SOME PROBLEMS OF SAMPLING IN THE FOREST SURVEY

By KAMEO MATUSITA, CHIKIO HAYASHI, MASATUGU ISIDA, HIROBUMI UZAWA, HIROSI HUDIMOTO, HIROTUGU AKAIKE AND TOSIO UEMATU

(Received Aug. 10, 1954)

1. Introduction

The methods and techniques of sampling are very useful in the forest survey and their application to it provides many interesting problems from the points of view of sampling theory itself as well as measure theory. In this paper we shall treat some of these problems in the light of the survey carried out under the auspices of Forestry Agency, Ministry of Agriculture and Forestry of the Japanese Government in 1952–1954.

In the forest survey, the most common method of sampling is area sampling. In sections 2 and 3 we shall treat some problems in relation to area sampling, and then in section 4 we shall refer to the sampling design of the survey of forest resources in the whole country of Japan. It is also an important problem in the application of sampling methods to improve the precision of estimation by taking advantage of previous information. In this connection, we shall in section 5 study the application of regression estimate to the forest survey. Finally, considering the importance of tree measurement in the forest survey, we shall in section 6 treat the analysis of errors in tree measurement.

In the preparation of the present paper thanks are due to Messers M. Motoo and K. Isii.

2. The error of traversing

In carrying out the area sampling in the forest survey, we cannot overlook non-sampling errors. Especially the evaluation of error in area measurement is very important. For, in the survey, we cannot do without area measurement, and in this measurement we cannot be free from the error. Taking consideration of the actual forest survey in Japan, where the traversing is commonly used, we shall treat the problem in the following.

The traversing is the following method of area measurement. First we set up many station marks on the circumference of the area, which is to be measured, and join every pair of neighboring station marks by a straight line. Then we get a polygon, vertices of which are station marks. Substituting the polygon for the given area, and proceeding along the circumference of the polygon, we measure the length, the angle of elevation, and the direction of every side of the polygon. From the measurement results thus obtained we can calculate the area of the polygon. If the results have some error, the area of the polygon thus calculated may also have some error. Let r, φ and θ represent generally the true values of distance, angle of elevation and angle of direction between two neighbouring vertices, respectively, and Δr, $\Delta\varphi$, $\Delta\theta$ errors of measurement of r, φ, and θ, respectively. We assume here that Δr, $\Delta\varphi$ and $\Delta\theta$ are random variables which have variances independent of the true values, that is, $V(\Delta r)\equiv\sigma_r^2$, $V(\Delta\varphi)\equiv\sigma_\varphi^2$, $V(\Delta\theta)\equiv\sigma_\theta^2$, and that there is no correlation between any two of them. First, we consider the case where the measurements of r, φ, and θ are unbiased, that is,

$$E(\Delta r)=E(\Delta\varphi)=E(\Delta\theta)=0\,.$$

Let r_k, φ_k, and θ_k be the true values of the k-th measurement and S be the true area of the polygon. Then, if we neglect error terms of higher order than Δr, $\Delta\varphi$ and $\Delta\theta$, the error ΔS of the area can be written as follows:

$$\begin{aligned}\Delta S \fallingdotseq & \sum_{k=1}^{n}(a_k\cos\varphi_k\sin\theta_k-b_k\cos\varphi_k\cos\theta_k)\Delta r_k \\ & +\sum_{k=1}^{n}(b_k r_k\sin\varphi_k\cos\theta_k-a_k r_k\sin\varphi_k\sin\theta_k)\Delta\varphi_k \\ & +\sum_{k=1}^{n}(a_k r_k\cos\varphi_k\cos\theta_k+b_k r_k\cos\varphi_k\sin\theta_k)\Delta\theta_k\end{aligned}\tag{1}$$

Here n is the number of station marks, and $(a_k, b_k)=1/2\mathfrak{x}_k+\mathfrak{x}_{k+1}+\cdots+\mathfrak{x}_n$, where $\mathfrak{x}_j$ is a vector of two dimensions, the co-ordinates of which are $r_j\cos\varphi_j\cos\theta_j$ and $r_j\cos\varphi_j\sin\theta_j$. From (1) we get $E(\Delta S)\fallingdotseq 0$, and

$$\begin{aligned}V(S)\fallingdotseq & \sum_{k=1}^{n}(a_k^2\cos^2\varphi_k\sin^2\theta_k+b_k^2\cos^2\varphi_k\cos^2\theta_k)\sigma_r^2 \\ & +\sum_{k=1}^{n}(a_k^2 r_k^2\sin^2\varphi_k\sin^2\theta_k+b_k^2 r_k^2\sin^2\varphi_k\cos^2\theta_k)\sigma_\varphi^2 \\ & +\sum_{k=1}^{n}(a_k^2 r_k^2\cos^2\varphi_k\cos^2\theta_k+b_k^2 r_k^2\cos^2\varphi_k\sin^2\theta_k)\sigma_\theta^2 \\ & -\sum_{k=1}^{n}a_k b_k\cos^2\varphi_k\sin 2\theta_k\sigma_r^2-\sum_{k=1}^{n}a_k b_k r_k^2\sin^2\varphi_k\sin 2\theta_k\sigma_\varphi^2 \\ & +\sum_{k=1}^{n}a_k b_k r_k^2\cos^2\varphi_k\sin 2\theta_k\sigma_\theta^2\end{aligned}\tag{2}$$

From (2) we can estimate the error ΔS of the area. In this case, for small n, we must make use of the Tchebycheff's inequality, but for large n, probably for $n \geqq 10$, we could assume that ΔS has asymptotically Gaussian distribution according to the central limit theorem.

Now, we shall give the relation between S and the relative error $\frac{\sqrt{V(\Delta S)}}{S}$ of its measurement by making use of numerical values of σ_r^2, ρ_φ^2, σ_θ^2 obtained by practical analysis. These numerical values are as follows:

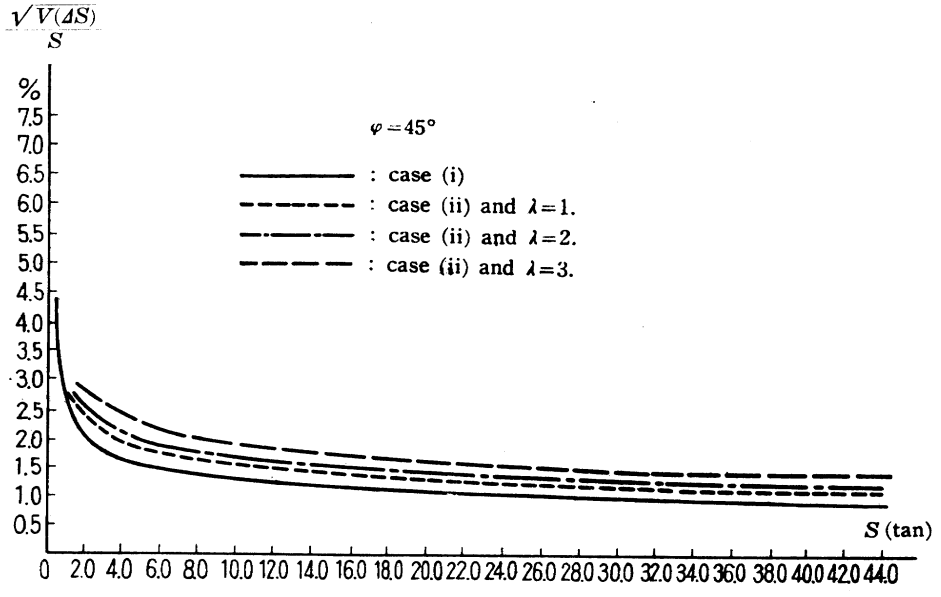
$$\sigma_r^2 = 0.05\,\mathrm{m}^2\,, \qquad \sigma_\varphi^2 = 1.05^{\circ 2}\,, \qquad \sigma_\theta^2 = 0.57^{\circ 2}\,.$$

Here we consider the following typical cases:

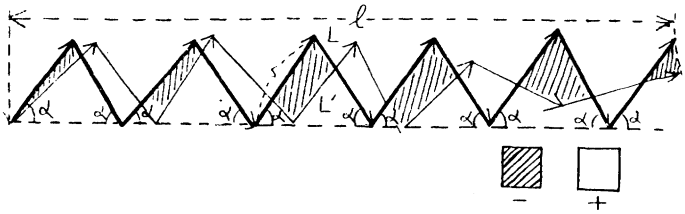
(i) the orthogonal projection of the polygon on the horizontal plane is a regular polygon with line segments of length 24 m and $|\varphi_1| = \cdots = |\varphi_n| \equiv \varphi$,

(ii) the orthogonal projection of the polygon on the horizontal plane is a rectangle such that the ratio of its two sides is $1 : \lambda$, $|\varphi_1| = \cdots = |\varphi_n| \equiv \varphi$, and the length of each side is multiple of 24 m.

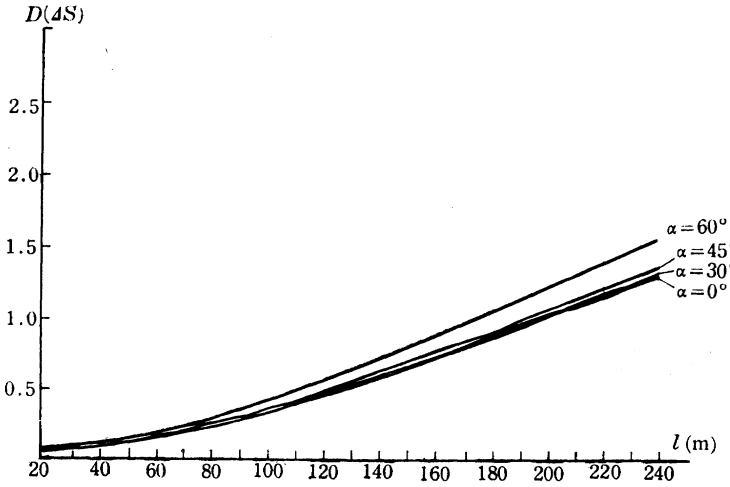
In practice we may assume that $\varphi \leqq 45°$, then the error ΔS attain its maximum when $\varphi = 45°$, and it suffices to study the case where $\varphi = 45°$s The relation between S and $\frac{\sqrt{V(\Delta S)}}{S}$ is given as follows.



Here "tan" is a unit of area in general use in Japan. To the case of a convex and central-symmetric area, too, the above relations would apply well.



In order to investigate the more complicated case of zigzag area, let us consider to estimate the area between L and L' in the above figure. The relation between l and $\sqrt{V(\Delta S)}$ turns out to be what the following figure shows, and it shows that the difference is small. Therefore the problem of a notch does not matter much.

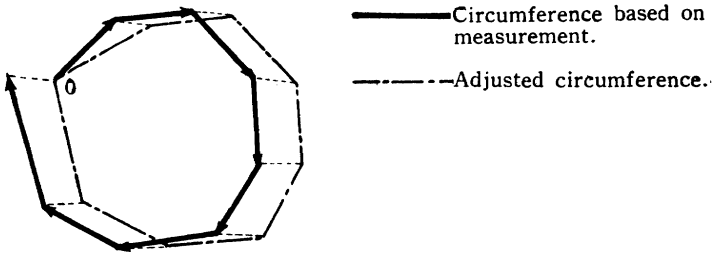


In the second place we consider the case where the measurements of r, φ have biases Δr, $\Delta\varphi$, the bias of θ is negligible, $|\varphi_1|=\cdots=|\varphi_n|=\varphi$, and Δr are independent of the size of r. Then the bias ΔS of S is given approximately as follows.

$$\Delta S=\sum_{k=1}^{n}(a_k\sin\theta_k-b_k\cos\theta_k)\Delta r'_k$$

where a_k or b_k is the same one as in (1), and $\Delta r'_k=\Delta r\cos\varphi-r_k\sin\varphi\Delta\varphi$. In the preceding cases (i) and (ii), it turns out that $\dfrac{\Delta S}{S}=2\dfrac{\Delta r'}{r'}$ where $r'=24$ m and $\Delta r'=\Delta r\cos\varphi-r\sin\varphi\Delta\varphi$.

Finally let us consider the effect of the adjustment of errors after the measurement. When we make a map according to the measured values, we get usually deviation between the start and end points. In that case we adjust the map by a parallel and length-proportional transformation of each polygon's side. (See the figure below.) We shall now make clear the influence of this adjustment from a statistical standpoint.



For simplicity, suppose that the area is a regular polygon with side-length r on the horizontal plane. Let S_1 and S_2 be the measured and adjusted values of S, respectively, and let ΔS_1, ΔS_2 be their errors. If we are free from any bias, we have $E(\Delta S_1)\fallingdotseq 0$ and

$$(3) \qquad V(\Delta S_1)\fallingdotseq \frac{nr^2}{8\sin^2\frac{\pi}{n}}\left\{\sigma_r^2\left(2+\cos\frac{2\pi}{n}\right)+r^2\sigma_\theta^2\right\}$$

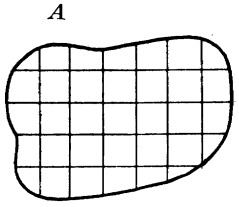
Now by some calculation, we obtain $E(\Delta S_2)\fallingdotseq 0$ and

$$(4) \qquad V(\Delta S_2)\fallingdotseq \frac{nr^2\sigma_r^2}{8\sin^2\frac{\pi}{n}}\left(1+\cos\frac{2\pi}{n}\right)$$

Then even if there is a any bias, by putting $E(S_1)=\overline{S}$, we have $E(\Delta S_1)\fallingdotseq E(\Delta S_2)\fallingdotseq \overline{S}$, and the variances are the same as above. From (3) and (4) we obtain $\sqrt{V(\Delta S_2)}\leqq \frac{2}{3}\sqrt{V(\Delta S_1)}$. It implies that, when the bias of measurement of S is small, the adjustment diminishes the error of the measurement of S. If the bias is not small, however, the error of the measurement of S cannot be made small, even if the variance of ΔS can be made small by adjustment. In the above case, we have assumed that the area is a regular polygon. In more complicated cases, the calculation will not be so easy, and it seems dangerous to draw any hasty conclusion from the above argument.

3. Determination of sampling unit

In carrying out the forest survey with sampling method, we encounter the problem of what form and what extent of sampling units are most efficient. Let us consider this problem, taking into account the cost of survey.



Let A be the area which is to be surveyed. First divide A into blocks as shown in the figure, and take every block as a sampling unit. If $X_1 \cdots X_N$ denote the volumes in these blocks, the total volume is $V=\sum_{i=1}^{N} X_i$, where N is the number of blocks. Let $x_1, \cdots, x_n$ be volumes of a simple random sample of size n from these N blocks, and put $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$. Then V will be estimated by $\hat{V}=N\bar{x}$. The coefficient of variation of this estimate is given by

$$\frac{N\sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}}}{V}=\sqrt{\frac{N-n}{N-1}}\frac{1}{\sqrt{n}}\frac{\sigma}{\overline{X}}$$

where $\overline{X}=\frac{V}{N}$ and σ^2 is the population variance among blocks, i.e., variance of the block population, and it gives the relative precision of the estimation.

It can be seen from this formula that the coefficient of variation depends on $\sigma/\overline{X}$, where $\sigma/\overline{X}$ varies according to the way of subdividing the area—the way in which we make the population of our concern. Therefore we must know at first how $\sigma/\overline{X}$ is expressed in terms of forms and extents of blocks which are sampling units, for the method of dividing the blocks depends only on them.

Considering the nature of our survey we shall here confine ourselves to the case where all blocks have almost the same form and equal area although the form may be taken in infinitely many ways. As for the form we take (1), (2), (3) in the following figure as the representatives.

(1) $\boxed{s}$ (2) $\boxed{s}$ (3) $\boxed{s}$

When the form is fixed the dividing method is determined by the area s common to all blocks, and thus we denote it simply by the area s. In this case $Z=\sigma/\overline{X}$ is a function of s: $Z=Z(s)$. Actually through the practical survey we first investigated the functional form of $Z=Z(s)$. We selected a part of forestry and studied every tree in it. In that part there were Japanese cypresses and larch-trees, but the majority were Japanese cypresses. Its extent was 5-*tan* and trees were 25 years old. In this case $Z(s)\fallingdotseq as^{-\lambda}$ was seen to be a good approximation and we got the following table

Method	a	λ
(1)	0.2310	0.3011
(2)	0.2095	0.2693
(3)	0.1981	0.2846

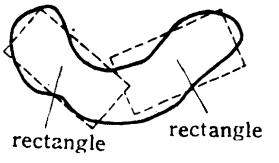
where s is measured by *tan*. Z decreases fast when s increases from 0 to 1, but slowly thereafter and decreases only a very little when s becomes greater than 2. Therefore, when we want to draw samples with the equal area, without taking into account the cost, it is desirable to take areas of extent of 1∼2 *tan* as sample units.

In practical cases, we usually try to approach the most profitable way of subdividing the area from the following two standpoints.

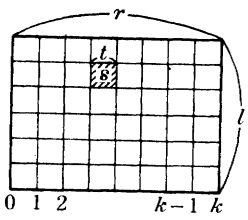
i) To make the precision as high as possible for a given total cost in the survey.

ii) To minimize the cost for the required precision.

In general the actual area is not a rectangle, but we can divide the area into parts which may be assumed to be approximately rectangles as in the right figure. Moreover, in the above mentioned survey, it was seen that the difference of functional form of $Z(s)=\frac{\sigma(s)}{\overline{X}(s)}$ among the above three forms of sampling unit produced little effect on the result. Therefore, it seems sufficient to solve the problem in the case where the area is a rectangle and the sampling unit is a square. In this case the way of subdividing the area is determined by the length t of the side of square. Now, let the area be one as shown in the left figure, and divide it into N sampling units.



We assume that the lower side of the area are divided into k intervals with the almost equal length, that is, $k=\frac{r}{t}$ is an integer. For the conveniences of calculation, however, we shall treat t as a continuous variable. When we consider (i) or (ii), it is necessary to get formulae for the precision and the cost in terms of n and t. As for the precision, it is given by



$$y=y(t, n)=\sqrt{\frac{N-n}{N-1}}\frac{1}{\sqrt{n}}\frac{\sigma(s)}{\bar{X}(s)}$$

Here we assume that $N\gg 1$ and $\frac{\sigma(s)}{\bar{X}(s)}\fallingdotseq as^{-\lambda}=at^{-2\lambda}$ holds in the general case. Then

$$\frac{N-n}{N-1}\frac{1}{n}\fallingdotseq\frac{1}{n}-\frac{1}{N}$$

therefore,

$$y=y(t, n)=\sqrt{\frac{1}{n}-\frac{1}{N}}at^{-2\lambda}$$

Now we shall consider the formula for the cost. The total cost may be considered to consist of the following.

a) The cost for the measurement of circumference of the area.

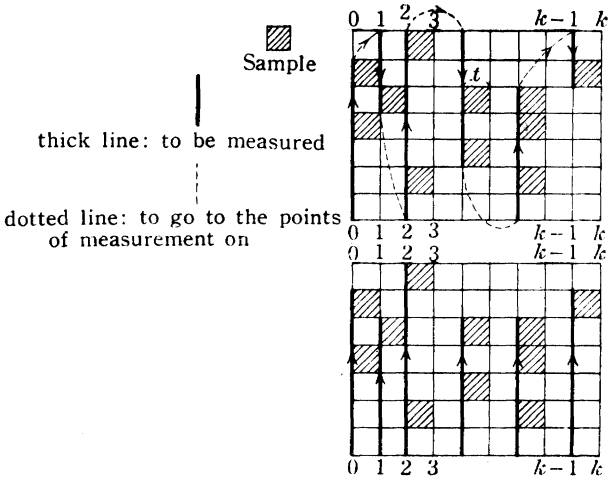
b) The cost for confirming the sample spots, that is, for reaching the sample-spots.

c) The cost for the measurement of circumference of the sample-spots and confirming them.

d) The cost to tally the sample-spots.

Assuming that the cost of the measurement of circumference is proportional to the length to be measured and the cost of tallying is proportional to the extent to be measured, we denote the cost per unit length and unit extent by α, β, respectively. Then for given t and n, the costs are $L\beta$ for (a), $4n\beta t$ for (c) and $n\beta t^2$ for (d). Let the cost for (b) be D. Then the total cost will be $C(t, n)=n\alpha t^2+4n\beta t+D+L\beta$. Let us consider the formula for D. For confirming the sample-spot, it suffices to determine the position of sample-spot relative to some station mark, and it seems to be convenient to take the station mark used for the circumference measurement of the area as a landmark. Assuming

that the stations of the circumference measurement are on the upper and the lower side of area, it is sufficient to measure lengths as is shown by the thick lines in figure 4. Since we are interested in the mean of these lengths for the random sampling, it is the same as measuring of thick lines in figure 5. Therefore the mean distance $\rho(n, t)$ to be measured for confirming the sample-spot is given by taking the mean of the sum of the largest distances between the lower side and the sample-spots on each row. Thus we have



$$\rho(t, n)=kl\left[1-\frac{k}{n+1}\left\{1-\left(1-\frac{1}{k}\right)^{n+1}\right\}\right]$$

$$D=\beta\rho(t, n)\,,$$

therefore, we obtain

$$C(t, n)=n\alpha t^2+4n\beta t+\beta\rho(t, n)+L\beta$$

As $\rho(t, n)$ has a very complicated form and is inconvenient for numerical calculation, we replace it by the following approximation formula which is the nearest straight line to the graph of $\rho(t, n)$, where t, k vary in the range appropriate for the practical purpose.

$$\rho(t, n)\fallingdotseq 0.16ln+2l$$

therefore we get

$$C(n, t)=n\alpha t^2+4n\beta t+\beta(0.16ln+2l)+L\beta$$

Using the above $C(n, t)$ and $y(n, t)$, the problems corresponding to (i) and (ii) are reduced to the following.

(i) To minimize $y(t, n)$ under the condition $C(n, t)\equiv C$, where C is a given total cost.

(ii) To minimize $C(t, n)$ under the condition $y(t, n)\equiv\varepsilon$ where ε is a given precision.

These problems are solved as follows:

The case (i)

Let $\overline{C}=C-L\beta$ be the total cost except for the circumference measurement.

Then t is given by

$$(A\alpha-\overline{C}+2l\beta)(1-2\lambda)t^2+2A\beta(1-4\lambda)t-0.32A\lambda l\beta=0$$

the sample size by

$$n=\frac{C-2l\beta}{\alpha t^2+4\beta t+0.16l\beta}$$

and the precision y by

$$y=\sqrt{\frac{1}{n}-\frac{t^2}{A}}\,at^{-2\lambda}$$

The case (ii)

t is given by

$$A\frac{\varepsilon^2}{a^2}(1-2\lambda)t^{4\lambda}-2A\frac{\varepsilon^2}{a^2}(4\lambda-1)t^{4\lambda-1}-0.32\lambda Al\frac{\varepsilon^2}{a^2}\beta t^{4\lambda-2}-2\beta t-0.16l\beta=0$$

$$n=\frac{A}{A\alpha t^{4\lambda}+t}$$

and

$$\overline{C}=n\alpha t^2+4n\beta t+\beta(0.16n+2)l$$

The values of α, β should be determined by taking into account the practical forest survey.

Actually we have obtained the following data from our survey.

$\lambda=0.2846$ $\quad a=1.414$ $\quad A=28400\,\mathrm{m}^2$ $\quad l=119$

$r=239\,\mathrm{m}$ $\quad \alpha=0.00194$ person hour/m^2

$\beta=0.015$ $\quad$ person hour/m

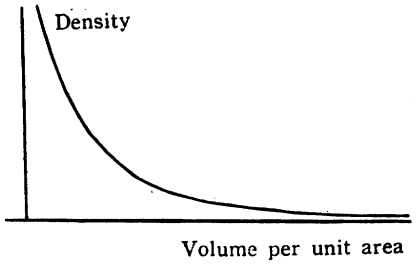
When using this data and $\overline{C}=42$ person hour—the standard cost of survey in Japan—, we get

$$S=1.4\tan \qquad n=8 \qquad y=5.2\%$$

4. A design of area sampling in forest survey

In the survey of forest resources in a comparatively wide region by the sampling method, it frequently happens that we cannot select any suitable sampling unit, and even if it is possible, our unit cannot be used practically by the reason of management or others. In such a case the area sampling method by topographical maps is convenient. In the following we shall state the outline of our survey plan of the forest resources in the whole country of Japan, which uses the topographical map of scale 1–50000.

As is seen from its nature, this survey has various survey items. Therefore, we did not consider the stratification about special items, and restricted ourselves only to the stratification by prefecture, which facilitates the drawing of the land spots. We also did not use the sub-sampling in consideration of the required precision and the nature of the results, and drew a sample directly from the whole country. Existing units, such as the administrative districts etc., were not suitable for sampling units. For, it was very difficult to collect the primary data, and the extents and boundaries of those units were very obscure. Thus we divided the whole country into blocks with a fixed area, suitable for the survey, and took those blocks as sampling units. We used a simple random sampling and drew a sample from these units. For the precision of 5%, the sample size is determined by $0.05 \geqq \frac{2\sigma}{\sqrt{n}\,\overline{X}}$, where $\overline{X}$ is the mean of volumes per unit and σ^2 is the variance of volumes per unit. Now the distribution of volume per unit area may be considered to be of L-type, as is shown in the figure. If the density of this distribution is e^{-x}, we have $\frac{\sigma}{\overline{X}}=1$. It seems that $\frac{\sigma}{\overline{X}}$ is usually between 1 and 1.5. Thus, to determine the sample size, we took 1.3 as the value of $\frac{\sigma}{\overline{X}}$[1]. Then the above formula is reduced to



1) As the results of the actual survey we got the following values.

$$\overline{X}=8.93\ \mathrm{m}^3$$
$$\sigma=14.05\ \mathrm{m}^3$$
$$\frac{\sigma}{\overline{X}}=1.57\ .$$

$$0.05 \geqq 2\frac{1.3}{\sqrt{n}}$$

namely

$$n \geqq 2704$$

and the sample size 3000 is sufficient for the case. We allocated 3000 sample spots to each prefecture proportional to its area, i.e.,

$$k_i = \frac{a_i}{A} n ,$$

where n: the number of all sample spots (3000),

a_i: the whole area of each prefecture,

$A = \sum a_i$: the whole area of Japan,

k_i: the number of sample spots allocated to each prefecture.

It is seen that as the area of the sampling unit becomes larger, the precision of the estimation becomes higher. However, we must consider the cost of survey at the same time, and it does not give advantage to take very large areas without consideration. Taking the preceding consideration about the sampling unit, it seems optimum to take 1～3 tan as the area of sampling unit. Moreover, since there are some survey items concerning the character of land, such as the state of utilization and the topographical feature of the land, the adoption of large unit makes it very difficult, or at least practically impossible to characterize each unit clearly, because the conditions vary much within the large unit. For this reason, too, the area of the unit is desirable to be small. Under these considerations we decided that the area of each sampling unit should be 50 m × 30 m = 1.5 tan. However, it is impossible, in practice, to divide the whole country into small blocks with area 30 m × 50 m and to draw a simple random sample of 3000 from them. We actually drew the sample in the following way. In the first place, considering the lattice, which divided the whole country into as many blocks of equal area, as sufficient for the purpose, we drew 3000 lattice points from this lattice, with equal probability. Then at each lattice point, which was drawn, we set up the 30 m × 50 m block in a certain way, the lattice point being taken as a cardinal point. As a lattice we took the circles of longitude and the parallels of latitude respectively at the interval of l'. (The extent of each area divided by it was almost $1.8 \times 1.6\,\mathrm{km}^2$.) The area

of the block, generated by the lattice of latitudes and longitudes lines, is not constant. However, we made the stratification with respect to the prefecture, and drew the sample spots within each prefecture. The error arising from the inequality of the block was negligible in practice. Now the lattice of latitude and longitude lines can be chosen in infinitely many ways. We must, of course, select one of them at random. But, for the practical reason, we used the longitude and latitude lines which had been printed on the existing 1–50000 maps. We believe that this selection does not cause any statistical bias.

In order to settle the spot in the field, and to carry out the measurement we used the following method.

1. Marking the lattice point, which belongs to the survey spot, on the topographical map with scale 1–50000, and, using the sketch maps and areal photographs, to investigate the landmark, which is convenient for deciding the position of the spot to be surveyed, and the route to reach it on.
2. For use of the traverse measurement, sticking the topographical map, marked with the lattice point, on the plate, to examine the declination of a magnetic needle in the neighbourhood of the spot.
3. Standing at a point of a wide field of vision in the neighbourhood of the lattice point, and using the triangular points, peaks and saddles of the mountains or other land marks, to decide this point by trigonometry.
4. From this point, to decide the lattice point by the trigonometry and the traverse measurement.
5. Measuring the horizontal distance 50 m to the north and 30 m to the east from the lattice point, to decide the survey block, and to study every tree and other items.

The errors of position of the lattice point, determined by the above mentioned method, was less than 100 m.

This fact was sufficient for the method to serve our purpose, and the work was accomplished by comparatively small labour. The estimation formula for the total volume of the whole country is given by

$$v_T = \frac{A}{a} \sum_{i=1} v_i / n$$

where A: area of the whole Japan

a: area of a sample survey spot (30 m × 50 m)

v_i: accumulation of i-th survey spot.

n: number of survey spots.

The variance σ_{v_T} is given by

$$\frac{A^2}{a^2}\frac{\sigma_v^2}{n}$$

where σ_v^2 is the variance of volumes among the sampling units. We can therefore calculate the length of confidence interval l, with confidence coefficient 95%, neglecting the effect of the stratification by prefecture, and neglecting the error in measurement, that is:

$$l=\pm 2\frac{A}{a}\frac{\sigma_v}{\sqrt{n}}$$

Now in the measurement of a, if there is error ε_i of measurement at the i-th spot, the variance σ_{v_T} of estimate v_T is given approximately by

$$\sigma_{v_T}^2=\frac{A^2}{a^2n^2}\left(\sigma_v^2+\frac{\sigma_a^2}{a^2}\overline{V}^2\right)$$

where $\overline{V}$ is the population mean of the sampling units.

As for the estimation of ratio, the variance of the estimate can be calculated in quite a similar way.

5. The use of the regression estimate

If we have some available information about a given forestry, we can make the estimation for the forestry more effectively by making use of them than in the case where we have not any information about it. For example, if the aero-photograph is available, we can improve the precision of the estimation of volume by the following method: on one hand, we estimate the volume of the forestry by using the aero-photograph, and on the other hand, making a small part complete survey, we determine the regression coefficient between the volume of stand and its estimate by means of the aero-photograph, and then using this regression coefficient we correct the estimate of the volume of the forestry derived from the aerophotograph. The improvement of the precision by means of experts' excellent intuitive judgement is similarly considered. In this case, for the ordinary regression estimate, the double sampling is required. Linear

regression estimation is carried out in general as follows[1]. Let n be the sample size, v the mark of a sample drawn in the survey, and x the mark known by the previous survey. Drawing the regression line, we carry out the estimation by

$$\bar{v}_L=\bar{v}+b(\overline{X}-\bar{x})$$

where $$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(v_i-\bar{v})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \qquad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i, \qquad \bar{v}=\frac{1}{n}\sum_{i=1}^{n}v_i,$$

$$\overline{X}=\frac{1}{N}\sum_{i=1}^{N}x_i,$$

and N is the size of the parent population.
If

$$N\gg 1, \qquad E(\bar{v}_L)=\overline{V}$$

and, assuming that the variance about the regression line is constant, we have

$$\sigma^2_{\bar{v}_L}\fallingdotseq\frac{\sigma_v^2(1-\rho^2)}{n}\left(1+\frac{1}{n}\right)$$

where $$\sigma_v^2=\frac{1}{N}\sum_{i=1}^{N}(v_i-\overline{V})^2,$$

and ρ is the correlation coefficient between v and x in the population. Moreover, if $n\gg 1$, then $\bar{v}_L$ is approximately normally distributed (see [3]). In this case, if we can use the previous information, say, $x(1),\cdots,x(R)$, for x, we got

$$\sigma^2_{\bar{v}_L}\fallingdotseq\frac{\sigma_v^2(1-\rho^2)}{n}\left(1+\frac{R}{n}\right)$$

where ρ is the multiple correlation coefficient between v and $x(1),\cdots,x(R)$. In the forest survey, the eye-measurement is used as auxiliary information. Now, the error of eye-measurement of the volume apt to be great in the stand with large volume, and it is 0 for a stand of volume 0. It seems most likely that the standard deviation σ_ε of the eye-measurement error of the volume is proportional to the value x^2. Namely,

1) In the following the ordinary assumptions about the linear regression estimation are assumed to be satisfied.

$$\sigma_e^2=\lambda x^2$$

Then, the approximation formula for the variance of the estimated value is given by

$$\sigma_L^2=\frac{\sigma_v^2(1-\rho^2)}{n}+\frac{\lambda}{n^2}\left(\overline{X}^2+2\overline{X}\frac{\mu_3}{\mu_2}-\frac{\mu_4}{\mu_2}\right)$$

where μ_i is the i-th moment of x. Though this variance is greater than that in the ordinary case, for a sufficiently large sample size, the 2nd terms on the right hand in the above formula can be neglected, and this variance may be reduced to the oridinary one. In general, it can not be determined, for what value of n, the second term can be neglected. In practice, however, the value of λ is known to be less than 0.1, and the second term is sufficiently small.

In the forest section consisting of comparatively wide stands, the standard deviation of eye-measurement is sometimes given by

$$\sigma_e^2=\lambda(x+d)^k$$

In this case, the variance of the estimation is

$$\sigma_L^2=\frac{\sigma_v^2(1-\rho^2)}{n}+\frac{\lambda}{n^2}\left\{(\overline{X}+\alpha)^k+k(\overline{X}+\alpha)^{k-1}\frac{\mu^3}{\mu_2}-\cdots\right\}$$

Now, we shall consider the case where we can not obtain the population mean of the eye-measurement of volume. The estimation formula $v_i=v+b(\overline{X}-\bar{x})$ contains $\overline{X}$, but the calculation of $\overline{X}$ is too laborious to be carried out, when N is large. In such a case, we draw another sample of size m independently of the previous one, and substitute $\bar{x}_m=\frac{1}{m}\sum_{i=1}^{n}x_i$ for $\overline{X}$. Then the variance of the estimate is given by the approximation formula

$$\sigma_L^2=\frac{\sigma_v^2(1-\rho^2)}{n}\left(1+\frac{1}{n}\right)+\beta^2\frac{\mu_2}{m}$$

If $m\gg n$, it can be reduced to the ordinary case. Thus if we know the coefficient of each term, we can replace $\overline{X}$ by $\bar{x}_m$. This method is a sort of double sampling.

When we got the eye-measurement x_0 for the volume of a given stand, we estimate the true volume of the stand by $\hat{v}$,

$$\hat{v}=\overline{V}+b(x_0-\overline{X})\ .$$

Then the variance of errors of the estimation is approximately given by

$$\sigma^2_{\bar{v}-v}=\frac{\sigma_v^2(1-\rho^2)}{n}\left\{1+\frac{(x_0-\overline{X})}{\mu_2}\right\}$$

Although all the foregoing variances are calculated by approximation, it is sufficient for the practical purpose to use this order of approximation.

In the following we shall consider the case where the regression curve is not a straight line.

If many different groups of surveyers cruise the same forest section, or if the state of forest sections varies from region to region, the regression curve is sometimes not a straight line. In this case we stratify the data into the groups of the same nature. Namely, we adopt the following stratifications.

1. The stratification with respect to the groups of surveyers.
2. The stratification with respect to the eye-measurements of volume.
3. The stratification with respect to the age class, the species of trees and forestry.—and so forth.

In calculation, assuming that the regression curve is a straight line for each stratum, we use the linear regression estimate for each stratum, and consider the sum of them as an estimate of the whole volume. Then the variance is given by

$$\sigma_L^2=\frac{1}{N^2}\sum_{i=1}^{k}N_i^2\sigma_{Li}^2$$

where k is the number of strata,

N_i is the number of units in each stratum,

and σ_{Li}^2 is the variance of the regression estimate in each stratum.

6. Tree measurement

It is obviously most fundamental in the forest survey to measure trees, but it seems that the measurement is often carried through without much care and reflection. In view of such a circumstance and of the primary importance of tree measurement, we shall take up here some items, and study the errors in tree measurement and their effect to the calculation of the volumes of tree. Let the diameter of the tree at the breast height be d, the height of the tree h, and the volume of the tree v. Then v is a function of d and h. Thus,

$$v=f(d, h)$$

It is an interesting problem to study the form of this function itself. But we shall here confine ourselves to the examination of the forms of the function, which are commonly used.

(1°) Measurement of the diameter of the tree at the breast height.

In practice, the diameter of the tree at the breast height is measured for the purpose of obtaining the sectional area of the tree at the breast height. If the cross section of the tree is a circle, the meaning of the diameter is obvious. But, otherwise, the term, 'diameter of the tree at the breast height', is rather ambiguous. Considering the practical measurement, we shall here call the distance between two parallel tangents of the cross section at the breast height "the diameter of the tree at the breast height", or simply, "diameter". In what follows, we shall consider the most reasonable method of the measurement of diameter for the purpose of obtaining the sectional area of the tree at the breast height. Usually, the following three methods are practically used. That is:

a) to employ the mean value of measured maximum and minimum diameters,

b) to employ the mean value of a diameter in any direction and that at the right angle to it,

c) to employ one diameter in any direction.

Now, assuming the form of the cross section of the tree at the breast height is an ellipse, we shall study the above three cases.

Case a)

Let a and b be the major and the minor axis of the ellipse, respectively. Then the sectional area d of the tree at the breast height is:

$$d=\pi\left(\frac{a+b}{2}\right)^2$$

$$=\pi ab+\frac{1}{4}\pi(a-b)^2 .$$

Therefore, d is always over-estimated by $\frac{1}{4}\pi(a-b)^2$, as the area of the ellipse is πab. Moreover, when we denote the errors in measurement of $2a$ and $2b$ by ε_a and ε_b, respectively, the area of cross section of the tree at the breast height is given as follows:

$$d=\pi\left(\frac{a+b+\frac{\varepsilon_a}{2}+\frac{\varepsilon_b}{2}}{2}\right)^2$$

where we assume that ε_a and ε_b are quite accidental errors, that is, we assume

$$E(\varepsilon_a)=0\ ,\qquad E(\varepsilon_b)=0\ ,\qquad E(\varepsilon_a\varepsilon_b)=0\ ,$$
$$E(\varepsilon_a^2)=E(\varepsilon_b^2)=\sigma^2\ .\qquad (E(\varepsilon)\text{ denotes the mean value of }\varepsilon)$$

Then, taking into consideration the errors, we have, in the sense of average,

$$d=\pi ab+\frac{1}{4}\pi(a-b)^2+\frac{1}{8}\pi\sigma^2$$

Therefore d is always over-estimated by $\frac{1}{4}\pi(a-b)^2+\frac{1}{8}\pi\sigma^2$.

Case b)

If the radius measured first in an arbitrary direction is deflected by θ from the major axis, the sectional area at the breast height is

$$d=\pi\left\{\frac{1}{2}\left(\sqrt{a^2\sin^2\theta+b^2\cos^2\theta}+\frac{\varepsilon_1}{2}+\sqrt{a^2\cos^2\theta+b^2\sin^2\theta}+\frac{\varepsilon_2}{2}\right)\right\}^2$$

where ε_1 and ε_2 are the errors in the measurements of diameters. If we take the mean value of d with respect to θ and ε, and consider that θ is a random variable uniformly distributed in the interval from 0 to $\pi/2$, we have approximately

$$E(d)=\pi ab+\frac{3}{8}\pi(a-b)^2+\frac{1}{8}\pi\sigma^2$$

and the difference from πab becomes $\frac{3}{8}\pi(a-b)^2+\frac{1}{8}\pi\sigma^2$.

Case c)

In this case we have

$$d=\pi\left(\sqrt{a^2\sin^2\theta+b^2\cos^2\theta}+\frac{\varepsilon}{2}\right)^2.$$

If we take the mean value of d with respect to θ and ε, we obtain, as in the case b),

$$E(d)=\pi ab+\frac{1}{2}\pi(a-b)^2+\frac{\pi}{4}\sigma^2.$$

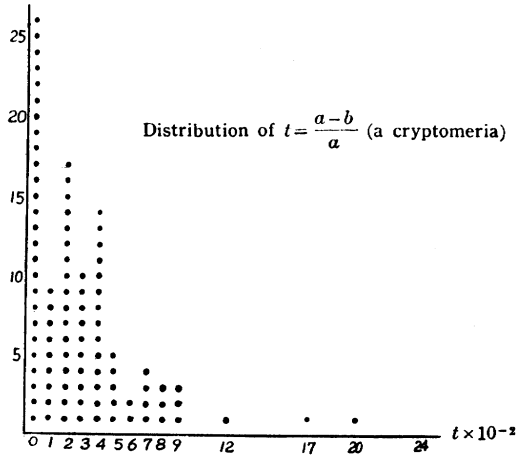
Therefore, $E(d)$ is always over-estimated by $\frac{1}{2}\pi(a-b)^2+\frac{\pi}{4}\sigma^2$.

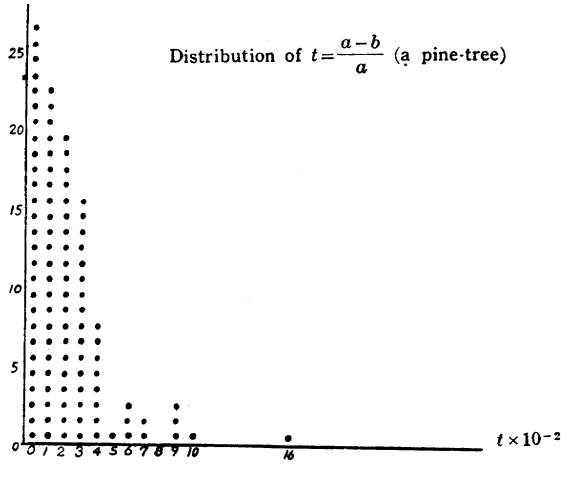
The relative bias in case a), b) and c) is shown as follows;

Case	Measurement is to be done for	The relative bias which has the effect on sectional area	Characteristic of bias
a)	the major and minor axis	$\frac{1}{4}\frac{t^2}{1-t}+\frac{1}{8}s^2$	Always positive
b)	the two arbitrary diameters at right angles	$\frac{3}{8}\frac{t^2}{1-t}+\frac{1}{8}s^2$	Always positive
c)	one diameter in any direction	$\frac{1}{2}\frac{t^2}{1-t}+\frac{1}{4}s^2$	Always positive

Here we put $t=\frac{a-b}{a}$ and $s^2=\frac{\pi\sigma^2}{\pi ab}$.

We shall investigate the magnitude of t in order to compare these three methods. Two examples of the distribution of t are shown in the following illustrations.





It seems that the value of t is distributed with the J-type approximately. This is, of couse, the result of investigation only on some forestries, so that it is difficult to draw a general conclusion from this result. But there would not be a great error in saying that the mean value of t is between 0.02 and 0.04. In these examples we cannot see the existence of correlation between t and a.

Now, let us consider which of methods a), b), and c) is the best. If the errors of measurement are very small, it is clear that the best method is a). When we measure a tree which has the sectional area nearly

of the shape circle, we can disregard the terms in t. Thus it is not useless to measure in two directions and make the errors of measurement half by the method b). (In this case, there is the possibility to find misreadings of scale.) As the terms in t is proportional and the terms in s is inversely proportional to the sectional area, we can say that the terms in t are of importance for big trees, and the terms in s are of importance for slender ones.

Our discussions thus far have been based on the assumption that the section of a tree is an ellipse in shape. Against this, there may well be the opinion that our assumption is not always true and sectional areas of trees are very complicated. However, the measurement of diameter in a random direction is well applied whenever the shape of the section is ellipse-like convex. In the case of concave shape, other instruments are necessary.

(2°) The errors in the measurement of height.

Usually, the measurement of height is apt to be far ruder than that of the diameter at the breast height. In this section we shall study the influence of the errors in measurement of the height on the calculation of the volume.

We use Terazaki's equation (cf. see [4]) to calculate the timber volume of the tree. This equation is in general use in Japan, and is given as follows.

$$v=cg\,10^{ah-b/h}$$

where v: timber volume,
g: area of the section at the breast height,
h: height of tree,
a, b, c: constants.

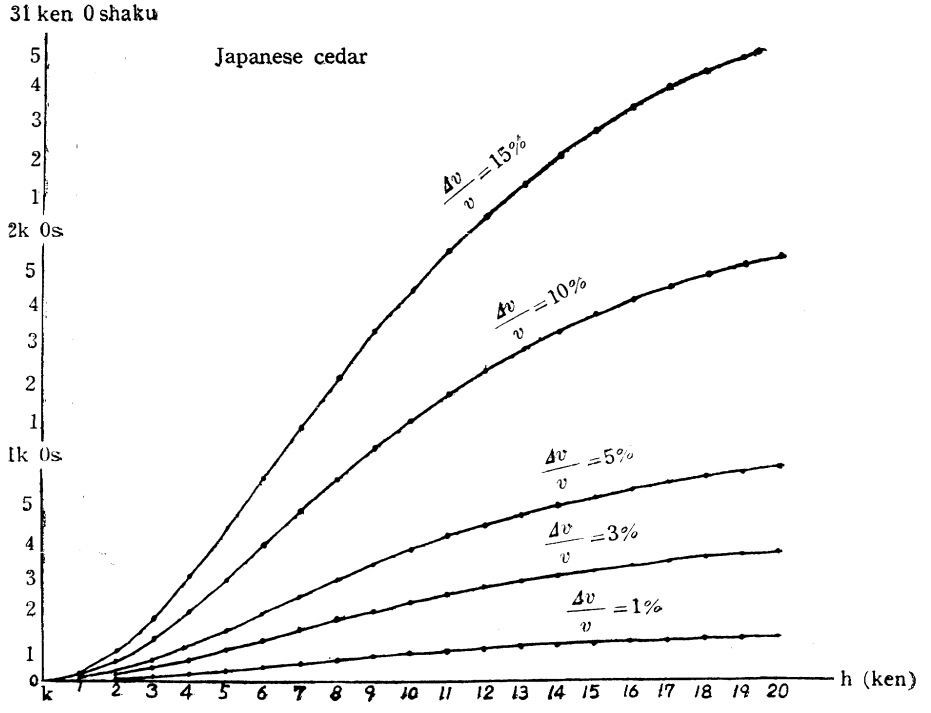
We shall now calculate the error of v, assuming that h has an error Δh. Putting

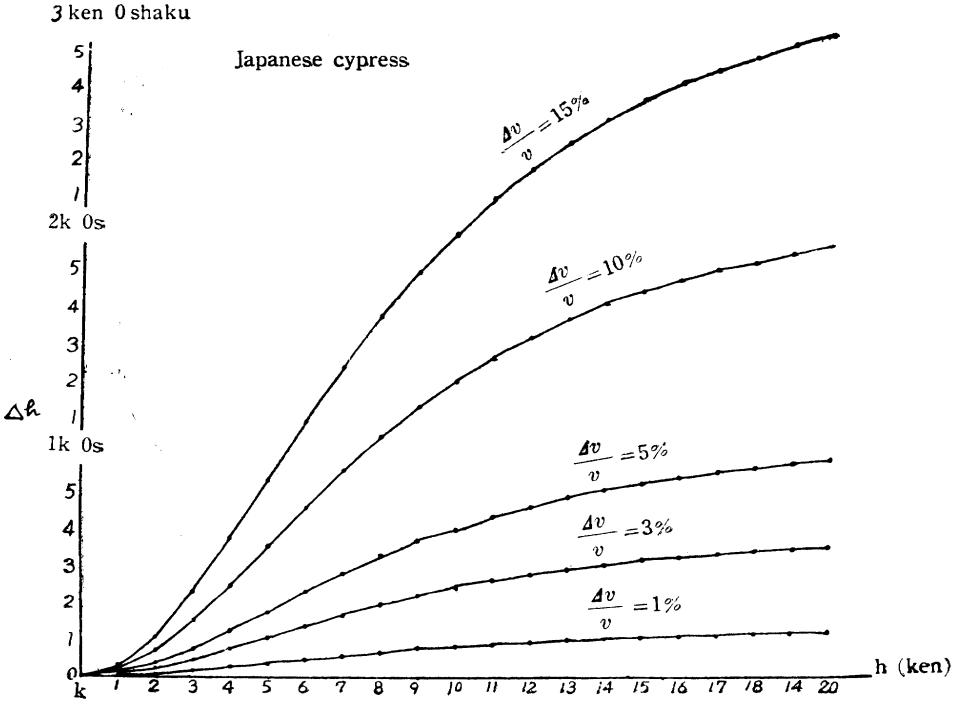
$$\frac{cg\,10^{a(h+\Delta h)-b(h+\Delta h)}}{cg\,10^{ah-b/h}}=K,$$

we obtain the following approximative formula:

$$\frac{b}{h}\varepsilon^2-\left(ah+\frac{b}{h}\right)\varepsilon+\log K=0,$$

where ε: $\Delta h/h$ (relative error in height)
K: $1+\Delta v/v$ (Δv is the error in volume)


31 ken 0 shaku
Japanese cedar
2k 0s
1k 0s
$\frac{\Delta v}{v}=15\%$
$\frac{\Delta v}{v}=10\%$
$\frac{\Delta v}{v}=5\%$
$\frac{\Delta v}{v}=3\%$
$\frac{\Delta v}{v}=1\%$
h (ken)


3 ken 0 shaku
Japanese cypress
2k 0s
1k 0s
Δh
$\frac{\Delta v}{v}=15\%$
$\frac{\Delta v}{v}=10\%$
$\frac{\Delta v}{v}=5\%$
$\frac{\Delta v}{v}=3\%$
$\frac{\Delta v}{v}=1\%$
h (ken)

Using this formula, we shall give the graphs for Japanese cedar and Japanese cypress.

We can see from these graphs that the errors we usually suffer in measuring the tree height have a considerable effect on the timber volume.

In practical surveys, height is usually measured with the eye, but we must notice the fact that the accuracy of eye-measurement is, even by an expert, far worse than is generally believed. Moreover, these errors have tendency to under estimate higher trees and also have some connection with the relative size of trees in the order of measurement. These facts increase difficulty in dealing with the errors in eye-measurement. After all, the proper volume can not be obtained unless the way of overcoming the above difficulties is found out.

7. The problem of non-sampling error (or response error)

The non-sampling errors are considered to occur in the process of every survey. Some of the problems were discussed in previous sections. Here we consider in what way we take them up in the system of the sampling errors. The non-sampling errors may be considered as the bias in one case and the variance or mean square error in another case. And, we may treat them by the same method as in the multi-stage sampling. Concerning the details refer to the literatures [5] and [6].

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

1. K. Matusita, C. Hayashi, M. Isida, H. Fudimoto, H. Akaike, H. Uzawa, T. Uematu, Some statistical problems in forest survey (in Japanese), *Proceedings of the Institute of Statistical Mathematics* vol. 1, No. 2, 1954.

2. Uematu, Statistical treatment of errors in measurement — on the errors in measurement by traverse method in forest survey (in Japanese) *Proceedings of the Institute of Statistical Mathematics*, Vol. 2, No. 2, 1954.

3. M. Isida, A remark on regression estimate, *Annals of the Institute of Statistical Mathematics*, Vol. 4, 1952.

4. W. Terazaki, Investigation on form-height tables for the principal conifers and some broad-leaved trees in Japan and bases on which they may be constructed. *Bulletin of the Forest Experiment Station*, 1915.

5. C. Hayashi, *How to Design Sampling Survey* (in Japanese), Tokyo University Press, 1951.

6. C. Hayashi, H. Aoyama, M. Isida, S. Nisihira, Y. Taga, M. Tutumi, H. Akaike, T. Taguchi, The statistical research on the national character of the Japanese. (in Japanese) *Proceedings of the Japan Statistical Association*, 1953.