Course 3

Text-only version

Introduction to sampling

The interactive version of this lesson is available free of charge at https://elearning.fao.org/

In this course

This course explains the general aspects of sampling in forest inventories.

About the course

This course covers the general aspects of sampling in forest inventories, and aims to introduce the basic concepts and characteristics of a sampling study, as well as provide an overview of the most important components of a national forest inventory (NFI).

Disclaimer: This course is not intended to adequately train experts in the sampling statistics that are needed to plan, analyze, report and correctly interpret sample-based estimates from an NFI.

Who is this course for?

The course is targeted mainly for those who engage in sampling and analysis phases of an NFI, but can be taken by anyone with an interest in the subject. Specifically, this course targets:

1. Forest technicians responsible for implementing their country's NFIs
2. National forest monitoring teams
3. Students and researchers, as curriculum material in forestry schools and academic courses
4. Youth and new generations of foresters

Course structure

There are three lessons in this course.

Lesson 1: About sampling	This lesson introduces the basic concepts and terms associated with statistical sampling. It provides an overview of the relevant characteristics of a sampling study and explains the basics of sampling for a non- expert audience.
Lesson 2: Design elements of a sampling study	This lesson presents the basics of design elements of sampling studies as they are relevant to NFIs, and the concepts to consider while preparing a sampling strategy. It also explains how to calculate the associated sample size.

Lesson 3: Estimation design	This lesson looks into the methods and formulae needed to derive unbiased estimates from the data collected following a certain sampling strategy.

About the series

This course is the third in a series of eight self-paced courses covering various aspects of an NFI. Here's a look at the complete series:

Course	You will learn about
Course 1: Why a national forest inventory?	Goals and purpose of an NFI and how NFIs inform policy- and decision-making in the forest sector.
Course 2: Preparing for a national forest inventory	The planning and work required to set up an efficient NFI or a National Forest Monitoring System (NFMS).
☞ **Course 3: Introduction to sampling**	**(You are currently studying this course)**
Course 4: Introduction to fieldwork	Considerations for fieldwork, plot-level variables and tree-level measurements.
Course 5: Data management in a national forest inventory	Information gathering and data management for NFIs.
Course 6: Quality assurance and quality control in a national forest inventory	QA and QC procedures in forest inventory data collection and management.
Course 7: Elements in data analysis	Typical approaches/calculations in data analyses and related topics.
Course 8: National forest inventory results: Reporting and dissemination	NFI reporting and the importance of reporting in the context of REDD+ actions.

Lesson 1: About sampling

Lesson introduction

This lesson introduces the basic concepts and terms associated with statistical sampling.

It also provides an overview of the relevant characteristics of a sampling study and explains the basics of sampling for a non-expert audience.

Learning objectives

At the end of this lesson, you will be able to:

1. Describe the importance of sampling in forest inventories
2. Define the rationale of statistical sampling.
3. Explain the basic concepts and terminology associated with sampling
4. Explain the importance of accuracy and precision during the estimation process

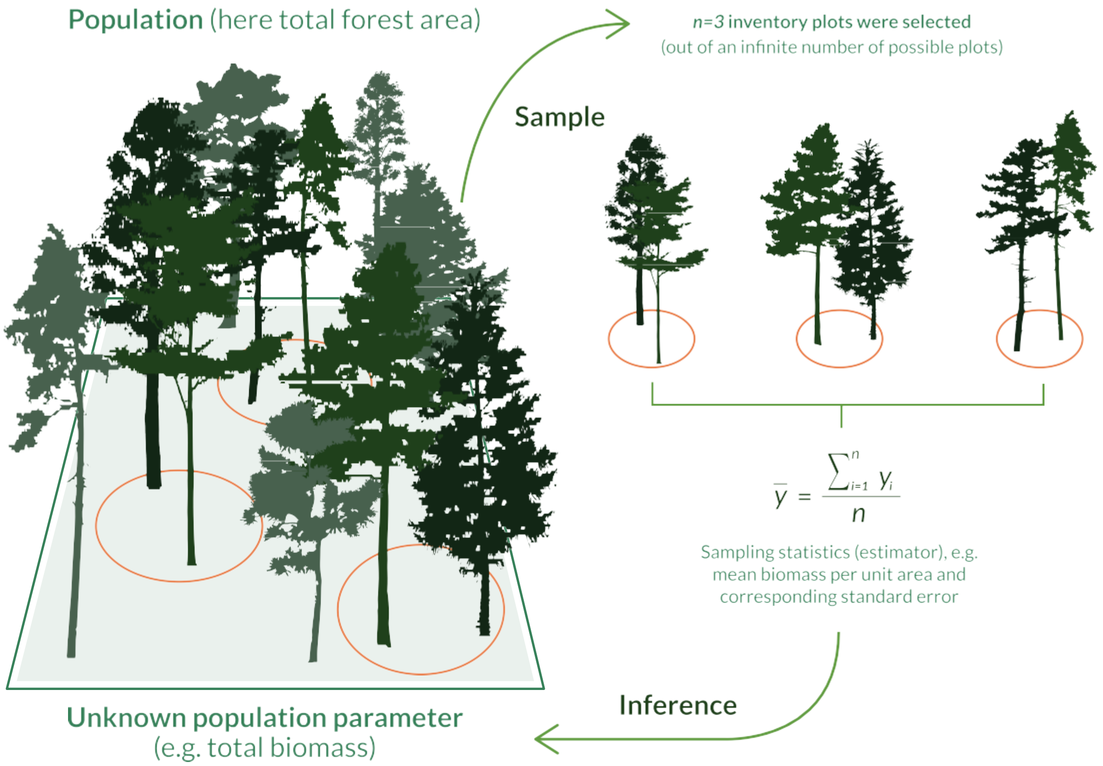
Why is sampling necessary?

Before we start learning about some of the relevant aspects of statistical sampling, let's take a step back and briefly think about the fundamental rationale of sampling studies in general.

Why is sampling such a fundamental concept in the context of forest inventories and monitoring?

The answer to this question is very simple: When we look at the field assessment of core variables, it is **neither feasible nor efficient to observe all the elements** in the forest area of a country. Instead, experts must make inferences about the current state and change of target variables by making observations on relatively small subsets or "samples" of the total forest area—referred to as sample plots.

We may imagine sampling to be similar to opening small windows that allow us to look at parts of the population in order to get an impression of the whole.


Population (here total forest area)
Sample
n=3 inventory plots were selected
(out of an infinite number of possible plots)
$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
Sampling statistics (estimator), e.g. mean biomass per unit area and corresponding standard error
Inference
Unknown population parameter
(e.g. total biomass)

A closer look at these samples of the total area, reveals that the credibility of results of a sampling study is influenced by **the way we select these samples**, the **methods we use to obtain the single observations** and the **applied estimation techniques and calculations**. These are also the three design elements of a sampling study, which need to be planned along statistical considerations—something we will consider in greater detail in the next lesson.

What is statistical sampling?

The selection process ensures that the sampling elements may be considered representative for the population. When sampling follows the rules of statistics, we call it **statistical sampling**. Statistical sampling is largely determined by randomization (and by the absence of subjective or arbitrary considerations), which means that by applying a random selection, we guarantee that each element in the population has a defined and known probability of being selected.

Other criteria for selection, such as **fairness** or **objectivity,** are not enough. Because the selection probabilities play a central role in statistical sampling, these techniques are also called **probabilistic sampling**. By that, representativeness of the sample is guaranteed, and unbiased estimators (i.e. statistically correct estimation approaches) are available for most of the common sampling and observation designs.

Subjective selection of the “most representative” population elements is not statistical sampling and does not allow statistical estimation nor inference.

Imagine sending out experts with the task of finding the “most representative” plot in a forest area (with regard to tree density, species mixture, slope, soil conditions, and so on). It is quite obvious that an estimate we derive from such a plot would exclusively refer to the expert’s choice (while another expert would very likely arrive a a different choice).

While an expert-based guesstimate may be good and close to the target population value, everything depends on the expert and no objective methodological approach is defined that could possibly be repeated by someone else. Statistical sampling, on the contrary, is transparent in all its methodological steps.

Did you know?

A lot of statistical sampling techniques were invented and presented in the context of forest inventories. While sample plots were already in wide use in forestry in the 19th century, a more formalized technique of statistical sampling for large populations was developed—and gradually accepted—as a methodology to produce valid results around 1900 only: in 1895, the Norwegian statistician A.N. Kiaer presented a sampling approach that was then called 'the **representative method**', where '**representativeness**' played a central role.

Forest inventory statisticians at that time made significant contributions to the analysis of systematic line sampling. The first NFI that was rooted in statistical sampling was implemented in 1919-1930 in Norway. This was followed by other Nordic European countries in the early 1920s: Finland in 1921–1924 and Sweden in 1923–1929.

Statistical soundness is one of the major characteristics in statistical sampling, as applied in forest monitoring. It is only by adhering to the principles of statistical sampling that the chosen inventory design can convincingly be defended, when—for example—doubts about the results are expressed.

Deriving inferences from a sample

Descriptive statistics deal with the quantitative characterization of a population of interest, or the domain about which such descriptive statements shall be produced. Sampling aims to derive inferences/conclusions about the total population from a limited number of selected sampling elements. In forest inventories, these elements are typically sample plots, which are subsets of the total forest area.

From the analysis of collected observations on the target variables of these sample plots, we can derive a statistical estimate of the unknown true population parameter. For example: from the biomasses per plot of the *n* sample plots we may produce an estimate of the biomass per hectare for the entire population. It is intuitively clear that we cannot expect that such an estimate is equal to the true value—it is an approximation, and it will vary whenever we take another sample following the same inventory design.

Note

The true values in a population are called **parameters** while the estimates produced from sampling studies are referred to as statistics. The true mean value of a population, the parametric mean, is estimated from the mean value in the sample.

It is important to have this distinction clear**: the true parameters will never be known but estimated by the sampling statistic**. The true value is a constant, one fixed value. The sampling statistic (= the estimated value) is a random variable that can take on many different values—depending on which sample had been selected— and follows a certain distribution.

Let's look at some examples of how the definitions provided above could be applied to a forest inventory for biomass.

1. The population of—for example, trees—is determined by an area, represented by an infinite number of dimensionless point centers where sample plots could be selected.
2. The sample consists of a certain number of plots (sample size) that has been selected following the sampling design.
3. The true value—or population parameter—of, for example, mean biomass, would be the mean biomass estimated over all possible infinite sampling locations in the area. Since we deal only with the current plot design, the true value remains unknown.
4. Using a suitable plot and estimation design, we can derive an unbiased estimate from our sample at hand.

Estimator and estimate

Whenever we talk about an **estimator** in statistical sampling, we mean the calculation algorithm or formula that we use to produce an estimation. In order to produce statistical estimates, the estimator needs to reflect: **the underlying selection process** of sampling elements; and **the way in which the single observations were obtained from the sampling element**.

What is the underlying population concept in forest inventories?

When sample plots are the “sampling elements” that are being selected, the next question that we arrive at is “what is the population then”? In general terms, a population is defined as the set of all sampling elements that theoretically can be selected. In forest inventory, we commonly use sample plots whose location is being selected by selecting a sample point. Then, the population is defined by all possible sample points within the area of interest. How many are those?

The number of points in any area is infinite. Sample points are selected from a continuum, and we call this an infinite population. The population is then 'the total number of possible sample plots in the defined area under study', where these sample plots are installed around the selected sample points.

However, since many variables of interest are aggregates of measurements on single trees found on these sample plots, they will only vary once the composition of included trees changes.

Therefore, for such variables, we can refine the population concept and say: 'the population is composed of all mutually exclusive clusters of trees that have a positive probability to be included jointly by the defined plot design'. The size of this population is not infinite, there is only a finite number of options for joint inclusions of spatially dispersed trees.

Limitations for conclusions

From a sample we can derive conclusions/inferences only about the part of the population that has a positive probability to become part of the sample. We call this part of the population the **sampling frame**. In the best case, the sampling frame comprises the complete population of interest, but in reality, it usually excludes some parts of the forest area for various reasons like lack of accessibility or risk of entry.

All our estimates refer exclusively to the set of sampling elements that are in the sampling frame and we need to make sure that the sampling frame is covering the maximum possible of the whole population.

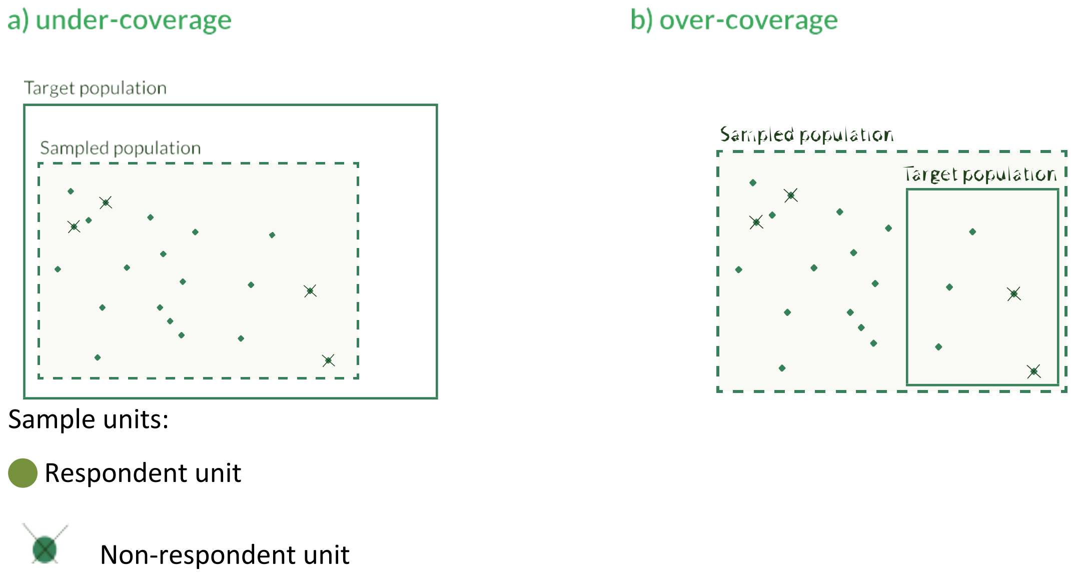
Population vs. sampling frame

Imagine a country does not use a biophysical forest definition—based on quantitative and qualitative criteria—but an administrative or legal definition of 'forest land'. If the selection of sampling elements is limited to this sampling frame, we cannot derive any conclusions about trees and biophysical forests that occur outside of the defined forest land. All conclusions would refer to the forest area inside the defined forest land exclusively. Hence, both population and sampling frame need to be clearly defined and mentioned during the reporting and interpretation of results.

In addition, there might be some points in the sampling frame, which cannot be accessed due to denied access or security reasons. Such missing observations are called **non-response**. The difference between **sampling frame** and **non-response** is: the sampling frame defines the sampling elements that can supposedly be selected and measured. But it may happen that some sample points turn out to be inaccessible—which are the non-response points, and different techniques are available to deal with this problem. Commonly, the rates of non-response are relatively low in NFIs. While there are imputing techniques to make model-based predictions of the would-be observations of such non-response plots, in NFIs, they are usually ignored, and the sample size is reduced.

The diagram below shows that a target population (area inside solid line) may often not coincide with the sampled population (area inside dotted line). Example A is under-coverage, very typical in NFIs, where certain areas of the population have been previously classifled as, for example, not accessible.

Example B is over-coverage, more rare in NFI contexts, but possible if the target population of interest is a particular subpopulation of the country, whereas the sampling was originally designed for the whole country. In both cases, some sampling units were accessible (respondent) and some inaccessible (non-respondent).



Basic concepts of sampling

Before we go deeper into some practical considerations about sampling in the NFI context, you should familiarize yourself with some relevant concepts and some central terminology. Even if statistics tend to become complex and sometimes not easy to digest, a lot of what is expressed in complex formulae is in fact relatively easy to understand with some basic math—and is often also quite intuitive.

In the following section, we will concentrate on several statistical concepts and touch only on those that are relevant to NFIs. However, there is much more to learn about forest inventories than just these few concepts.

Some important concepts and terminology

When we take a sample from a population (or from the sampling frame) there is not one single result: every selection of a new alternative sample will deliver a different estimate that is equally valid as all others.

As we are not able to determine the one and unique **true value** (called the **parameter**) of the population from a sample, the estimate that we derive always carries uncertainty. Only if the selection of samples follows statistical criteria, and the applied formulae, or the estimators, are correct, will we be able to determine the margin of this uncertainty.

In fact, when we determine this margin, this is also an estimate. A typical measure of uncertainty is the **confidence interval**, which defines an interval around the estimated value in which we expect the true value with a defined probability.

What is sample size?

Sample size refers to the number of **independently selected** observations (observed sampling elements) that are drawn from the sampling frame. Here, the term "independent" means: the selection of one element has no effect on the selection of another. Such a selection process takes place if single sample elements are selected randomly.

In forest inventories, however, this is rarely the case, because samples are collected at fixed intervals. It is important to note that "independent selection" as described here, should not be confused with "the independence of variables," which is a completely different concept.

More information on how to determine the sampling size for various forest inventory designs will be provided in Lesson 2.

What is the difference between sampling intensity and sample size?

Sampling intensity refers to the proportion of the sampling frame that is observed. **Sample size**, on the other hand, is about the absolute number of (independently) selected sampling elements.

Sampling intensity is defined as the fraction of population of sampling elements that come into the sample. However, as such a concept is not applicable to the infinite population, we define sampling intensity in forest inventories via area: it is the fraction of the sampled area (= the sum of all plot areas) from the total area over which the population is defined.

What does population variance mean?

The population variance **quantifies the variability in the population**. It is a characteristic of the population of sampling elements. That is, for each population element, there is one value for a specific target variable, such as biomass per hectare. The parametric population variance is the true variance of all these values. And this true (parametric) variance can be estimated from a sample.

How does population variance differ from error variance?

This distinction is a key element to understand a large portion of NFI relevant sampling statistics. While the population variance is an estimate of the variability among population elements (observations from inventory plots), the error variance is a property of the sample. That means it quantifies the expected variation among repeated estimates of the same target variable (e.g. mean biomass per area).

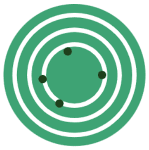
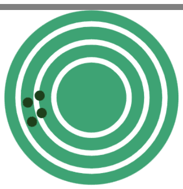
Let’s suppose an NFI is carried out repeatedly for a thousand times, every time with a new selection of plots. In this case, the variation among all the single means is an estimate of the error variance. This information is important for judging the quality of a sample as it gives an answer to the question “what would happen if we repeated our sample again and again - would we come up always with quite similar results, or would we expect that repeated inventories would lead to widely varying results?

In the latter case, we would say, that our estimate is less precise, and in the former that our estimate is precise. Usually, we do not report the error variance but its square root: the **standard error**. This is one of the most relevant statistics estimated from a sample, as it quantifies the precision of estimation. The reason that the square root is reported and not the error variance, is very simple: the standard error comes in the same units like the estimate itself, and is, therefore, much easier to understand.

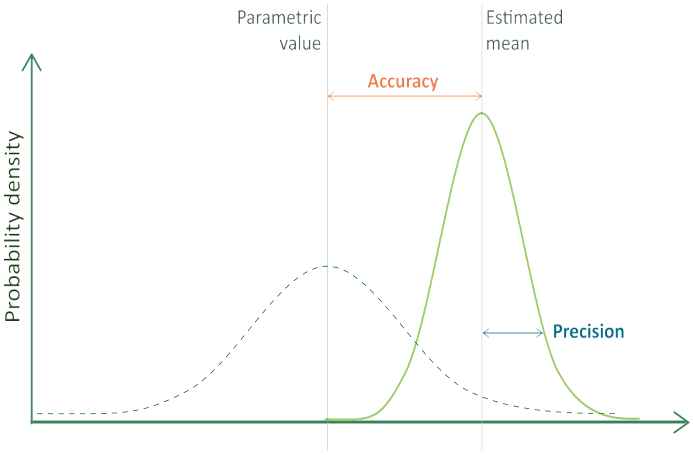
It is intuitively clear that trust and credibility or certainty in the results depend on this error variance. If there is no information given about the error variance, a user of the information may conclude that a single inventory alone is not enough to build on (since the next will likely produce a different estimate).

Accuracy and precision

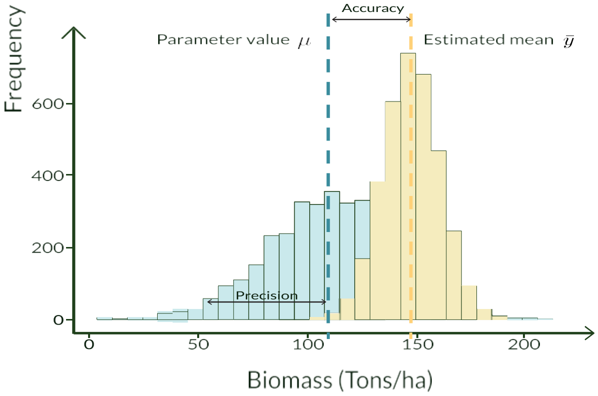
We have referenced the concept of precision previously, but let us now highlight the relevance and meaning of both accuracy and precision as we use it in forest inventory sampling. Let's understand the concepts better with the help of a dartboard example.

Low precision, high accuracy	**Low accuracy, high precision**
Imagine you throw 4 darts. The distribution of hits around the centre is an expression of your accuracy. (averaging the hits will result in a position close to the centre).	Continuing with the dartboard example, the spread of single hits is an expression of your precision (repeated throws are close together).
	

As you see in the graph below, we can graph a distribution over all collected observations (which are here single values on the x-axis). The y-axis values indicate the relative frequency of observations for the respective values. While the solid distribution leads to a relatively high precision (narrow distribution) of the estimated mean y bar, as in the figure, it is not very accurate (i.e. it is biased) because of its deviation from the true parameter µ. On the contrary, the dashed line distribution results in a very accurate estimate of the mean (here identical to the parametric value µ), but relatively low precision.



Let's now look at two examples where 3 500 sampling plots measured the mean biomass per ha. The yellow histogram depicts a distribution with relatively high precision (narrow distribution) of the estimated mean: , but low accuracy (i.e., it is biased) because of its deviation from the true parameter. On the contrary, the blue histogram reflects a very accurate estimate of the mean (here identical to the parametric value µ), but relatively low precision, because of the wider width of the distribution.



Looking back at the previous graph, remember that the sampling statistic calculated from one sample is but one estimate of the true population biomass, based on the one set of selected plots. If we (imaginarily) repeat the estimation of the sampling statistic with a different selection under the same design, we will produce different estimates of the population biomass. The distribution of these estimated means represents the relative frequency of these different estimates of the population biomass.

The width of this distribution is an expression of the variability (or dispersion) around the estimated mean value (y bar). If the dispersion of these values is low and they are relatively close together, we can conclude that repeated alternative samples would likely result in similar estimates. Therefore, the width of this distribution also allows a statement about precision (see graph above).

On the other hand, **accuracy**—or correctness—is the deviation of expected value from repeated samples from the true population parameter—this deviation is also called **bias** or **estimator bias**. Since the true value remains unknown, the size of this deviation cannot be quantified from the sample itself. It is rather a property of the applied estimator and an expression of a systematic error that cannot be compensated by increasing sample size.

The only way to guarantee 'unbiased' estimates is a mathematical proof that the sampling design and applied methods allow correct estimates (design-unbiased) or empirical simulations (in case the estimation relies on model application).

> *i* Keep in mind: In sampling studies we have no information about the true population value (target), we just have the sample at hand (darts). We are blind in regard to the position of the centre, and accuracy can be guaranteed only by using unbiased estimators.

What are the possible reasons for biased estimates? Let's find out.

Selection bias	A non-statistical selection was used, and it is not guaranteed that the sample is representative (e.g. a subjective selection of plots close to the road).
Observer bias	Observations or measurements are systematically wrong (e.g. DBH always measured in 1 m height instead of 1.3 m).
Estimator bias	A systematically wrong calculation (e.g. applying constantly a wrong plot expansion factor, such that all plot observations are too high).
Model bias	In case of model-based or model- assisted sampling techniques, but also in case of modelled observation (e.g. application of wrong biomass models), a potential bias of the model will directly affect the estimation bias.

Note

The limited meaning of sampling intensity regarding precision of estimation

In inventory guidelines or even government regulations we sometimes find thresholds for the sampling intensity (minimum area proportion) that should be sampled (e.g. at least 3 percent of the forest area). However, this sampling intensity has very little meaning for the resulting precision of estimates. Precision depends on sample size. Have a closer look at the estimators presented at the end of this lesson and you will see that “sampling intensity” cannot be found in any of the formulae.

Point and interval estimates

Usually, the estimated value alone is not sufficient information for a proper interpretation or for reporting and decision-making. Remember, we have not observed everything but derived an estimate from a sample. If we report about an estimated mean (e.g. mean volume or biomass per unit area), which we call a **point estimate**, this information alone does not allow any judgement about the quality (or reliability, credibility, or certainty) of this estimate.

We would also need additional information about the estimated precision of such point estimates so that we inform about its quality. Such information is given in terms of an interval around the estimated mean, in which we would expect the true value with a certain probability—and this is what we call an **interval estimate**.

Quick tips!

Reporting estimates

When reporting estimates, it is good practice to say, “*from our sampling study we estimate the growing stock to be 200m^3/ha ± x*” and not “from our sampling study we conclude that the growing stock is 200m³/ha." The fact that we are dealing exclusively with estimates is also evident from the fact that we accompany our estimates of mean values (point estimates) with estimates of precision of this estimate (interval estimates).

Immediate questions that may arise here are:

- From how many independent observations (plots) was this mean estimated?
- How much did these single observations vary (= population variance)?
- What is the expected variation of this mean if we were to (virtually) repeat the sample many times under the same design (= error variance)?

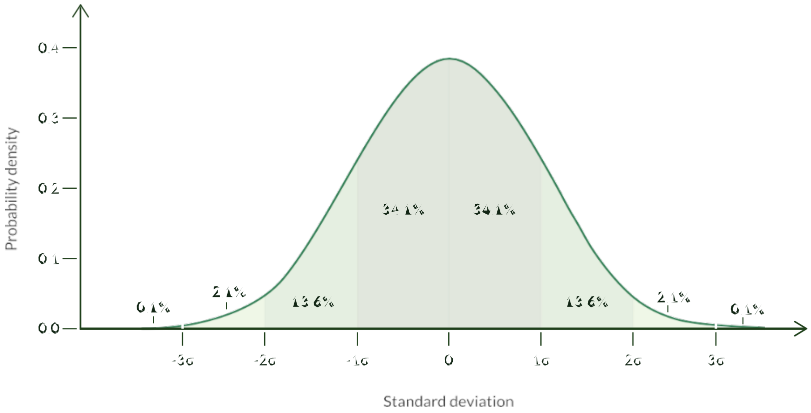
All of the above-mentioned questions are influencing the width of the so-called **confidence interval** around an estimated mean. This confidence interval is a probabilistic statement from which we can learn in which interval around the estimated mean we expect to find the true (unknown) population parameter with a defined probability.

This, however, is only possible if we assume a certain distribution of estimates, and this is the point where an interesting property of statistical samples comes into play.

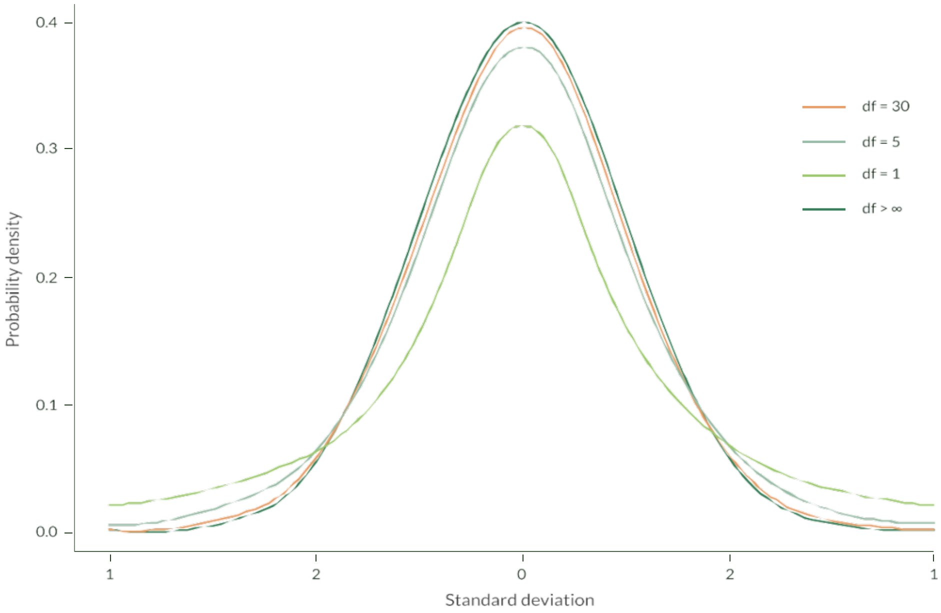
Distribution of samples

A very interesting character of samples allows the definition of this interval: the estimates from repeated sampling tend to follow a normal distribution. This holds for larger samples exceeding a sample size of 30, which in sampling statistics is taken as a rough threshold that distinguishes small and large samples; estimated mean values from smaller samples follow the Student's t-distribution. For large samples, we can use the normal distribution to determine the upper and lower limits of the interval in which we expect the true value with a defined probability (e.g. 95 percent).

The graphs below depict normal distribution and slightly different student's t-distribution. Both allow deriving an interval in which we expect the true parametric value with a defined probability.



Graph illustrating normal distribution of samples. Diagram from Wikipedia, author M. W. Toews, under the Creative Commons License.



Confidence intervals

As part of the estimation process, we would like to assess the level of confidence we have on our estimates. This is reflected by how close the estimate would be to the true parameter, for every sample taken. If for all possible samples the estimates were very close to the true population parameter, we would have high confidence in our estimates. To assess it, we often use the confidence intervals.

Formally, we may state that the probability P, that the true parameter, μ, is within a lower bound and upper bound is x%. The higher is that probability, the higher our confidence in our estimate is. For example, in the specific case of the estimate of the mean, our estimated confidence intervals (expressed in the same units as the mean estimate) will define our bounds as:

$$\bar{y} - C.I. \leq \bar{y} \leq \bar{y} + C.I.$$

where the confidence interval C.I. is defined by Student's t-distribution value and the standard error of the estimate:

$$\mathrm{C.I.} = \mathrm{t}\ \mathrm{S}_{\bar{y}}$$

Commonly, 95 percent confidence intervals are given. The origin of these 95 percent confidence intervals goes far back in the history of statistics and there is no perfectly convincing argument in favor of the error probability of 5 percent. Another confidence interval (such as 90 percent) could just as well be used, as long as this is clearly stated.

The standard error of estimates

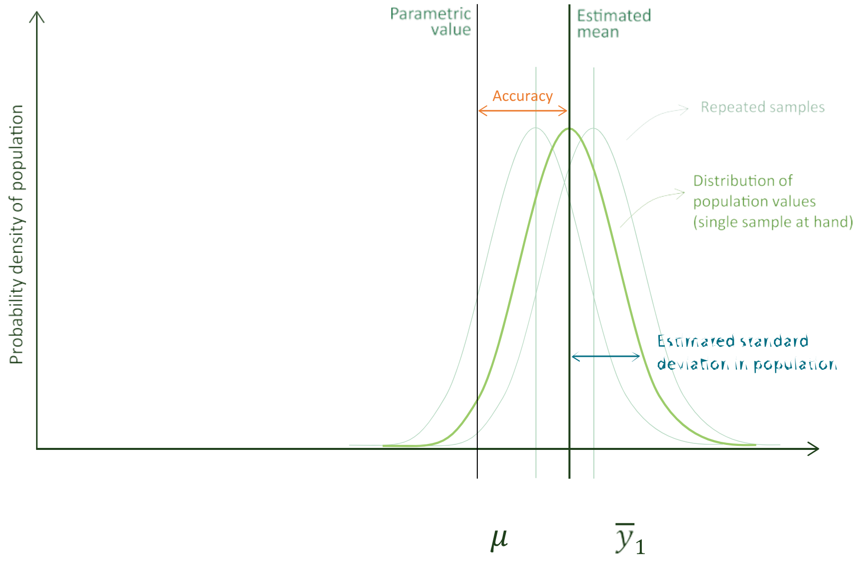
While looking at the single sample at hand (the one inventory that we have carried out), how can we derive an expectation about the variation of all other possible samples under the same design from the same population? In practice, of course, we cannot repeat the NFI many times.

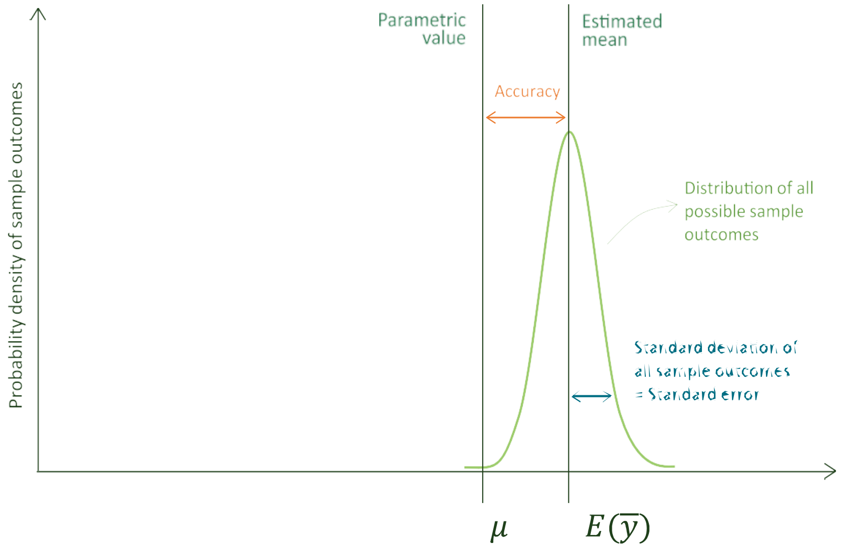
Well, we have learned that we can draw conclusions about the variability of (imagined) repeated samples from the single sample that we have at hand. The measure of this variability is the **error variance**.

The so-called **standard error** is the square root of this error variance. In other words: it is the estimated standard deviation of all possible sample outcomes. The standard error is the most frequently reported measure of precision of estimation. Contrary to the error variance, it comes with the same units as the estimated statistic. It is therefore easier to interpret than the error variance.

The following figure might help to disentangle these two different perspectives. In the upper graph we can see the distribution (variability) of population elements (e.g. plot values) from a single sample (bold line).

However, this single sample at hand is only one out of many possible samples (light green) that we could potentially draw. In the lower graph you see the distribution of all potential sample outcomes around the "expected value" and the standard error is the standard deviation of this distribution.


Probability density of population
Parametric value
Estimated mean
Accuracy
Repeated samples
Distribution of population values (single sample at hand)
Estimated standard deviation in population
μ
$\overline{y}_1$


Probability density of sample outcomes
Parametric value
Estimated mean
Accuracy
Distribution of all possible sample outcomes
Standard deviation of all sample outcomes = Standard error
μ
$E(\overline{y})$

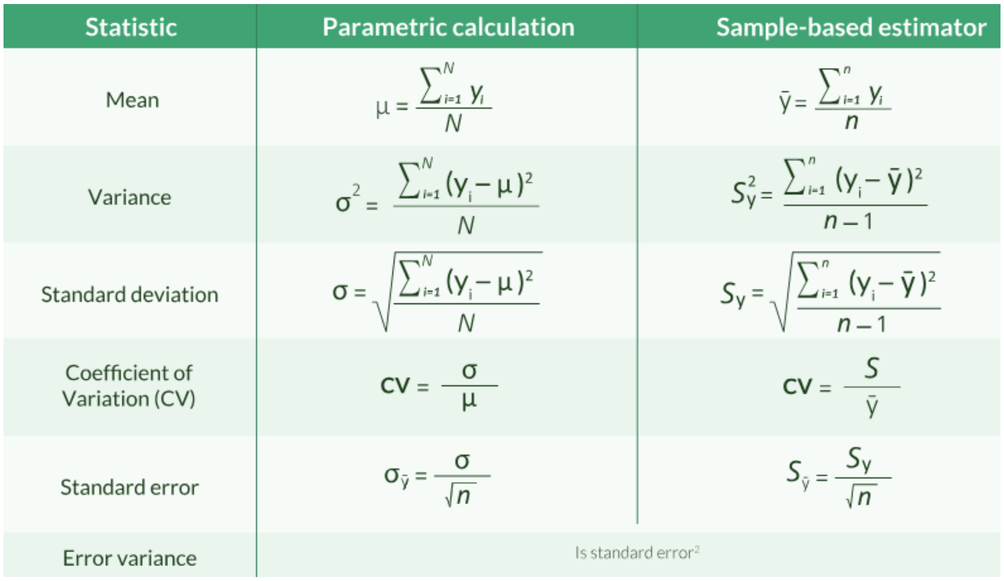
Estimation under simple random sampling (SRS)

We have arrived at the last segment of this lesson. In this section, we look at some more detailed explanations that will be discussed in the next lesson, and consider some estimators for **simple random sampling (SRS)**.

Simple random sampling refers to an independent random selection of each single sampling element. It means we assume an unrestricted random selection of sampling locations in a forest area. Unrestricted random sampling means that all sampling elements have the same selection probability. This is the basic foundation of sampling statistics and very appropriate for explaining estimators, because it is straightforward enough to determine the selection probabilities, which here are equal for all elements.

Even if rarely applied in forest inventory, this sampling design (or selection procedure) is fundamental for all statistics, because existing estimators are quite simple, and the characteristics of statistical sampling can be explained easily and is useful to mention for the sake of completeness.

In the following table you see on the left-hand side the calculation formula for the parametric (true) population value, which remains unknown, and on the right-hand side you see the corresponding estimated (sample-based) population value. Observe that the concept behind the error variance in the table was already explained earlier (Lesson 1, Basic concepts of sampling, Some important concepts and terminology, How does population variance differ from error variance).



Statistic	Parametric calculation	Sample-based estimator
Mean	$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$	$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N}$	$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N}}$	$S_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$
Coefficient of Variation (CV)	$CV = \frac{\sigma}{\mu}$	$CV = \frac{S}{\bar{y}}$
Standard error	$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$	$S_{\bar{y}} = \frac{S_y}{\sqrt{n}}$
Error variance	Is standard error^2	

These estimators given here are those for SRS. In later lessons, you will learn that they become slightly more complex as soon as we consider other sampling designs.

Here, and also in the following lessons, we assume **sampling without replacement** and **sampling from an infinite population -** and therefore ignore the so-called **finite population correction (fpc).** For more details on fpc, refer to **Gottingen University's wiki**, or consult any textbook on statistical sampling.

Summary

Before we conclude, here are the key learning points of this lesson.

- Experts must derive inference and draw conclusions about the current state and change of target variables by making observations on relatively small subsets or “samples” of the total forest area— referred to as sample plots.
- When sampling follows the rules of statistics, we call it “statistical sampling”. Statistical soundness is one of the major characteristics in statistical sampling, as applied in forest monitoring.
- An “estimator” in statistical sampling refers to the calculation formula that we use to produce an estimation.

Lesson 2: Design elements of a sampling study

Lesson introduction

This lesson introduces you to the basic design elements of sampling studies, as they are relevant to NFIs, and the concepts to consider while preparing a sampling strategy.

It also shows you how to calculate the associated sample size.

Remember that working through this lesson will not make you an expert in any of the techniques described here but will enhance your awareness of general concepts. As with other lessons in this course, this is only a ‘primer’ for learners who do not have a sound basis in statistics, which is an indispensable requirement for a comprehensive understanding of statistical sampling.

Learning objectives

At the end of this lesson, you will be able to:

- Describe the three technical design elements of a sampling study.
- Describe sampling design.
- Identify the types of sampling designs.
- Explain the rationale and approach of stratification.
- Describe plot/observation design.
- Summarize the concept of slope correction.

Three design elements of a sampling study

The planning of any sampling study can be broken down into three basic technical design elements that provide a framework for sampling projects. Remember that in order to prepare a sampling study, all three design elements need to be considered to their fullest extent. Let's look at what each of these mean.

Sampling design

Sampling design answers the question “How are sampling elements selected?” In forest monitoring, the sample points selected within the inventory area are theoretically infinitely small, hence considered dimensionless. These points define the position of the sample plot(s).

Observation design

Observation design, also known as plot design or response design, addresses the question “How are the observations taken on each sampling element?” Observation design is defined by the rules which guide how sample trees are included into the sample plot, with reference to the dimensionless sample point.

Estimation design

Estimation design answers the question “How are estimates calculated, and which statistical estimators are to be used?” This is the set of estimators, or formulae to be used for the given sampling and plot design. In sampling and plot design, you are free to choose designs that are "optimal" or fit your goals best. However, you are not free to choose the estimators. This is because they need to match the selected sampling and plot designs. Usually, there are only a few such estimators.

Remember that in this lesson, we will concentrate on typical sampling and plot designs in NFIs. Estimation designs will be covered in the next and final lesson of this course.

Determining sample size

One of the aspects defined in the sampling design is the number of sampling elements (plots) that should be observed. This is also called sample size. From a purely statistical perspective, there are two major criteria that determine the required sample size for a defined target precision for a given inventory situation:

①The variability in the population i.e. the population variance. This can be estimated from a pilot study or taken from previous inventories/inventories in comparable areas. We refer here to the population of sample plots and the population variances will be different for different plot designs for the same forest area.

② The desired target precision, which is a matter of definition. Commonly, precision is defined as half the width of the target confidence interval.

What happens when there is no prior knowledge or previous inventory information?

In absence of data from prior inventories or estimates of variability of the targeted variable, a pilot study can help obtain the relevant information. Since the estimated variance always refers to the specific plot design used, a relatively small number of plots could be distributed to different typical forest types found in a country. Such a pilot study may then deliver estimates on the population variance (even though probably not very precise ones).

It might also be that similar forest types can be found in neighboring countries, from which estimates of the population variance may be available to inform our sampling design.

When not even this information is available, forest statisticians may then have to rely on alternative information, often not based on probabilistic designs, like expert opinion or literature reviews. For a random sample, the sample size is as follows:

$$n = \frac{t^2 * S^2}{A^2} = \frac{t^2 * (CV\%)^2}{(e\%)^2}$$

where A refers to the confidence interval, in absolute value, that we aim to achieve in our estimates (as percentage e% if expressed as relative to the mean), t is the corresponding value of the Student's t-distribution and S^2 (usually pre-estimated from pilot studies or previous information) is the sample variance of the variable of interest, like volume per ha. CV percent is the coefficient of variation in that previous information, expressed in percentage as relative to the mean. The next exercise shows a practical example to calculate the sampling size.

Practice exercise

We want to calculate how many plots would be needed to estimate forest carbon stock with a precision of 10% (referring to the 95% confidence interval). Several studies have shown AGB values of around 100 t/ha with a standard deviation of 70 t/ha (CV%=70).
How many plots should be measured if we assume simple random sampling?

In order to calculate this, we need the corresponding value of the t-distribution for an error probability of 5% (or 0.05, two sided). However, to determined that t-value, we need to know the sample size – which is actually searched. Therefore, we first need to assume a sample size and then do an iterative calculation. We may start with a t- value of 2 in the first iteration – which corresponds to a large sample size of larger than 30 and get: $2^2*70^2/10^2$ = 196.

By referencing the **T-Table** for this sample size of n=196 (from the first iteration), we have arrived at a t-value of ~1.97 - and the above estimate can be newly calculated to $1.97^2*70^2/10^2$ ~ 190. Note that this estimate of the required sample size is only valid for SRS.

In reality, however, our resources are limited and only a certain number of field plots are feasible. In this case, we try to achieve the most precise result with the given budget. As you have learned already, increasing sample size will increase precision—we would therefore strive to plan for as many plots as possible under the given plot design and practical restrictions.

What is my target variable?

A forest inventory can only be optimized towards one single target variable (for which the precision should be maximized for the given resources). Frequently the stand basal area, which is highly correlated to volume and biomass, is used as the target variable. However, consideration of multiple purposes requires compromises in sampling and plot designs, and it may be that the sample size which optimizes precision of estimation for basal area, is not optimal for other variables.

Sampling design

Up to now, we have seen the basic concepts of sampling and considered the three elements of sampling design. Let’s now dive into some sampling design options.

The sampling design defines the selection process for the sampling elements, that is, how the sampling elements are selected, and how many (sample size). The result of such a selection process is a list of all coordinates of sampling locations.

Here we will limit ourselves to exclusively touch upon some typical sampling designs used in the NFI context. Do remember that we already visited before the SRS (check Lesson 1, Estimation under simple random sampling) as a mostly theoretical design that in practice is rarely used in NFIs, but was useful to establish a simple reference against which to compare the following options.

Systematic sampling – the most common sampling design in NFIs

Using a systematic grid of sampling locations is the standard sampling design in NFIs. Such a systematic sample has the advantage that the forest area is evenly covered by the sampling locations, and it ensures that all locations maintain a minimum distance from each other. It leads to a “proportional allocation” of sampling locations over the forest types that occur. And, as it evenly covers the whole area of interest, one may expect that such a sampling grid yields a “representative” sample of the population.

Theoretical considerations and numerous simulation studies have shown that systematic sampling virtually always yields higher precision than SRS, given the same number of observation points.

We may explain that by the fact that systematic sampling evenly covers the whole population so that all conditions are about uniformly covered; another reason is that in systematic sampling, neighboring sample points do always have a defined distance and cannot be very close together: in forests and many other natural populations, plots that are close together are usually more autocorrelated than distant ones and that is inefficient.

Note

Sample size refers to the **number of independently selected sampling elements**, where independently means selected by randomization. Since all sampling locations in a systematic grid are flxed once a starting point and grid orientation have been selected, a systematic sample based on one randomization is only one (sample size = 1). From one single independent observation, however, we cannot derive an estimate of variance and therefore neither an estimate of precision!

Looking at the variance estimator for SRS, if the sample size is n=1, then the denominator n-1 will be zero and, therefore, the variance of the variable of interest $S^{2/y}$ is not defined:

$$S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Frequently, the SRS estimator is used to calculate the error variance of a systematic sample. It is known that such error variance will overestimate the true error variance, so we say that the SRS estimator is here a **conservative estimator**.

That means that the true precision is higher than what we estimate with the SRS estimator - but we cannot say to what extent it is more precise. This underestimation of precision will also affect all other estimates, like the required sample size.

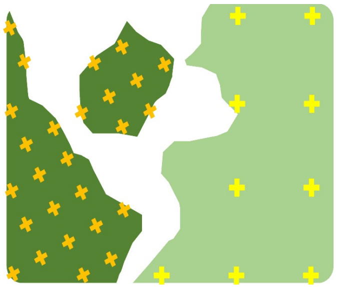
Stratification

We have already learned in earlier lessons that our aim is to **narrow** the distribution of observations as much as possible because this will increase the precision of estimates.

What else can we add to make the population 'more homogeneous' in order to increase precision?

Stratification aims to sub-divide the population into more homogeneous sub-populations. We call these sub- populations **strata** (singular: stratum). In each stratum, an independent sample is taken. When we use simple random sampling in each stratum, we call the design stratified random sampling. That is: we do not introduce here a completely new sampling design, but apply SRS independently in each stratum; the new thing is actually, how we combine at the end the strata-wise estimates to arrive at a compounded total of all strata.

To have more precision in this design, the stratification should be "homogeneous within the strata and heterogeneous among the strata."



Look at the diagram—the total forest area is subdivided into two different strata (light and dark), and we assume that each of them be more homogeneous than the total area and that they differ clearly in their mean values. In the sampling design, they are treated as independent sub-populations and different systematic grids are used.

There are many ways in which a population can be sub-divided into sub-populations: stratification criteria is, for example: forest types, or growth regions with homogenous site conditions. Sometimes administrative boundaries are also used, however, this does not necessarily lead to more homogeneous sub-populations or enhanced precision.

However, it may be used to facilitate inventory implementation or ensure that more precise per administrative units estimates can be delivered.

Calculation of sample size and allocation of samples to strata

When sample size is determined in stratified sampling, two questions must be answered, how many samples altogether, and how to distribute/allocate the samples to the strata.

The required sample size does always depend on the allowed error at a given error probability and on the variability within the population; in stratification we deal with a number of sub-populations, and we must consider that the sub-population variances differ among strata.

As the strata usually have different sizes, these different sub-population variances must be weighted when calculating the total sample size. If there is a number of *L* strata denoted by the subscript *h*, and each stratum has the size (for example in terms of area) N_h, the weight of each stratum is given by N_h/N.

But designing an inventory may also imply dealing with constraints in terms of the costs of the inventory. So while one may want to allocate sampling plots to minimize variability, one may also want to think on the total cost incurred in the inventory, where C_h is the cost per sampling unit in the stratum h. Then, total sample size can be calculated as:

$$n = \frac{t^2 \sum \frac{N_h^2 S_h^2}{C_h}}{N^2 A^2}$$

Where *A* is the allowable error, expressed as half the width of the target confidence interval. The allowable error is a matter of definition. Similar to the sample size estimation under SRS, provided before, *S* and *A* can be substituted by relative expressions: *CV*(%) and *e*(%).

After having calculated the total sample size, these samples must be allocated to the different strata. To do so, one may consider three strata characteristics, either individually or together:

1. **The stratum size**: the larger a stratum the more samples would be allocated.
2. **The variability in the stratum**: the more variable a stratum is the more samples would be allocated.
3. **The cost per sampling unit**: the larger the cost, the lesser samples would be allocated.

Proportional allocation	Neyman allocation	Optimal allocation with cost-minimization
Allocating samples according to stratum size alone	Considering the size of the strata and the variability inside the strata for allocation	In this option, cost implications (c) are also included in addition to stratum size and variability within the strata
$n_h = n \frac{N_h}{N}$	$n_h = n \frac{N_h S_h^2}{\sum_{h=1}^{L} N_h S_h^2}$	$n_h = n \frac{\frac{N_h S_h^2}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{N_h S_h^2}{\sqrt{C_h}}}$

Note

Remember that **any sampling technique can be applied per stratum**. There may also be different sampling techniques used in the different strata. It is important that for each stratum the point and interval estimate of the target variables can be produced, in the best case, unbiasedly.

In fact, **the main characteristic of stratified sampling is that it consists of several independently implemented sampling studies**. The only new thing is that one needs to find out how to eventually combine the estimates that come from the *L* different strata so that estimates for the whole population can be generated.

Post-stratification

We can also stratify the inventory after an unstratified sample was implemented (e.g. by deriving separate estimates for different forest types). This is called post-stratification and can be considered as a kind of data grouping for analyses. However, such post-stratified analysis needs to be done carefully, as the estimation does not strictly following the sampling design anymore. For example: the data grouping for analysis must not be done along with the target variable, by for example, forming three equally wide groups (post-strata) of low, medium and high values; such an approach would be entirely mistaken, even though it will lead to high (but untrue) precision values!

Before doing a post-stratified analysis with estimates of precision of estimation, you should consult a sampling expert to avoid unnecessary mistakes and mistaken inferences and conclusions: all estimators that are recommended in textbooks for post-stratified analyses come with some assumptions that need to be observed.

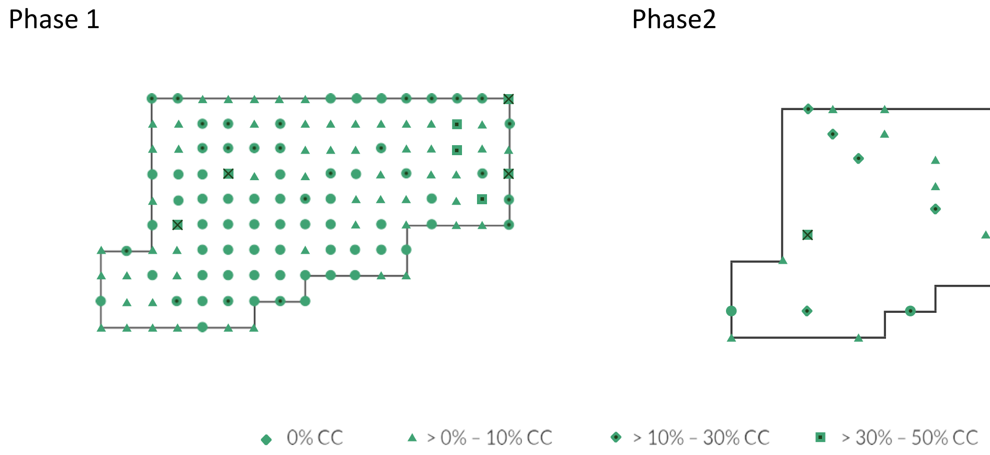
Two-phase or double sampling

In double sampling, a new feature is introduced—the use of **ancillary variables**, also known as **auxiliary variables** or **co-variables**. In order to increase the precision of estimation of the target variable, it is essential to understand the correlation between the target and ancillary variables. The idea is to collect a relatively large—but low cost—first phase sample to obtain information of such ancillary variables, such as through remote sensing.

Then, in the second phase, a smaller sample is selected where both the target and the ancillary variable are observed. This usually incurs a much higher cost per plot—let us understand this better with an example. When estimating forest biomass, it is possible to use a first phase estimation of the ancillary variable from remote sensing imagery and determine a vegetation index around many sample points. This is fast and cheap and can also be automatically done.

Then, in the second phase, a much lower sample size of sample plots is installed in the field where measurements are taken and plot biomass estimated. This is far more expensive than a sample in the first phase. For all field plots, not only the target variable (biomass, in this example) is determined, but also the ancillary variable (vegetation index, in this example) from the remote sensing data is observed.

The second phase data pairs of target and ancillary variable are then used to establish a model between target and ancillary variable, from which the estimates can be produced. The most common models used here are the simple ratio between both variables or a regression model. This would lead then to double sampling with the ratio estimator and double sampling with the regression estimator, respectively.



It might have become clear by now, that **the higher the positive correlation between the two variables**, the more efficient the estimator will be in terms of precision. That is, for an efficient ancillary variable, we always search for a variable that is highly positively correlated to the target variable. At the end, this is, of course, also a cost consideration, because the introduction of a first phase also increases the cost of the inventory.

In the next lesson on estimation design, you will see how this eventually improves the precision of estimating the target variable.

Double sampling is a very efficient way of seizing a (cheap to observe) ancillary variable to improve the precision of the estimation of a (more expensive to observe) target variable.

Double sampling for stratification

Double sampling is also relevant in the context of stratification. There are inventory cases where it is known or assumed that stratification may increase precision, but sometimes, it is not possible to clearly delineate the stratum boundaries (e.g. in remote sensing imagery), since they are 'fuzzy' or continuous transitions rather than strict lines. Besides, such a delineation takes a lot of time and requires robust prior knowledge!

So far, for stratified sampling, we have assumed that the strata will be defined prior to sampling, and we therefore also refer to it as **pre-stratification**. By doing so, we assume that such prior definition of the strata is error-free; that is: the size of the strata and their weights in the estimators are not considered a source of error.

In **double sampling for stratification** (DSS) or **two-phase sampling for stratification**, the strata do not need to be defined before sampling, but are defined during the sampling process and the strata sizes are estimated.

The two phases in DSS are as follows: **a relatively large first phase sample is selected** (frequently in remote sensing imagery, as this is quite inexpensive) and for each sample point it is determined to which stratum it belongs. That is: the ancillary variable that is observed in the first phase is **stratum**; in NFIs this could possibly be **forest type**.

In the second phase, a stratified sub-sample of the first-phase plots is selected and these plots are visited to observe the—relatively expensive—target variable, which is frequently done in the field. The allocation of the total sample size to the strata can be done along the same strategies as with pre-stratification: uniform, proportional to size, proportional to size and variability, or proportional to size, variability and cost. The decision about such an allocation will need to be done from available information about the expected variability and cost per plot in the strata that were distinguished.

Given the same sample sizes in the second phase sample and in normal pre-stratification, DSS will be less precise than pre-stratification. The reason is that in DSS the strata sizes are being estimated from the first phase sample and such size-estimation carries a sampling error which propagates into the total error. This can also be seen with the estimators for DSS, which are not given here but can be found in sampling textbooks.

Did you know?

Can we also use a remote sensing classification to separate strata?

Yes, we can, and it is often done like this. Imagine, for example, a remote sensing-based classification in different forest types, for which we expect differences in forest biomass. However, similar to the above- mentioned visual interpretation, every classification will have errors. Since the stratum area estimates come with errors, we need to account for that additional source of uncertainty in the estimator!

Quick tips!

Never “invent” a new sampling or plot design while ignoring the issue of deriving an unbiased statistical estimator! The accuracy of an estimator depends on a careful reflection of the selection and inclusion process. You can easily run into unsolvable statistical traps by just making small changes on the plot or sampling design.

For example, a simple rule like “extend the sample plot if a certain condition is fulfilled” can lead to unexpected statistical problems (the resulting inclusion of probabilities of trees cannot be calculated easily)! Other rules, such as **shift plots that overlap the forest boundary completely into the forest** are violating the definition of the population. They are simply wrong and might lead to biased estimates.

Plot or observation design

The sampling design outlines how sample points are selected, while the plot design outlines how the trees to be sampled are chosen around the selected point. The question is, which objects (e.g. trees) should be included at each sampling location around the sample point?

As in practically all design/planning steps for an NFI, while optimizing or adapting the plot design to the specific forest conditions, one needs to carefully think about how to allocate the limited resources (time, budget and personnel) in the most efficient way. **Efficiency** can be seen as the **relation between costs and resulting precision of estimates**. If the resources are not sufficiently considered, this may compromise the sustainability of a permanent NFI.

Capturing variability as a major goal in plot-design planning

From a purely statistical point of view in plot design optimization, we aim at capturing **a maximum of variability within each single plot.** The rationale behind this is that we then make the variability between the plot small. And that translates into a narrow distribution of plot observations around the estimated mean (*compare Lesson 1 of this course*), which means: higher precision of estimation.

A relevant concept in this context is called **spatial autocorrelation**, which we also observe in forest populations—this means that objects—in this case, plots—that are closer together tend to have higher correlated observations.

High correlation means that from knowing the value of the first object one may quite well predict the value of the second object at a given spatial distance. If that is the case, the measurement of the second observation is not very efficient as it does not bring much of additional information; it may even be a waste of money.

Considering the relevance of spatial autocorrelation in inventory design planning leads to some conclusions regarding plot design, as well as sampling design:

✓ It is good to have a certain distance between sample plots. Sample plots that are spatially close together are not efficient.

✓ It is good to have a plot design that covers a larger area so that within-plot observations exhibit lesser autocorrelation:

a) Therefore, given the same area, elongated strip plots are statistically more efficient than circular or square plots; and

b) Another option to raise efficiency for a given plot area is to subdivide the plot into spatially disjointed sub-plots at a certain distance from each other: this is what we call “cluster-plots”.

With these two options, please remember that not only statistical but also cost considerations are relevant. The per-plot cost will be higher for elongated strip plots or cluster plots as compared to a compact plot shape of the same area: therefore, in practice, these optimization considerations need always to balance statistical and cost criteria.

Typically, this spatial correlation drops after 50-200m (depending on forest type and management).

Fixed area and nested sample plots

The most basic plot design is the fixed area plot. The shape and size of these fixed area plots might be different according to the specific inventory purpose and forest conditions. In general, circular plots are more common than rectangular ones in forest monitoring, while in ecological surveys the square shape is more common—and sometimes the term '**quadrat**' is used in ecological surveys instead of 'plot'.

From a theoretical point of view, any plot shape is admissible; yet it is crucial to carefully consider the inventory purpose and forest conditions when selecting the appropriate plot design, and to balance the cost and practical considerations with the need for accurate data collection.

Note

The plot (or tree) expansion factor

Many variables of interest are area-related, such as 'number of trees per hectare'. This means that, for example, when one doubles the area of one plot, the number of trees found is also expected to double on average.

In order to expand or upscale the observation to the typical reporting unit of one hectare, such area-related, per-plot observations need to be multiplied with an expansion factor resulting from the relation 1 ha/plot area.

Area-related variables are commonly those associated with quantitative direct measurements, such as volume, biomass, number of trees or regeneration density.

Trees of different diameter classes normally appear with different densities (in natural forests, for example, there are many more small trees than large trees). Large trees carry much of the forests biomass, but they are low in number. If we then use a relatively large plot to be sure that on average we have some large trees within the plot area, we would need to measure a huge number of small trees. That is: one single plot area is, therefore, usually not efficient.

A common solution here is to use a so-called **nested plot design**, in which sub-plots of different sizes are nested, so that trees of different size class are being observed on different sub-plot areas. Here it is important to stick to a strict terminology in order to avoid confusions: the whole thing (the combination of all sub-plots) is the plot—while the different nested shapes of different size are the sub-plots.

The nested sub-plots are not assessed one after the other, but we check (in one sweep, typically clockwise) for each tree whether it is included or not. They will all end up in the same data table.

Note

Accounting for unequal probabilities

The inclusion probability in a forest inventory is the probability that a tree is included in a sample. Because effectively the sampling units are plots, based on areas, this probability is in effect the inverse of the expansion factor.

Hence, a design with sub-plots will lead to unequal inclusion probabilities, and we need to reflect this in the estimation design: the plot expansion factor will be larger for the smaller plots, where the smaller trees are observed!

Since sub-plots have different sizes and areas, trees will have different expansion factors according to the sub-plot where they were tallied. We therefore need to calculate the correct expansion factor for each tree individually, based on its dbh or its affiliation to a subplot and its corresponding area. Expansion factors standardize then all the areas to a single per-hectare basis.

The video '**How to assess (nested) fixed area forest inventory plots**' explains how to establish and assess nested fixed area sample plots in the field: **https://www.youtube.com/watch?v=IA-PflXW9_k&t=2s**

Slope correction

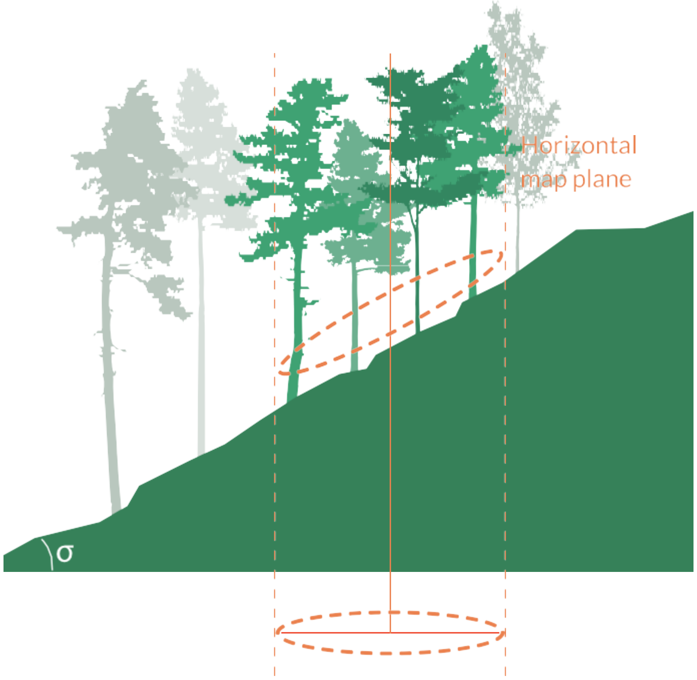
The area to which all observations and estimates refer is the map area, or the horizontal projection of the terrain into the map plane. Whenever it is not possible to directly measure horizontal distances when measuring the plot (by using modern electronic instruments), and distances are measured along the slope, the horizontal projection area of this plot is smaller than the intended plot area and the distances measured from the plot center to the trees are larger than the projected horizontal distances (except we measure exactly along the contour lines).

To ensure an equal plot area in the horizontal map plane, which is a prerequisite to derive unbiased area related estimates, the oblique plot area that constitutes the plot in the field needs to be enlarged depending on the slope angle.

When using circular fixed area plots, they become ellipses when projected into the slope. To establish these plots on the slope there are essentially two options:

1. Either a electronic distance meter is used that directly measures horizontal distance: then, automatically the right trees within the defined horizontal distance (radius) are included. An elliptical plot is established without the need to specifically lay it out.
2. Or—which is the traditional approach—one calculates the (larger) are of the slope-projected ellipse and establishes at the slope a circle with exactly that area. To do so, one needs to slope-correct the nominal plot radius in the horizontal plane with the factor below to obtain the larger radius of the circle that will be laid out at the slope.

$$\frac{1}{\sqrt{\cos \alpha}}$$



Video resource

How to correct for slope and how to deal with plots at the forest boundary

https://www.youtube.com/watch?v=InPERYNxQ0E&t=1s

In case that such slope correction was omitted during plot establishment, the observations obtained from this plot may be corrected afterwards (since plots have unequal sizes in the horizontal projection depending on slope). Since the actual horizontal area is then smaller than intended, the result would need to be multiplied with the correction factor 1/cos α. However, slope angle needs to have been measured; otherwise, a correction is not possible.

In most NFIs, slope correction is usually considered for slope angles > 10 percent, which is one of those conventions with which forest inventories work in practice Also, in the presence of gentle slopes of < 10 percent, distance measurements can often be taken horizontally by manual leveling.

Note

Slope correction applies to any plot design and must always be considered in advance, the corrections are quite simple for circular fixed area plots. The same principles of slope correction do, of course, also apply to square and rectangular plots.

And for these two plot shapes the slope correction is more laborious: for square plots, the corners of a larger effective plot area need to be marked on the slope so that the projected plot area corresponds to the nominal area. For elongated rectangular plots we usually walk along the central line and measure trees to the right and left in a defined distance: here, both directions of the plot need to be slope corrected: the long line along wish we walk and the measurements to the right and left.

Sampling with cluster plots

In NFIs, locating and travelling to the sampling locations is a major cost factor, particularly when the road network is poor. The sampling grid is usually sparse and the distances between plots are large. This is why we want to assess as much information as possible once the team is out on a plot location. This calls for large plots.

However, we learned that—because of spatial autocorrelation—it is good to have observations in some spatial distance to each other, so that, instead of establishing one large plot per sample point, NFIs usually opt in favor of establishing so-called **cluster plots**: the individual large plots are subdivided into sub-plots each that are laid out at some spatial distance.

What is resulting are subplots that are arranged at some geometric pattern (for example corners of a square or an L-shape). The set of sub-plots forms the plot and it is good to not confuse plots and sub-plots. Plots are the core sampling elements and the number of plots corresponds to the sample site; not the number of sub-plots.

The spatial distribution and distance between the subplots are arranged such that a cluster plot can “capture” much more variability compared to a single compact plot of the same area size. For planning a cluster plot design, we need to decide certain characteristics:

1. The number of subplots per cluster;
2. The distance among the subplots;
3. Size and shape of subplots; and
4. The spatial arrangement of subplots.

Some considerations about (sub-) plot shape and size

We have already concluded that each single sample plot should capture as much variability as possible in order to increase overall precision of estimation. If we had an unlimited number of resources at our disposal, the same number of larger plots would always be better than the same number of smaller plots.

On the other hand, under limited resources, we need to decide whether to use a larger number of smaller plots or a smaller number of larger plots. Increasing plot size involves increasing marginal precision costs for every additional tree measured, while increasing the number of plots implies smaller and smaller gains in precision in exchange of added walking time in between plots. However, from a certain plot size onwards, the effect of an increased sample size on precision will be more important than increasing plot size!

Did you know?

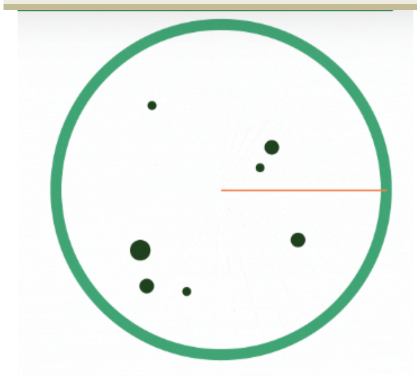
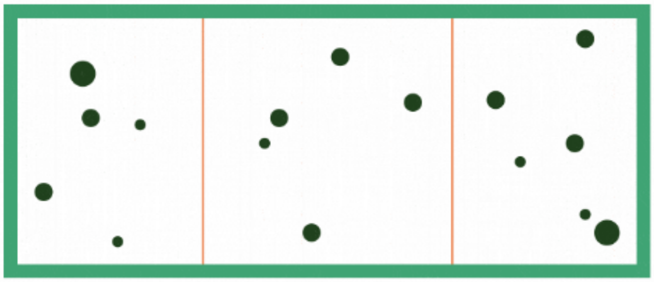
Plot size and statistical efficiency.

The marginal information gain we could expect from measuring one additional tree per plot is decreasing with every new tree. Imagine we have already measured 99 trees on a sample plot, would you expect that we learn something new from measuring the 100th tree? Probably not, because it will only add redundant information.

On the other hand, the assessment costs on the plot will increase linearly with each additional unit of observed area (or number of trees). **But where should we draw the line?**

Experience and empirical studies both suggest that assessing more than 15-20 trees per (sub-) plot is not efficient anymore. It is then better to invest the resources in increasing sample size instead (better more small plots than fewer large plots)!

There are several practical and statistical arguments dictating the shape of plots (or subplots), and traditions and common standards in different areas of the world must also be considered. The following general guidance applies on the same plot/subplot areas of different shapes:

Circular sample plots are:	Long rectangular plots:
	
• easy to implement – practically everything can be measured from the centre; • relatively easy regarding slope correction; but • very compact and likely to capture less variability.	• need more work in plot marking (e.g. by marking with a tape the central transect and walking along it); • have, on average, more border trees to be checked; • intersect, on average, more often with boundaries between forest types and need more consideration of border corrections; • are more time-consuming for slope correction; will likely capture more variability; and • are good when visibility is low (understorey is too dense), as only short distances to the right and left of the central line can be observed.

On the number of subplots per cluster

Clustering of subplots into one joint observation will always be less efficient than selecting the same number of sub-plots as independently selected plots over the whole inventory region. Sampling with cluster-plots is a compromise used to reduce travelling costs and observe larger plot areas at every single sampling location, while reducing redundancy caused by spatial autocorrelation by distributing the subplots spatially.

Therefore, the same arguments hold as for the design planning of single plots: increasing the observed area means increasing costs, while the standard error will be reduced down to a certain limit, beyond which there is barely any further reduction.

Thus, investing more and more time and effort into one single plot has, from a certain point onwards, no significant effect on precision. Usually, there is no strong effect on precision after a number of 3-5 subplots (depending on spatial variability), and measurement of more plots per cluster becomes inefficient.

Reality check

Feasibility as a guiding argument

We can derive a lot of statistical considerations from sample and plot size, but in the end, the most important argument is feasibility. In most cases, we are forced to look at the available resources and make the best out of them.

For planning purposes, it would be beneficial if, on average, a whole cluster plot could be measured by a single field team in one day. This will affect the number of subplots that are feasible and their size, considering relatively small subplots (in many forest inventories statistical considerations lead to plot sizes that include around 15-20 trees in average).

It is a common situation in NFIs that much time is needed to reach the sample point and to walk from subplot to subplot. We may consider these walking times as inefficient with respect to measuring our target variables: frequently the larger part of the time in the field is used for such inefficient walking, Then, one can easily imagine that more than 4-5 subplots will in many cases already become a challenge in terms of time consumption.

Summary

Before we conclude, here are the key learning points of this lesson.

- The planning of any sampling study can be broken down into three basic technical design elements - sampling design, observation/plot design and estimation design.
- One of the aspects defined in the sampling design is the number of sampling elements (plots) that should be observed.
- A forest inventory should usually be optimized towards one single target variable (for which the precision should be maximized for the given resources). Usually the stand basal area, which is highly correlated to volume and biomass is used as the target variable.
- In forest inventories, the sampling design defines how the samples are selected from the population and what the sample size is.
- Stratification means “subdividing” the total population (forest area) into more homogeneous sub-populations that we call “strata” (singular “stratum”).
- The plot design defines what is being done at each sample point; it does also define the rules how to include the sample trees that will be observed.

Lesson 3: Estimation design

Lesson introduction

In this lesson we will look into estimation design, which consists of the methods and formulae we apply to derive unbiased estimates from the data collected from a sampling design and a plot design.

Learning objectives

At the end of the lesson, you will be able to:

1. Describe the basic estimators for common sampling approaches.
2. Explain the importance of applying the correct estimator.

Estimation design

Let’s begin this lesson by looking at some typical estimation designs. Some of these alternatives are specific to the sampling design used, and others can be applied based on different sampling designs.

In some cases, we also have the freedom to apply different estimators to data collected under a given sampling design. For example, we can include ancillary data with a **ratio estimator** (discussed in later sections of this lesson). We might also derive an estimate without considering the ancillary variable if it doesn't help to produce a more precise estimate.

It is then up to the data analyst to decide which estimator to use. If multiple alternative estimates can be produced, the choice is usually the estimation design that leads to the higher precision (which is equivalent to “smaller standard error of the estimates”).

Design-based, model-assisted and model-based inference

In Lesson 1, we looked at the term '**inference**'. Sometimes, the terms **inference** and **estimation** are used interchangeably, because each estimation means making inferences about true population values. Some inventory experts, therefore, prefer to talk in general about inference when they refer to estimation, because inference implies more than only estimation: it also refers to the purpose of estimation. Let us now look at the three inferential paradigms: **design-based**, **model-based** and **model- assisted** inference.

Design-based inference	Model-based inference	Model-assisted inference	
We make no assumptions about the (spatial) structure of the population. We assume this structure as unknown, and aim at estimating characteristics of this flexed population. Unbiasedness is guaranteed from the sampling- and plot design exclusively i.e. from randomization.	The population is seen as a realization of a stochastic process, and assumptions about the underlying process or model can be considered during the estimation. The assumption is that we are looking at only one out of many possible populations (that make up a superpopulation). Since no model can describe this population perfectly, uncertainty will remain even after a full census, and comes from the 'quality' of the model used – not from the sampling design.	A model is used in support of design-based estimation, lying somewhere between design-based inference and model-based inference. This means that even if the model was not well-specified, this will not introduce bias, but will affect the precision of the estimation. Examples are the ratio and regression estimator that make use of simple models during estimation by establishing a relationship between an ancillary and the target variable.	**Population assumptions**
The validity of estimates (unbiasedness) depends exclusively on the sampling design (selection of sample plots, randomization).	The validity of the estimate depends entirely on the validity of the model	Field observations of the target variable plus auxiliary variables for the plots are considered.	**Validity of estimates**
Remote sensing or auxiliary data is not integrated in the estimation phase, but maybe in the planning phase, for example for stratification. Estimates are produced from plot observations of target variables alone.	Field observations are used to establish a relationship (model) to ancillary variables which are usually remotely sensed indices. Then the model is used to predict the target variable from a wall-to-wall coverage of these indices.	Validity of estimates depends on the sampling design – but precision of estimation can be increased by integrating the additional information that comes from the ancillary variable.	
	Example: For every pixel of a satellite image a model predicts biomass/ha, the statistics are later derived as aggregate of the pixel values.	**Example**: Instead of estimating biomass directly, a ratio between biomass and e.g. NDVI is estimated, where the NDVI values serve as ancillary variable and are available for the whole forest area (population).	

Estimation with cluster plots

Contrary to many textbooks on sampling, we do not refer to cluster sampling as a sampling design of its own, but to sampling with cluster plots, that is, we look at it as a plot design, since this is more consistent with the terminology we use for plot designs.

However, the meaning of both is the same: a single sampling element consists of several sub-elements, which are selected jointly in a single randomization step. Since subplots in a cluster plot are not selected independently from each other, **sample size refers to the number of selected clusters and not to the number of subplots**. The cluster plot can be considered as a single 'funny-shaped' plot where the funny shape comes from the spatially disjointed arrangement of the plot.

For simple random sampling of cluster plots: when the sub-plot observations are aggregated at the cluster level = plot level (only one value, either a mean or total per cluster), the subsequent estimation can follow the same estimators as introduced in lesson 1 (simple random sampling: SRS). However, it often happens that we need to consider clusters of different sizes (= different numbers of subplots), since not always all subplots are inside the target population. Then the ratio estimator would be a choice, using the size of the cluster (the number of sub-plots) as an ancillary variable.

It may, however, also be of interest to do an analysis per subplot within the clusters; this will not change anything regarding the results of point and interval estimate, but it allows additional analyses of the spatial structure of the forests and of the efficiency of the cluster plot design.

Show me the math

Estimation with cluster plots

Clusters may have an equal or unequal number of sub-plots (m) for all clusters. In this section, we present the estimator only for the situation of cluster-plots with equal number under random sampling. For clusters with unequal sizes, cluster reweighting is required. The estimated mean per sub-plot can then be calculated from the estimated mean per cluster and the mean number of subplots per cluster as follows:

$$y = \frac{\bar{y}_{cl}}{\bar{m}}$$

and the estimated error variance of the mean per sub-plot is:

$$v\hat{a}r_{cl}(\bar{y}) = \frac{1}{\bar{m}^2}\frac{S^2_{y_i}}{n}$$

Where y_i are observations per subplot. The estimated variance per cluster can be derived by calculating the variance over the per-cluster-observations with the known estimator for SRS:

$$S^2_{y_i} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_{cl})^2}{n-1}$$

The estimated total results, as usual, come from multiplying the mean with the total. In cluster sampling, one may take the cluster mean and the number of clusters (*N*), or the mean per sub-plot and the number of sub-plots (*M*):

$$\tau = N * \bar{y}_{cl} = M * \bar{y}$$

And the respective error variance for the total can be derived as:

$$var(\hat{\tau}) = N^2 v\hat{a}r(\bar{y}_{cl}) = M^2 v\hat{a}r(\bar{y})$$

Efficiency of sampling with cluster plots—Intra-cluster correlation

The similarity of observations within a cluster can be quantified by means of the **Intra-cluster Correlation Coefficient (ICC)**, sometimes referred to as the Intraclass Correlation Coefficient. The higher this correlation is, the more redundant the observations are from different sub-plots and the information gain becomes smaller.

Such an analysis is very instructive when it comes to understanding and analyzing the performance of cluster sampling for populations with different spatial autocorrelation structure, as this is directly mirrored in the intra-cluster correlation coefficient. When the ICC is high, one may consider making the distances between sub-plots larger (which increases the walking time and costs, of course) or to reduce the number of sub-plots.

However, one should consider that ICCs can be different for different variables and an optimal cluster-design for one variable is not necessary equally optimal for another. It is common practice to use basal area as a **guiding variable** in these optimizations, because it correlates well to various other tree variables (e.g. biomass).

For cluster plots consisting of several sub-plots, it turns out that if:

- ICC = 0 (observations uncorrelated), there is no difference in the performance of cluster sampling of n clusters and SRS with n*m subplots. We aim at keeping ICC low. But getting ICC close to zero is impossible in practice, as the distance between the sub-plots would turn out to be too long.
- ICC < 0 (negative correlation between subplots), sampling with *n* cluster plots would be more efficient than with *n**m* independently selected sub-plots. This situation is very unlikely in forest inventories (because of spatial autocorrelation).
- ICC > 0 (some redundancy inside the clusters) is the most typical case. Sampling with cluster plots is less efficient than independent selection of single plots.

However, if inventory costs are included, the overall efficiency is likely higher because of reduced travel.

Stratified sampling

You have learned that stratification aims at subdividing the total population into more homogeneous sub- populations, in which independent sampling studies are implemented. When combining the single estimates from different strata we need to remember that these strata have different sizes. Therefore, we need to weigh all stratum estimates with the respective relative sizes of the strata. In forest monitoring, stratum size is usually given in terms of area, and the sum of all strata weights would be 1 (i.e. equal to the total area).

Stratified sampling is not introducing a new sampling design but what is new is the framework used to integrate estimates from different strata into one estimate for the total area. Stratified sampling therefore actually introduces a variation of estimation design: combining independent estimates from L strata into one single estimate for the total population.

Show me the math

In the following, we use the notation *h* as an index for a stratum and *L* for the total number of strata.

Then, an unbiased mean over multiple strata can be estimated as a weighted sum. The weights here are the area proportions of the different strata *N_h/N:*

$$y = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_h$$

The estimator for the error variance is:

$$\hat{var}(\bar{y}) = \sum_{h=1}^{L} \left\{ \left(\frac{N_h}{N} \right)^2 \hat{var}(\bar{y}_h) \right\} = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h}$$

The square root of this error variance is the standard error.

The total is derived as:

$$\hat{\tau} = N\bar{y} = \sum_{h=1}^{L} \frac{N_h}{N} \hat{\tau}_h = \sum_{h=1}^{L} N_h \bar{y}_h$$

And the error variance of the estimated total is:

$$v\hat{a}r(\hat{\tau}) = v\hat{a}r(N\bar{y}) = N^2 v\hat{a}r(\bar{y})$$

Efficiency of stratified sampling

Statistical considerations reveal that stratification is the more efficient in increasing precision of estimating the mean the more different the strata means are. These positive effects (i.e. better overall precision), tend to become smaller with increasing number of strata.

From a statistical point of view, the formation of more than six strata usually has no significant effect on improving the precision of estimation. However, there might be more than just statistical arguments for forming the strata. The question is also whether a post-stratification might be more indicated in such case.

Double sampling for stratification

Double sampling for stratification was already mentioned in Lesson 2. It is a two-phase sampling design to estimate the sizes of strata (that cannot be delineated or pre-defined easily). Since the stratum areas (and weights) are estimated from the first phase sample, the sampling error of estimating these areas needs to be considered when estimating mean and variance for the whole population.

Show me the math

Estimation in double sampling for stratification

Assuming that the stratum weights are estimated from the first phase sample (denoted by the apostrophe) as

$$w'_h = \frac{n'_h}{n'}$$

And an unbiased mean can be estimated as

$$\bar{y} = \sum_{h=1}^{L} w'_h \bar{y}_h$$

Ignoring finite population correction and assuming *n'* is large, the respective error variance would be estimated as

$$\hat{var}(\bar{y}) = \sum_{h=1}^{L} \left({w'_h}^{2} * \frac{{s_h}^{2}}{n_h} + w'_h * \frac{(\bar{y}_h - \bar{y}')^{2}}{n'} \right)$$

This variance estimator looks very similar to the estimator in stratified random sampling – except for the last term in the brackets: there, an error component is added that comes from the fact that the strata sizes are only estimated and not known

The **Collect Earth tool**, part of FAO's Open Foris system, is useful in this context and has been applied many times. It was designed to make use of available and georeferenced satellite imagery and aerial images from Google Earth, Bing and others, for a visual interpretation of sampling locations or plots.

With the help of this tool, a high number of points can be visited and visually classified into different strata. Later the area size of strata can be estimated as proportion of sample points per stratum. A variance of this estimate can be derived and incorporated into the estimation as shown above.

The ratio estimator—utilizing quantitative ancillary information

There are situations in forest inventory sampling in which the value of the target variable is known (or suspected) to be well correlated to another variable (called a co-variable, ancillary variable or auxiliary variable).

If such an ancillary can be observed on the plot without too much effort and costs (e.g. by remote sensing analysis), it will be efficient to also observe it, and utilize the correlation to the target variable to eventually improve the precision of estimating the target variable. This is where the ratio estimator is applied.

Note

Imagine a classification of satellite imagery was done to produce a continuous prediction of **crown cover percent** for a forest area with varying crown density. Assuming a high correlation to plot volume or biomass, this would be a case where the ratio estimator is applied. In closed forest areas with complete forest cover in all places, this would not make any sense, because there, crown cover percent would not vary but be constantly 100 percent, so that the correlation to biomass between the ancillary variable **crown cover percent** and the target variable **biomass** would be close to zero.

Instead of estimating standing biomass per unit area from the field plots directly, the ratio estimator uses a detour: we estimate a ratio, r, of the two means, which gives us **biomass/crown cover percent**, and in the following, we use the known crown cover to derive an estimate of biomass. Mean biomass could then be estimated as **r*Mean crown cover percentage**.

Another typical case for the ratio estimator is if a certain proportion of large plots (or cluster plots) are sloping over beyond the boundaries of the inventory region and are only partly inside the target population. In that case, the plot area inside the forest is not identical for all plots, and the ratio estimator may be applied, with plot area as an ancillary variable to account for that. In fact, we can then assume that the plot area will be highly correlated with the stock variables (including basal area, volume, biomass, carbon and number of trees) recorded on the plot area.

For an estimate of precision, we need to know the parametric value (mean) of the ancillary variable (here, mean crown cover percentage over the total forest area, or total forest area to be inventories in the plot-size example).

Show me the math

Estimation with the ratio estimator

The parametric ratio between target variable *y* and ancillary variable *x*

$$R = \frac{\mu_y}{\mu_x}$$

is estimated based on the sample from

$$r = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

The estimated variance of this estimated ratio is:

$$var(r) = \frac{1}{n}\frac{1}{{\mu_x}^2}\frac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1}$$

The estimated total is derived from:

$$\tau_y = r\tau_x$$

with an associated error variance of:

$$var(\hat{\tau}_y) = {\tau_x}^2 v\hat{a}r(r)$$

Given the estimated ratio, *r* , the mean of the target variable could be estimated as:

$$y_r = r\mu_x$$

This estimated mean carries an estimated variance of:

$$var(\bar{y}_r) = {\mu_x}^2 v\hat{a}r(r) = \frac{1}{n}\left\{{s_y}^2 + r^2{s_x}^2 - 2r\hat{\rho}s_x s_y\right\}$$

Show me the math

Estimation design with the regression estimator

While the ratio estimator models the relationship between the target and an ancillary variable, the regression estimator uses a regression model with both intercept and slope coefficient. Remember: in the ration estimator, the intercept is assumed to be zero. The mean is estimated from the regression estimator as follows:

$$y_L = \bar{y} + b\left(\mu_x - \bar{x}\right)$$

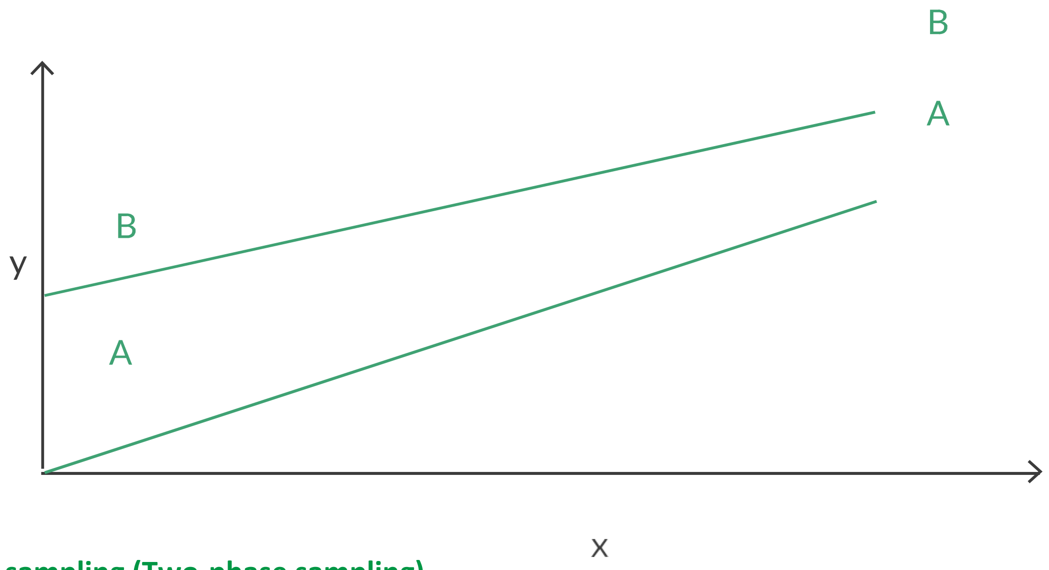
The estimated variance of this estimated mean is given with:

$$var(\bar{y}_L) = \frac{1}{n}\frac{1}{n-2}\left\{\sum_{i=1}^{n}(y_i - \bar{y})^2 - b^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}$$

Ratio vs. regression estimator

The ratio estimator uses a simple fixed ratio, which means that the target variable, y, will be zero if the ancillary variable, x, is zero. However, there are situations where this is not correct. Imagine we can find small trees on those plots, for which no crown cover was detected in the remote sensing images (for example, due to low spatial resolution). In this case, a regression line with an intercept coefficient which is not forced to be zero (as with the ration estimator) would be more appropriate; if, for example, crown cover percentage is zero, there might still be considerable biomass on the ground. Here, the regression estimator is using a simple linear model.

In both cases, ratio- or regression estimates, the overall efficiency depends on the correlation between target and ancillary variables, which should be highly positive. Sometimes it turns out that this correlation is relatively low and that the expectations were too high, after, for example, very expensive remote sensing imagery was purchased.



Double sampling (Two-phase sampling)

For the ratio and regression estimator, the parametric mean or the parametric total of the ancillary variable needs to be known. If such information is not available, one may estimate these values from a sample.

This is exactly what double sampling is about, also referred to as two-phase sampling: in the first phase sample, the ancillary variable is estimated, usually with a large sample of a variable that can be observed relatively easily and inexpensively, and which is known to be highly positively correlated to the target variable.

Then, in the second phase sample, a smaller sample is taken of the target variable, which is frequently a variable that is much more expensive or much more difficult to observe. A relationship between a target and an ancillary variable can then be established (either a simple ratio or a regression, which would be double sampling with the ratio estimator, or the regression estimator, respectively).

Here, the stronger the positive correlation to the ancillary variable, the smaller the required sample size in the second phase, when the more complex/expensive/difficult target variable is observed.

In the following, we address dependent phases, where the second phase sample is a subset of the first phase (and not an independently selected sample). The presented estimators are for SRS exclusively.

Show me the math

Estimation in double sampling

For double sampling with the ratio estimator, the mean of *y* can be estimated as:

$$\bar{y}_{md.r} = \frac{\bar{y}}{\bar{x}}\bar{x}' = r\bar{x}'$$

With an estimated variance of the estimated mean of:

$$v\hat{a}r(\bar{y}_{md.r}) = \frac{S_y^2 + r^2 S_x'^2 - 2rS_{xy}}{n} + \frac{2rS_{xy} - r^2 S_x'^2}{n'} - \frac{S}{N}$$

And for the regression estimator, the mean is estimated as:

$$\bar{y}_{md.reg} = \bar{y} + b(\bar{x}' - \bar{x})$$

With an estimated variance of the mean of:

$$v\hat{a}r(\bar{y}_{md.reg}) = \frac{S_y^2}{n}\left\{ 1 - \frac{n' - n}{n'}\hat{\rho}^2 \right\}$$

Where ρ is the estimated coefficient of correlation between *x* and *y*.

For both cases, the error variance of the total is calculated, as usual, as:

$$v\hat{a}r(\hat{\tau}) = N^2 v\hat{a}r(\bar{y}$$

The overall efficiency of double sampling depends on the relation of costs between observing phase 1 and 2 samples and on the correlation between the two variables. In fact, we strive to exploit the ancillary variable as much as possible, to be able to reduce the number of (costly) second phase samples. The higher the correlation and the more expensive the observations in the second phase, the smaller the phase two sample.

Did you know?

Choosing between alternative estimators

Depending on the applied sampling design, there might be alternative estimators that could be applied. For example, an estimate might be produced on field samples alone, or consider additional auxiliary variables. Or a post-stratification may or may not be applied to data. In such situations, producing different, valid estimates, with alternative estimators, should result in the same mean, but different estimates of precision. In case that multiple unbiased estimators are available, we would prefer the one producing the smallest standard error of estimates.

Summary

Before we conclude, here are the key learning points of this lesson.

- For **design-based inference** we make no assumptions about the (spatial) structure of the population. We assume this structure as unknown, and we aim at estimating characteristics of this fixed population.
- In **model-based inference**, field observations are used to establish a relationship (model) to ancillary variables which are usually remotely sensed indices. Then the model is used to predict the target variable from a wall-to-wall coverage of these indices.
- In **model-assisted inference**, a model is used in support of design-based estimation, lying somewhere in the middle of design-based inference and model-based inference.
- When using cluster plots, a single sampling element (plot) consists of several sub- elements

(sub-plots), which are selected jointly. Since subplots in a cluster plot are not selected independently from each other, sample size refers to the number of selected clusters, not the number of subplots.

- The similarity of observations within a cluster can be quantified by means of the Intra-cluster Correlation Coefficient (ICC), also referred to as the Intra-class Correlation Coefficient.
- Stratified sampling is not a new sampling design, but a framework to integrate independently generated sample-based estimates from different strata into one estimate for the total population, that is: it is rather a variation of estimation design.