A Study on Sampling for Estimating Tobacco Disease Incidences

Hong Nai Park*

Introduction

For crops that are planted in a lattice layout, sampling designs can be made to take advantage of this regular arrangement. In order to select which tobacco plants to be examined in a survey to estimate disease loss in tobacco a method of, so called, bent plots was devised based on the regularity of plantings in the tobacco fields. We will first describe this sample selection and measurement method and then provide estimators and their bias and variance properties.

The Method of Bent Plots

A tobacco field is commonly composed of a number of parallel row-groups each consisting of four rows of evenly spaced plants. Between row-groups is an unplanted sled row. The four plants in a line roughly perpendicular to the row-group will be called a plant group. Sometimes there is a boundary row-group of only two plants and if so, this row-group will be deleted from the survey. Similarily next to the two boundaries that are roughly perpendicular to the row-groups one finds plants in incomplete plant-groups and these will be deleted also. Incomplete plant groups in the field interior are not deleted. The first step in applying the method is to sketch the outlines of

*Department of Computer Science and Statistics, Seoul National University

the field and show each row-group as a numbered line in the field along with rough counts of plant-groups in the rows. If the field is perfectly rectangular, then only one or two row-groups need be counted but more counting may be required if the row-groups differ in length. However, rough counts are all that is needed so at most four row-groups may need to be counted in the field.

A number say, H, of cross cutting subdivisions or strata are next made on the sketch map using sled rows and numbered plant-groups as boundaries. We have used $H=4$, a two-by-two design, and also $H=9$, a three-by-three subdivision. Of course, $H=1$ is available for free. The plant-groups in each stratum are counted and put into serpentine order by establishing an ordering direction along each row-group that is shown by arrows on the sketch map. A random number in each stratum is used to select one particular plant-group to serve as maker plant-group for the selected plot.

The plot itself consists of the next LK plant-groups along the serpentine ordering. Thus the plot may be in more than one row-group as the ordering reaches a boundary and proceeds back along the adjacent row-group. This create, so called, bent plots. By convention the last plant-group is joined in order to the first. The two integers L and K along with H are the design parameters that completely determine the selection and measurement procedures.

The enumerator locates the maker plant-group in the field and finds the direction of the ordering. He counts K plant-groups along the ordering direction and at the Kth plant-group he makes measurements on these four plants individually and then scans them. He goes K more, measures the four plants of the $2K$th plant group, and and K more, etc. until arriving at the KLth plant group from the marker plant-group. He then walks back to the

marker plant-group and scans all the plants in the whole plot as he goes. Thus there are two intensities of measurement: examination of individual plants and scan. Several variables are routinely recorded for both levels of measurement.

This is the method as it was used in pilot surveys. However, there are a few elaborations that we will be considering in the theory that might be mentioned here. One may wish to select more than one plot in each stratum and to vary K and L from one to another plot. In particular, one may wish to scan only in some plots and do individual plant measurements on only a subsample of plots. Each K plant groups may be denoted a subplot and some data may be recorded by subplot as a further variation in the method.

Estimates and the Variances

It will be denoted the scan data as X vales and the individual plant data as Y-measurements. The notational scheme for population values is: Y_{fijqr} =Average of our individual plant scores for the rth plant-group of gth subplot in the jth whole plot of the Lth stratum in the fth field.

y_{fijqr}=Average of four individual plant scores for the rth plant-group of the qth subplot in the jth whole plot of the ith stratum in the fth field.

$\bar{y}=\sum_{f}^{N}\sum_{i}^{H}\sum_{j}^{M}\sum_{q}^{L}\sum_{r}^{K} y_{fijqr}/NHMLK$: population mean.

In estimating population mean, a regression estimate in double sampling is considered. The LK plant-groups within a whole plot is regarded as the first sample and the L plant-groups from the first sample is regarded as the record sample.

A regression estimate of the population mean is defined as

$$\bar{y}_{lr}=\frac{1}{NHMLK}\cdot\frac{N}{n}\sum_{f}\frac{H}{h}\sum_{i}\frac{M}{m}\sum_{j}LK\hat{\bar{y}}_{fij}$$

$$=\frac{1}{nhm}\sum\sum\sum L\hat{\bar{y}}_{fij}$$

where $\hat{\bar{y}}_{fij}$, a regression estimate of a subplot within jth whole plot in the ith stratum of the fth field can be written

$$\hat{\bar{y}}_{fij}=\hat{\bar{y}}_{fji}+b\ (\bar{x}'_{fij}-\bar{x}_{fij})$$

where $\bar{x}'_{fij}$ is based on the first sample, $\bar{x}_{fij}$ is based on the second sample.

The variance of $\bar{y}_{lr}$ can be derived as

$$V(\bar{y}_{er})=\frac{1}{(NHML)^2}\left[N^2\frac{N-n}{N}\frac{S^2_f}{n}+\frac{N^2}{n}H^2\frac{H-h}{H}\frac{\sum S^2_{fi}}{n}\right.$$
$$\left.+\frac{N}{n}\frac{H}{h}M^2\frac{L-l}{L}\frac{\sum\sum S^2_{fij}}{l}+\frac{N}{n}\frac{H}{h}\frac{M}{m}L^2\sum\sum\sum S^2\hat{\bar{y}}_{fi}\right]$$
$$=\left(1-\frac{n}{N}\right)\frac{S_F{}^2}{n}+\left(1-\frac{h}{H}\right)\frac{S^2_S}{nh}+\left(1-\frac{m}{M}\right)\frac{S^2_W}{nhm}$$
$$+\left[\frac{S_E{}^2\rho^2}{nhmlk}+\frac{S_E{}^2(1-\rho^2)}{nhml}\right]$$

Since

$$S^2\hat{\bar{y}}_{fij}=\frac{S_E{}^2\rho^2}{lk}+\frac{S_E{}^2(1-\rho^2)}{l}$$

$$S^2_f=\frac{\sum(y_i-\bar{y})}{N-1}\qquad S^2_{fi}=\frac{\sum(y_{fi}-\bar{y}_f)}{H-1}\text{...etc.}$$

$$S^2_F=\frac{(\bar{\bar{y}}_f-\bar{\bar{y}})}{N-1}\qquad S_S{}^2=\frac{(\bar{\bar{y}}_{fi}-\bar{\bar{y}}_f)^2}{H-1}\text{...etc.}$$

In the formula, S^2_F, S^2_S, S^2_W, S^2_E are the variance component of field, strata, whole plot, plant-group respectively.

If we assume $h=H$, $m=1$, $l=L$, then variance can be reduced M.

$$V(\bar{y}_{er})=\left(1+\frac{n}{N}\right)\frac{S^3_F}{n}+\left(1-\frac{1}{M}\right)\frac{S_W{}^2}{nH}+\frac{S_E\rho^2}{nHLK}\frac{S^2_E(1-\rho^2)}{nHL}$$