STANDARD ERRORS IN FOREST AREA

Joseph McCollum[1]

ABSTRACT.—I trace the development of standard error equations for forest area, beginning with the theory behind double sampling and the variance of a product. The discussion shifts to the particular problem of forest area—at which time the theory becomes relevant. There are subtle difficulties in figuring out which variance of a product equation should be used. The equations developed may be extended to other areas of forest inventory.

What follows is a development of standard error equations. Key topics include double sampling and the variance of a product. Applications to calculation of the variance of forest area follow.

Although the following equations seem very abstract at first glance, they are highly relevant to the development of standard error equations in computation of forest area.

Goodman (1960) pointed out that if $Z = XY$, then

$$\sigma_Z^2 = \mu_X^2\sigma_Y^2 + \mu_Y^2\sigma_X^2 + \sigma_X^2\sigma_Y^2 \qquad (1)$$

for X and Y independent, where

σ_A^2 denotes the variance for the subscripted variable A,

μ_A denotes the true mean for the subscripted variable A, and

A denotes X, Y, or Z, as appropriate.

Goodman (1962) showed that if the means and variances are not known, then an unbiased estimator is

$$s_Z^2 = \overline{X}^2 s_Y^2 + \overline{Y}^2 s_X^2 + \frac{n_X n_Y - n_X - n_Y}{n_X n_Y} \cdot s_X^2 s_Y^2 \qquad (2)$$

for X and Y independent, where

$\overline{A}$ denotes the sample mean for variable A (also represented by m_A),

s_A^2 denotes the sample variance for variable A,

n_A denotes the number of observations for variable A, and

A denotes X, Y, or Z, as appropriate.

If we let $X = \overline{X}$ and $Y = \overline{Y}$ then,

$$s_{\overline{Z}}^2 = \overline{X}^2 s_{\overline{Y}}^2 + \overline{Y}^2 s_{\overline{X}}^2 + \frac{n_X n_Y - n_X - n_Y}{n_X n_Y} \cdot s_{\overline{X}}^2 s_{\overline{Y}}^2 \qquad (3)$$

If $n_X = n_Y = n$, then

$$s_Z^2 = \overline{X}^2 s_Y^2 + \overline{Y}^2 s_X^2 + \frac{n-2}{n} \cdot s_X^2 s_Y^2 \qquad (4)$$

Let us see how these equations work with real data. Suppose you have independent Bernoulli variables that take on the value 1 with probability 0.5 and zero otherwise. You have decided to take two samples of each variable. The experiment may give results in 16 possible ways, shown in table 1.

On the other hand, suppose that you were not interested in $Z = XY$, but rather in $W = \overline{X} \cdot \overline{Y}$.

If $W = \overline{X} \cdot \overline{Y}$, then $n_X = n_Y = 1$, and

$$s_W^2 = \overline{X}^2 s_{\overline{Y}}^2 + \overline{Y}^2 s_{\overline{X}}^2 - \cdot s_{\overline{X}}^2 s_{\overline{Y}}^2 \qquad (5)$$

where

$s_{\overline{A}}^2$ denotes the standard error of the mean for variable A (also represented by $s^2(m_A)$).

n_A denotes the number of observations used to form the sample of W (which is one observation), and

k_A denotes the number of observations used to form the sample of variable A.

The k_A does not appear in the equation directly. Again, the experiment may give results in 16 possible ways, shown in table 2.

In table 2, $k_X = 2$ and $k_Y = 2$. However, $k_W = n = 1$. At first, it seems strange that an experiment with one observation could have a variance at all. However, suppose we think of the 16 possible results of the experiment as 16 possible worlds. Equation (5) gives the variance of $W = \overline{X} \cdot \overline{Y}$ among those 16 possible worlds.

[1] Computer Specialist, USDA Forest Service, Southern Research Station, 4700 Old Kingston Pike, Knoxville, TN 37919.

Table 1.—*Bernoulli variables demonstrating Goodman's 1962 equations*

	X_1	X_2	Y_1	Y_2	Z_1	Z_2	$s^2(Z)$	$s^2(mz)$	Eq. (2)	Eq. (3)
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0
4	0	0	1	1	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0	0
6	0	1	0	1	0	1	0.5	0.25	0.25	0.125
7	0	1	1	0	0	0	0	0	0.25	0.125
8	0	1	1	1	0	1	0.5	0.25	0.5	0.25
9	1	0	0	0	0	0	0	0	0	0
10	1	0	0	1	0	0	0	0	0.25	0.125
11	1	0	1	0	1	0	0.5	0.25	0.25	0.125
12	1	0	1	1	1	0	0.5	0.25	0.5	0.25
13	1	1	0	0	0	0	0	0	0	0
14	1	1	0	1	0	1	0.5	0.25	0.5	0.25
15	1	1	1	0	1	0	0.5	0.25	0.5	0.25
16	1	1	1	1	1	1	0	0	0	0
Mean	0.5	0.5	0.5	0.5	0.25	0.25	0.1875	0.09375	0.1875	0.09375

Table 2.—*Bernoulli variables demonstrating Goodman's 1960 equations*

	X_1	X_2	Y_1	Y_2	m_X	m_Y	$s^2(m_X$	$s^2(m_Y)$	Eq. (5)	$Z=m_Xm_Y$
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0.5	0	0.5	0	0
3	0	0	1	0	0	0.5	0	0.5	0	0
4	0	0	1	1	0	1	0	0	0	0
5	0	1	0	0	0.5	0	0.5	0	0	0
6	0	1	0	1	0.5	0.5	0.5	0.5	0.0625	0.25
7	0	1	1	0	0.5	0.5	0.5	0.5	0.0625	0.25
8	0	1	1	1	0.5	1	0.5	0	0.25	0.5
9	1	0	0	0	0.5	0	0.5	0	0	0
10	1	0	0	1	0.5	0.5	0.5	0.5	0.0625	0.25
11	1	0	1	0	0.5	0.5	0.5	0.5	0.0625	0.25
12	1	0	1	1	0.5	1	0.5	0	0.25	0.5
13	1	1	0	0	1	0	0	0	0	0
14	1	1	0	1	1	0.5	0	0.5	0.25	0.5
15	1	1	1	0	1	0.5	0	0.5	0.25	0.5
16	1	1	1	1	1	1	0	0	0	1
Mean	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.25	0.078125	0.25
Var	0.25	0.25	0.25	0.25	0.125	0.125				0.078125

In that same table, we see that the variance of $W = \overline{X} \cdot \overline{Y} = m_X \cdot m_Y$ is on average equal to Goodman's estimator of 1960. Goodman showed this result in 1960; he pointed out the irony of subtracting the third term for the estimator but adding it in the theoretical case. Also, in this table, I would like to point out that if instead of 0,1 variables there were –1,+1 variables, Goodman's estimator of 1960 would produce a negative estimate of the variance—an impossibility—in 4 out of 16 cases. Nonetheless, the estimator is still unbiased. This result may make Goodman's estimator unusable to some analysts. There are at least several alternatives; one is to set the estimate equal to zero if it is negative; however, the estimator would no longer be unbiased. Another possibility would be to take a sample size so large that the size of the third term would be insignificant. One last option would be to realize that the variable of interest has a high coefficient of variation.

At this point, the reader might wonder when Goodman's equations can be put to practical use. Two examples follow. The first example illustrates the use of Goodman's equation of 1962. It involves the combination of two data sets that have been partially matched; Goodman's equation of 1962 requires paired data. It involves the calculation of an observation times an observation. The larger data set is collected by photo-interpreters. The smaller data set is a subset of the larger one, which is collected by field crew. The assumption is that the photointerpreters' accuracy rate for the points that the field crew does not check is the same as their accuracy rate for the points that the field crew does check.

The second example is an illustration of Goodman's equation of 1960. It involves the calculation of the variance of an unknown proportion—it is an average times an average. The mean and variance of each component is known, and thus a variance of the product may be computed.

APPLICATION TO FOREST AREA

Goodman's equations may be used to derive the traditional standard error equations of forest area.

A forest inventory begins with a photointerpretation phase, known as phase 1. Aerial photographs are scored with dots, and the photointerpreters call each dot either "forest" or "nonforest." A phase 1 estimate is the proportion of dots believed to be forested. The photointerpreter is to consider only the immediate neighborhood of the dot, not the entire area nearest a dot.

Among the dots on the aerial photographs are 1) regular plots, which field crews visit, measure the conditions and trees on the plots and 2) intensification plots, which field crews visit, call either "forest" or "nonforest," but make no other measurements.

The phase 2 estimate is the Bayesian expectation of forested land based on the photointerpreter calls and the field crew calls. This result typically has a much lower variance than the phase 1 estimate. Suppose there are n_{ij} plots called class i in phase 1 and class j in phase 2. If there are two classes, there would be $n_{\cdot 1} = n_{11} + n_{21}$ forested plots and $n_{\cdot 2} = n_{12} + n_{22}$ non-forested plots.

Let

$d = \sum_{i=1}^{I} d_i$ denotes the total number of dots,

$P_i = \frac{d_i}{d}$ denotes the phase 1 estimate for class i,

$n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$ denotes the total number of plots in class j,

$p_{ij} = \frac{n_{ij}}{n_{\bullet j}}$ denotes the proportion of plots in class j of phase 2 that were in class i of phase 1.

$P_{ij} = P_i \cdot p_{ji}$ denotes the adjusted proportion of the landscape believed to be in class i of phase 1 and class j of phase 2, and

$P_{\bullet j} = \sum_{i=1}^{I} P_{ij}$ denotes the phase 2 estimate for class j,

where

I is the number of classes.

There is then a tabulation such as the following:

d_1	=	1200	d_2	=	1300
P_1	=	0.48	P_2	=	0.52
n_{11}	=	45	n_{12}	=	3
n_{21}	=	5	n_{22}	=	57
$n_{\bullet 1}$	=	50	$n_{\bullet 2}$	=	60
p_{11}	=	0.90	p_{12}	=	0.05
p_{21}	=	0.10	p_{22}	=	0.95
P_{11}	=	0.4320	P_{21}	=	0.0480
P_{21}	=	0.0260	P_{22}	=	0.4940
$P_{\bullet 1}$	=	0.4580	$P_{\bullet 2}$	=	0.5420

In this case, at the end of phase 1, the proportion of area believed to be forested would be 48 percent. At the end of phase 2, the proportion is adjusted to 45.8 percent.

What is a confidence interval on this estimate?

If the user is interested in a confidence interval on the proportion of forested land of a randomly chosen acre of ground, then what is desired is $Var(P_{\bullet 1})$. However, more commonly, what is desired is a confidence interval on the mean proportion across the landscape, or $Var(\overline{P}_{\bullet 1})$.

$$Var(\overline{P}_{\bullet 1}) = Var(\overline{p_{11}P_1}) + Var(\overline{p_{12}P_2}) + 2Cov(\overline{p_{11}P_1}, \overline{P_{12}P_2}) \quad (6)$$

The factors in each term are observations; therefore Goodman's equation of 1962, or equation (3) in this paper, should be used.

$$Var(\overline{p_{1k}P_k}) \approx P_k^2 \frac{p_{1k}(1-p_{1k})}{n_{\bullet k}} + p_{1k}^2 \frac{P_1(1-P_1)}{d} = Q \quad (7)$$

Equation (7) says that the variance of a product of independent variables is approximately equal to the square of the mean of the first factor times the variance of the second factor plus the square of the mean of the second factor times the variance of the first factor. Kish (1965) offered this approximation.

$$Var(\overline{p_{1k}P_k}) = Q + \frac{n_{\bullet k}d - n_{\bullet k} - d}{n_{\bullet k}^2 d^2} p_{1k}(1-p_{1k})P_1P_2 \quad (8)$$

Equation (8) is an implementation of equation (3).

$$Cov(\overline{p_{11}P_1}, \overline{p_{12}P_2}) = \frac{-p_{11}p_{12}P_1P_2}{d} \quad (9)$$

The factors p_{11} and p_{12} are independent of the others; P_1 and P_2 sum to 1.

By combining equations (5), (6) (first where k = 1 and then where k = 2), and (8),

$$Var(\overline{P}_{\bullet 1}) \approx P_1^2 \frac{p_{11}(1-p_{11})}{n_{\bullet 1}} + P_2^2 \frac{p_{12}(1-p_{12})}{n_{\bullet 2}} + \frac{P_1P_2}{d}(p_{11}-p_{21})^2 = V \quad (10)$$

According to Dr. James Rosson (USDA Forest Service, personal communication), DeLury developed this equation (circa 1950).

Two additional terms from equation (8) may be added for an exact answer, which is

$$Var(\overline{P}_{\bullet 1}) = V + \frac{n_{\bullet 1}d - n_{\bullet 1} - d}{n_{\bullet 1}^2 d^2} p_{11}(1-p_{11})P_1P_2 + \frac{n_{\bullet 2}d - n_{\bullet 2} - d}{n_{\bullet 2}^2 d^2} p_{12}(1-p_{12})P_1P_2 \quad (11)$$

In the example, the standard error of the mean is equal to 2.63 percent if both phase 1 and phase 2 are considered; it is nearly 4.75 percent if only the phase 2 plot coordinates are considered.

This method works for a sample. Card (1982) offered an alternate method for a full census (such as what might be involved in remote sensing). Remote sensing is the analysis of digital satellite images. Every point on the landscape cannot be photointerpreted; rather, a finite sample must be taken. However, a digital image may be analyzed wall-to-wall. The caveat is that it may be analyzed at the resolution of the data. There may be surface water, nonforested land, and forested land in one pixel; the satellite sensor will record an average of sorts for that pixel.

VARIANCE ESTIMATES OF A KNOWN PROPORTION

The amount of area in a particular State (A_t) is a legally defined amount. The proportion of forested land is not known; let it be represented by φ. If one desires to estimate the number of acres of forested land (A_{ft}), then

$$E(A_{ft}) = A_t\varphi \quad (12)$$

The variance on this estimate is

$$\sigma^2(A_{ft}) = A_t^2 \cdot \frac{\varphi \cdot (1-\varphi)}{n_t} \quad (13)$$

The standard deviation is

$$\sigma(A_{ft}) = A_t \sqrt{\frac{\varphi \cdot (1-\varphi)}{n_t}} \quad (14)$$

To get the answer in terms of a proportion, divide both sides by A_t

$$\frac{\sigma(A_{ft})}{A_t} = CV(\varphi_t) = \sqrt{\frac{\varphi \cdot (1-\varphi)}{n_t}} \quad (15)$$

One might call this quantity the coefficient of variation on the proportion of forest for the total.

Suppose one is interested in only a portion of a State (A_s). Suppose $p = A_s/A_t$. If the State's land is relatively homogeneous, then

$$E(A_{fs}) = A_s\varphi \quad (16)$$

The variance on this estimate is

$$\sigma^2(A_{fs}) = A_s^2 \cdot \frac{\varphi \cdot (1-\varphi)}{n_s} \quad (17)$$

The standard deviation is

$$\sigma(A_{fs}) = A_s \sqrt{\frac{\varphi \cdot (1-\varphi)}{n_s}} \quad (18)$$

Divide both sides by A_s to get the coefficient of variation

$$\frac{\sigma(A_{fs})}{A_s} = CV(\varphi_s) = \sqrt{\frac{\varphi \cdot (1-\varphi)}{n_s}} \quad (19)$$

If we divide equation (19) by equation (15), we get

$$\frac{CV(\varphi_s)}{CV(\varphi_t)} = \frac{\sigma(A_{fs})}{\sigma(A_{ft})} \cdot \frac{A_t}{A_s} = \sqrt{\frac{n_t}{n_s}} = \sqrt{\frac{1}{p}} \quad (20)$$

which is what FIA reports have published for many years.

VARIANCE ESTIMATES OF AN UNKNOWN PROPORTION

There are other aspects of forest inventory where the proportion under consideration is unknown. Although it is a variable no longer collected, wetland status was a subjective call. Photointerpreters did not attempt to delineate wetland timberland from other timberland. Quality assurance crews did check a number of plots a second time, but cross-tabulation data were not kept. If such data had been kept, the double-sampling equations could be used.

The first temptation would be to use a single-sampling equation. Suppose that one-half of the land is believed to be forested in a particular county, with a standard error of 2 percent. Also suppose that 10 percent of the forested plots were judged to be wetland by the field crews. There are 100 plots in the county, and the average acreage expansion factor is 6,000.

Under a single-sampling model, $\varphi = 0.05$, the standard error of which is 0.021794; thus one can say with 67 percent confidence that between 16,923 and 43,076 acres are wetland timberland.

However, if acreage expansion factor in thousands of acres (X) and wetland status (Y) are independent, then equation (5) may be used. It is generally impossible to assign an exact number of acres to particular plots; instead averages are used. For instance, if there are 300,000 plus or minus 6,000 forested acres in the county, and there are 50 forested plots, then on average there are 6,000 acres per plot. From equation (20), the standard deviation on 6,000 acres and one plot is 848.5.

However, unless more information is known (such as the total number of acres in particular ownership classes), then every plot is assigned an acreage expansion factor of 6,000—some don't get 5,900 while others get 6,100.

In this case, $\overline{X} = 300$, $s_X^2 = 36$, $\overline{Y} = 0.1$, $s_Y^2 = 0.09$, and $s_W^2 = 300^2 \cdot 0.09 + 0.1^2 \cdot 36 - 0.09 \cdot 36$, or 8,097. There are 50 observations, so the standard error of the mean is 12,725. One can then say with 67 percent confidence that in this county, there are between 17,275 and 42,725 acres of forested wetland—in this case there is only a small improvement over the single-sampling equation.

CONCLUSIONS

Although the long-term plan is to do phase 1 estimates by remote sensing, it was a useful exercise to document equations that could be used to estimate standard errors in forest area. They may be adapted to other topics in forest inventory, such as extent to which quality control should be done and the development of standard error equations for the number of

trees, volume, biomass, as well as growth, removals, and mortality.

LITERATURE CITED

Brown, M. and others. 2001. Wetland timberland statistics for the South Atlantic States. Resour. Bull. SRS-62. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 52 p.

Card, D. 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering and Remote Sensing. 48: 431-439.

Goodman, L. 1960. On the exact variance of products. Journal of the American Statistical Association. 55: 708-713.

Goodman, L. 1962. The variance of the product of K random variables. Journal of the American Statistical Association. 57: 54-60.

Kendall, M.G.; Stuart, A. 1963. The advanced theory of statistics. London, England: Charles Griffin and Company, Ltd. 433 p.

Kish, L. 1965. Survey sampling. New York, NY: John Wiley & Sons, Inc. 643 p.

Rosson, J.F. 2000. Forest resources of east Texas, 1992. Resour. Bull. SRS-53. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 70 p.

Schweitzer, C. 2000. Forest statistics for Tennessee, 1999. Resour. Bull. SRS-52. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 78 p.