Environmental Statistics

Chapter 3: Sampling and Monitoring Networks

Session 2020/2021

University of Glasgow

University of Glasgow

We will cover

- Sampling and monitoring - general
  - Statistical sampling strategies
    - Simple random sampling
    - Stratified random sampling
  - Analysing data from these strategies – how and comparisons
  - How many samples do we need?
- Designing monitoring networks
  - BACI
- Note: Some of this will be revision – remember to set what we are learning in the context of environmental data

University of Glasgow

We will cover

- **Sampling and monitoring - general**
  - **Statistical sampling strategies**
    - **Simple random sampling**
    - **Stratified random sampling**
  - Analysing data from these strategies – how and comparisons
  - How many samples do we need?
- Designing monitoring networks
  - BACI
- Note: Some of this will be revision – remember to set what we are learning in the context of environmental data

Statistical Sampling
What and why?
Revision

- Why?

University of Glasgow

Statistical Sampling
What and why?
Revision

- A process that allows *inferences* about properties of a large collection of things (*the population*) to be made based on observations on a small number of individuals belonging to the population (*the sample*).

- **Why bother?**

  Valid statistical sampling techniques increase the chance that a set of specimen is collected in a manner that is *representative* of the population.

Statistical sampling allows a *quantification of the precision* with which inferences or conclusions can be drawn about the population.

University of Glasgow

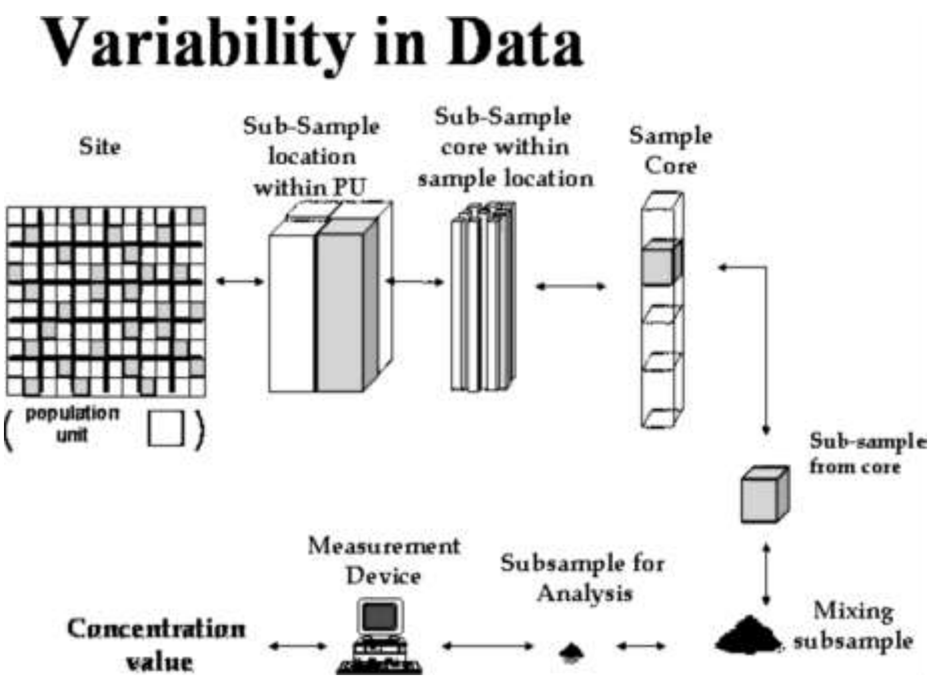
What is statistical sampling? Variation

Example:

- Soil or sediment samples taken side-by-side
  - from different parts of the same plant, or
  - from different animals in the same environment,

exhibit <u>different</u> activity densities of a given radionuclide (measured in Bequerel (Bq).

- The distribution of values observed will provide an estimate of the *variability* inherent in the <u>population of samples</u> that, theoretically, could be taken.



From Gilbert and Pulsipher (2007)

University of Glasgow

Potential sources of statistical variation

Sources include (but are in no way limited to…)

- Natural background levels inherent in the environmental system (random variation)
- Time: cyclical patterns, seasonal patterns, day of week
- Physical/Chemical/Biological: topography, hydrogeology, meteorology, action of tides,
- Spatial: Distance / direction / elevation / area
- Species diversity; sex, age, mobility

University of Glasgow

5 stages to create a sampling experiment

- **Stage 1**: Define the objectives
- **Stage 2**: Summarize the environmental context.
  - Expected behaviour and environmental properties of the compound of interest in the population members
- **Stage 3**: Identify the target population
- **Stage 4**: Select an appropriate sampling design
- **Stage 5**: Implement and summarise

Stage 1: Know your objectives.

It could be …

- **Description** of

  A characteristic of interest (usually the average, but could also be the variability or a high percentile – we'll get to quantile regression later ☺),

  Temporal or Spatial patterns of a characteristic

- Detecting **temporal or spatial trends**

- **Quantification of contamination** above a background or specified intervention level

- Assessing **environmental impacts** of specific facilities, or of events such as accidental releases

University of Glasgow

Stage 2/3: Representativeness

An essential concept is that the taking of a sufficient number of individual samples should reflect the population.

Representativeness of environmental samples is difficult to demonstrate.

Think about taking a sample of water from a river

- Is the water well mixed?
- What depth is it from?
- How wide is the river?

Usually, representativeness is considered justified by the procedure used to select the samples

University of Glasgow

Stage 3: What is the population?

- The population is the set of all items that could be sampled, such as
    - all fish in a lake,
    - all people living in the UK,
    - all trees in a spatially defined forest,
    - all 20-g soil samples from a field.

- Appropriate specification of the population includes a description of its <u>spatial extent</u> and perhaps its <u>temporal stability</u>

University of Glasgow

Stage 3: What are the sampling units?

In some cases, sampling units are
- discrete entities (i.e., animals, trees),
- but in others, the sampling unit might be investigator-defined, and arbitrarily sized.

Example: Technetium in Shellfish

The objective here is to provide a measure (the average) of technetium in shellfish (e.g. lobsters for human consumption) for the West Coast of Scotland.

- Population is all lobsters on the west coast
- Sampling unit is an individual animal.

University of Glasgow

Stage 4: Sampling Schemes

- Simple random sampling
- Stratified random sampling
- Systematic sampling
- Spatial Sampling
    - Grid sampling
    - Transect sampling

University of Glasgow

Stage 5: Statistical analysis

- First what is the objective and how is it expressed in terms of a 'population' parameter?
  - e.g. estimate the average (most common), or a population proportion

- Second what sampling strategy have you adopted?
  - e.g. random sampling or systematic?

- Often the simplest case is to imagine that the sample statistic offers a reasonable estimate of the population parameter

Simple Random Sampling (SRS)

Simple random sampling

The sampling frame
Simple Random Sampling

Population of N units
(N 1m² areas),

Use simple random sampling to select n of these units.

Generate n random digits between 1 and N,

1	2	3	4	**5**				9
							17	
				23		**25**		
			31		**33**			
					42			**45**
46					**51**			**54**

10 random digits:
5, 17, 23, 25, 31,
33, 42, 45, 46 ,51

University of Glasgow

Simple random sampling

Every sampling unit in the population is expected to have an **equal probability** of being included in the sample.

The first step requires complete **enumeration of the population members** (in a list or a **sampling frame**).

In the simple random-sampling scheme, one generates a set of random digits that are used to objectively identify the individuals to be sampled and measured.

University of Glasgow

Sample mean and variance - SRS

- For a sample of n observations $y_1, \ldots\ldots yn$, the sample mean is:

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

- What is the total number of possible samples of size n that could be collected from a population of size N?

N^n *With replacement*

$\binom{N}{n} = \frac{N!}{n!\,(N - n!)}$ *Without replacement*

In practical terms, what is the more likely option?

University of Glasgow

Simple Random Sampling
What is the Sampling Variability?

- There are fewer options of taking samples without replacement than there are with replacement.
- There is less variability among the possible samples selected without replacement than the possible samples selected with replacement

- In practice, we are often interested in a specific population of finite size.

- This is addressed in two ways:

1. a change in the estimate of the variance
2. a finite population correction factor (FPC) when calculating the variability of the mean

University of Glasgow

Simple Random Sampling
What is the Sampling Variability?

$$s^2 = \frac{\sum\left(y_i - \bar{y}\right)^2}{n-1}$$

- The variance in the mean of all samples, $\bar{y}$, equals the estimated population variance, s^2 , divided by the sample size, n, times the FPC

$$FPC = \frac{N-n}{N} = 1 - \frac{n}{N}$$

$$Var(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = \frac{s^2}{n}(1 - f)$$

University of Glasgow

Simple Random Sampling
What is the Sampling Variability?

$$FPC = \frac{N-n}{N} = 1 - \frac{n}{N}$$

What does the FPC represent?

The correction factor <u>reflects the proportion of the population</u> that remains unknown.

As the sample size *n* approaches the population size, *N*, the FPC factor approaches zero, and the amount of variation associated with the estimate also approaches zero.

University of Glasgow

A numerical example

Observed data from the 10 randomly selected sampling units: 276, 281, 281, 278, 277, 274, 277, 283, 283, 282

Sample mean =

$$\frac{276 + 281 + 281 + 278 + 277 + 274 + 277 + 283 + 283 + 282}{10} = \mathbf{279}\ \text{Bq kg}^{-1}$$

Sample variance = $\frac{9+4+4+1+4+25+4+16+16+9}{10-1} = \frac{92}{9} = \mathbf{10.222}\text{Bq kg}^{-1}$

Sample standard deviation = $\sqrt{10.22}$ = 3.192 Bq kg^{-1}

The FPC is 1-(10/100) = 0.9

The sample variance (of the mean) is therefore **$3.192^2[(1\text{-}0.1)/10] = 0.92$**

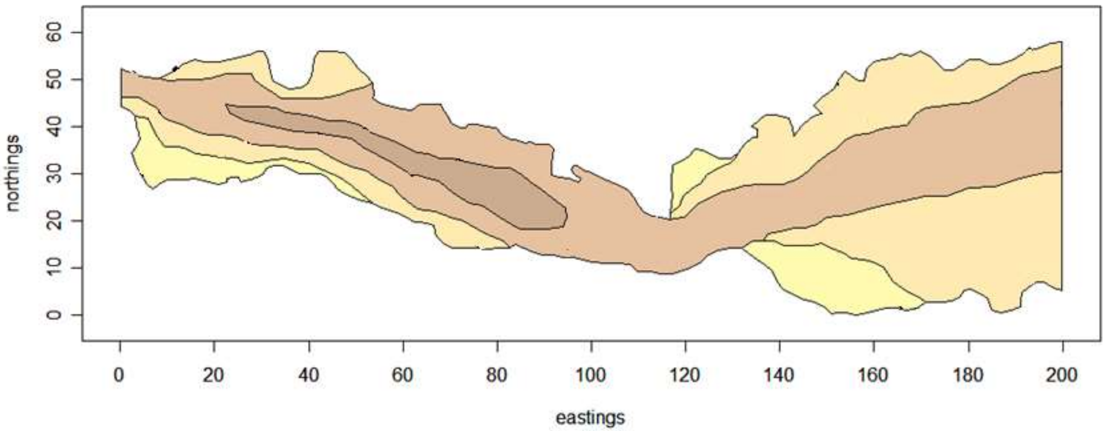
And the random sampling error is **$\sqrt{0.92} = 0.96$.**

Stratified Sampling

Example: ^{60}Co activity in sediment of an estuary

- Cobalt-60 (**^{60}Co**): synthetic radioactive isotope of cobalt
- half-life > 5.000 years
- produced in nuclear reactors.
- aim: estimate the inventory of ^{60}Co in the sediments of an estuary

We know that ^{60}Co is particle reactive and we have a map of sediment type in the estuary.



How would we make use of this information?

University of Glasgow

Stratified Random Sampling

In stratified sampling, the population is divided into two or more strata that individually are more homogeneous than the entire population.

- Assume there are N units in the overall target population
- Divide these N units into L non-overlapping strata such that the variability within each stratum is less than the variability over the entire population

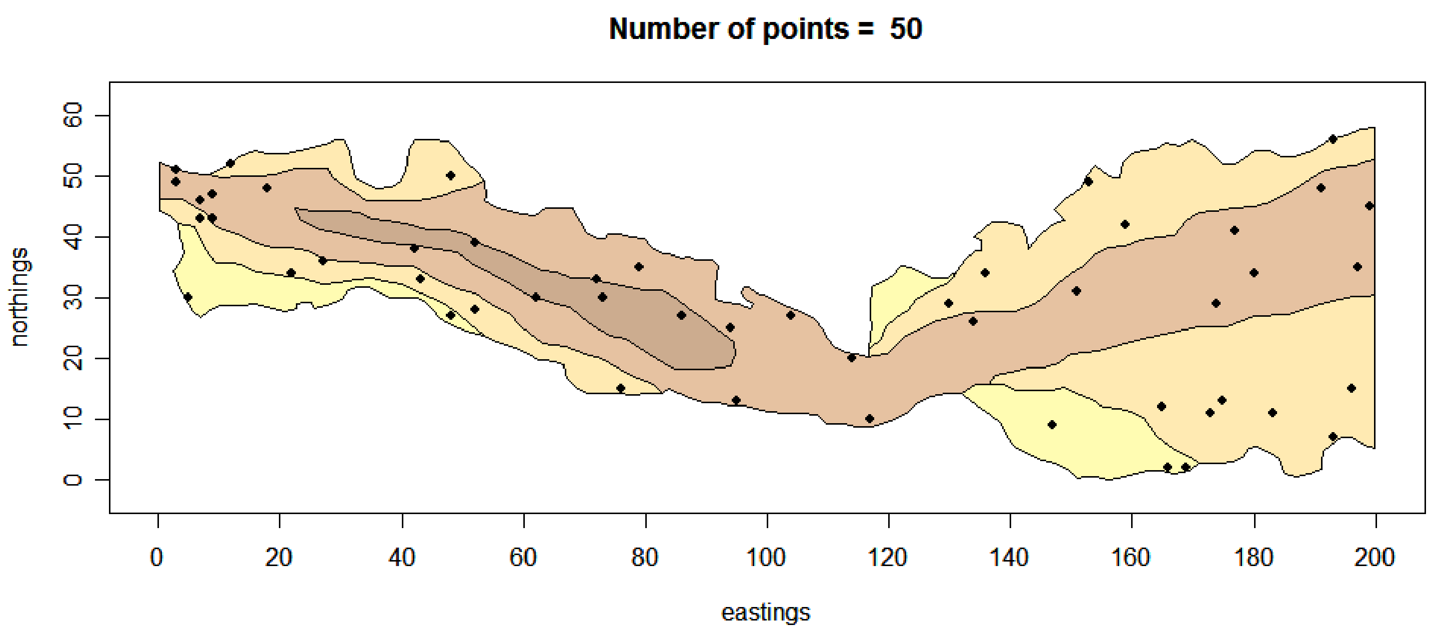
A sampling method is used to estimate the properties of each stratum. To analyse the whole population we need to combine these estimates correctly.

Frequently the proportion of sample observations in each stratum is similar to the stratum proportion in the population (referred to as **proportional allocation**)

University of Glasgow

Example: ^{60}Co activity in sediment of an estuary

If there is knowledge of different strata over the sampling domain (such as soil type), the use of a stratified sample would be recommended and a random sample of locations would be selected within each stratum.



University of Glasgow

Stratified Random Sampling

Stratified population, divided into L non-overlapping strata of sizes $N_1, \ldots, N_L$

From each stratum, we have the mean $\bar{y}_l$, $l=1, \ldots, L$

The overall estimated population mean (i.e. sample mean) is given by A_c

Sample mean

$$A_c = \frac{\sum_l \left(N_l \bar{y}_l\right)}{N}$$

University of Glasgow

Stratified Random Sampling - Stratified RS

- Let $W_l = N_l/N$ be the *l*-th stratum weight. These are assumed to be known prior to sampling.

- From each stratum, we have the subpopulation variances s_l^2 and the overall variance

$$Var(A_c) = \sum_l \left[W_l^2 \frac{s_l^2}{n_l} (1 - f_l) \right]$$

$$where \quad f_l = \frac{n_l}{N_l}$$

University of Glasgow

Comparing SRS and Stratified Sampling

One reason for stratification is to **increase the efficiency** of our estimators

Both are unbiased estimates of the population mean

Now let us consider their variances, if we assume that the stratum sizes are large enough (n_l / N_l) is negligible

University of Glasgow

Stratified Sampling: Inventory

We might be interested in the **inventory** in each stratum (e.g. the total pollutant in each stratum)

The inventory in each stratum is

$$I_l = N_l \mu_l$$

Where μ_l is the population mean of stratum l *(estimated by* $\bar{y}_l$*)*

and the overall inventory is

$$I = \sum_{l=1}^{L} N_l \mu_l$$

Derive the variance of the overall inventory.

Stratified Sampling: Inventory

Derive the variance of the overall inventory.

$$var\,(I) = var\left(\sum N_l \mu_l\right)$$

$$= \sum N_l^2 \mathrm{var}(\mu_l)$$

$$= \sum N_l^2 \frac{\sigma_l^2}{n_l}\left(1 - \frac{n_l}{N_l}\right)$$

In practice we use s_l^2 in place of σ_l^2

University of Glasgow

We will cover

- **Sampling and monitoring - general**
  - **Statistical sampling strategies**
    - **Simple random sampling**
    - **Stratified random sampling**
  - Analysing data from these strategies – how and comparisons
  - How many samples do we need?
- Designing monitoring networks
  - BACI
- Note: Some of this will be revision – remember to set what we are learning in the context of environmental data