

Sampling methods used in the studies of natural resources

Danutė Krapavickaitė

Vilnius Gediminas Technical University
Statistics Lithuania

Workshop of Baltic-Nordic-Ukrainian Network
on Survey Statistics 2012
August 24-28, Valmiera, Latvia

Outline

1. Peculiarities and objectives of the surveys of natural resources
2. Features of the natural environmental populations
3. Sampling populations in space
4. Sampling and estimation for continuous population
5. Mobile populations. Detectability and sampling
6. Spatial prediction or kriging
7. Sources of error in natural and environmental studies
8. Conclusions
9. References

Peculiarities of the survey of natural resources

Natural (environmental) resources: **air, water, soil** and **biota**, that sustain in our environment

Objective of the survey: status, condition, extent of resource, amount of resource

Target population:

- discrete and finite – lakes and wetlands with well defined units,
- 1, 2 or 3 dimensional continuum – forest, water in a lake.

Survey type: one time survey or long-term monitoring to access change or trend

Addressing both objectives requires **balance** of repeatedly visiting sites to assess trend and adding new sites to assess status.

Traditional surveys of small systems

Current environmental issues: global warming, contamination of surface and ground water by pollutants, extensive landscape alteration are not localized

Large scale studies need **environmental sampling methodology** to make regional, continental, global environmental conclusions.

Objectives of the environmental survey

Traditional survey methodology directed to estimate mean or total of the population.

Environmental survey often has more general object:

- estimating **distribution function** or **quantiles**,
- **properties** of the population **in various classes**, for example: proportion of lakes satisfying some criteria

Common objectives:

- to characterize **status** of some resource,
- change or **trend** in that status

They have conflicting design criteria:

- status is best assessed by sampling as **much of the resource** as possible,
- trend is best detected by observing the same resource **over time**

Difficulties to survey the environmental populations

1. Environmental populations exist in **spatial context** (response has spatial pattern and structure)
2. **Ancillary information** is almost always available (from satellites, aerial photography)
3. Environmental resources are **expensive, time consuming** to sample (remote locations, laboratory costs for analysis, needs for permission)
4. Difficult to obtain accurate **sampling frames** (substantial portion of nontarget elements, fail to cover entire population)
5. Environmental sampling connected with **political, economical considerations** \Rightarrow statistical considerations represent only 1 part of sampling designs. Multiple groups, agencies, organizations have interests in the results of the survey. Meeting those interests + maintaining statistically-based design = challenge.
6. With the aim to estimate status and trend of the resource may be needed to **regroup, recombine, expand the sample** in order to address evolving objectives.

Defining sampling frames in space

Historically, many environmental samples have been chosen for convenience or subjectively to be representative. Probab. aspects:


1. **Nontarget units** have to be eliminated in sampling, this complicate estimation procedures. No easy way to compensate for units in the target population omitted from the list.
2. Area sampling design based upon a single area sampling frame. **Survey area is partitioned** into a set of mutually exclusive and exhaustive areas.
3. **A grid** can be used to frame a population which is distributed over some spatial extent and then sampling at each or around each grid point.
4. With availability of **GIS**, electronic representations of maps are becoming more common for environmental populations.
5. The **spatial information** is needed to allocate sample point on a landscape. There is available richer **landscape information** which can be **attached** to the geographical coordinates: eco-region, land use, soil topology and so on ⇒ Joint evaluation of multiple resources.

Sampling for population in space

SRS is rarely used, it is inefficient compared to the methods that utilize knowledge about population characteristics and structure.

1. Multistage designs. Features:

1.1 Spatially constrained designs. Observations of elements that are near one another contain redundant information. \Rightarrow samples that are well dispersed over the population tend to lead to more precise estimates of population attributes than samples without spatial control. Techniques to achieve spatial control: area sampling, spatial stratification, systematic and grid-based sampling, spatially balanced designs.

1.2 Spatially balanced designs (Yates, 1946). **Definition.** A sample s , $s \subset \mathcal{U}$, for estimation of mean μ_y of variable y is said to be balanced over auxiliary variable x , if the x -values (known beforehand) are chosen so, that the sample mean of the x -values is equal to the true population mean μ_x of x : $\bar{x} = \mu_x$. The intuition: the auxiliary variable is correlated with the unknown response. By balancing over the auxiliary variable \Rightarrow approximate balance over the response variable. 

2. **Stratified designs** can improve accuracy of the estimates in the case of appropriate stratification.

- 2.1 **Systematic** sampling, spatial dispersion achieved. For 2-dimensional resource (forest) 2-dimensional grid placed on the map of resource, points selected systematically in the center of grid cells or as intersections of the grid lines. Shortcomings: If the features of landscape influence nonresponse, it may be high; systematic samples are inflexible; no unbiased variance estimator.
- 2.2 **Spatially stratified** samples. Strata are defined to be disjoint areas with few sampled units per stratum. Equal amount of resource selected in each stratum for equiprobable design; equal sample size in the stratum, amount of resource per stratum vary for not equiprobable design.
- 2.3 **Random tessellation stratified** designs (RTS). A grid is placed randomly over the domain (population). 1 point is selected randomly in each cell. Compromise between SRS and SYS. Shortcoming: no unbiased variance estimator.

3. Other designs

3.1 Adaptive sampling. Natural populations frequently exhibit **clustering**: individuals of the same type or species tend to group together. Suppose a regular square grid has been placed over the domain of some clustered population. Suppose grid cell area is substantially smaller than the average cluster size.

(i) An initial **sample of grid cells** is selected. The cells in the initial sample are visited, and the response is recorded for each cell.

(ii) If the response meets some criteria (number of observed elements is greater than some constant), the **adjacent cells are added to the sample**.

This sequence of observations is continued until no newly observed cells meet the criteria.

Thompson shows how to obtain some modified weights that permit unbiased estimates of the total.

3.2 **Capture-recapture** sampling. Estimation of size of the wild population (wildlife, fish). **Assumption.** (i) *Closed population* – no additions, no removals during the observation period. (ii) *Each individual has a constant and equal probability to be captured.* **Sampling scheme.** (i) From N size population *initial sample* of size M is selected, marked and then released. (ii) A *subsequent sample* records number m of marked individuals recaptured from the first sample as well as the total number of sampled individuals n . We expect: $\frac{m}{n} \approx \frac{M}{N}$, and estimate the population size

$$\hat{N} = n \frac{M}{m}, \quad \widehat{Var} \hat{N} = \frac{Mn(M-m)(n-m)}{m^3} \quad (\text{Deming})$$

Chapman (1951) modified this estimator:

$$\hat{\hat{N}} = (n+1) \frac{M+1}{m+1} - 1, \quad \widehat{Var}(\hat{\hat{N}}) = \frac{(M+1)(n+1)(M-m)(n-m)}{(m+1)^2(m+2)}$$

$\hat{p} = \frac{m}{M}$ is estimator of detection prob. in the second sample.

In practice capture-recapture sampling is rarely used, because assumptions don't hold: population is not closed, its size is different at each occasion. Method has been modified.

Sampling and Estimation for Continuous Populations

Problem of estimating attributes of continuums, objects or entities that do not naturally divide into smaller discrete units.

For example, plants and animals, their organs, landscapes, lakes, the atmosphere, span of time.

Approach: to treat each entity as a continuous population of points and to define the total amount of attribute possessed by the entity as **an integral of a continuous attribute density function**.

Let us find a volume of a cucumber fruit of length L . Let $\rho(x)$ denote the cross-sectional area of the cucumber at point x , $0 \leq x \leq L$. The volume of a fruit:

$$t_\rho = \int_0^L \rho(x) dx.$$

$\rho(x)$ – cross-sectional area – **attribute density** at x . If $\rho(x)$ is known $\Rightarrow t_\rho$ will be found by integration. If $\rho(x)$ is not known \Rightarrow Monte Carlo is used for numeric integration. Numbers x_s are chosen between 0 and L , $s = 1, 2, \dots, n$, $\rho(x_s)$ is measured and...

In general, let t_ρ denote the definite integral of a continuous function $\rho(x)$, measurable in $[a, b]$, $\rho(x) \geq 0$, i. e.

$$t_\rho = \int_a^b \rho(x)dx = \tau_\rho(b) - \tau_\rho(a), \quad \tau_\rho(u) = \int_a^u \rho(x)dx.$$

$x_s = ?$ Selection of a particular point x , at which to measure $\rho(x)$, is done according to the **probability density** (selection density)

$$f(x) > 0, \quad x \in [a, b]; \quad f(x) = 0, \quad x \notin [a, b].$$

Let $F(x) = \int_{-\infty}^x f(x)dx$ is a point selection distribution function,

$$F(x_s) - F(a) = \int_a^{x_s} f(x)ds = u_s.$$

Inverse transform method. Let $u_s \sim U([0, 1])$. Then solve

$$F(x_s) - F(a) = u_s \tag{1}$$

with respect to x_s and use x_s for estimator $\hat{t}_{\rho_s} = \rho(x_s)(b - a)$.

Crude Monte-Carlo sampling

(a) **Crude Monte-Carlo sampling** method uses constant probability density function over the interval of integration:

$$f(x) = \frac{F(b) - F(a)}{(b - a)} = \frac{1}{b - a}, \quad x \in [a, b].$$

Substituting it into (1) we obtain:

$$\int_a^{x_s} \frac{1}{b - a} dx = u_s, \quad \frac{x_s - a}{b - a} = u_s, \quad x_s = a + u_s(b - a) \sim U(a, b).$$

This selection formula for x_s is **continuous analog of SRS with replacement** in finite population.

Estimator of volume

Let us select x_1, x_2, \dots, x_n as said in (1).

Then estimators for volume t_ρ :

$$\hat{t}_{\rho_s} = \frac{\rho(x_s)}{f(x_s)}, \quad s = 1, 2, \dots, n.$$

$$\hat{t}_\rho = \frac{1}{n} \sum_{s=1}^n \hat{t}_{\rho_s} = \frac{1}{n} \sum_{s=1}^n \frac{\rho(x_s)}{f(x_s)},$$

$$\text{Var}(\hat{t}_\rho) = \frac{1}{n} \left(\int_a^b \frac{\rho^2(x)}{f(x)} dx - t_\rho^2 \right),$$

$$\widehat{\text{Var}}(\hat{t}_\rho) = \frac{1}{n(n-1)} \sum_{s=1}^n (\hat{t}_{\rho_s} - \hat{t}_\rho)^2,$$

unbiased, $n > 1$.

$f(x) = \frac{1}{b-a}$, $x \in [a, b]$ for crude Monte Carlo.

Importance sampling

(b) Let $f(x)$ is any positive density function for $x \in [a, b]$ and $f(x) = 0$ for $x \notin [a, b]$. Inverse transform method (1) requires to select x_s such that

$$\int_a^{x_s} f(x)dx = u_s, \quad u_s \sim U[0, 1]$$

or $F(x_s) - F(a) = u_s$.

If $g(x)$ is a measurable function on $[a, b]$, then let

$$G = \int_a^b g(x)dx,$$

take $f(x) = g(x)/G$, $x \in [a, b]$, and use \hat{t}_ρ to estimate volume t_ρ . $g(x)$ may be a model for $f(x)$. Sampling of x_s according to (1) with any density function is called importance sampling.

Importance sampling is a **continuous analog of with replacement sampling with probabilities proportional to size.**

Special cases of importance sampling

There are some kinds of sampling used in forestry which can be considered as special cases of importance sampling.

- **Bitterlich** sampling. The points (trees) in 2-dimensional space are selected at random, and population elements in a circle of radius proportional to point tree diameter at the breast height are included into the sample.
- **Line-intersect** sampling. The direction as a line is selected in the 2-dimensional space at random. All objects (example: bushes of berries) intersecting with this line are selected. The larger the object – the higher its probability to be selected.

Mobile populations. Detectability and sampling

In the surveys of most bird species is unlikely that every bird in a selected plot will be detected. In a survey of large mammals some animals may remain unsighted. In a survey of fish or other marine species not every individual in the path of a net is caught. In a survey of human populations some individuals sampled may remain undetected also.

Definition. *The probability p that an object in a selected unit is observed (seen, heard, caught or detected by some other means) is called as detectability.*

1. Let **detectability** for a given species is equal to a **known constant** probability p for any animal in a region A , y – number of animals observed (random), t – total number of animals. $Ey = tp$.

Then $\hat{t} = \frac{y}{p}$ is unbiased estimator of the number of animals (unknown):

$$y \sim Bin(p, t), \text{Var}(y) = tp(1 - p).$$

$$\text{Var}(\hat{t}) = t \frac{1-p}{p}, \widehat{\text{Var}}(\hat{t}) = y \frac{1-p}{p^2} \text{ unbiased.}$$

2. **Detectability** p unknown, **estimated** by double sampling, capture-recapture, multiple regression or other method. Let \hat{p} is approximately unbiased for p . Then $\hat{t} = \frac{y}{\hat{p}}$ – estimator of the number of animals, ratio of 2 estimators. By Taylor linearization

$$Var(\hat{t}) \cong t \frac{1-p}{p} + \frac{t^2}{p^2} Var(\hat{p}).$$

3. The area divided into the N plots. n size **SRS of plots**, **known constant detectability**. \bar{y} – mean of the animals detected in the selected plots, then

$$\hat{t} = \frac{N\bar{y}}{p}, \quad Var(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n} + \frac{N}{n} t \frac{1-p}{p}.$$

4. The area divided into the N plots. **SRS of plots and estimated const. detectability.** \bar{y} – mean of the animals detected in the selected plots, \hat{p} – approximately unbiased and uncorrelated with \bar{y} . Assume, $\widehat{Var}(\hat{p})$ is available. Then

$$\hat{t} = \frac{N\bar{y}}{\hat{p}}$$

is not unbiased, but may be approximately unbiased. By Taylor's theorem for $\mu = t/N$

$$\begin{aligned} Var(\hat{t}) &\cong \frac{N^2}{p^2} (Var(\bar{y}) + \mu^2 Var(\hat{p})) \\ &= N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n} + \frac{N}{n} t \frac{1-p}{p} + \frac{t^2}{p^2} Var(\hat{p}), \\ \widehat{Var}(\hat{t}) &= \frac{N^2}{\hat{p}^2} \left(1 - \frac{n}{N}\right) \frac{s^2}{n} + \frac{N}{n} \hat{t} \frac{1-\hat{p}}{\hat{p}} + \frac{\hat{t}^2}{\hat{p}^2} \widehat{Var}(\hat{p}). \end{aligned}$$

5. Unequal probability sampling design, unknown unequal detection in the plots...

Line transect sampling

Line transect sampling is used **to survey animal or plant species**. An observer moves along a selected line and notes the location relative to the line of every individual of the species detected.

It typically occurs in such surveys, that more individuals are detected close to the line than far away from it, because probability of detection is higher near the line than far from it.

A line transect is characterized by a detectability function giving the probability that an animal at a given location is detected. In most situations, the probability of detection can be expected to decrease as distance from the transect line increases.

Separate cases further

1. **Narrow strip method** Let L denote the length of the transect, w_0 – maximum distance from the line to which detectability is assumed perfect, consider it as a constant.

y_0 number of animals detected within the narrow strip.

Density D of the number of animals per unit area is estimated:

$$\hat{D} = \frac{y_0}{2w_0L}.$$

If the region area equals to A , then number of animals in the area is estimated by

$$\hat{t} = A\hat{D} = \frac{Ay_0}{2w_0L}.$$

2. **Random sample** of transects, estimated detectability function...

Spatial Prediction or Kriging

In geostatistical studies is desired to **predict the amount** of ore or fossil fuel that will be found at the site.

The prediction may be based on values at other sites in the region, and these sites may be irregularly spaced in the region. Let y_i means ore, pollutant or animal abundance at a location as a random variable.

From observed values y_1, \dots, y_n at n sites a_1, \dots, a_n we wish to estimate or predict the value y_0 at a new site a_0 . y_1, \dots, y_n are **correlated**.

y_0 is viewed as a random variable, inference problem is referred as a **prediction in space**. The objective is to find predictor \hat{y}_0 using n observed values that is unbiased for y_0 : $E\hat{y}_0 = E(y_0)$ and which minimizes the mean square prediction error:

$$E(y_0 - \hat{y}_0)^2 \rightarrow \min$$

May be $\hat{y}_0 = \sum_{i=1}^n b_i y_i$ for constants b_1, \dots, b_n . The spatial prediction is called **kriging in geostatistics**. It is a **model-based prediction approach in survey sampling** with aux info.

Sources of errors in environmental studies

Some sources of errors are not unique for environmental studies, but they are most commonly observed in environmental situations than in the other types of data.

1. The only **sources of data are not located at the site of interest**. Frequent measurement needs installation of expensive equipment, which already exist in present locations. Example. Water pollution measurements are taken near the outfall from industrial facilities, but the concern is effect on drinking water at people's home.
2. Design identifies sample location, but data collecting individuals may **not be able to access the sites**. Location may be in the middle of the river rapids, or on private property, where the land owner refuses to provide permission.

3. **Seasons may effect** the location being sampled. Living species move locations across seasons. Location of fish varies greatly by time of the year. The results of specific species may depend on the time of a year at which data are collected.
4. The need to collect physical samples from **inconsistent physical material** is a source of unique error in environmental surveys. For example, collecting samples from municipal dump sites to measure the presence of toxic materials requires developing procedures to assure a adequate samples of materials from a combination of computer parts, furniture and miscellaneous waste.

Conclusions

- Topic discussed at our previous events by
Termeh Shafie, Estimators for snowball sampling, Vilnius 2010
Steven Thompson, Adaptive sampling, Norfällsviken 2011,
Anton Grafström, Spatially balanced sampling, Norfällsviken
2011, Nordstat 2012
- Fields of applications of sampling methods, where survey
statisticians may direct their efforts, show initiatives and find
new jobs.

References

1. David A. Marker and Don L. Stevens Jr. Sampling and Inference in Environmental Surveys. In: *Handbook of Statistics 29. Vol. 29A Sample Surveys: Design, Methods and Applications*. Eds.: D. Pfeffermann, C. R. Rao. Amsterdam: Elsevier, p. 487-512, 2009.
2. Timothy G. Gregoire, Harry T. Valentine. *Sampling Strategies for Natural Resources and the Environment*. Chapman & Hall/CRC. 2008.
3. Thompson S. K. *Sampling*. New York: John Wiley & Sons. 1992, 2002.
4. Thompson S. K., Seber G. A. F. *Adaptive Sampling*. New York: John Wiley & Sons. 1996.