CHAPTER 2

DESIGN-BASED SAMPLING AND INFERENCE

ANNIKA KANGAS

University of Helsinki, Finland

2.1 BASIS FOR PROBABILITY SAMPLING

The target of sampling is usually a finite population of N elements called sampling units. A sample s is a subset of this population with size n. A sample can be any subset of the population, but usually a random sample is used.

For sampling to fulfil the requirements of random sampling, it is enough that 1) a set S_n of all samples s of size n that it is possible to obtain can be defined; 2) each sample has a known probability $p(s)$ of being selected; 3) the probabilities are non-zero and the sum of these probabilities is one $\sum_{s \in S} p(s) = 1$, and 4) the sample s is selected according to the probabilities $p(s)$. The units are selected independently, i.e. selection of any one unit does not affect the selection of others. No other requirements are needed. The probabilities $p(s)$ then define the sampling design (Särndal et al. 1992, p. 8). Lund and Thomas (1989) provide a good overview of various sampling designs used in forest and stand inventories. It is worth noting, however, that systematic sampling does not fulfil the above requirements of independent selection, and this will affect inferences based on this sampling design (section 2.4.).

Another important probability measure is the inclusion probability π_i. This measures the probability of each sampling unit i entering the sample s. The inclusion probability and selection probability are connected (see section 2.2). When these probabilities are known, the sample statistics of interest can be calculated. The most general estimators that apply to all kinds of sampling design are those based on

A. Kangas and M. Maltamo (eds.), Forest Inventory – Methodology and Applications, 13–38.

arbitrary inclusion probabilities. An estimate $\hat{T}$ for total value T of some interesting variable y in the population can be calculated with the Horwitz-Thompson estimator as

$$\hat{T} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}, \tag{2.1}$$

where π_i is the inclusion probability of unit i. The variance estimator for the Horwitz-Thompson estimator is

$$\mathrm{var}(\hat{T}_{HT}) = \frac{1}{2} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2, \tag{2.2}$$

where π_{ij} is the probability of units i and j both being included in the sample at the same time, provided all the inclusion probabilities are above zero. All the estimators developed for different sampling designs can be derived from these general formulae.

Although the variance of the values of y in the population affects the estimates, the variance of an estimator in design-based inference is not statistically dependent on the distribution of y (Gregoire 1998). The expected value of an estimator and the variance of the estimators are based on the variation in the estimates (i.e. values of the estimators) between all the possible samples s in the set S_n. Since all the randomness comes from the selection of the sampling units, not from the population itself, the values of y in the population are treated as fixed but unknown (for a different situation, see Chapter 3). This also means that design-based inference is independent of the potential spatial correlation between the sampling units (Gregoire 1998). It is enough that the units are not correlated in terms of their selection.

One estimator of the mean value of y in the population is the sample mean, the expected value of which can be calculated as

$$E(\hat{\bar{y}}_s) = \sum_{S_n} \hat{\bar{y}}_s p(s). \tag{2.3}$$

This is the weighted mean of all possible sample means, weighted with the probability $p(s)$ of selecting each sample s. An estimator is design-unbiased, i.e. unbiased under a certain sampling design, if and only if its expected value coincides with the true population value. The bias of an estimator for the mean value is then defined as (see Schreuder et al. 1993 p. 21)

$$B(\hat{\bar{y}}_s) = E(\hat{\bar{y}}_s) - \bar{Y}, \tag{2.4}$$

where $\overline{Y}$ is the true population mean. The variance of the estimator is

$$V(\hat{\bar{y}}_s) = E\left\{\hat{\bar{y}}_s - E(\hat{\bar{y}}_s)\right\}^2 = \sum_{S_n}\left(\hat{\bar{y}}_s - E(\hat{\bar{y}}_s)\right)^2 p(s) \tag{2.5}$$

and the mean square error (MSE) of the estimator is

$$MSE(\hat{\bar{y}}_s) = E\left(\hat{\bar{y}}_s - \overline{Y}\right)^2 = \sum_{S_n}\left(\hat{\bar{y}}_s - \overline{Y}\right)^2 p(s) = V(\hat{\bar{y}}_s) + \left\{B(\hat{\bar{y}}_s)\right\}^2. \tag{2.6}$$

More generally, the expected value, bias and variance can be defined in the same way for any estimator $\hat{Y}_s$ based on observed values y_i from sample s (Särndal et al. 1992 p. 40).

In typical sampling situations the population is easy to define and finite, whereas in forest inventories the population may be infinite and is often difficult to define. In many cases the population to be inventoried is assumed to be that of sample plots, i.e. the sampling unit is a sample plot (Shiver and Borders 1996 p. 59). This is justified by the fact that the interest lies in the forest characteristics per unit area, such as volume per hectare and so on. Consequently, the size of the population is often assumed to be the number of similar-sized sample plots that will fit into the area, i.e. the total area divided by the plot area. This definition is the easiest to operate with.

Such a definition is not adequate on all occasions, however. For instance, when circular sample plots are used it is not possible to divide the area into mutually exclusive plots that cover the whole of it. In point (or plotless) sampling with a relascope or angle gauge, the size of the sample plot is zero, so that the number of potential sampling units per unit area in infinite, as is the size of the population. When the aim is to estimate the forest area, the population is defined based on plots or points.

In addition to stand-level characteristics, tree-level characteristics such as the mean diameter or number of stems may be of interest, so that the most natural population would be the population of trees. On some occasions the trees may also be the primary sampling units, e.g. with sampling proportional to size in a stand, for example for relascope sampling. If the sampling units are trees, the size of the population is practically never known. One definition that would be adequate in many situations is that the population consists of trees but the sampling unit is a plot.

2.2 SIMPLE RANDOM SAMPLING

Simple random sampling, SRS, can be done either with or without replacement. Sampling with replacement means that each unit can be selected several times. This method is not very important in practice, but it is of theoretical importance as many formulae for this design are very simple. The probability *p(s)* of selecting a given

sample s of size n out of the population of size N is $p(s) = 1/N^n$ (Särndal et al. 1992 p. 50), as there are N^n samples of size n that can be drawn from the population. In this case, the inclusion probability can be calculated as one minus the probability of not drawing a certain unit i, i.e. $\pi_i = 1-(1-1/N)^n$, and the probability of selecting two units i and j is $\pi_{ij} = 1-2(1-1/N)^n + (1-2/N)^n$ (Särndal et al. 1992 p. 50).

In sampling without replacement, on the other hand, each unit can be selected only once and the selection and inclusion probabilities are not quite as easy to calculate as in the earlier case. The number of possible samples is nevertheless $\frac{N!}{(N-n)!n!}$ and the probability of each of these being selected is its inverse

$$p(s) = \frac{(N-n)!\,n!}{N!}. \tag{2.7}$$

The inclusion probability for any unit i is $\pi_i = n/N$ and the probability of selecting two units i and j is $\pi_{ij} = \left(n(n-1)/N(N-1)\right)$ (Särndal et al. 1992 p. 66).

The estimators for the mean and its variance can then be derived from these probabilities with (2.1) and (2.2). Although the Horwitz-Thompson estimator is for the total value, the estimators for the total value $\hat{T}$ and mean $\hat{\bar{y}}$ are related according to

$$\hat{T} = N\hat{\bar{y}} \tag{2.8}$$

and their variances according to

$$\mathrm{var}(\hat{T}) = N^2\,\mathrm{var}(\hat{\bar{y}}), \tag{2.9}$$

assuming in both cases that the population size N is known.

One estimator for the population mean in SRS is the sample mean

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{2.10}$$

where y_i is the value of the variable of interest for unit i. For sampling without replacement, an estimator for its variance is

$$\mathrm{var}(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}, \tag{2.11}$$

while that for sampling with replacement is

$$\text{var}(\hat{\bar{y}}) = \frac{s_y^2}{n}, \tag{2.12}$$

where s_y^2 is the sample variance of y. The true variance of the sample means could be calculated if the population variance S_y^2 were known, but usually it is not. Therefore, estimators of the sampling variances are given in this chapter and not formulae for the true sampling variances. The formula for sampling with replacement (eq. 2.12) can also be used if the population is assumed to be infinite or very large.

The standard error of the mean is

$$s_e = \sqrt{\text{var}(\hat{\bar{y}})}\ . \tag{2.13}$$

This describes how much the sample means from different samples vary around the true mean. In the case of design-based sampling, the standard error can be interpreted as implying that the sample mean deviates less than $\pm 1.96 s_e$ from the true mean in 95 samples out of 100 selected. This is based on the assumption that the distribution of sample means is normal. The statements concerning the accuracy of sampling are correspondingly based on the assumption of repeated sampling.

The proportion of a certain class i can be estimated from

$$\hat{p}_i = \frac{n_i}{n}, \tag{2.14}$$

where n_i is the number of sampling units belonging to class i. Its variance is estimated as

$$\text{var}(\hat{p}_i) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}_i(1 - \hat{p}_i)}{n-1}\ . \tag{2.15}$$

2.3 DETERMINING THE SAMPLE SIZE

The number of units to be selected is obviously limited by the budget. However, the minimum amount of units that should be selected depends on the requirements on the accuracy of the estimator. The sample size n can be calculated from the probability that the deviation of the sample mean from the true mean μ is less than a given d with probability $1-\alpha$, $P(|\bar{y} - \mu| \leq d) = 1 - \alpha$.

If it is assumed that the sample means follow a normal distribution, an equation (for sampling without replacement)

$$d = z_{\alpha/2}\sqrt{\frac{N - n}{N}}\frac{S}{\sqrt{n}} \tag{2.16}$$

can be obtained from the probability statement, where $z_{\alpha/2}$ is the critical value for the normal distribution, i.e. the value above which a normally distributed value is located with a probability *α/2*. Then, *n* is

$$n = \left(\frac{z_{\alpha/2}}{d}\right)^2 \left(\frac{N - n}{N}\right) S^2 . \tag{2.17}$$

In practice, the sample size is first solved for an infinite population (to avoid *n* on both sides of the equation):

$$n_0 = \left(\frac{z_{\alpha/2}}{d}\right)^2 S^2 \tag{2.18}$$

and then, based on this, for finite populations as

$$n = \frac{n_0}{\left(1 + \frac{n_0}{N}\right)} . \tag{2.19}$$

The equation requires knowledge of the population variance S^2, which is typically unknown. It can be estimated, however, from previous surveys or a pilot study. If the estimate for the variance is calculated from a sample, Student's t-distribution is used instead of the normal distribution and a corresponding critical value, $t_{\alpha/2}$, is used.

In the case of proportions, an upper bound for the sample size is obtained by assuming p_i to be 0.5, which gives the maximum variance.

2.4 SYSTEMATIC SAMPLING

In systematic sampling, every k^{th} unit is typically selected into the sample. This means that there has to be a predefined order among the sampling units. It also means that the number of possible samples is only *k*. The predefined order is typically easy in a forest inventory, as the plots are always perfectly ordered with respect to their coordinates. In forest inventory, the number of possible samples may be infinite, if point sampling is applied. If the plots have fixed size, and they are not allowed to overlap, the size of the plot defines the number of possible samples. Furthermore, when the first unit is selected, the selection of the other sampling units follows automatically. Thus the units are not independently selected, and no design-based estimators exist for the standard errors of systematic sampling.

In theory, standard errors can be calculated for systematic sampling from

the variance between all the k samples. It can be proved that the standard error depends on the inner correlation ω:

$$\omega = 1 - \frac{n\sigma_w^2}{(n-1)\sigma^2} = \frac{\sigma_b^2 - \sigma^2/n}{(n-1)\sigma^2/n}, \tag{2.20}$$

where

$$\sigma_w^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \bar{y}_j)^2}{nk} \text{ and } \sigma_b^2 = \frac{\sum_{j=1}^{k}(\bar{y}_j - \bar{y})^2}{k}.$$

Since $\sigma^2 = \sigma_w^2 + \sigma_b^2$, the variance of the mean is

$$var(\hat{\bar{y}}) = \frac{\sigma^2}{n}[1 + (n-1)\omega] = \sigma_b^2. \tag{2.21}$$

Therefore, the larger the within-sample variance σ_w^2 is compared with the total variance σ^2, the smaller the standard error of systematic sampling. A heterogeneous sample represents the population better. On the other hand, the smaller the between-sample variation σ_b^2 is, the smaller the standard error. If ω is negative, systematic sampling is more efficient than SRS. Unfortunately, (2.21) cannot be used as an estimator for variance if only one sample is measured; it only can be used for theoretical analysis.

In many cases SRS estimators are also used in systematic sampling. This is reasonable if the order of the units is completely random, but if there is a trend in the population, the SRS standard error overestimates the standard error of systematic sampling. On the other hand, if there is periodical variation in the population, systematic sampling may be highly inefficient (Särndal et al. 1992 p. 82).

Apart from using SRS estimators, the standard error of a systematic sample can be calculated 1) by taking several small samples and determining the variation between them (Chapter 10), 2) by using approximate formulae (Chapter 10), or 3) by using formulae from stratified sampling (section 2.5). The sample is then divided into several strata along the trend.

In a forest inventory, there may be a trend within any one forest stand if the site index increases from one side to the other, for example. There is also a large-scale trend in a north-south direction in Finland due to changes in climate conditions. Periodical variation within a stand might be due to the ditch network, and large-scale periodic variation could be due to hills etc.

Example 2.1 Heikki Surakka

In an inventory of a 100-hectare forested area in Southern Finland the sample plots were laid out on a square grid where both the line interval and the plot interval was 100 metres. Altogether 102 circular and point sample plots were measured. If the average diameter at breast height was less than 8 cm, the trees were measured on a circular sample plot of radius 2.52 metres (area 20 m^2). Otherwise point sampling was used with a basal area factor of 2, and if there were also understorey trees, they were measured on a circular plot of size 20 m^2 as well. An estimate for stem volume per hectare was calculated for each sample plot.

SRS estimators can always be used for population means and totals, and in this example, an SRS estimator was also used for the sampling variance. The mean stem volume per hectare was

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{102}\sum_{i=1}^{102} y_i = 193 m^3/ha \,,$$

where y_i is the stem volume per hectare of plot i. In order to calculate the standard error of the mean, we first have to determine the population and sample sizes (N and n). In general, if we had only circular plots or fixed-area plots of any other shape, then n could be simply determined as the number of sample plots and N as [total area] divided by [sample plot size], i.e. the number of sample plots located and shaped so that whole area is covered with no overlapping. As the size of a point sample plot is variable, it is impossible to determine N and n accurately, but we can estimate an approximate sampling ratio $f=n/N$:

$$f = \frac{\sum_{i=1}^{n} a_i}{A} = \frac{\sum_{i=1}^{102} a_i}{100.0} = 0.0200 \,,$$

where a_i is the area of the circle from which the basal area median tree is counted as belonging to plot i and A is the total area. We can then calculate the sample variance, which is an estimator of the population variance:

$$s_y{}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i{}^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right) = \frac{1}{101}\left(\sum_{i=1}^{102} y_i{}^2 - \frac{\left(\sum_{i=1}^{102} y_i\right)^2}{102}\right) = 12601\left(m^3/ha\right)^2 .$$

The standard error of the mean stem volume per hectare is

$$s_{\hat{\bar{y}}} = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_y{}^2}{n}} = \sqrt{(1 - 0.0200)\frac{12601}{102}} = \sqrt{121.07} = 11.0 m^3 / ha \, .$$

In this case, the sampling ratio is so small that the finite population correction factor *1 - n/N* can be ignored.

The estimate for the total stem volume is

$$\hat{T} = A\hat{\bar{y}} = 100.0 \cdot 193.25 = 19325 m^3$$

and its standard error

$$s_{\hat{T}} = \sqrt{\mathrm{var}\left(\hat{T}\right)} = \sqrt{A^2 \, \mathrm{var}\left(\hat{\bar{y}}\right)} = \sqrt{100{,}0^2 \cdot 121.07} = \sqrt{1210700} = 1100 m^3 \, .$$

The confidence interval for the true population mean is

$$\left(\hat{\bar{y}} - z_{(\alpha/2)} s_e ; \hat{\bar{y}} + z_{(\alpha/2)} s_e\right),$$

where $z_{(\alpha/2)}$ is a value from the normal distribution with a confidence level α. Thus the 95% confidence interval for the true mean stem volume per hectare would be

$$(193.25 - 1.96 \cdot 11.00; 193.25 + 1.96 \cdot 11.00) = \left(172 m^3 / ha; 215 m^3 / ha\right).$$

Example 2.2

The proportion of the population that is of a certain character is often a matter of interest, for example the proportion of a given tree species or a given site type. Let us assume that we now want to know the proportion of mineral sites in this 100-hectare inventory area. A decision has to be made for every sample plot regarding its soil class, i.e. it is either a mineral site, spruce swamp or pine bog. The estimate for the proportion of mineral sites is

$$\hat{p}_{ms} = \frac{n_{ms}}{n} = \frac{81}{102} = 0.79 \, ,$$

where n_{ms} is the number of mineral site sample plots and n is the total number of plots. Thus mineral sites make up 79% of the inventory area and mires 21%.

The standard error is estimated as follows:

$$s_{\hat{p}_{ms}} = \sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}_{ms}\left(1-\hat{p}_{ms}\right)}{n-1}} = \sqrt{\left(1-0.0200\right)\frac{0.79412\cdot\left(1-0.79412\right)}{102-1}}$$

$$= \sqrt{0.0015864} = 0.040$$

Table 2.1 Plot data for the inventory area.

Plot ID	Soil class[a]	Volume m^3/ha	Basal area m^2/ha	Plot area[b] m^2	Stratum[c]	Plot ID	Soil class	Volume m^3/ha	Basal area m^2/ha	Plot area m^2	Stratum
1	1	155	26	71	2	52	1	236	34	90	2
2	1	242	32	118	2	53	1	217	34	83	2
3	1	108	18	65	2	54	3	157	16	310	3
4	2	269	26	335	2	55	3	135	22	75	2
5	1	114	18	74	2	56	3	284	32	235	3
6	1	93	16	64	2	57	1	33	2	20	1
7	1	201	32	88	2	58	1	74	10	126	3
8	1	80	12	115	1	59	2	430	40	317	3
9	1	66	14	37	1	60	1	340	30	361	3
10	1	363	34	316	3	61	1	315	28	359	3
11	1	171	22	163	2	62	3	93	18	42	2
12	1	217	26	135	2	63	3	23	4	93	2
13	1	36	13	20	1	64	1	45	5	20	1
14	1	176	24	118	2	65	1	360	42	159	3
15	1	278	32	178	3	66	1	181	18	209	3
16	1	210	22	267	3	67	1	330	30	467	3
17	1	20	3	20	1	68	2	224	34	84	2
18	1	347	32	405	3	69	2	209	30	106	3
19	1	260	32	177	3	70	1	371	38	208	3
20	1	164	14	406	3	71	1	248	34	107	2
21	1	149	26	62	2	72	1	247	38	80	2
22	2	25	6	20	1	73	1	445	38	385	3
23	1	407	44	212	3	74	1	130	20	85	2
24	2	330	32	280	3	75	1	223	22	256	3
25	2	368	36	286	3	76	1	408	38	448	3
26	1	114	14	173	3	77	1	241	24	289	3
27	1	221	18	491	3	78	1	89	16	60	1
28	1	310	26	406	3	79	1	278	30	219	3
29	1	85	19	20	1	80	1	355	30	445	3
30	1	344	34	276	3	81	3	66	8	240	3
31	3	288	32	213	2	82	1	247	26	230	3
32	1	141	20	154	3	83	1	136	22	67	2
33	1	224	24	235	3	84	1	166	22	147	2
34	1	297	28	278	3	85	1	151	24	75	2
35	1	212	22	271	3	86	1	164	22	118	2
36	3	227	26	184	2	87	1	119	28	26	2
37	2	208	18	491	3	88	1	169	24	105	2
38	1	263	30	224	3	89	1	0	0	20	1
39	1	0	0	20	1	90	2	164	22	104	2
40	1	242	24	240	3	91	1	112	20	59	2
41	1	392	34	357	3	92	1	63	6	388	1
42	1	255	24	342	3	93	1	109	10	489	2
43	2	196	22	199	3	94	3	36	8	31	1
44	1	130	20	86	3	95	1	140	22	80	2
45	1	0	0	20	1	96	1	215	22	299	2
46	1	339	32	275	3	97	1	64	6	427	1
47	1	386	36	304	3	98	1	59	12	40	1
48	2	224	22	332	3	99	1	103	16	83	2
49	2	255	30	177	3	100	1	130	14	256	2
50	1	124	18	123	3	101	1	37	4	296	1
51	1	195	20	272	3	102	1	18	2	491	1

[a]Soil class: 1=Mineral site
2=Spruce swamp
3=Pine bog

[b]Plot area: In the case of point sample plot it is the area of the circle from which the basal area median tree is counted as belonging to plot, otherwise 20 m^2.

[c]Stratum: 1=Open areas, seedling stands and stands of seed trees
2=Middle aged stands
3=Mature stands

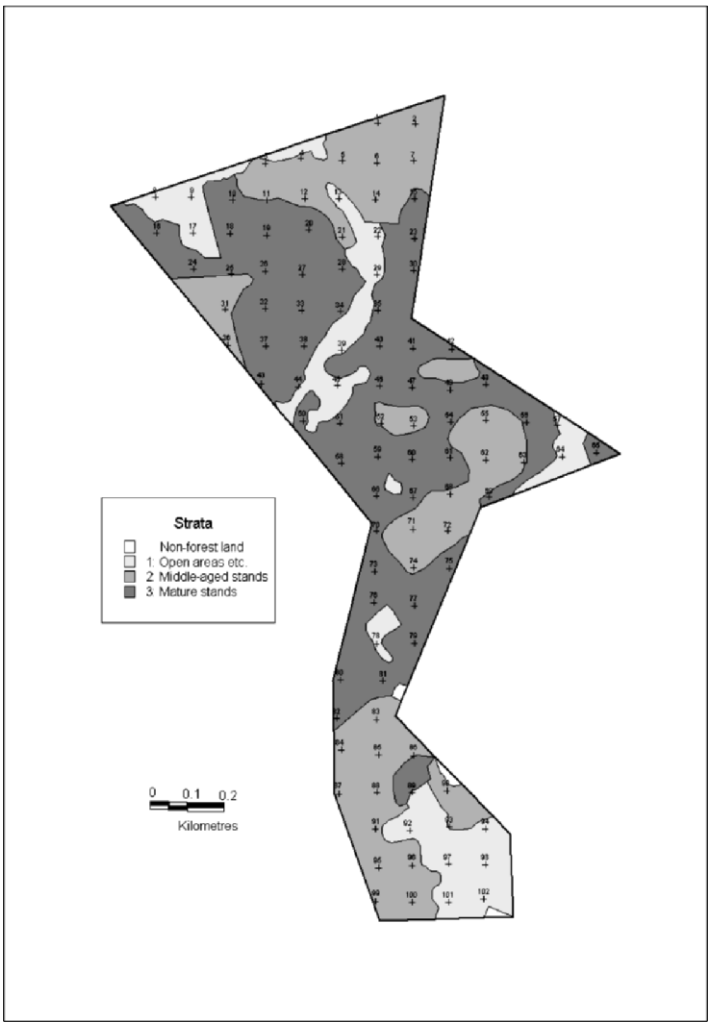

Figure 2.1 The inventory area, strata and sample plot locations.

2.5 STRATIFIED SAMPLING

In stratified sampling there exists certain auxiliary information according to which the population can be divided to homogeneous groups or strata. Stratified sampling is in most cases more efficient than SRS, meaning that the standard errors are smaller. Each stratum can be interpreted as a small sub-population, for which the estimates are calculated using suitable estimators. Typically, selections within strata are performed using SRS, but systematic sampling, for instance, can also be used. The population values are then obtained as weighted averages of the sub-population values as

$$\hat{\bar{y}}_{STR} = \sum_{h=1}^{L} W_h \hat{\bar{y}}_h , \tag{2.22}$$

with an estimator of variance

$$\text{var}(\hat{\bar{y}}_{STR}) = \sum_{h=1}^{L} W_h^2 \text{var}(\hat{\bar{y}}_h) = \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h} - \sum_{h=1}^{L} W_h \frac{s_h^2}{N} , \tag{2.23}$$

where L is the number of strata, W_h is the proportion of stratum h and s_h^2 is the sample variance within stratum h :

$$s_h^2 = \sum_{i=1}^{n_h} \frac{(y_{hi} - \bar{y}_h)^2}{n_h - 1} . \tag{2.24}$$

Stratified sampling generally becomes more efficient with increasing homogeneity within the strata, as the weighted averages of small variances are obviously smaller than those of large variances, although the allocation of sampling units to strata also has an effect on the variance.

The allocation of sampling units to strata h can take place in several ways, being either constant, proportional, Neyman (optimal) or optimal with respect to costs. Constant allocation means that, a constant number of units is selected from each stratum. In proportional allocation, the proportion of selected units $f = n_h/N_h$ is similar in each stratum h, while in Neyman allocation the number of units selected depends on both the size of the stratum and the variation within it. This method is more efficient than the former ones if the variation varies among strata, meaning that it gives the smallest standard error for a given n. The sample size in each stratum is then

$$n_h = n \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} . \tag{2.25}$$

If the measurement costs vary between the strata, this can be accounted for by choosing

$$n_h = n\frac{W_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h/\sqrt{c_h}} \tag{2.26}$$

where c_h is the measurement cost in stratum h. This allocation gives the smallest standard error for a given budget.

The stratification can also be performed after the sample has been selected (=post-stratification). In this case it cannot be used for allocating the sample optimally, but the estimators of stratified sampling can be used. This could be useful if post-stratification is less costly than stratification before sampling for some reason. In the case of known stratum sizes and proportional allocation, post-stratification is almost as efficient as "normal" stratification (Särndal et al. 1992 p. 265). If the stratum sizes are not known, this will introduce additional error (see Chapter 14).

Example 2.3 Heikki Surakka

The same 100-hectare area was then post-stratified with the help of aerial photographs. Three strata were defined:

Stratum	*A*	*n*
Open areas, seedling stands and stands with seed trees	18.0	18
Middle-aged stands	33.3	35
Mature stands	48.7	49

The mean stem volumes per hectare for each stratum are

$$\hat{\bar{y}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i = \frac{1}{18}\sum_{i=1}^{18} y_i = 42 m^3/ha$$

$$\hat{\bar{y}}_2 = 167 m^3/ha \text{ and}$$

$$\hat{\bar{y}}_3 = 268 m^3/ha$$

and the standard errors of the mean stem volumes per hectare are

$$s_{\hat{\bar{y}}_1} = \sqrt{\left(1-\frac{n_1}{N_1}\right)\frac{s_{y_1}{}^2}{n_1}} = \sqrt{(1-0.01147)\frac{827.35}{18}} = \sqrt{45.437} = 6.7 m^3/ha$$

$$s_{\hat{\bar{y}}_2} = 10.0 m^3/ha \text{ and}$$

$$s_{\hat{\bar{y}}_3} = 13.3 m^3/ha\ .$$

Total stem volumes for each stratum are

$$\hat{T}_1 = A_1 \hat{\bar{y}}_1 = 18.0 \cdot 41.976 = 756 m^3$$

$$\hat{T}_2 = 5557 m^3 \text{ and}$$

$$\hat{T}_3 = 13033 m^3$$

and the standard errors of the total stem volumes are

$$s_{\hat{T}_1} = \sqrt{\text{var}\left(\hat{T}_1\right)} = \sqrt{A_1{}^2\ \text{var}\left(\hat{\bar{y}}_1\right)} = \sqrt{18.0^2 \cdot 45.437} = \sqrt{14722} = 121 m^3$$

$$s_{\hat{T}_2} = 334 m^3 \text{ and}$$

$$s_{\hat{T}_3} = 649 m^3\ .$$

The mean stem volume per hectare for the whole area is

$$\hat{\bar{y}}_{str} = \sum_{h=1}^{3} W_h \hat{\bar{y}}_h = 0.180 \cdot 41.976 + 0.333 \cdot 166.84 + 0.486 \cdot 267.67 = 193 m^3/ha$$

and its standard error

$$s_{\hat{\bar{y}}_{str}} = \sqrt{\sum_{h=1}^{3} W_h{}^2\ \text{var}\left(\hat{\bar{y}}_h\right)} = \sqrt{0.180^2 \cdot 45.437 + 0.333^2 \cdot 100.27 + 0.487^2 \cdot 177.72}$$
$$= \sqrt{54.734} = 7.4 m^3/ha.$$

The total stem volume for the whole area is

$$\hat{T}_{str} = \sum_{h=1}^{3} \hat{T}_h = 755.6 + 5557.1 + 13033.4 = 19346 m^3$$

and its standard error

$$s_{\hat{T}_{str}} = \sqrt{\sum_{h=1}^{3} \mathrm{var}\left(\hat{T}_h\right)} = \sqrt{14722 + 111246 + 421372} = \sqrt{547340} = 740 m^3.$$

Example 2.4

In this example we will demonstrate how to determine the sample size. The question is derived from the previous examples. How many sample plots would be needed in normal systematic sampling to have the same standard error of the mean stem volume per hectare as in stratified sampling?

First we determine the allowable deviation of the sample mean from the population mean. The standard error of the mean stem volume per hectare in stratified sampling was 7.4 m^3/ha. If we use a 95% confidence level, the confidence interval and the allowable deviation will be $\pm 1.96 \cdot 7.4$ m^3/ha.

To determine the sample size, the population variance should be known. As it is not known, it has to be estimated from the sample.

The sample size needed for an infinite population is

$$n_0 = \left(\frac{t_{\alpha/2}}{d}\right)^2 s_y{}^2 = \left(\frac{1.960}{1.960 \cdot 7.3983}\right)^2 \cdot 12601 = 230\,,$$

but for finite populations, the size of the population should be known. We can estimate this by dividing the total area by the average plot area:

$$\hat{N} = \frac{A}{\bar{a}} = \frac{100.0}{0.01961} = 5098\,.$$

The sample size for a finite population is now

$$n = \frac{n_0}{\left(1 + \frac{n_0}{N}\right)} = \frac{230.22}{\left(1 + \frac{230.22}{5098.3}\right)} = 220\,.$$

2.6 CLUSTER SAMPLING

Cluster sampling is used when the population can be divided to separate groups. In forest inventory these are typically groups of sample plots located near each other or groups of trees located near each other. (Each sample plot could also be interpreted as a cluster of trees if the mean values for trees were of interest.) In cluster sampling the clusters are the basic sampling units. In one-stage cluster sampling, all the units within a selected cluster are measured, while in multi-stage cluster sampling another sample is selected from within the cluster. The sampling units at different stages vary.

Cluster sampling is not usually as efficient as the other selection methods given a fixed size of sample n. This is because the sampling units in one cluster may be correlated, i.e. the new information resulting from measuring a new unit is less than it would be if the units were independent. The usefulness of cluster sampling is based on cost efficiency: it is usually possible to measure more units with the same budget when they are located in clusters. In a forest inventory a cluster design will reduce walking distances in the forest. It is also typical for the clusters to be laid out in a systematic fashion, the groups of plots forming a line or a rectangular of a certain size, and for this reason the definition of a cluster is also somewhat more complicated in forestry than for clusters formed by families, classes or schools as in the social sciences.

The estimator for the population mean is the mean of the cluster means:

$$\hat{\bar{y}}_{CLU} = \sum_{\alpha=1}^{a} \frac{\hat{\bar{y}}_{\alpha}}{a}, \tag{2.27}$$

where a is the number of clusters selected and $\hat{\bar{y}}_{\alpha}$ is the mean in cluster α. If the clusters are of different sizes, this formula might be biased. The bias occurs if the variable of interest is dependent on the cluster size, e.g. if it has larger values in larger clusters. The mean estimator should then be calculated as a weighted mean of the clusters. If y is independent of cluster size, the results are unbiased, although equal size is assumed (Cochran 1977). The variance estimator of the mean is

$$\text{var}(\hat{\bar{y}}_{CLU}) = \left(1 - \frac{a}{A}\right) \sum_{\alpha=1}^{a} \frac{(\hat{\bar{y}}_{\alpha} - \hat{\bar{y}}_{CLU})^2}{a(a-1)}, \tag{2.28}$$

where A is the total number of clusters. The efficiency of cluster sampling increases as the variation between cluster means decreases, i.e. the more homogeneous the clusters are. This, on the other hand, depends on the inner heterogeneity of the clusters: the larger the amount of the population variation that is within-cluster variation, the better. The principle is similar to that of systematic sampling presented in section (2.4). This can be expressed using the intra-cluster correlation ω

$$\varpi = \frac{\sigma_b^2 - \sigma_w^2/(B-1)}{\sigma^2}, \tag{2.29}$$

where B is the size of a cluster. The variance in cluster sampling can then be presented as (Cochran 1977, Tokola and Shrestra 1999)

$$\text{var}(\hat{\bar{y}}_{CLU}) = \left(1 - \frac{a}{A}\right)\frac{S^2}{aB}\left[1 + (B-1)\varpi\right]. \tag{2.30}$$

Thus the smaller the intra-cluster correlation is, the smaller the variance.

In two-stage cluster sampling, the variance of the mean is larger, because the second-stage sample also contains sampling error. The variance is

$$v(\hat{\bar{y}}_{CLU}) = \left(1 - \frac{a}{A}\right)\frac{s_b^2}{a} + \left(1 - \frac{b}{B}\right)\frac{a}{A}\frac{s_w^2}{ab}, \tag{2.31}$$

where

$$s_b^2 = \frac{1}{a-1}\sum_{\alpha=1}^{a}(\hat{\bar{y}}_\alpha - \hat{\bar{y}}_{CLU})^2 \text{ and}$$

$$s_w^2 = \frac{1}{a(b-1)}\sum_{\alpha=1}^{a}\sum_{\beta=1}^{b}(y_{\alpha\beta} - \hat{\bar{y}}_\alpha)^2,$$

and where B is the population size within a cluster and b is the corresponding sample size.

2.7 RATIO AND REGRESSION ESTIMATORS

In a stratified inventory information on some auxiliary variables is used both to plan the sampling design (e.g. allocation) and for estimation, or only for estimation (post-stratification). Stratification is not the only way to use auxiliary information, however, as it can be used at the design stage, e.g. in sampling proportional to size (section 2.8). It can also be used at the estimation stage in ratio or regression estimators, so that the standard error of the estimators can be reduced using information on a variable x which is known for each sampling unit in the population. The estimation is based on the relationship between the variables x and y. In ratio estimation, a model that goes through the origin is applied. If this model does not apply, regression estimator is more suitable. The ratio estimator for the mean is

$$\hat{\bar{y}}_{rat} = \frac{\bar{y}}{\bar{x}} \bar{X} = r\bar{X} \,, \tag{2.32}$$

where $\bar{X}$ is the mean of a variable x in the population and $\bar{x}$ in the sample . Ratio estimators are usually biased, and thus the root mean square error (RMSE) should be used instead of the standard error. The relative bias nevertheless decreases as a function of sample size, so that in large samples (at least more than 30 units) the accuracy of the mean estimator can be approximated as (Cochran 1977 p. 155)

$$\text{var}(\hat{\bar{y}}_{rat}) \cong \left(1 - \frac{n}{N}\right) \sum_{i=1}^{n} \frac{(y_i - rx_i)^2}{n(n-1)} \,. \tag{2.33}$$

The ratio estimator is more efficient the larger the correlation between x and y relative to the ratio of the coefficients of variation. It is worthwhile using the ratio estimator if

$$corr(x, y) > \frac{1}{2} \frac{CV(x)}{CV(y)} \,. \tag{2.34}$$

The (simple linear) regression estimator for the mean value is

$$\hat{\bar{y}}_{reg} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x}) \,, \tag{2.35}$$

where $\hat{\beta}$ is the OLS coefficient of x for the model, which predicts the population mean of y based on the sample means. In a sampling context, the constant of the model is not usually presented, but the formula for the constant, $\hat{\alpha} = \bar{y} + \hat{\beta}\bar{x}$, is embedded in the equation. The model is more efficient the larger the correlation between x and y. The variance of the regression estimator can be estimated as

$$\text{var}(\hat{\bar{y}}_{reg}) = \left(1 - \frac{n}{N}\right) \sum_{i=1}^{n} \frac{\left[(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\right]^2}{n(n-2)} \,. \tag{2.36}$$

Example 2.5 Heikki Surakka

There were also data for the same 100-hectare area that contained only basal area measurements. These had been collected from a very dense grid with a basal area factor of 1. The sample covered the area so well that the estimates (mean and total basal areas) can be regarded as true, as if every tree included in the area had been measured. We will next use the basal area as an auxiliary variable and determine the

mean stem volume per hectare by ratio estimation. There is a very high correlation between stem volume and basal area, and the relationship is almost linear and goes through origin.

First we calculate the ratio between the estimates:

$$r = \frac{\bar{y}}{\bar{x}} = \frac{193.25}{22.373} = 8.6376m \ .$$

As the true mean basal area was slightly smaller than the estimate for the mean basal area, the ratio estimate for mean stem volume per hectare is smaller than that obtained without ratio estimation:

$$\hat{\bar{y}}_{rat} = r\bar{X} = 8.6376 \cdot 22.254 = 192m^3 / ha$$

Its variance estimate is

$$\begin{aligned} \mathrm{var}\left(\hat{\bar{y}}_{rat}\right) &= \left(1 - \frac{n}{N}\right)\sum_{i=1}^{n}\frac{\left(y_i - rx_i\right)^2}{n(n-1)} = (1 - 0.0200)\sum_{i=1}^{102}\frac{\left(y_i - 8.6376 \cdot x_i\right)^2}{102 \cdot (102-1)} \\ &= 23.280\left(m^3 / ha\right)^2 \end{aligned}$$

and the standard error estimate

$$s_{\hat{\bar{y}}_{rat}} = \sqrt{23.280} = 4.8m^3 / ha \ .$$

The ratio estimate for total stem volume is

$$\hat{T}_{rat} = rT_x = 8.6376 \cdot 22.254 \cdot 100.0 = 19197m^3$$

and its standard error is

$$\begin{aligned} s_{\hat{T}_{rat}} &= \sqrt{\left(1 - \frac{n}{N}\right)T_x{}^2\frac{1}{\bar{X}}\sum_{i=1}^{n}\frac{\left(y_i - rx_i\right)^2}{n(n-1)}} \\ &= \sqrt{(1 - 0.0200) \cdot \frac{(22.254 \cdot 100.0)^2}{22.254}\sum_{i=1}^{102}\frac{\left(y_i - 8.6376 \cdot x_i\right)^2}{102 \cdot 101}} = \sqrt{232802} = 482m^3 \end{aligned}$$

2.8 SAMPLING WITH PROBABILITY PROPORTIONAL TO SIZE

The basic properties of sampling with arbitrary probabilities (2.1) can also be utilized in sampling with probability proportional to size (PPS), such as sampling with a relascope. It is then assumed that unit *i* is selected with the probability kx_i, where *k* is a constant and *x* is a covariate (diameter of a tree in relascope sampling). PPS sampling is more efficient the larger the correlation between *x* and *y*. For perfect correlation the variance in the estimator would be zero (Schreuder et al. 1993 p. 46). PPS sampling might even be less efficient than SRS, however, if the correlation were negative. This could be the case when multiple variables of interest are considered simultaneously, for example, when correlation with one variable (say volume) might give efficient estimates but the estimates for other variables (say health and quality) might not be so good.

In practice, PPS sampling can be performed by ordering the units, calculating the sum of their sizes (say $\sum x_i$), and calculating $\sum x_i / n$. The probability of a unit *i* being selected is then $x_i \big/ \sum x_i$ and a cumulative probability can be calculated for the ordered units. A random number *r* is then picked and each unit with a cumulative probability equal to (or just above) *r, r+1, r+2,...r+n-1* is selected for the sample. Every unit of size greater than $\sum x_i / n$ is then selected with certainty.

2.9 NON-LINEAR ESTIMATORS

The simple variance estimators presented in the above sections are not applicable to non-linear estimators. A typical example of a non-linear estimator is a ratio of two estimators, $\hat{Y}_1 / \hat{Y}_2$. Although the mean value in the whole sample is a linear estimator, the mean in any sub-population is a ratio estimator, because the number of sample units in the sub-population, n_s, is a random variable having a variance that needs to be accounted for.

In such situations, the non-linear estimator needs to be linearized in order to be able to derive an (approximate) formula for the variance estimator. The ratio estimator $g(\hat{\mathbf{Y}}) = \hat{Y}_1 \big/ \hat{Y}_2$ (where $\hat{\mathbf{Y}}$ is the vector of estimators) can be linearized using Taylor series expansion. The variance can then be estimated as

$$\text{var}(g(\hat{\mathbf{Y}})) = \text{var}\left(\sum_{j=1}^{2} \frac{\partial g(\hat{\mathbf{Y}})}{\partial y_j} (\hat{Y}_j - Y_j) \right), \qquad (2.37)$$

giving

$$\mathrm{var}(g(\hat{\mathbf{Y}})) = \frac{1}{\hat{Y}_2^2}\left(\mathrm{var}(\hat{Y}_1) + \left(\frac{\hat{Y}_1}{\hat{Y}_2}\right)^2 \mathrm{var}(\hat{Y}_2) - 2\frac{\hat{Y}_1}{\hat{Y}_2}\mathrm{cov}(\hat{Y}_1, \hat{Y}_2) \right). \qquad (2.38)$$

The Taylor series approach applies in the general case. In most simple cases, however, separate linearization is not required, since the estimators already presented for the variance of a ratio estimator can be used directly (and can be derived using 2.37, for instance: compare 2.38 with 2.41). An example of a non-linear estimator is the case where whole stands (or compartments) are sampling units. Then, as the stands are of different sizes, the sampled area is not known before sampling but is also a random variable. The proportion of the area fulfilling a certain condition may be estimated as

$$\hat{R} = \sum_{i=1}^{n} A_i z_i \Big/ \sum_{i=1}^{n} A_i , \qquad (2.39)$$

where z_i is an indicator variable with the value 1 if the condition is fulfilled and zero otherwise, and A_i is the area of the stand i. The standard error of this estimator can then be approximated as (Cochran 1977)

$$S_{\hat{R}} = \sqrt{\frac{1}{\mu_x^2}\frac{S_u^2}{n}\left(\frac{N-n}{N}\right)} , \qquad (2.40)$$

where

$$S_u^2 = \frac{\sum_{i=1}^{n} A_i^2 z_i^2 + \hat{R}^2 \sum_{i=1}^{n} A_i^2 - 2\hat{R}\sum_{i=1}^{n} A_i^2 z_i}{n-1} \qquad (2.41)$$

and $\mu_x = \frac{A_T}{N}$ is the mean area of the sampling units. Assuming that the mean area is estimated with the sample mean, this formula can be simplified to (Heikkinen, personal information)

$$S_{\hat{R}} = \sqrt{\frac{n}{\left(\sum_{i=1}^{n} A_i\right)^2} \frac{(1-\hat{R})^2 \sum_{z=1} A_i^2 + \hat{R}^2 \sum_{z=0} A_i^2}{n-1} \frac{N-n}{N}} . \qquad (2.42)$$

This means that it is enough to separate the areas of stands fulfilling the condition and those not fulfilling it.

2.10 RESAMPLING

In many cases the capacity of modern computers can be utilized to estimate the sampling variances. There are several methods that work in quite a similar fashion, e.g. jackknife and bootstrap methods. These work as follows:

1) Draw K replicate samples of size n from the original sample (of size n) with replacement.
2) For each replicate, calculate the estimator of interest (e.g. mean or ratio).
3) Estimate the variance of this estimator from the variance between the estimates from replicate samples.

The replicate samples have to be drawn using the original design, i.e. with SRS for simple random sampling, by strata for stratified sampling etc., which means that these simple resampling estimators are not useful for systematic sampling. They nevertheless make variance estimation easy for complex sampling designs and for non-linear estimators.

In Bootstrap method, at least about 100, preferably more than 500 replicates are drawn. The estimator is then

$$\mathrm{var}(\hat{\theta}_{BOOT}) = \frac{\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta}_{BOOT})^2}{K - 1}, \tag{2.43}$$

where

$$\hat{\theta}_{BOOT} = \frac{\sum_{k=1}^{K}\hat{\theta}_k}{K} \tag{2.44}$$

is the mean of the estimates from replicate samples.

For systematic sampling, parametric bootsrap may be an option. In parametric bootstrap, the distribution F of the sampling units is estimated based on the sample data. The distribution could be, for example, normal distribution. The bootstrap samples are then sampled from the estimated distribution. After that, the parametric bootstrap proceeds similarly as the simple bootstrap.

In the jackknife method, jackknife samples $x_{(i)}$ are taken, defined as samples with the i^{th} observation left out, e.g. $x_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. $\hat{\theta}_{(i)}$ is then the i^{th} jackknife replication of the estimator $\hat{\theta}$. From these jackknife replications, pseudo-values are calculated as

$$\theta_i^{(p)} = n\hat{\theta} - (n-1)\hat{\theta}_i . \tag{2.45}$$

The jackknife variance can then be estimated as

$$\text{var}(\hat{\theta}_{JACK}) = \frac{\sum_{i=1}^{n}(\hat{\theta}_i^{(p)} - \hat{\bar{\theta}})^2}{n(n-1)}, \tag{2.46}$$

where

$$\hat{\bar{\theta}} = \frac{\sum_{i=1}^{n}\theta_i^{(p)}}{n}$$

is the mean of the replicate pseudo-values. The original sample estimator could also be used instead (e.g. Pahkinen and Lehtonen 1989). The same variance estimator could also be written without using pseudo-values as (Efron and Tibshirani 1998)

$$\text{var}(\hat{\theta}_{JACK}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\bar{\theta}}_i)^2 . \tag{2.47}$$

In this formula the differences between the jackknife resamples are assumed to be small relative to the variation between the pseudo-values, and therefore the value is inflated by *(n-1)/n* (Efron and Tibshirani 1998).

2.11 SELECTING THE SAMPLING METHOD

Optimal data acquisition can be considered from several points of view. Traditionally, it has been understood as the sampling design giving minimum variance for certain estimates, e.g. mean volume, with a given budget. The inventory costs can be assumed to include fixed costs that are similar for each sampling design, costs per cluster (in a cluster design), costs per plot, and costs per sample tree. The total costs can then be expressed as a function of the number of clusters (m) and plots (n) measured, e.g.

$$C = C_f + C_m m + C_n n . \tag{2.48}$$

If the costs differ between sub-populations, this should also be accounted for in the cost function.

The variance can also be expressed as a function of the number of plots and clusters, even though this is non-linear. If there are several variables of interest, either one variable is chosen or the variances of all of them are combined in some way, e.g. using a weighted sum. Burkhart et al. (1978) suggested that the largest variance or the variance of the most important variable should be used to determine the sample size, while Scott and Köhl (1993) used the accuracy relative to the

desired level of accuracy and averaged across all the variables. The problems can then be presented as a single (non-linear) optimization problem:

$$\begin{aligned} &\textit{Minimize } S = \frac{1}{K}\sum_{k=1}^{K}\frac{s_k(m,n)}{S_k} \\ &\textit{Subject to} \\ &C_t = C(m,n) \end{aligned} \quad , \tag{2.49}$$

where K is the number of variables of interest, S is the desired level of accuracy and s is the actual level of accuracy as a function of number of clusters and plots, C_t is the given cost level and C is the actual cost as a function of number of clusters and plots. There may also be more restrictions. A non-linear optimization problem can be fairly difficult to solve, however, and linear optimization is not applicable (Scott and Köhl 1993).

Expressing the variance of the estimator as a function of the number of plots requires information on the population variance S^2. This is typically obtained from a previous study or from a small preliminary sample. In some cases it is possible to anticipate the population variance mathematically, assuming the locations of trees in the area to follow a known random process such as a Poisson process (Mandallaz and Ye 1999).

Another approach is to minimize the cost function at a given precision level. Constraints can then be given separately for all the variables of interest, e.g. the maximum variance level as

$$s_k < S_k \, . \tag{2.50}$$

It is also possible to minimize the utility function, which is the weighted sum of the inventory costs and MSE (Päivinen 1987). The problem then becomes a non-constrained optimization, which is easier to solve. The problem of weighting the costs and accuracy remains, however.

In some cases it is not necessary to compare methods in an optimization problem of the kind presented above, as the cost-effectiveness of the designs can be compared using the relative efficiency of the alternatives (provided they reflect the same costs). The efficiency of alternative A relative to B can be defined as the variance of alternative A divided by the variance of alternative B (Scott and Köhl 1993, Pahkinen and Lehtonen 1989):

$$DEFF = V(\hat{\bar{y}})_A / V(\hat{\bar{y}})_B \, . \tag{2.51}$$

For a cluster sampling design, for example, the *DEFF* coefficient, assuming a constant number of clusters and constant cluster size, can be derived from formula (2.27) as $DEFF = \left[1 + (B-1)\varpi\right]$, where ω is the intra-cluster correlation. In the case of cluster sampling the latter can be defined as (Cochran 1977)

$$\varpi = \frac{\sigma_b^2 - \sigma^2}{(B-1)\sigma^2}, \tag{2.52}$$

where σ_b^2 is the between-cluster variance, σ^2 is the total variance, and B is the size of the cluster. This enables different cluster shapes such as an L-shaped or square tract, or different plot distances within a cluster, to be compared (Tokola and Shrestra 1999). Similar problems can also be solved using a model forest, e.g. based on a satellite image, in which different designs can be compared (Päivinen 1987).

Another point of view is to optimize the intervals between subsequent forest inventories so that the information is always fresh enough for decision making at minimum cost. In such cases, the database can be updated in terms of forest growth by means of growth and yield models. Silvicultural measures can be ascertained from the forest owner or from aerial images, for instance (Anttila 2002, Hyvönen and Korhonen 2003).

It is evident, however, that the traditional approach based on the mean square errors of the estimates does not necessarily produce any information regarding the usefulness of the measured information for decision-making purposes. This aspect has been studied using cost-plus-loss analysis, in which the expected losses due to non-optimal decisions caused by inaccurate data are added to the total costs of the forest inventory (Hamilton 1978, Burkhart et al. 1978). Ståhl et al. (1994), for example, analysed whether it is more profitable to make accurate inventories at long intervals or moderately accurate inventories at shorter intervals.

The hardest part of cost-plus-loss analysis is to define the expected losses. Holmström et al. (2003), when studying the usefulness of different inventory methods for decision-making, defined the average loss in terms of the net present value (NPV) in the next 5-10 years, where the optimal NPV was taken to be the maximum value without any restrictions. This analysis suggested that extensive field sampling methods were worthwhile in the case of mature stands where the optimal treatment was to be expected in the near future. This kind of approach is a simplification of the true situation, however, as all decisions can be revised. The errors may therefore be non-symmetric in the sense that cuttings proposed for too early a stage can be postponed (provided the necessity can be observed in the field), but those proposed for too late a period cannot be transferred to an earlier occasion.

REFERENCES

Burkhart, H.E., Stuck, R,D., Leuschner, W.A. and Reynolds, M.A. 1978. Allocating inventory resources for multiple-use planning. Canadian Journal of Forest Research 8:100-110.

Cochran, W.G. 1977. Sampling techniques. 3rd edition. Wiley, New York.

Efron, B. and Tibshirani, R.J. 1998. An introduction to the Bootstrap. Monographs on Statistics and Applied Probability 57. Chapman & Hall / CRC. 436 p.

Gregoire, T. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Canadian Journal of Forest Research 28:1429-1447.

Hamilton, D.A. 1978. Specifying precision in natural resource inventories. In: Integrated inventories of renewable resources: proceedings of the workshop. USDA Forest Service, General technical report RM-55:276-281.

Holmström, H., Kallur, H. and Ståhl, G. 2003. Cost-plus-loss analyses of forest inventory strategies based on kNN-assigned reference sample plot data. Silva Fennica 37:381-398.

Loetsch, F., Zöhrer, F. and Haller, K. 1973. Forest Inventory. Vol I-II 436+ 469 p. BLV Verlagsgesellschaft.

Lund, H.G. and Thomas, C.E. 1989. A primer on stand and forest inventory designs. Gen. Tech. Rep. WO-54. Washington, DC; USDA Forest Service. 96 p. http://www.fs.fed.us/rm/ftcol/publications/outofprint/wo_54.pdf

Mandallaz, D. and Ye, R. 1999. Forest inventory with optimal two-phase, two-stage sampling schemes based on the anticipated variance. Canadian Journal of Forest Research 29:1691-1708.

Pahkinen, E. and Lehtonen, R. 1989. Otanta-asetelmat ja tilastollinen analyysi. Gaudeamus. 286 p.

Päivinen, R. 1987. Metsän inventoinnin suunnittelumalli. Summary: A planning model for forest inventory. University of Joensuu, Publications in Sciences N:o 11. 179 p.

Särndal, C-E., Swensson, B. and Wretman, J. 1992. Model assisted survey sampling. Springer-Verlag. 694 p.

Schreuder, H.T., Gregoire, T.G. and Wood, G.B. 1993. Sampling Methods for Multiresource Forest Inventory. John Wiley & Sons. New York. 446 p.

Scott, C.T. and Köhl, M. 1993. A method for comparing sampling design alternatives for extensive inventories. Mitteilungen der Eidgenössischen Forschunganstallt für Wald, Schnee und Landschaft. Band 68, Heft 1.

Shiver, B.D. and Borders, B.E. 1996. Sampling techniques for forest resources inventory. John Wiley & Sons. 356 p.

Ståhl, G. 1994. Optimizing the utility of forest inventory activities. Swedish University of Agricultural Sciences, Department of Biometry and Forest Management. Report 27.

Ståhl, G., Carlson, D. and Bondesson, L. 1994. A method to determine optimal stand data acquisition policies. Forest Science 40:630-649.

Tokola, T. and Shrestra, S.M. 1999. Comparison of cluster-sampling techniques for forest inventory in southern Nepal. Forest Ecology and Management 116:219-231.