Sampling for vine health certification in grape vine nursery stock

Neil McRoberts

Kari Arnold

University *of* California
Agriculture and Natural Resources

General comments on sampling

- It **does not** give definitive answers
  - Statistically designed sampling plans have known long-run performance but can under- or over-estimate disease in any specific case
- It **does not** always (ever?) reduce uncertainty
- It will almost always be constrained by money and/or time
- It should be done often and as early as possible in the propagation chain
- Do not overlook the value of visual inspection

Sampling propagated vines

Sampling the source material will be more efficient

Illustrating the scale of the problem
Suppose N = 5 mother vines
n = 10 budsticks from each = 50 propagated vines

Suppose we want to take Simple Random Sample (SRS) of m = 5 sticks

There are $\binom{50}{5} = 2{,}118{,}760$

ways to draw the sample. ***n^N=100,000 combinations have wood from all 5 mother vines so only 100,000/2,118,760 = 0.047 (5%) of SRS capture all 5 mother vines.***

Sampling propagated vines cont’d.

Sampling the source material will be more efficient

More realistic (but still tiny-size) problem
Suppose N = 50 mother vines
n = 100 budsticks from each vine

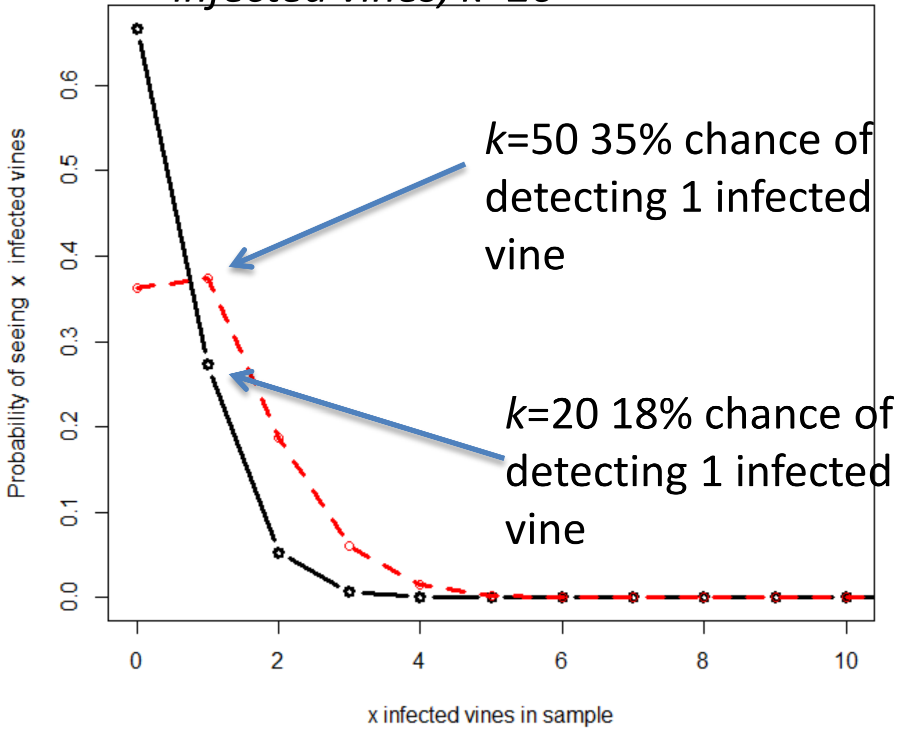
Suppose d = 1 infected mother vine = n*d = 100 infected daughter vines in n*N = 5000

We sample k = 20 vines off the truck using a SRS and send for testing. What is the probability we find x = 0,1, … k infected vines in the sample?

Hypergeometric distribution

$$\Pr(X = x) = \frac{\binom{n\cdot d}{x}\binom{n\cdot N - n\cdot d}{k - x}}{\binom{n\cdot N}{k}}$$

>65% chance of detecting no infected vines, k=20

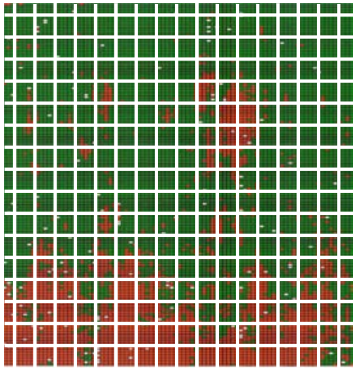


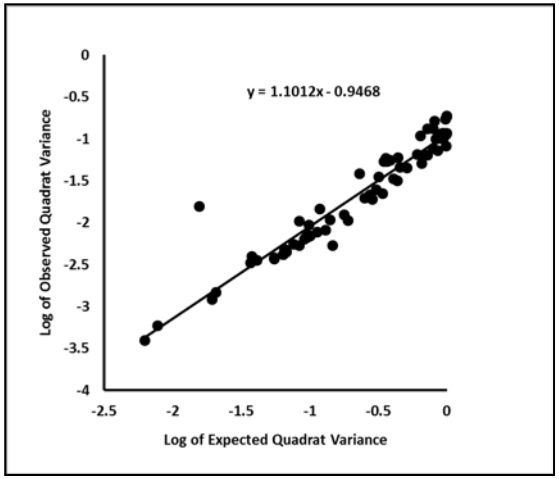
k=50 35% chance of detecting 1 infected vine

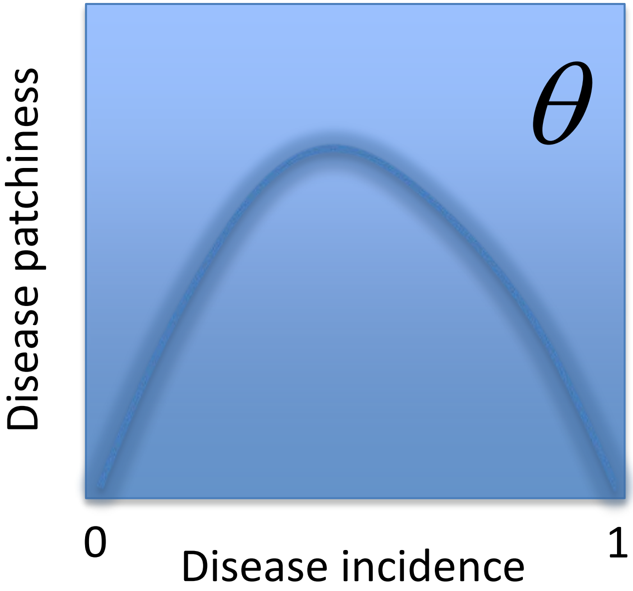
k=20 18% chance of detecting 1 infected vine

University *of* California
Agriculture and Natural Resources

Block sampling for disease incidence




y = 1.1012x - 0.9468
Log of Observed Quadrat Variance
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-2.5
-2
-1.5
-1
-0.5
0
Log of Expected Quadrat Variance



If you don’t find it, is it really not there?

$$\Pr(X = 0) = (1 + n\theta)^{-N\frac{p}{\theta}}$$

Probability of not detecting disease if true vine incidence is p, group size is n and N groups of tests are made

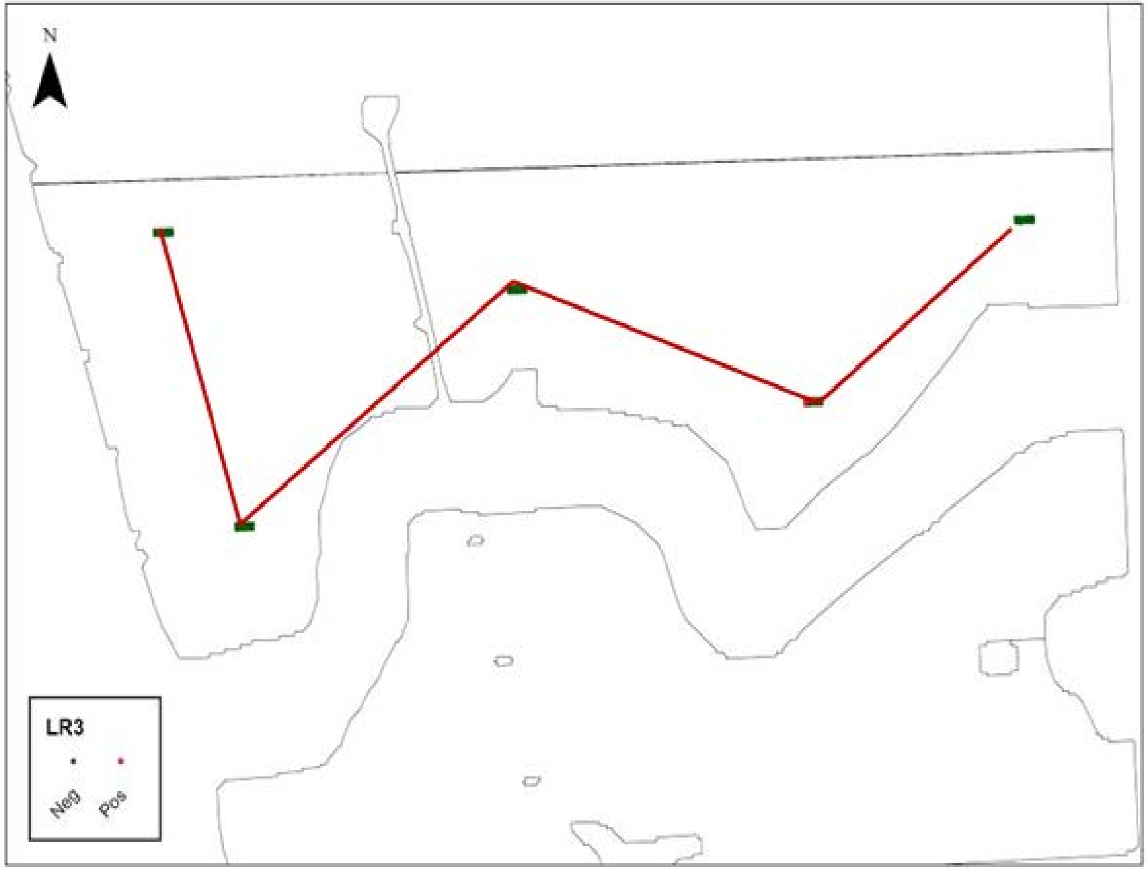
$$p = -\theta \cdot log(P)/N \ \cdot \log(1 + n\theta)$$

Maximum true vine disease incidence that could result in zero positives, given group size n, N groups, with probability P.

$$N = -\theta \cdot log(P)/p \ \cdot \log(1 + n\theta)$$

Sample size required to generate zero positives, given group size n and true disease incidence p, with probability P. Larger samples will give one or more positives

Case Study



Grower decided to test using this structure:

- 5 sets (quadrats)
- 10 samples (n=10) in each set

Row XXXXX
Row XXXXX

- Each vine individually tested
- "W" formation throughout field block
  - "X" works too

University of California
Agriculture and Natural Resources

Where are the positives?

GRBaV

15 positive of 50, approx. 15%

5 Quadrats of 10:

Quadrat	# Positive
1	3/10
2	2/10
3	0/10
4	0/10
5	10/10

GLRaV-3

5 positive of 50, approx. 5%

5 Quadrats of 10:

Quadrat	# Positive
1	1/10
2	0/10
3	0/10
4	0/10
5	4/10

GLRaV-3 in the given samples

BINOMIAL

Fit Statistics	
-2 Log Likelihood	17.2
AIC (smaller is better)	19.2
AICC (smaller is better)	20.5
BIC (smaller is better)	18.8

BETA-BINOMIAL

Fit Statistics	
-2 Log Likelihood	13.3
AIC (smaller is better)	17.3
AICC (smaller is better)	23.3
BIC (smaller is better)	16.5

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
p	0.1	0.04243	5	2.36	0.065	0.05	-0.00906	0.2091

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
p	0.09716	0.07138	5	1.36	0.2316	0.05	-0.08632	0.2806
alpha	0.3491	0.4129	5	0.85	0.4364	0.05	-0.7123	1.4105
beta	3.2439	4.3086	5	0.75	0.4854	0.05	-7.8316	14.3194
rho (intraclass corr.)	0.2177	0.2201	5	0.99	0.3681	0.05	-0.3482	0.7836

GRBaV in the given samples

BINOMIAL

Fit Statistics	
-2 Log Likelihood	43.9
AIC (smaller is better)	45.9
AICC (smaller is better)	47.2
BIC (smaller is better)	45.5

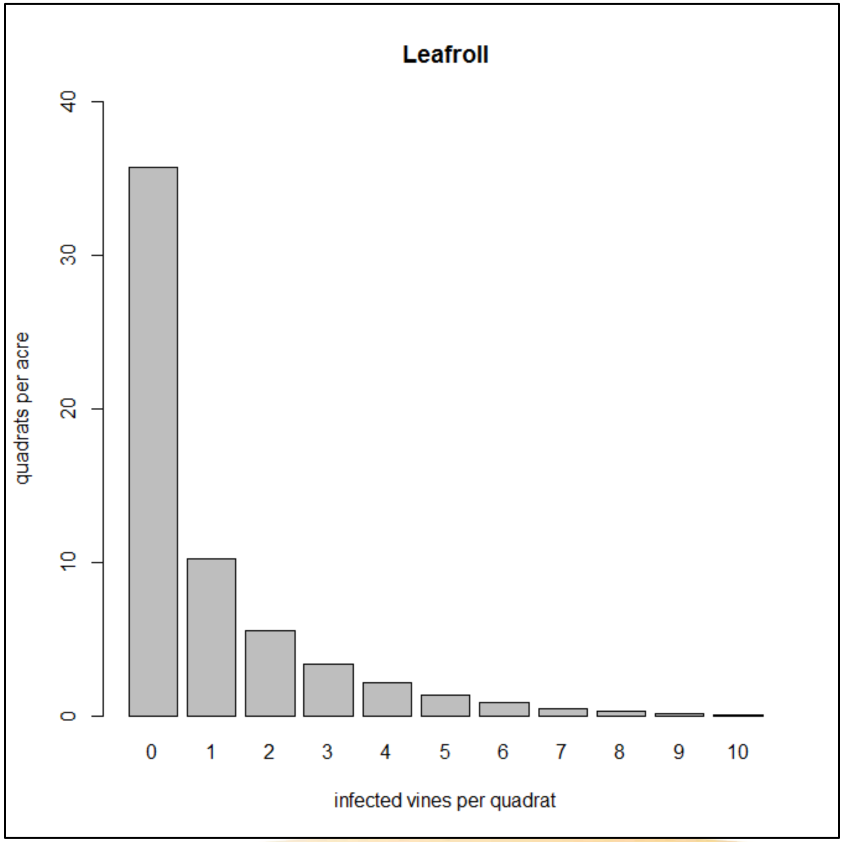
BETA-BINOMIAL

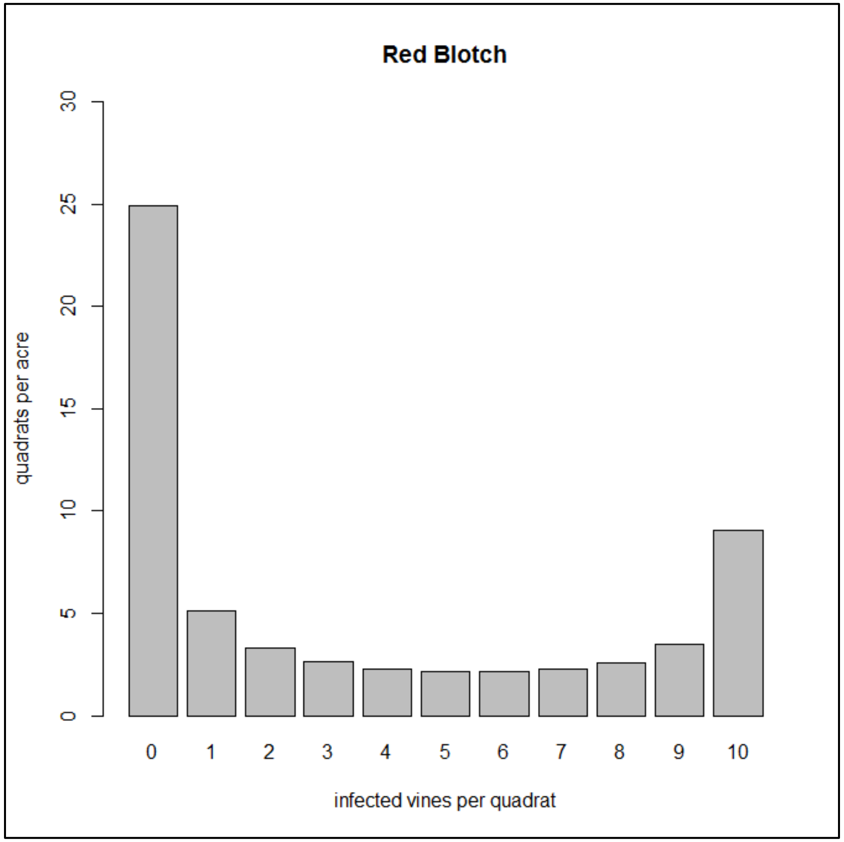
Fit Statistics	
-2 Log Likelihood	19.4
AIC (smaller is better)	23.4
AICC (smaller is better)	29.4
BIC (smaller is better)	22.6

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
p	0.3	0.06481	5	4.63	0.0057	0.05	0.1334	0.4666

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
p	0.3519	0.1738	5	2.02	0.0988	0.05	-0.09483	0.7986
alpha	0.1928	0.1709	5	1.13	0.3105	0.05	-0.2465	0.6321
beta	0.3551	0.3511	5	1.01	0.3582	0.05	-0.5474	1.2576
rho (intraclass corr.)	0.646	0.2017	5	3.2	0.0239	0.05	0.1277	1.1644

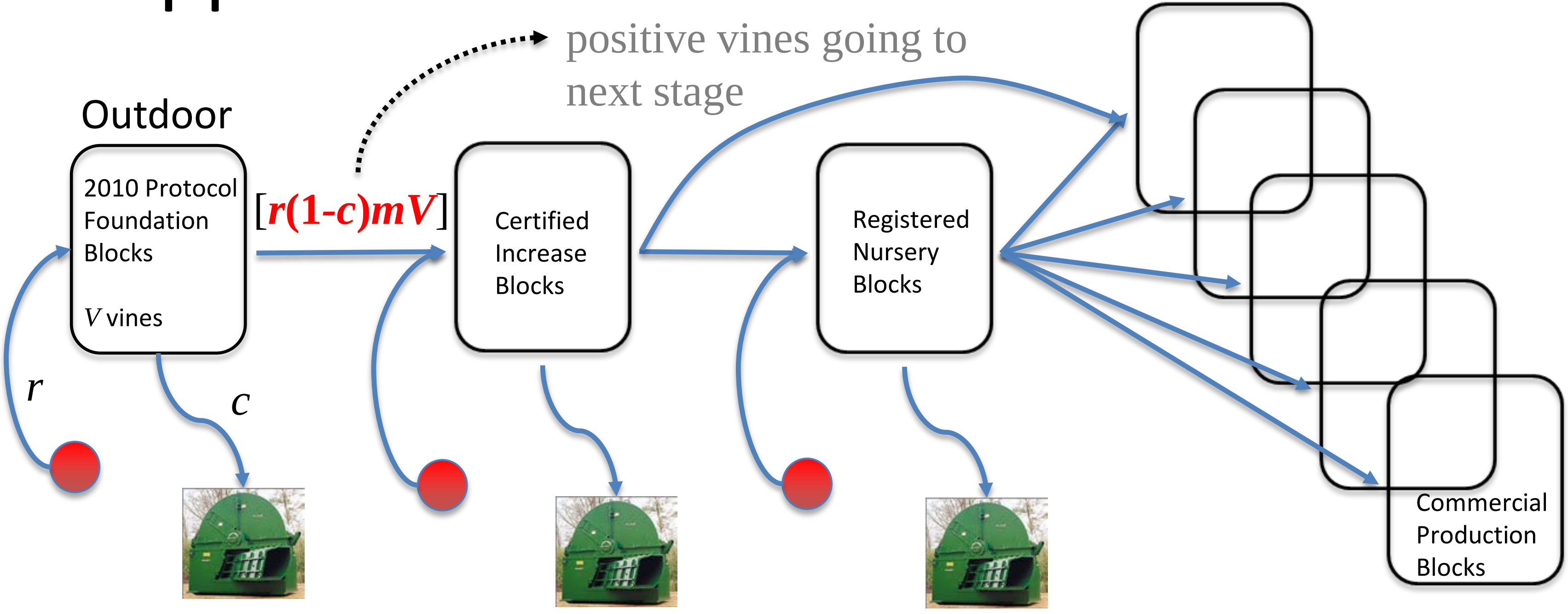
Potential Distribution





University *of* California
Agriculture and Natural Resources

The certification discussion and the future: realistic expectations are the key to happiness



Number of undetected positive vines going to next stage

Outdoor

2010 Protocol Foundation Blocks

V vines

$[r(1-c)mV]$

Certified Increase Blocks

Registered Nursery Blocks

Commercial Production Blocks

r

c

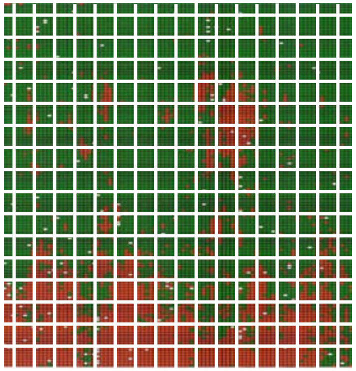
$c = d \times tpp$

d = probability of detection (sampling) = $f(n,N,p,\theta)$

tpp = diagnostic true positive proportion

r: background contamination rate

Sampling depends on spatial scale relationships



Assume composite samples of n vines each
In a simple world where disease has a random pattern

$$p_c = 1 - (1 - p_v)^n$$

The proportion of composites with at least one positive test as a function of the proportion of infected vines

$$p_v = 1 - (1 - p_c)^{\left(\frac{1}{n}\right)}$$

The proportion of infected vines as a function of the proportion of infected composites

$$\tilde{v}_v = \frac{p_c(1 - p_c)^{\left(\frac{2-n}{n}\right)}}{n^2}$$

Approximate variance in vine disease incidence based on composite disease incidence

What about composites?

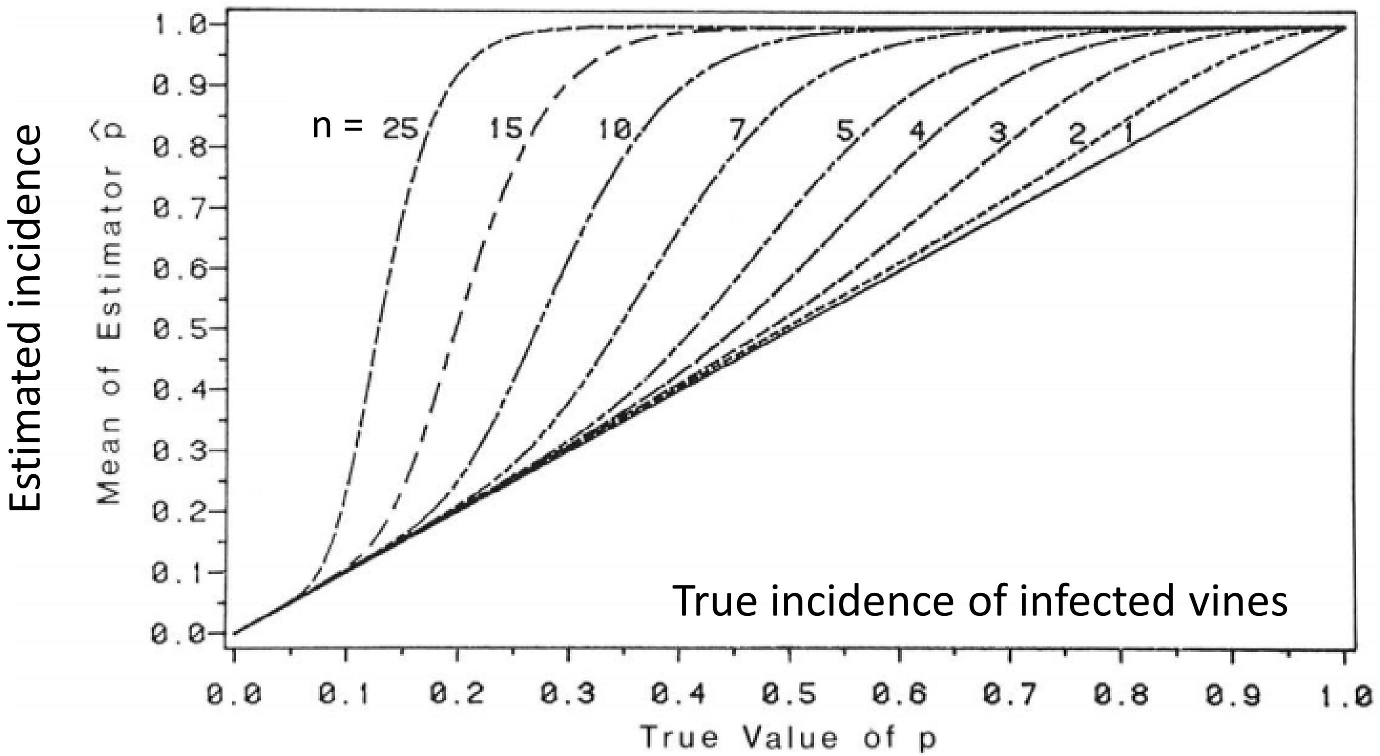


Fig. 1. Expected value (mean) of the maximum likelihood estimator ($\hat{p}$) of the infection rate or probability (p) of disease transmission by a single vector versus the true value of p for tests employing k = 1 to 25 vectors per test plant with N = 25 test plants.

Swallow (1985)
Phytopath. 75

Sample size calculation for composite sampling (***at low incidence***)

Guess of likely vine disease incidence

N = number of composites needed = sample size

$$N = \frac{(1-p_v)^2((1-p_v)^{-n}-1)}{n^2}\left(\frac{z_{\alpha/2}}{h}\right)^2$$

n = composite size (number of vines per group)

Desired confidence interval (precision)

For n ≤10 estimated vine incidence is not overly biased provided:
Assumption of “randomness” is met
True disease incidence is 40% or less

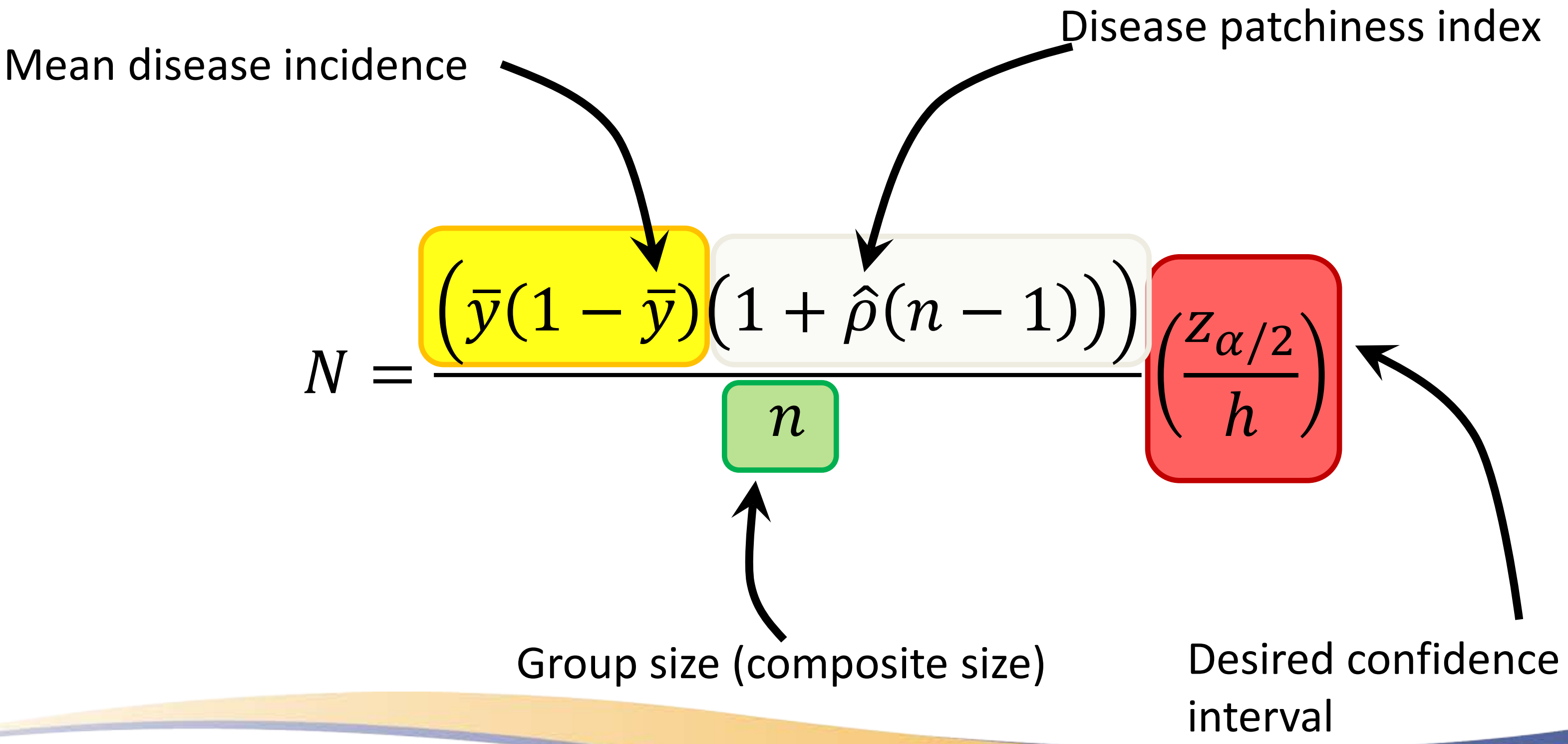
Certification and sampling: take home

Certification is based on sampling ***not a census***

Sampling is ***not perfect***

Sampling according to a known statistical model provides ***long-run known results***

The long-run known results are what certification "means"



$$N = \frac{\left(\bar{y}(1-\bar{y})\right)\left(1+\hat{\rho}(n-1)\right)}{n}\left(\frac{z_{\alpha/2}}{h}\right)$$

Mean disease incidence

Disease patchiness index

Group size (composite size)

Desired confidence interval

University *of* California

Agriculture and Natural Resources

Thank you:

AVF, CGRIC, IAB, CDFA

Questions?

University *of* California
Agriculture and Natural Resources