

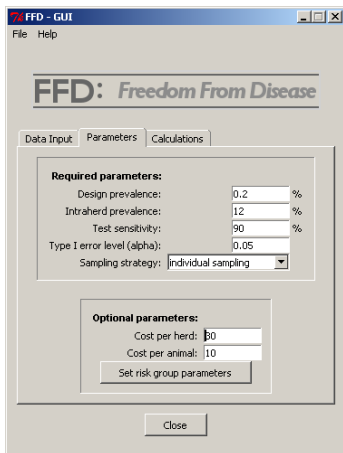
FFD: R-softwaretool for the design of risk based sampling schemes to substantiate freedom from disease

I. Kopacka, J. Hofrichter
Graz, September 8, 2011



Daten, Statistik
und Risikobewertung

Software package FFD



The screenshot shows the 'FFD - GUI' window with the 'Parameters' tab selected. The window has a menu bar with 'File' and 'Help'. The title bar says 'FFD - GUI'. The main area is titled 'FFD: Freedom From Disease'. Below the title are three tabs: 'Data Input', 'Parameters', and 'Calculations'. The 'Parameters' tab is active, showing two sections: 'Required parameters' and 'Optional parameters'. The 'Required parameters' section includes: 'Design prevalence: 0.2 %', 'Intraherd prevalence: 12 %', 'Test sensitivity: 90 %', 'Type I error level (alpha): 0.05', and 'Sampling strategy: individual sampling' (a dropdown menu). The 'Optional parameters' section includes: 'Cost per herd: 80' and 'Cost per animal: 10'. There is a button 'Set risk group parameters' below these. At the bottom of the window is a 'Close' button.

Parameter	Value	Unit
Design prevalence	0.2	%
Intraherd prevalence	12	%
Test sensitivity	90	%
Type I error level (alpha)	0.05	
Sampling strategy	individual sampling	
Cost per herd	80	
Cost per animal	10	

- Package for the open source software **R**
- Design and analysis of surveys to substantiate freedom from disease
- Two-stage sampling
- Finite populations
- Imperfect diagnostic test
- Risk-based sampling
- Graphical user interface (GUI), S4 classes, functions

Content

- Disease freedom
- Exact alpha-error
- Risk-based sampling

Disease freedom

Freedom from disease

Motivation: Brucella melitensis

EU-requirement:

Member states must show with a probability exceeding **95 %** that no more than $\pi_0 = \mathbf{0.2 \%}$ of the herds are infected.



⇒ Statistical test:

H_0 : prevalence $\pi = \pi_0$, H_A : $\pi < \pi_0$
 $\alpha = 0.05$.

Two sampling stages:

- Sample herds ($n_{herds} = ?$)
- Sample animals from the selected herds ($n_{animals} = ?$)

One-stage sampling [1]

Test setup:

T^+ ...number of test-positive individuals in the sample.

$T^+ = 0 \Rightarrow \text{reject } H_0$

$T^+ > 0 \Rightarrow \text{do not reject } H_0$

Diagnostic test might be imperfect.

Definition:

The **sensitivity** of a diagnostic test is defined as the probability of obtaining a positive test result, given the individual is diseased.

One-stage sampling [2]

Parameters: N ...population size

n ...sample size

Se ...sensitivity of the diagnostic test

H₀: d ...number of diseased in population ($d \approx N \cdot \pi_0$)

[Cameron, Baldock, 1998]

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

\Rightarrow choose smallest n with $P(T^+ = 0 | n, Se) \leq \alpha$.

Two-stage sampling

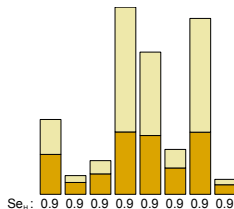


Diagnostic test on herd level = separate sampling scheme.

Sensitivity on top-level = confidence on lower level (assume constant).

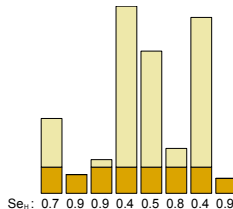
	herd level	animal level
N	no. of herds in pop.	no. of animals in herd
Se	herd sensitivity	sensitivity of diag. test
π	design prevalence	intra-herd prevalence
α	overall significance	1-herd sensitivity
n	no. of herds to test	no. of animals to test per herd

Sampling strategies [Ziller et al., 2002]



Individual Sampling:

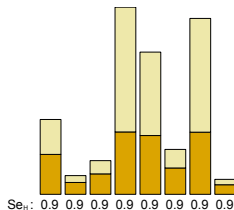
Constant herd sensitivity
(lower bound),
 $n_{animals}$ depends on herd
size.



Limited Sampling:

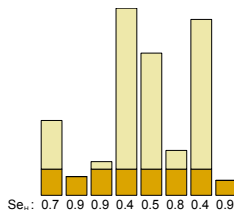
$n_{animals}$ fixed,
Herd sensitivity varies
(\Rightarrow use mean value).

Sampling strategies [Ziller et al., 2002]



Individual Sampling:

- Choose herd sens.
- Compute n_{herds}
- Compute $n_{animals}(N)$



Limited Sampling:

- Choose $n_{animals}$
- Compute mean herd sens.
- Compute n_{herds}

Exact alpha error

Alpha error

The computed herd sensitivity is an approximation:

- **Individual sampling:** Herd sensitivity is systematically under-estimated. Conservative and resource consuming.
- **Limited sampling:** Herd sensitivity is averaged. Depends on realization of sample:
 - “Too many” small herds: herd sensitivity under-estimated.
 - “Too many” large herds: herd sensitivity over-estimated.

⇒ Computed significance α is also an approximation, the true value depends on the chosen sample.

Exact alpha error [1]

Let $Se_H^{(1)}, \dots, Se_H^{(n)}$ be the herd sensitivities of a specific sample:

$$\alpha_{ex} = \sum_{y=\max(0, d-(N-n))}^{\min(d, n)} \frac{\binom{d}{y} \binom{N-d}{n-y}}{\binom{N}{n}} \frac{1}{\binom{n}{y}} \sum_{(I \subset \{1, \dots, n\} \wedge |I|=y)} \prod_{j \in I} (1 - Se_H^{(j)}).$$

- How many diseased herds are there in the sample?
- Which y of the n herds in the sample are diseased?
- What is the probability falsely classifying all diseased herds as being healthy?

If all herd sensitivities are the same, the formula above coincides with [Cameron, Baldock, 1998].

Exact alpha error [1]

Let $Se_H^{(1)}, \dots, Se_H^{(n)}$ be the herd sensitivities of a specific sample:

$$\alpha_{ex} = \sum_{y=\max(0, d-(N-n))}^{\min(d, n)} \frac{\binom{d}{y} \binom{N-d}{n-y}}{\binom{N}{n}} \frac{1}{\binom{n}{y}} \sum_{(I \subset \{1, \dots, n\} \wedge |I|=y)} \prod_{j \in I} (1 - Se_H^{(j)}).$$

- **How many** diseased herds are there in the sample?
- Which y of the n herds in the sample are diseased?
- What is the probability falsely classifying all diseased herds as being healthy?

If all herd sensitivities are the same, the formula above coincides with [Cameron, Baldock, 1998].

Exact alpha error [1]

Let $Se_H^{(1)}, \dots, Se_H^{(n)}$ be the herd sensitivities of a specific sample:

$$\alpha_{ex} = \sum_{y=\max(0, d-(N-n))}^{\min(d, n)} \frac{\binom{d}{y} \binom{N-d}{n-y}}{\binom{N}{n}} \frac{1}{\binom{n}{y}} \sum_{(I \subset \{1, \dots, n\} \wedge |I|=y)} \prod_{j \in I} (1 - Se_H^{(j)}).$$

- How many diseased herds are there in the sample?
- Which y of the n herds in the sample are diseased?
- What is the probability falsely classifying all diseased herds as being healthy?

If all herd sensitivities are the same, the formula above coincides with [Cameron, Baldock, 1998].

Exact alpha error [1]

Let $Se_H^{(1)}, \dots, Se_H^{(n)}$ be the herd sensitivities of a specific sample:

$$\alpha_{ex} = \sum_{y=\max(0, d-(N-n))}^{\min(d, n)} \frac{\binom{d}{y} \binom{N-d}{n-y}}{\binom{N}{n}} \frac{1}{\binom{n}{y}} \sum_{(I \subset \{1, \dots, n\} \wedge |I|=y)} \prod_{j \in I} (1 - Se_H^{(j)}).$$

- How many diseased herds are there in the sample?
- Which y of the n herds in the sample are diseased?
- What is the **probability falsely classifying all diseased herds as being healthy**?

If all herd sensitivities are the same, the formula above coincides with [Cameron, Baldock, 1998].

Exact alpha error [1]

Let $Se_H^{(1)}, \dots, Se_H^{(n)}$ be the herd sensitivities of a specific sample:

$$\alpha_{ex} = \sum_{y=\max(0, d-(N-n))}^{\min(d, n)} \frac{\binom{d}{y} \binom{N-d}{n-y}}{\binom{N}{n}} \frac{1}{\binom{n}{y}} \sum_{(I \subset \{1, \dots, n\} \wedge |I|=y)} \prod_{j \in I} (1 - Se_H^{(j)}).$$

- How many diseased herds are there in the sample?
- Which y of the n herds in the sample are diseased?
- What is the probability falsely classifying all diseased herds as being healthy?

If all herd sensitivities are the same, the formula above coincides with [Cameron, Baldock, 1998].

Exact alpha error [2]

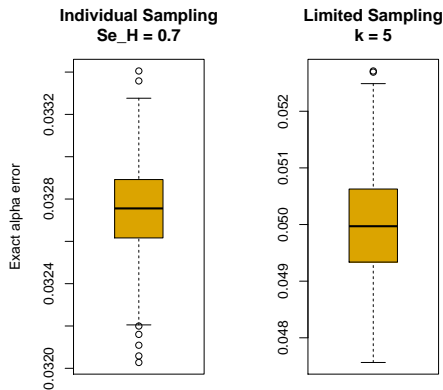
The exact alpha can be used for:

- A-posteriori validation of the sampling plan
- *Dynamic sampling*: Compute alpha-error inline during sampling and increase sample until alpha-error falls below threshold.

Simulation: Exact alpha error

Data/parameters:

15287 sheep herds in Austria, $\alpha = 0.05$, $\pi = 0.002$,
 $\pi_{IH} = 0.13$, $Se = 0.9$, 1000 iterations:



Simulation: Dynamic Sampling

Data/parameters:

15287 sheep herds in Austria, $\alpha = 0.05$, $\pi = 0.002$,
 $\pi_{IH} = 0.13$, $\text{Se} = 0.9$, 1000 iterations:

Strategy	n_{fix}	n_{dyn}		
		$q_{0.025}$	$q_{0.5}$	$q_{0.975}$
ltd, $k = 5$	2332	2305	2333	2360
ltd, $k = 7$	1965	1951	1965	1980
ind, $\text{Se}_H = 0.5$	2882	2245	2259	2274
ind, $\text{Se}_H = 0.7$	2058	1811	1818	1825
ind, $\text{Se}_H = 0.9$	1601	1574	1576	1577

Risk-based sampling

Risk factors

Not all herds have the same risk of being infected.

Possible risk factors are:

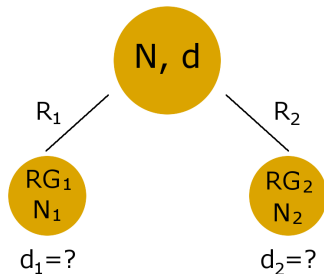
- lively trade
- import from abroad
- high herd density
- proximity to the border
- ...

Goal:

Reduction of sample size through targeted sampling of high-risk groups.

2 risk factors

Assumption: Population is divided into 2 risk groups RG_1 , RG_2 with risks R_1 , R_2 of contracting the disease.



If d_1 , d_2 is known \Rightarrow multiply the alpha-errors.

BUT: only the disease risks are known, d_1 , d_2 are random variables.

Number of diseased d_i

Setup: We pick a diseased individual from the population. What is the probability of it originating from risk group RG_1 ?

$$P(RG_1|D) = \frac{P(D|RG_1)P(RG_1)}{P(D)} = \frac{R_1 \cdot \frac{N_1}{N}}{P(D)}.$$

With

$$\begin{aligned}P(D) &= P(D \cap RG_1) + P(D \cap RG_2) \\&= P(D|RG_1) \cdot P(RG_1) + P(D|RG_2) \cdot P(RG_2) \\&= R_1 \cdot \frac{N_1}{N} + R_2 \cdot \frac{N_2}{N}\end{aligned}$$

we find

$$p_1 := P(RG_1|D) = \frac{R_1 N_1}{R_1 N_1 + R_2 N_2}.$$

Significance level

$\Rightarrow d_i$ are binomially distributed:

$$d_i \sim B(d, p_i), \quad \text{for } i = 1, 2$$

$$\text{with } p_i = \frac{R_i N_i}{R_1 N_1 + R_2 N_2}.$$

Note: Only relative values for R_i are necessary.

Alpha error:

$$\begin{aligned} P(T^+ = 0 | N_1, N_2, n_1, n_2, R_1, R_2, d) &= \\ &= \sum_{y_1 = \max(0, d - N_2)}^{\min(d, N_1)} \binom{d}{y_1} p_1^{y_1} \cdot (1 - p_1)^{d - y_1} \cdot \\ &\quad \cdot P_h(T^+ = 0 | N_1, n_1, d_1 = y_1) \cdot P_h(T^+ = 0 | N_2, n_2, d_2 = d - y_1). \end{aligned}$$

Sources

R-package FFD:

<http://ffd.r-forge.r-project.org/>

<http://cran.r-project.org/web/packages/FFD>

Literature:

A. G. Cameron and F. C. Baldock, *Two-stage sampling in surveys to substantiate freedom from disease*, Prev. Vet. Med. 34 (1998), pp. 18-30.

I. Kopacka and J. Hofrichter, *A-posteriori alpha-error determination for two-stage sampling strategies to substantiate freedom from disease*, Technical-Report No. 008 (07/2011), AGES-DSR/EPI, Graz.

M. Ziller, T. Selhorst, J. Teuffert, M. Kramer and H. Schlüter, *Analysis of sampling strategies to substantiate freedom from disease in large areas*, Prev. Vet. Med. 52 (2002), pp. 333-343.