals).

The package `FFD` provides tools to compute the number of herds, as well as the number of animals per herd that need to be tested using two different sampling schemes that were established in Ziller *et al.* (2002). These schemes are known as *individual sampling* and *limited sampling*. For individual sampling the number of animals that needs to be tested per herd depends on the herd size, while limited sampling uses a pre-fixed number of animals,

irrespective of the herd size. The advantages and disadvantages of the two sampling schemes will be discussed in the sections below.

2. The basics of two-stage sampling

A survey to substantiate freedom from disease can be regarded as a test that is being applied to an entire population. As for a diagnostic test, sensitivity and specificity can be determined for the survey. The sensitivity of a test is defined as the probability of achieving a positive test result, given that the true disease status of the tested individual is positive (individual is sick). Analogously the specificity of a test is the probability of achieving a negative test result, given that the true disease status of the tested individual is negative (individual is healthy).

A typical requirement of a trading partner might be that one must show with a probability exceeding 95 % that no more than 0.2 % of the population is diseased. The probability of establishing a prevalence of 0.2 % or lower reflects the uncertainty that is always present due to sampling of a subpopulation only and imperfect tests. This percentage is often referred to as the *confidence*, while the compliment (1-confidence) is referred to as the *significance* α. In our case $\alpha = 0.05$. The prevalence limit is often referred to as the *design prevalence*.

Now let's look at the survey in the context of sensitivity and specificity. The confidence of our survey is the probability of finding the disease in a diseased population, i.e., the confidence can be regarded as the overall sensitivity of the (statistical) test. The specificity of the test is the probability of classifying a population as free from the disease, given the population is truly free. This parameter is related to the *power* $(1-\beta)$ of the test. In surveys substantiating freedom from disease a common assumption is that of perfect specificity, i.e., that there are no false positives. This assumption on the one hand simplifies the computation but it is also practically founded. A positive result can have undesired economical implications. It can therefore be assumed that all positive results are thoroughly checked using multiple tests in order to rule out false positive results.

2.1. One-stage sampling

Let us begin by considering a one-stage sampling scheme for a finite population using an imperfect diagnostic test, e.g., let's say we consider a herd of animals, and we pick a certain number of animals at random. We then test these animals in order to determine if the entire herd is infected with a disease or not. In general terms that means that we test n individuals from a population with size $N \geq n$. If all tested individuals have a negative test result we classify the population as being free from the disease. If we find one or more individuals that test positive we classify the population as diseased. We have to reach a prescribed significance α, i.e., the probability of finding no testpositives, given the population is diseased must be smaller than (or equal to) our significance level α. In order to compute this probability we need to know

- the population size N,
- the sample size n,

- the prevalence π of the disease in the population (or the number of diseased individuals in the population $d = N \cdot \pi$) and
- the sensitivity Se and the specificity Sp of the diagnostic test.

The probability of finding no testpositives, given that at least d individuals are diseased in the population (denoted by $P(T^+ = 0|d)$) can then be computed using a modified hypergeometric formula due to Cameron and Baldock (1998a):

$$P(T^+ = 0|d) = \sum_{y=\max(0,n-N+d)}^{\min(d,n)} \frac{\binom{d}{y}\binom{N-d}{n-y}}{\binom{N}{n}} (1 - \text{Se})^y \text{Sp}^{n-y}. \tag{1}$$

In the case of perfect specificity the equation is simplified to

$$P(T^+ = 0|d) = \sum_{y=\max(0,n-N+d)}^{\min(d,n)} \frac{\binom{d}{y}\binom{N-d}{n-y}}{\binom{N}{n}} (1 - \text{Se})^y. \tag{2}$$

In order to compute the optimal sample size one must therefore find the smallest sample size n, using equation (2), that still satisfies $P(T^+ = 0|d) \leq \alpha$.

2.2. Two-stage sampling

The principles of section 2.1 can easily be extended to two-stage sampling; see Cameron and Baldock (1998b). Instead of testing a few animals out of a herd in order to determine the disease status of a herd we, e.g., "test" a few herds out of a larger population in order to determine whether a region or country is diseased or not. The only difference is that we cannot directly apply a diagnostic test to an entire herd. The so called *herd test* rather consists of again picking a certain number of animals out of the herd at random and testing those animals using a (possibly imperfect) diagnostic test. The sensitivity of the herd test is the probability of finding the disease, given the herd is infected. As again a herd is considered as diseased if at least one animal tests positive, the sensitivity of the herd test is given by

$$\text{Se}_{herd} = P(T^+ > 0|d) = 1 - P(T^+ = 0|d) = 1 - \alpha.$$

Individual sampling

One way to determine the number of herds to test is to fix a desired herd sensitivity, e.g., $\text{Se}_{herd} = 0.7$. That means that for each herd that is being tested we have a 70 % chance of finding the disease. For each herd size we can then apply the principles of section 2.1 to determine n, the number of animals we have to test, where N is the herd size, $\alpha = 1 - \text{Se}_{herd}$ and d is the number of diseased individuals in the herds, assuming the herd is infected. This number is related to the *intra herd prevalence* π_{IH}, via $d = N \cdot \pi_{IH}$. The intra herd prevalence is usually higher that the design prevalence, due to disease clustering on herd-level, and is mostly fixed by expert opinions, or determined by surveys.

Example 2.1 *We consider a population of herds where the biggest herd has no more than 300 animals. Using a herd sensitivity of* $Se_{herd} = 0.7$*, in intra herd prevalence of* $\pi_{IH} = 0.2$

and a diagnostic test with a sensitivity of 90 % *the necessary number of animals to test for each herd is given in table 1.*

Table 1: Sample size corresponding to the herd size

Herd size	**No. of animals to test**
1 - 3	entire herd
4 - 5	4
6	5
7 - 31	6
32 - 300	7

The number of herds to test can then again be computed by applying the principles of section 2.1, where n is the number of herds to sample, N is the number of herds in the population, α is the significance level of the survey, d is the number of diseased herds in the population according to the design prevalence and Se is the herd sensitivity.

Example 2.2 *We consider a population of 15000 herds, the biggest of which having no more that 300 animals. We need to prove with a confidence of 95 % (⇒ significance level* $\alpha = 5$ %*) that no more than 0.2 % of the herds are infected, i.e., the design prevalence is* $\pi = 0.002$*. The diagnostic test we are using has a sensitivity of 90 % and the intra-herd prevalence of the disease is* $\pi_{IH} = 0.2$*.*

The parameters above are all determined by the population, the infectivity of the disease and by regulations of the trading partner. The herd sensitivity, however, can be chosen (almost) freely and determines the number of herds and the number of animals per herd that need to be sampled. We fix $Se_{herd} = 0.7$*.*

The number of animals tested per herd is then given in table 1. The number of herds to be tested can be determined using (2) with $N = 15000$*,* $\alpha = 0.05$*,* $\pi = 0.002$ *and* $Se = 0.7$*. With the survey parameters above one needs to draw a sample of 2036 herds.*

Note that in the example above the herd sensitivity was fixed. We want to stress that, using the methodology above, every herd sensitivity chosen within a reasonable range yields a sampling scheme satisfying the prescribed significance level. The herd sensitivity merely determines the balance between the number of animals to test per herd and the number of herds to test and can, e.g., be chosen according to economical aspects.

Limited sampling

Another strategy used for two-stage sampling is *limited sampling*. With limited sampling a pre-fixed number of animals k (*sample limit*) is tested in each herd, irrespective of the herd size. If the herd has fewer animals then the entire herd is tested. With this approach the herd sensitivity, i.e., the ability to find a disease in a herd, is no longer constant over the population (as it is for individual sampling), but depends on the herd size. If, e.g., 7 animals are tested out of a herd of 300 then the probability of finding a diseased animal is significantly

smaller than when 7 animals are tested out of a herd of 10 animals. Hence, as opposed to individual sampling where the number of animals to test varies over the population, while the herd sensitivity is constant, for limited sampling the opposite is true. The number of animals to test is the same for every herd, but the herd sensitivity varies.

The herd sensitivity $\text{Se}_{herd} = 1 - P(T^+|d)$ can be computed for each herd using (2), where N is the herd size, $d = N \cdot \pi_{IH}$, Se is the sensitivity of the diagnostic test and $n = min(N, k)$.

In order to compute the number of herds to be tested we, however, require one fixed herd sensitivity and not - as it is the case here - a herd sensitivity depending on the herd size. What is usually done is that one uses the mean herd sensitivity

$$\text{Se}_{mean} = \sum_{j=1}^{N_{max}} \text{Se}_{herd}(N = j, k) \cdot P(N = j), \tag{3}$$

where N_{max} is the biggest herd size in the population, $\text{Se}_{herd}(N = j, k)$ is the herd sensitivity of a herd of size j using limited sampling with a sample limit k and $P(N = j)$ is the proportion of herds with size j in the population, i.e., the "probability" that a herd is of herd size j. Using the mean herd sensitivity as herd sensitivity the number of herds to be tested can again be determined using (2).

Similar to the herd sensitivity in individual sampling the sample limit determines the balance between the number of animals to test per herd and the number of herds to test, while maintaining a constant significance level.

3. Sample size calculation using S4 classes

The package `FFD` offers convenient tools to compute the sample sizes on herd and on animal level for individual and limited sampling using S4-classes. With these classes the survey parameters need to be specified once, creating an object of the class `SurveyData`. With this object different sampling strategies can conveniently be compared with respect to effectivity and costs and appropriate strategies can be evaluated and exported as html-files.

Furthermore, functions are available to evaluate (2), find the optimal sample sizes on herd and animal level, to evaluate herd sensitivities for limited sampling etc. These functions operate with conventional R-classes (vectors, data frames) and, while the use is not as convenient as with the methods for the S4 classes, they offer a greater flexibility.

3.1. Specifying the survey parameters

The following parameters/data are required in order to fix the sample size:

- `nAnimalVec`: A vector of herd sizes (=number of animals in a herd). Each component of the vector corresponds to a herd in the population,
- `designPrevalence`: The prevalence threshold in the population that the survey must establish,
- `alpha`: Significance level of the survey (= 1 - confidence),
- `intraHerdPrevalence`: The assumed prevalence of the disease within an infected herd,

- `diagSensitivity`: The sensitivity of the diagnostic test.

If it is desired to optimize the sampling strategy with respect to overall costs, parameters `costHerd`, `costAnimal`, describing the cost of each tested herd (excluding the cost per tested animal) and the cost of each tested animal, respectively. The overall costs are then computed using the simple model:

cost = number of tested herds * cost per herd + number of tested animals * cost per animal.

The cost per tested animal, e.g., contain the cost of drawing and analyzing the sample. The cost per tested herd could contain the travel costs of the vet etc.

All the survey parameters are packed into an S4 object of the class `SurveyData` using the constructor `surveyData()`. Additionally, further population data, such as herd identifiers, names and addresses of the owners etc. can be passed to the constructor in the form of a data frame, where each row of the data frame corresponds to a component of the vector `nAnimalVec`.

In the following example the data set `sheepData`, contained in the package `FFD`, is used. The data set contains simulated data resembling the sheep holdings in Austria.

```
> data(sheepData)
> mySurvey <- surveyData(nAnimalVec = sheepData$nSheep,
       populationData = sheepData, designPrevalence = 0.002,
       alpha = 0.05, intraHerdPrevalence = 0.2,
       diagSensitivity = 0.9, costHerd = 30, costAnimal = 7)
> summary(mySurvey)

Survey Parameters:
------------------
Design Prevalence:               0.002
Significance level:              0.05
Intra herd prevalence:           0.20
Sensitivity of diagnostic test:  0.90
Cost per tested herd:            30.00
Cost per tested animal:          7.00

Survey Data:
------------
Number of herds:             15287
Total number of animals:     224606
Number of animals per herd:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    4.00    8.00   14.69   17.00  249.00
Additional population data:
'data.frame':        15287 obs. of  3 variables:
 $ herdId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ state : int  7 7 6 3 8 7 3 7 4 3 ...
 $ nSheep: num  22 30 4 11 11 3 94 53 4 24 ...
```

Objects of the class `SurveyData` are the basic building blocks used in the package `FFD`, containing all the necessary data for the design of an appropriate sampling scheme using individual or limited sampling.

3.2. Individual sampling

With individual sampling the number of animals to test per herd in order to achieve a specified herd sensitivity depends on the herd size. The herd sensitivity, hence, determines the number of animals to test per herd, as well as the number of herd to test, while maintaining a constant overall significance level α. If a low herd sensitivity is chosen the number of animals to test per herd is low, while the number of herds to test might be rather high. If, however a high herd sensitivity is specified the number of animals tested per herd increases, while the number of herds to test decreases. If the cost per tested herd and the cost per tested animal is known a herd sensitivity might be chosen in order to minimize the overall costs of the survey.

Cost optimization

The package `FFD` provides the S4-class `IndSamplingSummary` and the function `indSampling-Summary()`, as a convenient tool to minimize the survey costs for individual sampling. The class constructor `indSamplingSummary()` takes an object of the class `SurveyData` and a step size for the herd sensitivities as an argument and computes the number of herds to test, the expected total number of animals tested based on the herd size distribution in the population, as well as the expected overall costs of the survey for a sequence of herd sensitivities. The herd sensitivities range from 0.1 to the sensitivity of the diagnostic test, the step size for the discretization is either specified by the user or a default value of 0.02 is used.

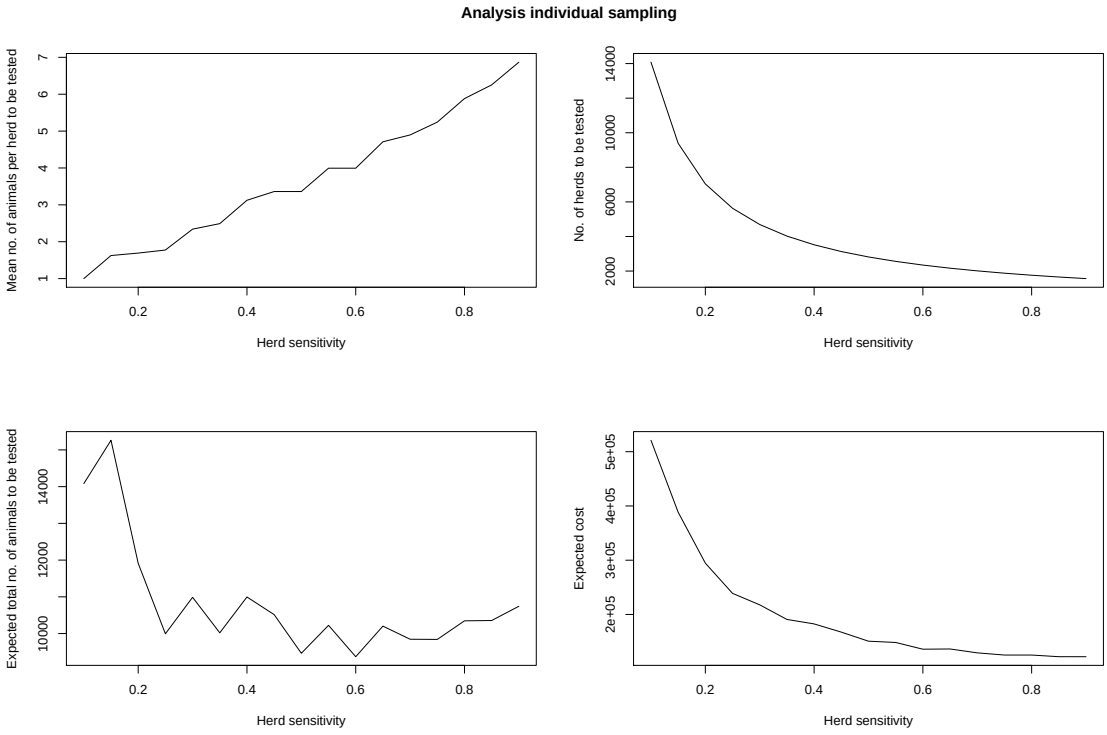
```
> myIndSamplingSummary <- indSamplingSummary(survey.Data = mySurvey,
       stepSize = 0.05)
> summary(myIndSamplingSummary)

INDIVIDUAL SAMPLING:

Survey Parameters:
------------------
Design Prevalence:               0.002
Significance level:              0.05
Intra herd prevalence:           0.20
Sensitivity of diagnostic test:  0.90
Cost per tested herd:            30.00
Cost per tested animal:          7.00

Survey Data:
------------
Number of herds:             15287
Total number of animals:     224606
Number of animals per herd:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
   1.00    4.00    8.00   14.69   17.00  249.00
Additional population data:
'data.frame':        15287 obs. of  3 variables:
 $ herdId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ state : int  7 7 6 3 8 7 3 7 4 3 ...
 $ nSheep: num  22 30 4 11 11 3 94 53 4 24 ...

Cost optimal sampling strategy:
--------------------------------
Herd sensitivity:                              0.90
Number of herds to test:                       1564
Expected total number of animals to test:  10743.38
Expected total costs of the survey:        122123.67
```

A plot of the object of class `IndSamplingSummary` can be created using `plot()`. The plot consists of (row-wise from top left to bottom right)

- the mean number of animals to test per herd plotted against the herd sensitivity,
- the number of herds to test plotted against the herd sensitivity,
- the expected total number of animals to test plotted against the herd sensitivity,
- the expected overall costs plotted against the herd sensitivity.

```
> plot(myIndSamplingSummary)
```



The summary of the object of class `IndSamplingSummary` can further be exported to an html-file using the method `HTML`. This method creates an html-file and a css-file containing the the data in the `IndSamplingSummary` object, as well as the diagnostic plots.

```
> HTML(myIndSamplingSummary)
```

The method further accepts the same arguments as the function `HTMLInitFile()` from the package `R2HTML`, e.g., `filename`, `outdir`, `CSSFile` and `Title`.

Parameters for a fixed herd sensitivity

If one has decided on an appropriate herd sensitivity, number of herds to test, the expected total number of animals to test, the expected costs and a lookup table containing the number of animals to test per herd depending on the herd size can be computed using the function `indSampling()` to create an object of the class `IndSampling`. The function takes two arguments, `survey.Data`, an object of the class `SurveyData`, and the herd sensitivity `herdSensitivity`. The computed parameters can again be displayed using the methods `show()`, `summary()` and `HTML()`.

For a herd sensitivity of 0.7 the parameters are:

```
> myIndSampling <- indSampling(survey.Data = mySurvey,
       herdSensitivity = 0.7)
> summary(myIndSampling)

INDIVIDUAL SAMPLING:

Survey Parameters:
------------------
Design Prevalence:               0.002
Significance level:              0.05
Intra herd prevalence:           0.20
Sensitivity of diagnostic test:  0.90
Cost per tested herd:            30.00
Cost per tested animal:          7.00

Survey Data:
------------
Number of herds:             15287
Total number of animals:     224606
Number of animals per herd:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    4.00    8.00   14.69   17.00  249.00
Additional population data:
'data.frame':        15287 obs. of  3 variables:
 $ herdId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ state : int  7 7 6 3 8 7 3 7 4 3 ...
 $ nSheep: num  22 30 4 11 11 3 94 53 4 24 ...
```

```
Sampling strategy:
------------------
Herd sensitivity:                           0.70
Number of herds to test:                    2011
Expected total number of animals to test:  9845.44
Expected total costs of the survey:        129248.09
Lookup table for the number of animals to test per herd:

   Herd size  |  No. of animals to test
   -----------------------------------
     1 - 3    |       entire herd
     4 - 5    |            4
       6      |            5
    7 - 31    |            6
   32 - 249   |            7
```

3.3. Limited sampling

For limited sampling a pre-fixed number of animals per selected herd (=the sample limit) is tested, irrespective of the actual herd size. The chosen sample limit determines the (mean) herd sensitivity and thus the sample size on a herd level. The sample limit and the number of herds act in a complementary fashion in the sense that low sampling limits result in a large number of herds to be tested and vice versa. If the cost per tested herd and the cost per tested animal is known the package can be used to find the cost optimal sample limit.

Cost optimization

The package `FFD` provides the S4-class `LtdSamplingSummary` and the function `ltdSampling-Summary()`, where the mean herd sensitivity, the number of herds to test, the expected total number of animals tested based on the herd size distribution in the population, as well as the expected overall costs of the survey is computed for a sequence of sample limits. The smallest considered sample limit is 1 animal per herd, the largest sample limit can be specified by the user via the argument `sampleSizeLtdMax`, or if no upper bound is specified, the largest herd size is used.

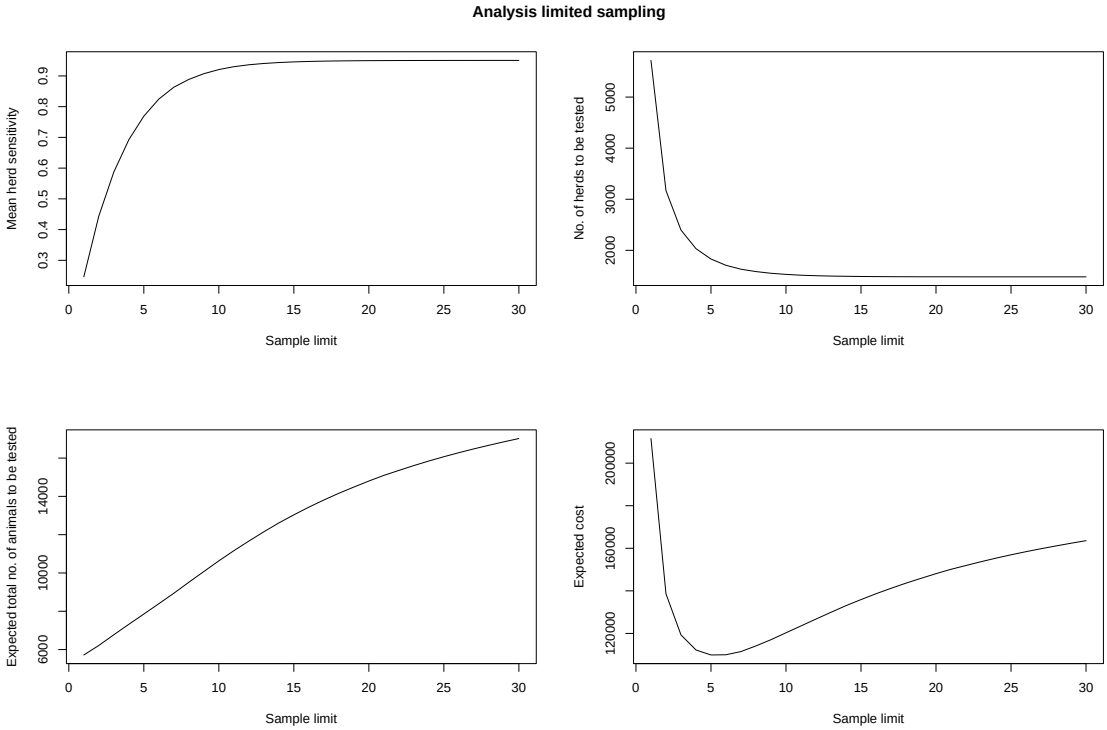
```
> myLtdSampleSummary = ltdSamplingSummary(survey.Data = mySurvey,
       sampleSizeLtdMax = 30)
> summary(myLtdSampleSummary)

LIMITED SAMPLING:

Survey Parameters:
------------------
Design Prevalence:              0.002
Significance level:             0.05
Intra herd prevalence:          0.20
```

```
Sensitivity of diagnostic test:  0.90
Cost per tested herd:            30.00
Cost per tested animal:          7.00

Survey Data:
------------
Number of herds:              15287
Total number of animals:      224606
Number of animals per herd:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    4.00    8.00   14.69   17.00  249.00
Additional population data:
'data.frame':        15287 obs. of  3 variables:
 $ herdId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ state : int  7 7 6 3 8 7 3 7 4 3 ...
 $ nSheep: num  22 30 4 11 11 3 94 53 4 24 ...

Cost optimal sampling strategy:
-------------------------------
Fixed number of animals to test per herd:  5
Mean herd sensitivity:                     0.77
Number of herds to test:                   1830
Expected total number of animals to test:  7854.38
Expected total costs of the survey:        109880.68
```

A plot of the object of class `LtdSamplingSummary` can be created using `plot()`. The plot consists of (row-wise from top left to bottom right)

- the mean herd sensitivity plotted against the sample limit,
- the number of herds to test plotted against the sample limit,
- the expected total number of animals to test plotted against the sample limit,
- the expected overall costs plotted against the sample limit.

```
> plot(myLtdSampleSummary)
```



The summary of the object of class `LtdSamplingSummary` can further be exported to an html-file using the method `HTML`. This method creates an html-file and a css-file containing the the data in the `IndSamplingSummary` object, as well as the diagnostic plots.

```
> HTML(myLtdSamplingSummary)
```

The method further accepts the same arguments as the function `HTMLInitFile()` from the package `R2HTML`, e.g., `filename`, `outdir`, `CSSFile` and `Title`.

Parameters for a fixed sample limit

If one has decided on an appropriate sample size the herd sensitivity, number of herds to test, expected total number of animals to test and expected costs can be determined using the function `ltdSampling()` to create an object of the class `LtdSampling`. The function takes two arguments, `survey.Data`, an object of the class `SurveyData`, and the sample limit `sampleSizeLtd`. The computed parameters can again be displayed using the methods `show()`, `summary()` and `HTML()`.

Let's say we have chosen the appropriate sample limit to be 7 animals per herd:

```
> myLtdSampling <- ltdSampling(survey.Data = mySurvey,
      sampleSizeLtd = 7)
> summary(myLtdSampling)

LIMITED SAMPLING:
```

```
Survey Parameters:
------------------
Design Prevalence:               0.002
Significance level:              0.05
Intra herd prevalence:           0.20
Sensitivity of diagnostic test:  0.90
Cost per tested herd:            30.00
Cost per tested animal:          7.00

Survey Data:
------------
Number of herds:             15287
Total number of animals:     224606
Number of animals per herd:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    4.00    8.00   14.69   17.00  249.00
Additional population data:
'data.frame':        15287 obs. of  3 variables:
 $ herdId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ state : int  7 7 6 3 8 7 3 7 4 3 ...
 $ nSheep: num  22 30 4 11 11 3 94 53 4 24 ...

Sampling strategy:
------------------
Fixed number of animals to test per herd:  7
Mean herd sensitivity:                     0.86
Number of herds to test:                   1630
Expected total number of animals to test:  8939.78
Expected total costs of the survey:        111478.48
```

4. Sample size calculation without classes

In order to provide sufficient flexibility `FFD` offers a set of tools that operate with traditional data types (mostly vectors and data frames). The basis of these tools is equation (1), the evaluation of which is implemented in the function `computePValue`. The function takes the population size, the sample size, the number of diseased individuals in the population, the sensitivity and the specificity of the test as arguments and returns the probability of finding no testpositive individuals, given that the disease is present in the population with the design prevalence:

```
> p.value <- computePValue(nPopulation = 15287, nSample = 1630,
      nDiseased = round(15287*0.002), sensitivity = 0.8633, specificity = 1)
> p.value

[1] 0.04997705
```

The optimal sample size is defined as the smallest sample size that still produces a probability smaller than a given significance level. This sample size can be evaluated using the function `computeOptimalSampleSize`. The function takes the population size, the design prevalence, the significance level, the sensitivity and the specificity of the test as arguments (the argument `lookupTable` will be discussed in section 4.1 on individual sampling) and returns the optimal sample size:

```
> nSample <- computeOptimalSampleSize(nPopulation = 15287,
       prevalence = 0.002, alpha = 0.05, sensitivity = 0.8633,
       specificity = 1, lookupTable = FALSE)
> nSample

[1] 1630
```

4.1. Individual sampling

For individual sampling the herd sensitivity is fixed and constant for every sampled herd. Hence the number of herds to test can be computed using `computeOptimalSampleSize`. The arguments are the number of herds in the population (`nPopulation`), the design prevalence of the survey (`prevalence`), the desired overall significance level (`alpha`) and the herd sensitivity (`sensitivity`). The specificity should be 1 and `lookupTable` is set to `FALSE`.

The number of animals to test for each herd using individual sampling can be computed using the function `computeOptimalSampleSize` by setting the switch `lookupTable` to `TRUE`. The function then produces a lookup table in the form of a matrix. The input arguments are the maximal herd size that should be included in the lookup table (`nPopulation`), the intra herd prevalence (`prevalence`), 1 - the desired herd sensitivity (`alpha`), the sensitivity of the diagnostic test (`sensitivity`) and the specificity of the diagnostic test (`specificity`), which should be kept at 1.

```
> lookupTable <- computeOptimalSampleSize(nPopulation = max(sheepData$nSheep),
       prevalence = 0.2, alpha = 0.3, sensitivity = 0.9,
       specificity = 1, lookupTable = TRUE)
> lookupTable

     N_lower N_upper sampleSize
[1,]       1       1          1
[2,]       2       2          2
[3,]       3       3          3
[4,]       4       5          4
[5,]       6       6          5
[6,]       7      31          6
[7,]      32     249          7
```

4.2. Limited sampling

For limited sampling the herd sensitivity depends on the herd size. The herd size is complementary to the significance level α of the herd test, i.e., herd sensitivity = 1 - α. The α-values of the herd test, as well as the mean α (= 1 - mean herd sensitivity) is computed via the function `computeAlphaLimitedSampling()`. The function takes a vector containing the herd sizes for each holding, the sample limit, the intra herd prevalence and sensitivity and specificity of the diagnostic test and returns a list with two elements. The first element `alphaDataFrame` is a data frame with columns `size` and `alpha` containing the alpha-errors (=1-herd sensitivity) for each herd size. The second element `meanAlpha` is the mean of the alpha values corresponding to the herd size distribution in the population:

```
> alphaList <- computeAlphaLimitedSampling(stockSizeVector = sheepData$nSheep,
       sampleSizeLtd = 7, intraHerdPrevalence = 0.2, diagSensitivity = 0.9,
       diagSpecificity = 1)
> str(alphaList$alphaDataFrame)

'data.frame':        173 obs. of  2 variables:
 $ size : num  1 2 3 4 5 6 7 8 9 10 ...
 $ alpha: num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.0325 0.0725 0.118 ...

> alphaList$meanAlpha

[1] 0.1367245
```

The number of herds to be tested can then be computed using `computeOptimalSampleSize`. The arguments are the number of herds in the population (`nPopulation`), the design prevalence of the survey (`prevalence`), the desired overall significance level (`alpha`), the mean herd sensitivity = 1 - `alphaList$meanAlpha` (`sensitivity`). The specificity and `lookupTable` should be kept at their default values.

5. A-posteriori calculation of the alpha-error

The calculation of the sample size on herd level, i.e., the number of herds to test is based on the herd sensitivity. For limited sampling the herd sensitivity depends on the size of each herd, hence a mean herd sensitivity is used for the sample size calculation. This, on the other hand, means that the overall significance level of the scheme depends on the chosen sample, i.e., if the sample contains a high proportion of very large herds the mean herd sensitivity in the sample is lower than the mean herd sensitivity in the population and hence the desired overall significance level is not met. If the sample contains a lot of very small herds, then the mean herd sensitivity in the sample exceeds that of the population and the significance of the sampling scheme falls below the desired significance level, i.e., the sampling scheme is "too thorough", more herds are being tested than necessary.

It is therefore of interest to compute the significance level of the sampling scheme after the sample has been drawn, i.e., to compute the a-posteriori alpha-error, which is the probability of finding no testpositives in the **given sample**, given that the disease is present at the design prevalence. The a-posteriori alpha-error can then be used to assess a given sample and to possibly modify it, i.e., reduce or extend it in order to meet the prescribed significance level.

The a-posteriori analysis of the alpha-error is also of interest when using individual sampling. With individual sampling the number of animals to test is computed for every herd size, in order to guarantee the "same" herd sensitivity for every herd. The herd sensitivity as a function of the number of animals tested is however a discrete function with jumps. The exact value can in most cases not be achieved and the desired herd sensitivity is taken as a lower bound, i.e., for each herd the number of animals to test is computed as the smallest number that achieves a herd sensitivity greater or equal to the desired value. E.g., for a herd of a given size the herd sensitivity when testing 4 animals might be 0.68 and for 5 animals it might be 0.74. If the desired herd sensitivity is 0.7 then 5 animals would be tested. The mean herd sensitivity for individual sampling therefore always exceeds the desired value, hence the number of herds to test is generally higher than necessary.

The package `FFD` offers tools to compute the a-posteriori alpha-error for a given sample. Furthermore a sampling scheme is implemented that dynamically updates the a-posteriori alpha-error during the sampling procedure and updates the sample size automatically in order to prevent under- or overestimation of the sample size.

5.1. Computation of the a-posteriori alpha-error using FFD

The a-posteriori error of a given sample can be computed using the function `computeAposterioriError()`. The function requires the population size, the number of diseased elements in the population according to the design prevalence and a vector of the herd-level alpha-errors of the herds in the sample (= 1 - herd sensitivity). Furthermore it can be specified if the a-posteriori error should be computed exactly or if an approximation should be used. The exact calculation is computationally costly due to combinatorial issues and is not recommended if there are more than 6 diseased elements in the population. The approximation comes very close to the exact value and is significantly more efficient.

The vector of herd-level alpha-errors can be generated using the function `computeAlpha()`. It takes the vector of herd sizes, the intra herd prevalence, the sensitivity of the diagnostic test as arguments, as well as parameters concerning the sample strategy: for `method == "limited"` the sample limit `sampleSizeLtd` must be specified, for `method == "individual"` the herd sensitivity `herdSensitivity` must be specified:

```
> sampleVec <- sample(sheepData$nSheep, 2550, replace = FALSE)
> alphaVec <- computeAlpha(nAnimalVec = sampleVec,
      method = "limited", sampleSizeLtd = 9,
      intraHerdPrevalence = 0.2, diagSensitivity = 0.9)
> system.time({
   errorExact <- computeAposterioriError(alphaErrorVector = alphaVec,
       nPopulation = 5000, nDiseased = 5, method = "exact")})

   user  system elapsed
   0.33    0.00    0.33

> errorExact

[1] 0.04461766
```

```
> system.time({
   errorApprox <- computeAposterioriError(alphaErrorVector = alphaVec,
       nPopulation = 5000, nDiseased = 5, method = "approx")})

   user  system elapsed
      0       0       0

> errorApprox

[1] 0.04461784
```

5.2. Sampling using S4 classes

The method `sample()` has been implemented for the classes `IndSampling` and `IndSampling`. It takes two arguments, the first argument `x` is an object of the class `IndSampling` or `LtdSampling` and the second argument `size` is a character string specifying the sampling strategy. For `size = "fixed"` the fixed number `x@nHerds` of herds is sampled using simple random sampling. For `size = "dynamic"` dynamic sampling is used. The method returns a list with two items: a vector of indices of the sampled herds corresponding to `x@surveyData@nAnimalVec` and the a-posteriori alpha-error of the sample:

```
> ## Fixed sampling:
> ##################
> sampleFixed <- sample(x = myIndSampling, size = "fixed")
> ## Sample Size:
> length(sampleFixed$indexSample)

[1] 2011

> ## Significance:
> sampleFixed$aPostAlpha

[1] 0.03027402

> ## Sample:
> head(sampleFixed$sample)

   herdId state nSheep
2       2     7     30
4       4     3     11
6       6     7      3
20     20     6      3
27     27     7     10
45     45     3      8
```

```
> ## Dynamic sampling:
> ####################
> sampleDynamic <- sample(x = myIndSampling, size = "dynamic")
> ## Sample Size:
> length(sampleDynamic$indexSample)

[1] 1741

> ## Significance:
> sampleDynamic$aPostAlpha

[1] 0.04998022

> ## Sample:
> head(sampleFixed$sample)

   herdId state nSheep
2       2     7     30
4       4     3     11
6       6     7      3
20     20     6      3
27     27     7     10
45     45     3      8
```

Index

References

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

Cameron AR, Baldock FC (1998b). "Two-stage sampling surveys to substantiate freedom from disease." *Preventive Veterinary Medicine*, **34**, 19–30.

Ziller M, Selhorst T, Teuffert J, Kramer M, Schlüter H (2002). "Analysis of sampling strategies to substantiate freedom from disease in large areas." *Preventive Veterinary Medicine*, **52**, 333–343.

Affiliation:

Ian Kopacka
Austrian Agency for Health and Food Safety (AGES)
Division for Data, Statistics and Risk Assessment
Department for EPI-VET
Beethovenstraße 8
A-8010 Graz, Austria
E-mail: ian.kopacka@ages.at
URL: http://www.ages.at/