A Simulation-Based Approach to Determine Sample Sizes in Stochastic Scenario Tree Models for Freedom of Disease

Henrik Stryhn, Adel Elghafghuf, Jette Christensen

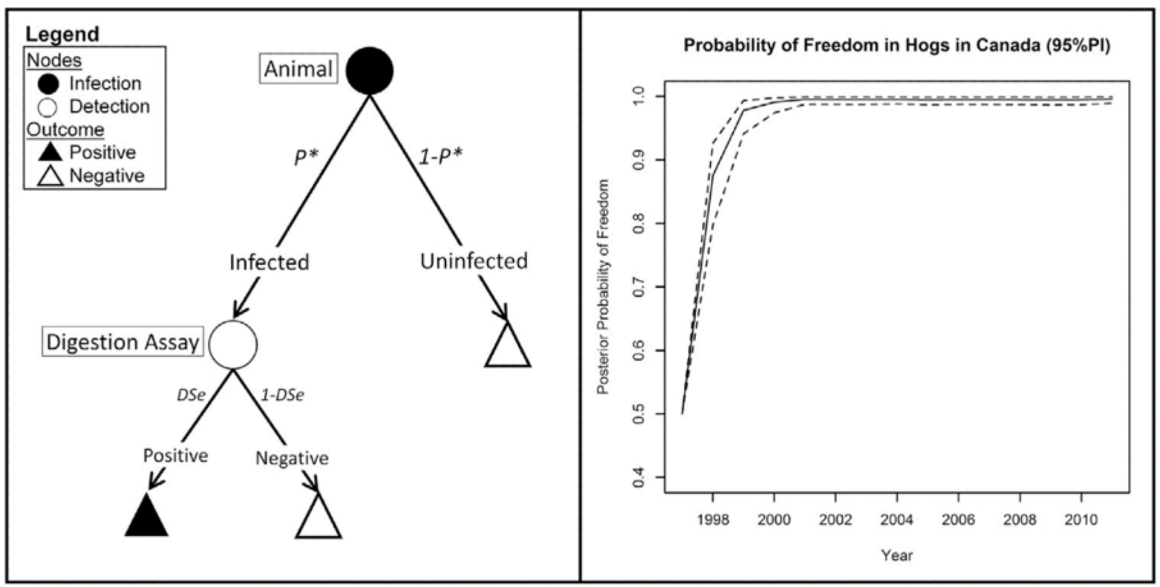
CVER - Centre for Veterinary Epidemiological Research
University of Prince Edward Island

Scenario Tree Models for Freedom of Disease

- Freedom of disease (in a population) $\sim$ prevalence below some threshold (P^*, the "design prevalence",
- The flow of sampling and testing of animals is organized in a scenario tree,
- When all animals test negative, compute Pr(prevalence $\leq P^*$), the "probability of freedom of disease" (PFD) — an important concept e.g. for trade.

Example from Vanderstichel et al. (2013):

(*Trichinella spiralis*, $P^* = 0.01\%$)



Objective (talk): to outline how sample sizes for stochastic scenario tree models (STMs) can be determined by simulation, and to illustrate the impact of key parameters in STMs.

KEY PARAMETERS OF SCENARIO TREE MODELS

- design prevalence: fixed value set by user/context for "acceptable" low level of disease,
  - 1-level model ("animals"): single value P^* for population prevalence,
  - 2-level model ("animals within herds"): values P_u^* and P_h^* for unit (within-herd) and between-herd prevalences, resp.,
- probability of "introduction" (P_{in}) of disease from one time step to the next: fixed or stochastic value, possibly time-dependent,
- initial PFD (P_0, at start-up of sampling/model): fixed or stochastic, sometimes set arbitrarily at 0.5,
- sampling or test parameters, e.g. diagnostic test sensitivity (DSe): fixed or stochastic.

Stochastic nodes/parameters are drawn from probability distributions, e.g. the commonly used PERT[1] distribution (a, b, c) for values within a bounded range (e.g. probabilities or DSe's):

- a beta distribution scaled from $(0, 1)$ to an arbitrary interval (a, c),
- the two parameters of the beta distribution are restricted to one, the PERT distribution's most likely value b, where $a < b < c$.

[1] PERT stands for program evaluation and review technique, a project management tool developed in the 1950s and 1960s with a statistical component for the duration of project phases.

Planning a Scenario Tree Model

...involves to ...

- decide about the time step for the model, e.g. yearly updates,
- determine the sampling design: which units to be sampled per time step,
- determine the structure (e.g., number of levels) and nodes of the tree (including any risk nodes) and their distributions (or fixed values),
- set values or distributions for the basic model parameters (P^*, P_{in}, P_0).

Sample size(s) are determined to meet a specified criterion, say PFD $\geq$ 0.90, but...

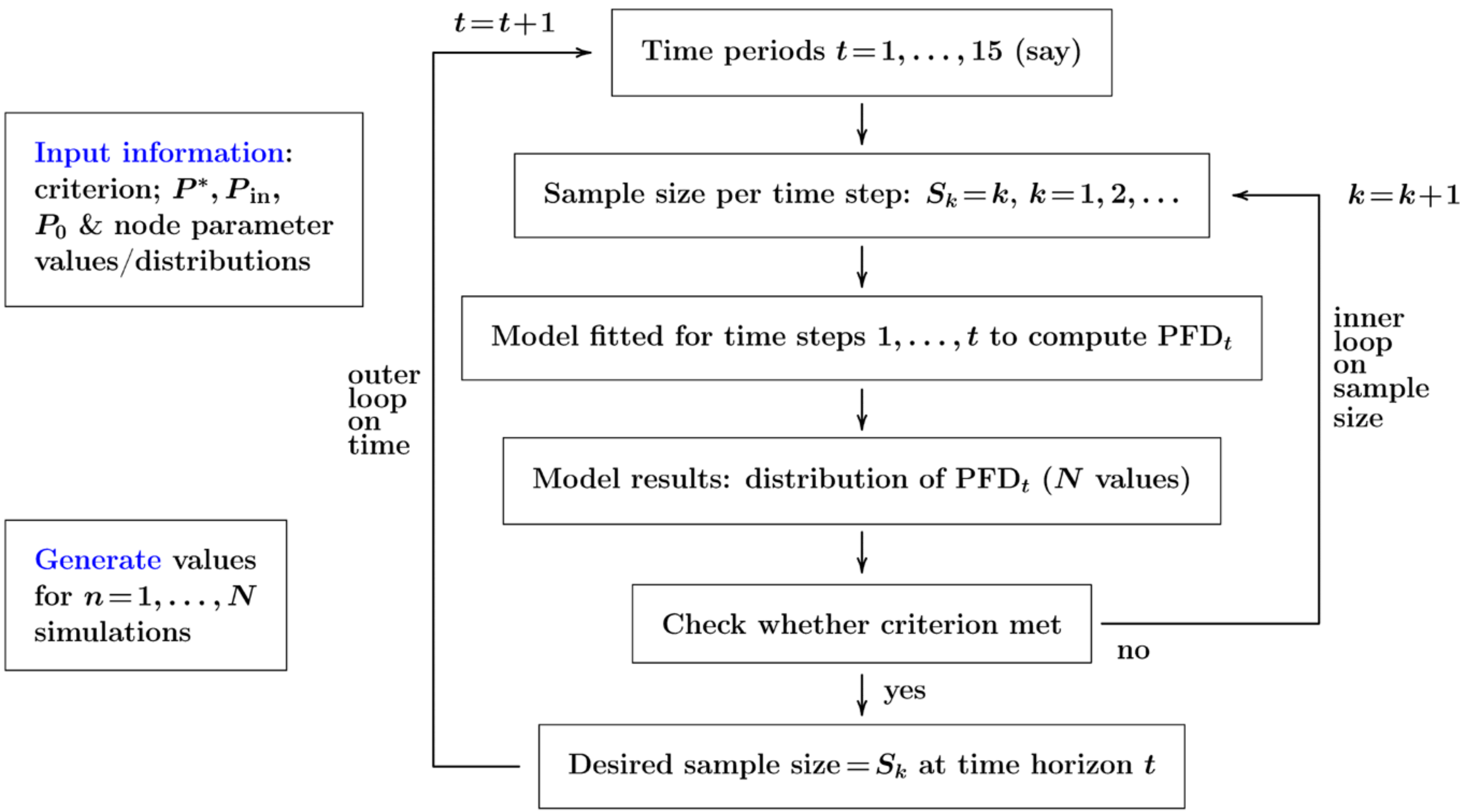
- PFD is random (for a stochastic STM), so need to focus on a feature of its distribution, e.g. the mean or a percentile, determined by simulation[2],
- PFD is time-dependent (calculated after each time step), so need to decide the time horizon for the criterion to be met[3]; note: it may be logistically infeasible to get a system up to a desired criterion in a single time step.

Additional consideration: the sampling will most naturally be designed for two phases: (i) start-up until criterion is met, (ii) maintenance of criterion.

[2] The number of simulations (N) will depend on the complexity of the distribution and the feature of interest; e.g., tail percentiles require larger N than the mean.

[3] If a criterion is to be met immediately (first time step), sample size formula and calculators exist, for both deterministic and simple stochastic trees; e.g., Cameron & Baldock (1998), Cannon (2001), Johnson et al. (2004), and Epitools at the Ausvet website: `https://epitools.ausvet.com.au/samplesize`.

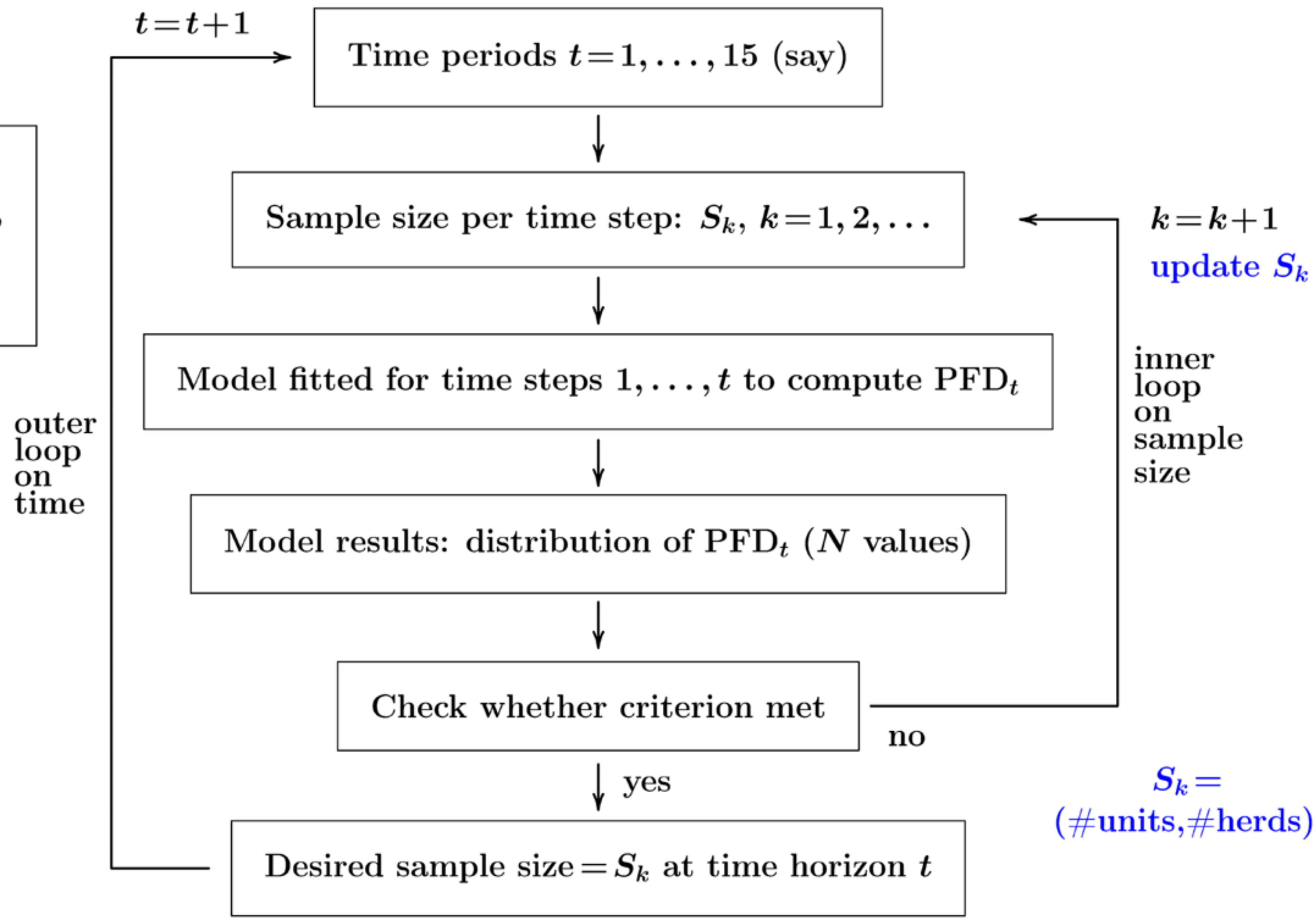
Sample Size Algorithm for 1-Level Stochastic Scenario Tree Model



Sample Size Algorithm for 2-Level Stochastic Scenario Tree Model

Input information: criterion; P^*_u, P^*_h, P_{in}, P_0 & node parameter values/distributions

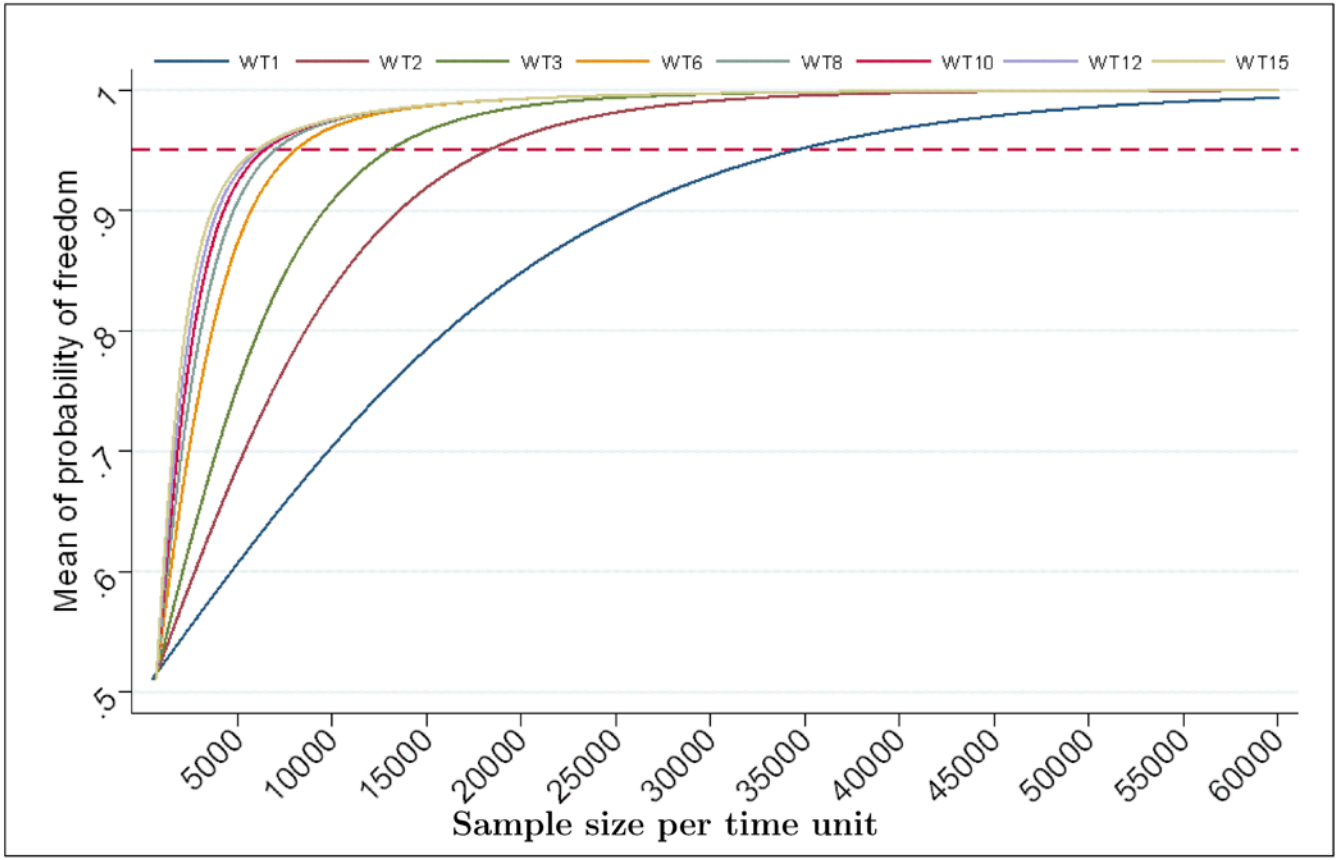
Generate values for $n = 1, \ldots, N$ simulations



First Results: PFD versus Sample Size (1-Level Model)

Model Settings:
$P^* = 0.0001$, $P_0 = 0.5$, $P_{\text{in}} \sim \text{PERT}(.001, .03, .07)$; **test node**: DSe $\sim$ PERT(.4, .95, .99).

(WT = waiting time for system's PFD)

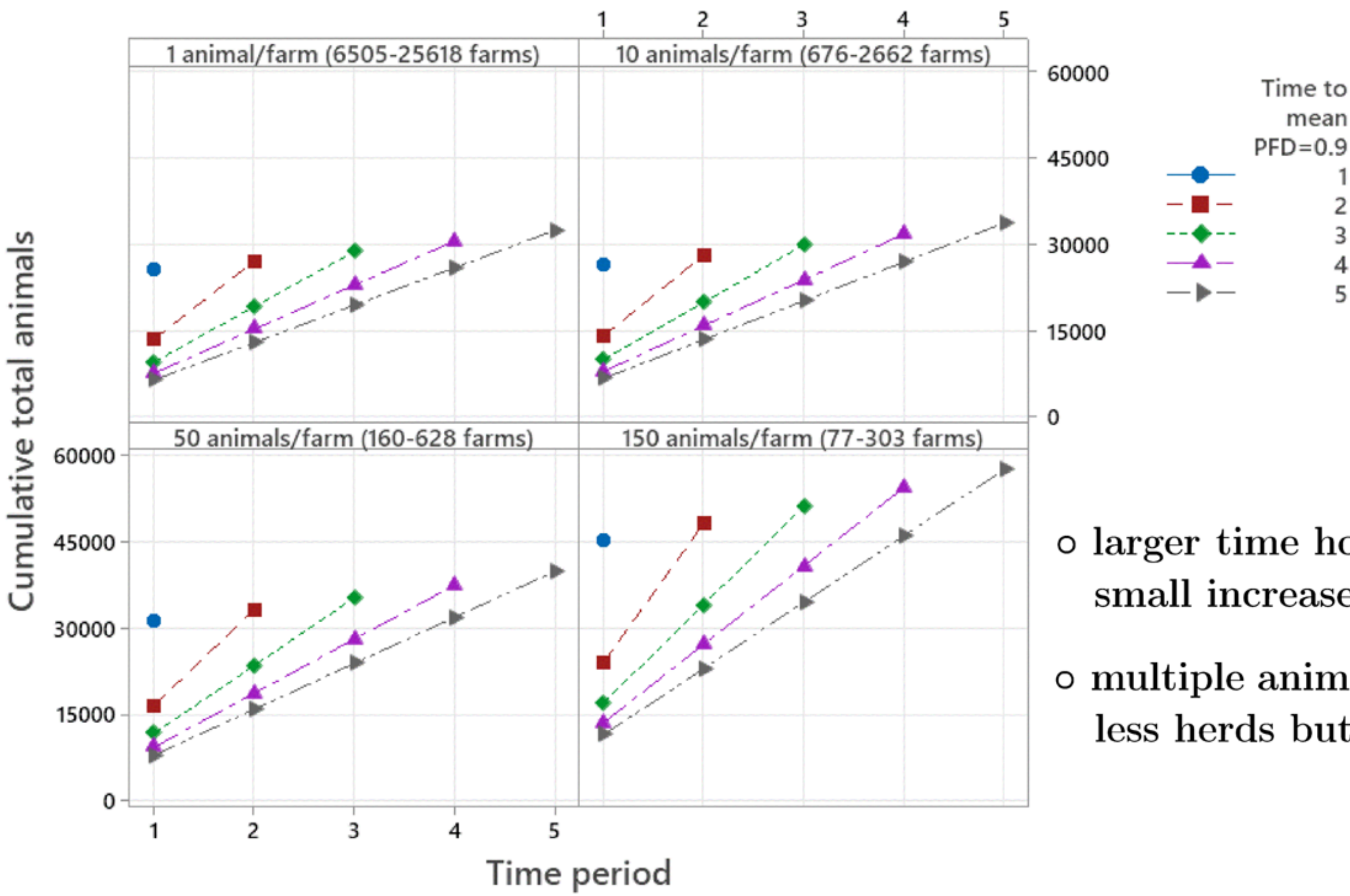


Findings:

- (mean) PFD increases with sample size and years sampled,
- high PFD may require large number of samples (total).

Next Results: Total Animals Required (2-Level Model)

Model Settings: $P_u^* = P_h^* = 0.01$, other settings unchanged.



- larger time horizons ⇒ small increases in # animals
- multiple animals per herd ⇒ less herds but more animals

Algorithm Details

- calculation of PFD used Bayesian updating formulas, which are now standard for scenario tree models (e.g., Martin et al., 2007), in particular

$$\mathrm{PFD}_t = \frac{\mathrm{PFD}_t^*}{\mathrm{PFD}_t^* + (1-\mathrm{PFD}_t^*)(1-\mathrm{SSe}_t)},$$

  where SSe is the system sensitivity, and PFD_t^* is the (posterior) PFD_{t-1} from the previous time step updated by the probability of introduction P_{in}[4], now having the role of a prior probability for the next time step,
- the search over sample sizes for a 1-level model may utilize that larger time horizons require fewer samples per time step,
- the search over sample sizes for animals and herds in a 2-level model can be implemented in different ways, e.g. by fixing first the # herds and searching for # animals, or vice versa,[5]
- simulations were based on $N = 10\,000$ iterations,
- all coding was done in R, using the `mc2d` library.

[4] This update (also referred to as temporal discouting) is a simple multiplication of PFD_{t-1} by $(1 - P_{\mathrm{in}})$, assuming independence between disease introduction prior to and during the time step.
[5] Note: Some settings with low # herds may not be able to meet the search criterion, regardless of # animals.

Maintaining High PFD; Adaptive Sampling

Fact: The requirements on sample size to start up a STM surveillance are typically far heavier than for maintaining an ongoing surveillance, [6]

$\Rightarrow$

Also of interest to know required sample sizes for a running system:

- simple approach: start with high value for P_0 (but ignores uncertainty in PFD distribution),
- adaptive sample size determination: for each time step, assume the minimal number of samples to meet the PFD criterion,
  - essentially the same algorithm, but the PFD distribution from the previous time steps can be used directly for the next time step,
  - two-dimensional searches for (# animals, # herds) may need further assumptions/restrictions on what is desirable.

Some findings for adaptive sampling (1-level scenario):

- adaptive sample sizes stabilize, quickly with high P_{in} and mean-based criteria,[7]
- adaptive sampling always ($t > 1$) requires more units than fixed horizon sampling.

[6] The trade-off between information from new and past samples is largely controlled by P_{in}.
[7] Equilibrium results for PFD exist when P_{in} and the system sensitivity are constant.

Concluding Remarks

Main message: exploring implications of sample sizes is relatively easy to do in stochastic scenario tree models once you know how to update such models
— intuitively because there is no variation in the actual data (all test results are negative).

Second main message: the approach is flexible enough to incorporate specific features of the scenario tree to be set up (e.g., its structure, its time horizon to meet the PFD criterion, time-varying parameters or sample sizes).

Our results largely confirmed general rules and expectations for the system's behaviour (e.g., about the gain of sampling herds relative to animals within herds), but the ability to simulate a system provides quantitative information not otherwise available.

You Have Just Seen . . .

A simulation-based approach to determine sample sizes in stochastic scenario tree models for freedom of disease
presented by Henrik Stryhn (http://stryhnstatistics.ca)

Thank you for your attention!