# Important formulae for surveillance

Evan Sergeant
evan@ausvet.com.au
AusVet Animal Health Services
July 2012
http://epitools.ausvet.com.au/docs/Important_formulae_for_surveillance.pdf

## *Representative surveillance for disease Freedom*

### Terminology:

In this document some specific terminology relating to unit, cluster and population values have been used to try and simplify the formulae presented. These terms are explained here:

*A **unit** is either an individual (animal, plant, etc) as part of a cluster, or a cluster as part of a larger population.*

*A **cluster** is a grouping level of individual units (animals, plants, fish, etc) at a higher level such as a herd, flock, tank, pen, farm, etc. Clusters usually are only considered at one level, but can occur at multiple levels (for example pens within farms within districts).*

***Unit sensitivity** is the sensitivity at the unit level for a particular analysis. For cluster-level analyses, unit sensitivity is the sensitivity of the test (or combination of tests) used, whereas for population-level analyses unit sensitivity is the cluster-level sensitivity for clusters sampled.*

***Population sensitivity** is a sensitivity calculated at some population or grouping level. Depending on the context, the population can be either a cluster of multiple individuals or a larger population comprising multiple clusters.*

***Component sensitivity** is a population-level sensitivity, usually at a country or regional level, calculated for one part (component) of a surveillance system which comprises multiple separate components or activities.*

***System sensitivity** is a population-level sensitivity, usually at a country or regional level, calculated from one or more components*

### Population sensitivity and Sample size

| | | |
|---|---|---|
| Binomial | Population sensitivity:<br>$SeP = 1 - (1 - SeU \times P^*_U)^n$<br>$= 1 - \prod (1 - SeU_i \times P^*_U)$ where SeU varies among units<br>Sample size:<br>$n = \log(1 - SeP)/\log(1 - SeU \times P^*_U)$ | Assumes:<br>• sampling with replacement or sample small (<10%) relative to population<br>• Specificity = 100% |

| | | |
|---|---|---|
| Hypergeometric approximation | Population sensitivity:<br>$SeP = 1 - (1 - SeU \times n/N)^d$<br>$SeP = 1 - (1 - SeU_{avg} \times n/N)^d$ where SeU varies among units<br>Sample size:<br>$n = (N/SeU)*(1 - (1 - SeP)^{1/(P^* \times N)})$ | Assumes:<br>• Sampling without replacement or where sample size is large relative to population<br>• Specificity = 100% |
| Exact | Population sensitivity:<br>$SeP = 1 - (1 - SeU)^d$<br>$SeP = 1 - (1 - SeU_{avg})^d$ where SeU varies among units | Assumes:<br>• Sampling of the entire population<br>• Specificity = 100% |

## Negative Predictive Value (or confidence of population freedom: PFree)

$PFree = (1 - PrInf)/(1 - PrInf \times SeP)$ or
$= PriorPFree/(1 - SeP \times (1 - PriorPFree))$

assuming specificity = 100%.

## Revising confidence of freedom in successive time periods

$PFree_t = 1 - [1 - PFree_{t-1} + PIntro_t - ((1 - PFree_{t-1}) \times PIntro_t)]$

## Equilibrium PFree

Maximum or minimum stable value for PFree for given combinations of SeP and PIntro

$PFree_{equ} = (1 - (PIntro / SeP)) / (1 - PIntro)$

Maximum or minimum value for PriorPFree (after discounting) for given combinations of SeP and PIntro

$PriorPFree_{equ} = 1 - (PIntro / SeP)$

## Design prevalence to achieve specified population sensitivity

Where cluster size is unknown (binomial):

$P^*_U = (1 - \exp((\log(1 - SeP))/n))/SeU$

Where cluster size is known (hypergeometric approximation):

$P^*_U = \log(1 - SeP)/\log(1 - SeU \times n/N)/N$

## Population sensitivity required to achieve desired PFree

SeP = (1 - PriorPFree/PFree)/(1 - PriorPFree)

where PFree is the target value and PriorPFree is the current prior value.

## Population sensitivity required to stay above specified threshold PFree

SeP = PIntro/(1 – Target PFree)

## Combining test sensitivities in series

(For example in a diagnostic process with multiple steps)

$SeU_{combined} = \prod Se_i$

## Combining component sensitivities in parallel, assuming independence

Calculates system sensitivity from multiple components, assuming independence (no overlap between units sampled) between components, for example different compartments or different clusters represented in the surveillance system.

$SeP = 1 - \prod (1 - CSe_i)$

## Updating cluster sensitivities between components where there is overlap

This assumes no independence between components, for example where the same clusters (herds or flocks, etc) are represented in multiple surveillance system components. The probability of infection for each cluster is adjusted between components and resulting component sensitivities are then combined as for assuming independence. For this example binomial calculations are used, but hypergeometric or exact could also be used if appropriate:

***Method 1: Adjusting effective probability of infection between components:***

1. Calculate SeC for each cluster [$SeC = 1 - (1 - SeU \times P^*)^n$] for each component.
2. Calculate posterior confidence of freedom and hence posterior probability of infection for each cluster for the first component (component order is a matter of convenience):

   $PFree_c = (1 - P^*)/(1 - P^*_c \times SeC)$ where $P^*_c$ is the cluster-level design prevalence

   $PostPInf_c = (1 - PFree_c)$
3. Calculate probability that each cluster has a negative test result and hence component sensitivity (CSe) for first component:

   $P(Neg) = 1 - P^*_c \times SeC$

   $CSe = 1 - \prod (P(Neg))$
4. Calculate P(Neg) for each cluster and CSe for the second component after substituting $PostPInf_h$ instead of P* in formula:

   $P(Neg) = 1 - PostPInf_h \times SeC_2$

$$CSe = 1 - \prod (P(Neg))$$

5. Repeat for as many components as necessary
6. Clusters start with P* at the first component in which they appear and then get updated as necessary
7. When all component sensitivities have been calculated, calculate overall system sensitivity (probability that one or more components will yield a positive result if the population is infected at the design prevalence), using independence formula.

$$SSe = 1 - \prod (1 - CSe_i)$$

***Method 2: Aggregating data between components:***

An alternative (often simpler) approach is to aggregate the data for each cluster to calculate single SeC values and then combine these values to calculate overall system sensitivity:

$$SeC = 1 - \prod((1 - P^* \times SeU_i)^{\wedge} n_i)$$

For where $SeU_i$ and $n_i$ are test sensitivity and sample size for each of the i components in the surveillance system.

## Key:

| Abbreviation/symbol | Meaning |
|---|---|
| n, N | Sample size and corresponding population size |
| d | Number of diseased elements in population |
| t | Time period |
| $P^*_U$ | Unit level design prevalence (individual or cluster) |
| Se | Test sensitivity |
| SeU | Unit level sensitivity (test sensitivity when calculating cluster/herd-level sensitivity or cluster/herd-level sensitivity when calculating population or component sensitivity) |
| SeP | Population sensitivity (can be cluster level or overall population level) |
| SeC | Cluster sensitivity |
| $SeC_i$ | Cluster sensitivity for the i-th cluster |
| $SeU_{avg}$ | Average unit sensitivity across all units (individuals or clusters) sampled |
| $CSe_i$ | Component sensitivity for the i-th surveillance system component |
| SSe | System sensitivity |
| PFree | Confidence of population freedom (= negative predictive value) |
| PriorPFree | Confidence of population freedom before undertaking current surveillance |
| PrInf | Prior probability of being infected = 1 – prior confidence of freedom |
| PostPInf | Posterior probability of being infected = 1 – posterior confidence of freedom (NPV) |

## *Risk-based freedom surveillance*

### Adjusted risk and effective probability of infection

$AR_L = 1/(RR \times PPr_H + PPr_L)$
$AR_H = RR \times AR_L$

or for multiple risk levels:

$AR_i = RR_i / \sum( RR \times PPr)$

EPI = P* × AR (for respective risk categories)
EPI > 1 is invalid – design prevalence and/or relative risk should be revised to ensure EPI < 1.

### Population sensitivity for simple, 1-stage, no risk factors, one factor affecting sensitivity

Assuming large population relative to sample size (binomial) and only two unit sensitivity values:

$$SeP = 1 - (1 - P^* \times SeU_H)^{n(h)} \times (1 - P^* \times SeU_L)^{n(l)}$$

n(h) and n(l) are sample sizes for high and low sensitivity groups, respectively; or

assuming small population:

$$SeP = 1 - (1 - SeU_{avg} \times n/N)^d$$

### Sample size for simple, 1-stage, one risk factor (2 levels), constant sensitivity

$USe = EPI_H \times SeU_H \times SPr_H + EPI_L \times SeU_L \times SPr_L$

$n = \log(1 - SeP)/\log(1 - USe)$

$SeU_H$ and $SeU_L$ are the mean values for SeU for high and low risk groups respectively.

### Population sensitivity for simple, 1-stage, one risk factor, one factor affecting sensitivity

$$SeP = 1 - (1 - EPI_H \times SeU_H)^{n(hh)} \times (1 - EPI_H \times SeU_L)^{n(hl)} \times (1 - EPI_L \times SeU_H)^{n(lh)} \times (1 - EPI_L \times SeU_L)^{n(ll)}$$

n(hh), n(hl),n(lh) and n(ll) are sample sizes for high risk & high sensitivity, high risk & low sensitivity, low risk & high sensitivity and low risk & low sensitivity groups, respectively.

## Sample size for simple, 1-stage, one risk factor, one factor affecting sensitivity

$LRSe = SPr_{LH} \times SeU_H + (1 - SPr_{LH}) \times SeU_L$

$HRSe = SPr_{HH} \times SeU_H + (1 - SPr_{HH}) \times SeU_L$

$USe = EPI_H \times HRSe \times SPr_H + EPI_L \times LRSe \times SPr_L$

$n = \log(1 - SeP)/\log(1 - USe)$

LRSe, HRSe are weighted average sensitivity in low and high risk samples respectively.

USe is the probability of a single randomly selected animal from the sample being positive, given the population is infected at the design prevalence.

$SPr_H$, $SPr_L$, $SPr_{LH}$ , $SPr_{HH}$ are proposed sample proportions from the high-risk sub-population, low-risk sub-population, high sensitivity group in high-risk sub-population and high sensitivity group in low-risk sub-population respectively.

## Key:

**See also key for representative freedom surveys**

| Abbreviation/symbol | Meaning |
|---|---|
| RR | Relative risk |
| AR | Adjusted risk |
| $PPr_H$, $PPr_L$ | Population proportions in high and low risk groups, respectively |
| $SPr_H$, $SPr_L$ | The proportion of the surveillance sample from the respective risk group |
| EPI, $EPI_H$, $EPI_L$ | Effective probability of infection and EPI in high and low risk groups. Probabilities of infection after adjusting design prevalence for group relative risks |
| $SeU_H$, $SeU_L$ | Sensitivity in high and low risk groups, respectively. May be test (animal) sensitivity or herd-sensitivity, depending on level at which being calculated. |
| USe | The probability of a single randomly selected animal from the surveillance sample being positive, given the population is infected at the design prevalence. |

## *Prevalence estimation*

### Apparent or seroprevalence

(assumes perfect test sensitivity and specificity)

Estimated prevalence: $P = x/n$

Asymptotic (normal approximation) confidence intervals:

$$CI = P \pm Z\sqrt{((P(1 - P)/n)}$$

Alternative (binomial, Wilson binomial) CI methods usually better, particularly as P approaches 0 or 100%.

Sample size: $n = (Z^2 \times P(1 - P))/e^2$

Assumes a large population. Where expected sample size is large (10%) relative to populations size use following adjustment:

$$n_{adj} = (N \times n)/(N + n)$$

### Estimated true prevalence

(allows adjustment for imperfect sensitivity and specificity)

$$TP = (AP + SP - 1)/(Se + Sp - 1)$$

**Note:** Method fails when Se + Sp = 1 due to division by 0.
TP may be negative if AP + Sp < 1 (Sp estimate is lower than suggested by the results).

Asymptotic (normal approximation) confidence intervals assuming known sensitivity and specificity :

$$CI = TP \pm Z\sqrt{[AP(1 - AP)/(n \times (Se + Sp - 1)^2)]}$$

Assumes Se and Sp known exactly (no uncertainty).
Lower CI may be <0 if TP is close to 0.

Sample size:

$$n = (Z/e)^2 \times (Se \times TP + (1 - Sp) \times (1 - TP)) \times (1 - Se \times TP - (1 - Sp) \times (1 - TP))/(Se + Sp - 1)^2$$

Asymptotic (normal approximation) confidence intervals assuming uncertain sensitivity and specificity :

$$CI = TP \pm Z\sqrt{[AP \times (1\text{-}AP)/(n \times (Se + Sp - 1)^2) + (Se \times (1\text{-}Se) \times TP^2)/(M \times (Se + Sp - 1)^2) + (Sp \times (1\text{-}Sp)*(1\text{-}TP)^2)/(R*(Se + Sp - 1)^2)]}$$

## *Key:*

| Abbreviation/symbol | Meaning |
|---|---|
| n, N | Sample size and corresponding population size (animal level) |
| P | Observed or expected prevalence (proportion) |
| x | Number of units with the characteristic of interest |
| Z | Z distribution value corresponding to desired confidence level<br>Z = 1.96 for 95%, 2.58 for 99% and 1.64 for 90% |
| e | Desired precision of estimate (± relative to estimate). Confidence interval width = 2e |
| $n_{adj}$ | Sample size adjusted for small population |
| TP | True prevalence estimate |
| AP | Apparent prevalence estimate |
| Se, Sp | Sensitivity and specificity of the test used |
| CI | Confidence interval |
| M | Sample size for estimating test sensitivity |
| R | Sample size for estimating test specificity |