Statistical Notes on

- **methods for sampling of potato fields for diseases and/or off-types;**
- **sample sizes required for achieving confidence that the true % of disease or off-types is less than a specified percentage**

23 March 2015

Notes prepared by Dave Saville, Principal Biometrician,
Saville Statistical Consulting Limited, Box 69192, Lincoln 7640, New Zealand
(Email: savillestat@gmail.com; phone: 64-3-345 5799)

based upon information supplied by Stephen Ogden & Champak Mehta,
New Zealand Seed Potato Certification Authority
Potatoes New Zealand Inc.
P. O. Box 10232, Wellington, New Zealand

Method of sampling a potato field

In the "UNECE Guide to Seed Potato Field Inspection" document dated 5 August 2014, Figures 1 and 2 on pages 10 and 11 show possible sampling patterns. I greatly prefer the method shown in Figure 1 to that shown in Figure 2.

The method could be implemented as follows. For example, suppose that a potato field has 200 rows. Then it could be divided conceptually into 10 sections of 20 rows. Within each section there are then 10 pairs of rows. The first pair of rows could be walked (inspecting the two plants on either side, i.e., both rows), for about one tenth of its length, then the inspector could move, in a perpendicular fashion, to the next pair of rows and inspect about one tenth of its length, and so on. By the end of the 10^{th} pair of rows, one tenth of the section of field would have been sampled, regardless of how accurately the inspector guessed "about one tenth of the length" of the rows. After repeating this procedure for the other 9 sections of field, one tenth of the plants in the field would have been sampled.

The number of plants inspected can then be calculated by dividing the number of plants in the field by ten.

If less intensive sampling is desired, the 200-row field could be divided conceptually into 5 sections of 40 rows, for example. This would mean 20 pairs of rows per section, with the inspector sampling about 1/20 of each row before moving to the next pair of rows.

The random component can be incorporated by randomly deciding upon a starting point by drawing a random number between 1 and 10 to decide whether the sampling starts in the 1^{st} pair of rows, 2^{nd} pair (up to 10^{th} pair). This effectively staggers, or shifts the pattern of sampling so there are 10 possible patterns.

The above method has the advantage that some sampling is carried out within each row of potatoes. Hence if virus has been spread down the entire length of a particular row, this should be detected.

The method can easily be adapted to differing numbers of rows in the field. For example, if there are 160 rows, then it could be divided conceptually into 8 sections of 20 rows, or into 10 sections of 16 rows.

Officially, no account should be taken of disease observed in non-sampled plants. However, if the tolerance is 0% for a particular disease, then any sighting of this disease would mean the field does not meet the tolerance, so such sightings need to be noted. To provide for departure from the sampling pattern to check out such sightings, the inspector could carry an electric fence standard so they can mark where they had got to in their sampling, so that they can then return to this spot and carry on sampling.

Statistical proof that field meets the specified tolerance

If a tolerance of, for example, 0.1% is specified for the disease level in a field, then this is 1 diseased plant out of every 1,000 plants in the field. And if 1,000 plants are independently sampled and the true disease level is 0.1%, the probability of finding disease in a plant follows a binomial distribution with parameter 0.001 (1% chance of disease). In this situation, there is a 37% chance that 0 diseased plants will be found in the sample of 1,000 plants, a 37% chance that 1 diseased plant will be found, and a 26% chance that 2 or more diseased plants will be found in the sample.

Similarly, if the true disease % is slightly greater than 0.1%, there is still a high chance of observing no disease in a sample of 1,000 plants. Therefore sampling 1,000 plants and finding no disease does NOT constitute proof that the true disease level in the field is 0.1% or less.

In the next section we show that for 95% confidence that the true disease level in the field is 0.1% or less, we must sample a minimum of 3,000 plants, and observe no disease.

Minimum sample size required for statistical proof

Derivation of formula

If the true % disease is exactly 0.1%, for example, then the probability that a single randomly selected plant is NOT diseased is 0.999. Hence the probability that "x" randomly, independently selected plants are ALL not diseased is 0.999^x (0.999 to the power of x).

Now we want to find the value of "x" for which 0.999^x is equal to 0.05.

The reason is that for sample sizes larger than x, we would then have a probability of $p<0.05$ of observing NO disease in all sampled plants (if the true % disease is exactly 0.1%), so such a result is statistically inconsistent with the idea that "the true % disease is exactly 0.1%" (or larger). That is, we have statistical proof that "the true % disease is less than 0.1%". In other wording, we are "95% confident" that "the true % disease is less than 0.1%".

To solve the equation $0.999^x = 0.05$,
we take logs (to any base) of both sides, yielding $\log[0.999^x] = \log[0.05]$
which can be simplified to $x \log[0.999] = \log[0.05]$
and hence $x = \log[0.05] / \log[0.999]$
or, **$x = \log[1 - 0.95] / \log[1 - 0.001]$**

where 0.95 corresponds to 95% confidence and 0.001 is the maximum allowable proportion of diseased plants (or off-type plants).

The answer, x, tells us the minimum sample size required for proving, with 95% certainty, that the true percentage of diseased plants is less than (or equal to) 0.1%. However, such proof exists only if NO diseased plants are found, out of the "x" plants sampled. If even one diseased plant is found, with a sample size of "x", then there is no statistical proof that the true percentage of diseased plants is less than (or equal to) 0.1%.

Clearly other confidence levels (such as 90% or 99%) can be substituted into the formula, and different tolerances can be substituted for 0.1% tolerance. When this is done, the results are as shown in Table 1 (which gives unrounded "x" values) or Table 2 (which gives rounded "x" values).

Specified maximum level of disease	Minimum sample size (along with NO disease in sampled plants) required for statistical proof that the true level of disease is less than the specified maximum, at confidence level:		
	90%	95%	99%
0%	A census (100% sample) of all plants is required for proving this is the case.		
0.01%	23,025	29,956	46,049
0.1%	2,301	2,994	4,603
0.2%	1,150	1,496	2,300
0.25%	920	1,197	1,840
0.5%	459	598	919
0.8%	287	373	573
1%	229	298	458
1.5%	152	198	305
2%	114	148	228
6%	37	48	74

Table 1: Minimum sample size (along with NO disease in sampled plants) required for statistical proof that the true level of disease is less than the specified maximum, at confidence levels of 90%, 95% and 99%.

Specified maximum level of disease	Minimum sample size (along with NO disease in sampled plants) required for statistical proof that the true level of disease is less than the specified maximum, at confidence level:		
	90%	95%	99%
0%	A census (100% sample) of all plants is required for proving this is the case.		
0.01%	23,100	30,000	46,100
0.1%	2,310	3,000	4,610
0.2%	1,150	1,500	2,300
0.25%	920	1,200	1,840
0.5%	460	600	920
0.8%	290	380	580
1%	230	300	460
1.5%	160	200	310
2%	120	150	230
6%	40	50	75

Table 2: *Rounded* minimum sample size (along with NO disease in sampled plants) required for statistical proof that the true level of disease is less than the specified maximum, at confidence levels of 90%, 95% and 99%.

"Larger than minimum" sample sizes

For 95% confidence that the true disease % in a field is ≤ 1%, for example, and for a sample of 300 independently sampled plants, there must be NO disease observed in the sample (Table 2).

However, if a larger sample is taken, then the allowable number of diseased plants may be greater than zero. For example, if the sample size is 1,000, then for 95% confidence that the true disease % in a field is ≤ 1%, there can be up to 4 diseased plants observed (or ≤ 0.4%).

As a second example, if the sample size is 10,000, then for 95% confidence that the true disease % in a field is ≤ 1%, there can be up to 83 diseased plants observed (or ≤ 0.83%).

Survey or census?

The above calculations assume the field is reasonably large; in such cases the statistical accuracy depends on the *number of plants sampled*, not the percentage of plants sampled. However, if a field has only 450 plants, for example, then the inspector may well inspect all plants, in which case results are exact, and not subject to statistical variation. This is called a census.

Lack of independence

In the above sample size calculations, independence of sampling is a basic assumption. If an inspector is observing disease in consecutive plants along two adjacent rows, this assumption is almost certainly violated. This means that the above sample sizes are likely to be gross under-estimates. Therefore higher levels are sampling are desirable.