Cluster Sampling to Identify Risk

Ned Jones
1-alpha Solutions
Wake Forest, NC 27587

Introduction

Sample units do not always lend themselves to being sampled one at a time but rather must be sampled in clusters. The need to sample clusters maybe the result of logistics, economics or convenience or any combination of these factors. In this situation identifying the risks presented by the commodity or other item requires a cluster sampling approach. Sampling clusters to detect risk can be confusing but the JMP Profiler provides a clear way forward and an easily understood picture.

Application

The application presented involves sampling inbound plants from a foreign origin for invasive pest species. The plants usually ship in boxes or bags. As the plants are put in the boxes and move in shipment they become inter-twined. This makes sampling individual plants difficult. Each plant species poses a threat from associated pest(s) based on country of origin. The plant/country risks are developed from scientific based risk analysis. A unique pest detection level is specified based on the risk presented. The question then becomes how many boxes should be inspected to assure the pest infestation is below a safe level. Or in other terms how large of a sample is required to ensure the shipment is safe.

Probability of an infested box

The consignment invoice provides the number of boxes, the total number of plants, genus species, and country of origin for a given shipment. From this information we can assign risk (low, medium or high) and P, detection values (10%, 5% or 1%, respectively). We calculate M, the number of plants per box. Given the assigned P we can calculate ,$\Pr(a_M > 0)$, the probability of $a_M > 0$, one or more pests in a box. The result is as follows:

$$Pr\ (a_M > 0) = 1 - (1 - P)^M \qquad (1)$$

This gives the probability of an infested box $Pr\ (a_M > 0)$ as a function of P the desired detection level and M the number of plants in the box. Note equation (1) is one minus the probability that $a_M = 0$. When the probability $a_M = 0$ is subtracted from 1 we get the probability sum of all possible positive $a_M,\ a_M > 0$. This provides an exact result.

Number of boxes to sample

The hypergeometric distribution is used to estimate the number of boxes to sample . This distribution requires four parameters, N, the number of boxes in the consignment, A, the number of boxes infested, n, the sample size and a_n, the number of infested boxes in the sample. The JMP hypergeometric function is coded as follows:

$$Hypergeometric\ Distribution\ (N, A, n, a_n) \qquad (2)$$

The parameters are fairly straight forward but we must estimate A as an equation (1) result multiplied by N and multiplied by the inspection sensitivity, $A = Se \cdot N \cdot Pr\ (a_M > 0)$. Using this estimate for A equation (2) provides a function for hypergeometric probability based on P the desired detection.

References

Cochran, W. G. (1977). Sampling Techniques, WILEY.

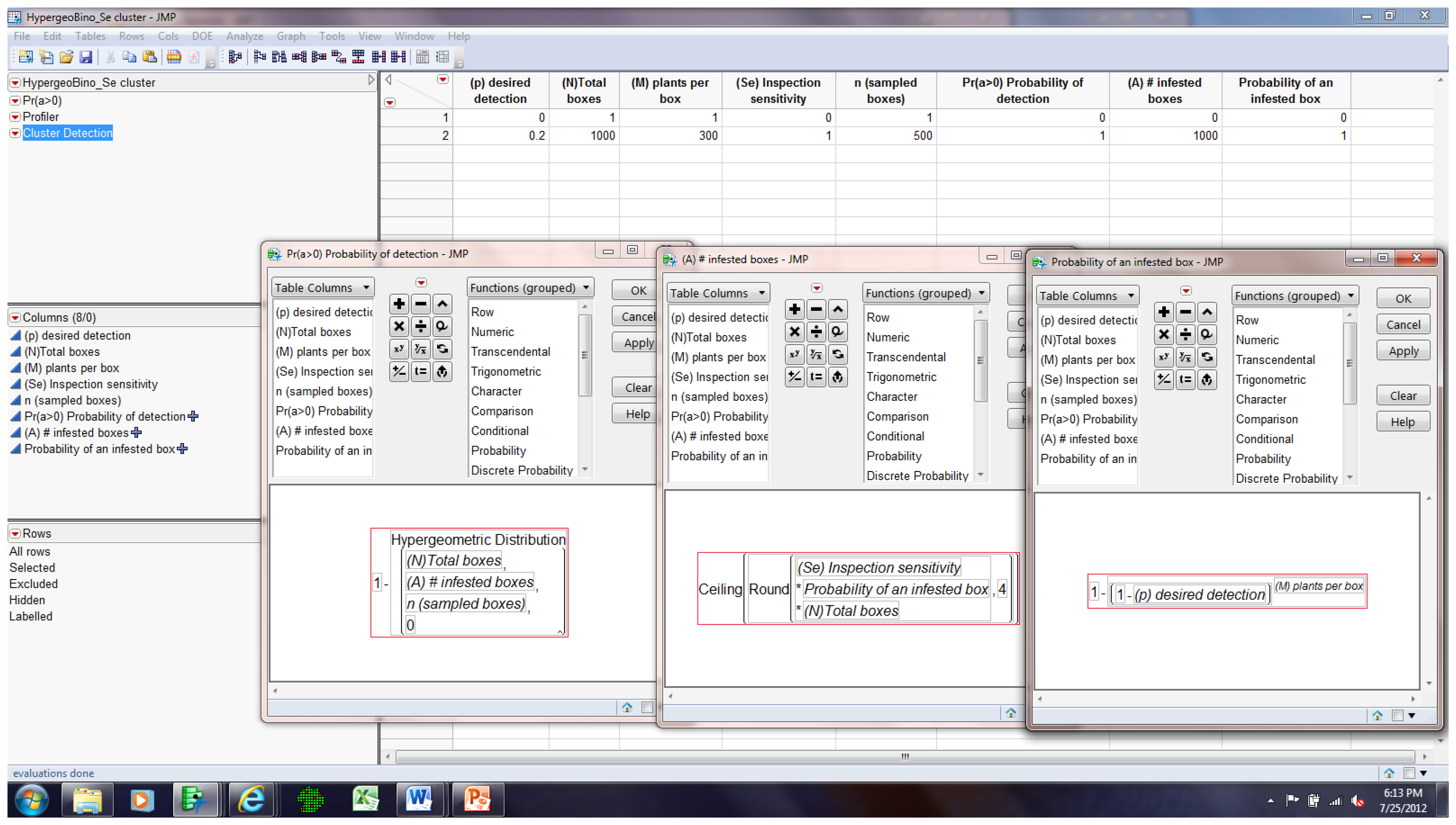
Gould, W. P. (1995). "PROBABILITY OF DETECTING CARIBBEAN FRUIT FLY (DIPTERA: TEPHRITIDAE) INFESTATIONS BY FRUIT DISSECTION." Florida Entomologist **78**(3): 502-507.

SAS_Institute_Inc. (2011). JMP.

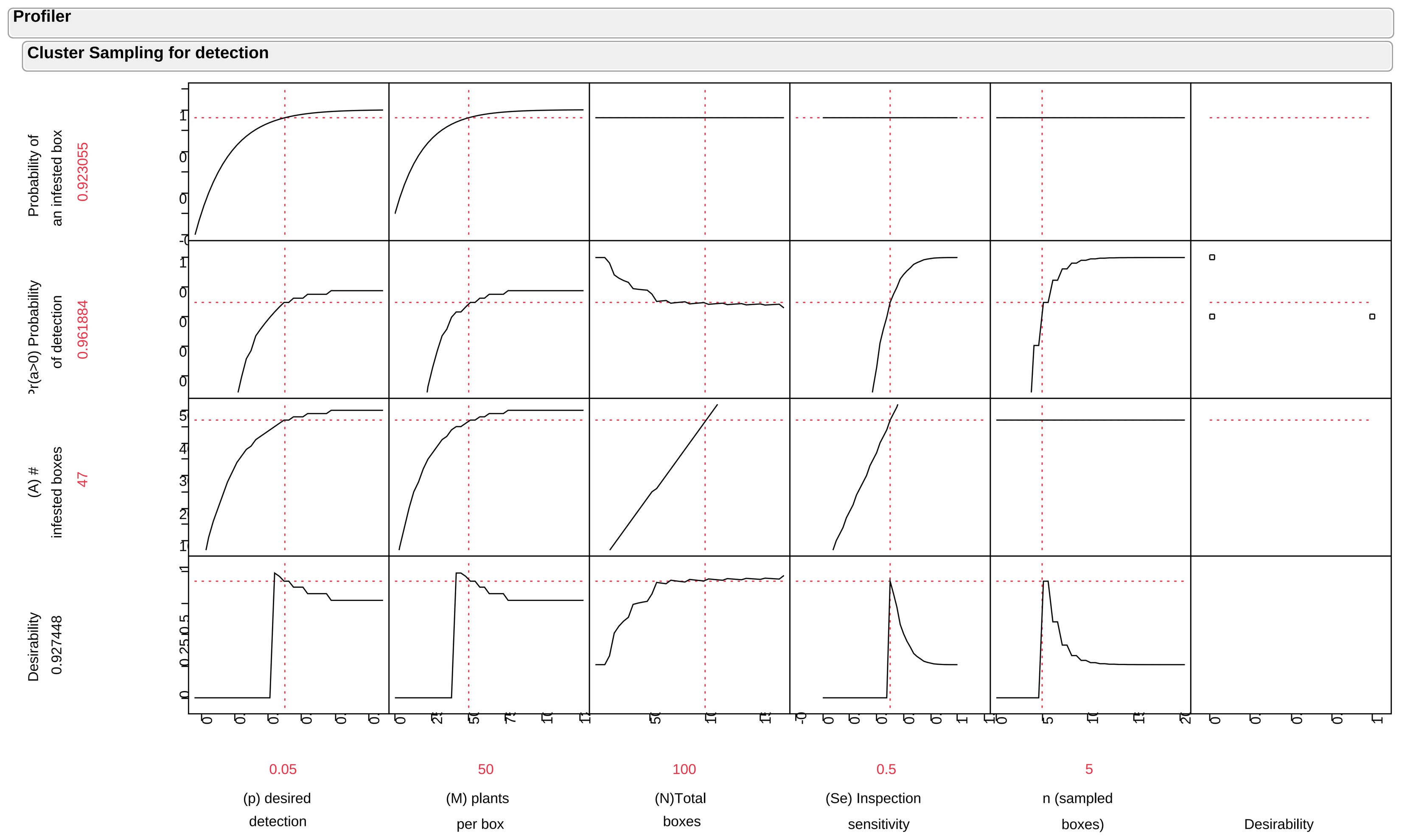
The Se is included to allow for inspection sensitivity which is the measure of how often an inspector identifies an infestation when it is actually present. When $a_n = 0$, the equation (2) provides the probability of finding no infested boxes in the sample. If we subtract this result from 1 the new result gives the probability of the cluster sample of n boxes detecting an infested box. The JMP formula is as follows:

$$1 - Hypergeometric\ Distribution\ (N, ceiling[Se \cdot 1 - (1 - P)^M \cdot N], n, 0) \quad (3)$$

Using an iterative approach the sample size can be adjusted until the probability of an infest box is 0.95 or greater; however, if this design is entered into JMP as follows: (see the image variables and equations)



Next select the 'Profiler' from under the 'Graph' pull-down menu and enter the three equation variables (columns) and click 'OK'. Then select 'Desirability Function' from under the second red triangle. Then double-click on the graph window to the far right for '$Pr(a > 0)$ *Probability of detection*'. Select 'match target' and high=1, middle=0.95 and low 0.9499. Double-click and select none for the other two far right graph windows. Then set the inputs P, N, M *and* Se as needed and adjust 'n sampled boxes' so that '≥ 0.95 an example result is as follows:



Discussion

- The sample design relies on the binomial and hypergeometric distributions.
- The desired detection level and the plants per box are applied in the binomial distribution. The result is the probability of an infested box, i.e. one or more infested plants in the box.
- The total number of boxes, probability of an infested box and inspection sensitivity are multiplied to estimate A, the number of infested box detections expected in the population. How, A is rounded has a direct effect on the actual detection level. Rounding down is the most conservative approach.
- The probability of detection uses the hypergeometric distribution. All the inputs affect this result. In a given sampling situation P, N, M *and* Se are fixed inputs. We adjust the number of sampled boxes to ensure a 0.95 (95%) or larger probability of detecting an infested box.
- The number of boxes sampled effects the probability of detection and nothing else (note the flat lines in the graph above sampled boxes).
- Combining these relationships in JMP and applying the JMP Profiler provides a clear picture of how the inputs affect the results. When the desirability function is applied to target 0.95 probability of detection (confidence) an easily understandable picture of the overall sample design is presented.