Environmental Statistics

Chapter 3: Sampling and Monitoring Networks

Session 2020/2021

University of Glasgow

Recalling…

- Why is thinking about a sampling strategy important?
- Why do we need to think about the variance of our estimators?

- Stratified sampling – what's that?
- Stratified sampling – why?
- Implications for choice of/ allocation of sample size(s)?

What we will cover

- Sampling and monitoring - general
  - Statistical sampling strategies
    - Simple random sampling
    - Stratified random sampling
    - Systematic 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

What we will cover

- Sampling and monitoring - general
  - Statistical sampling strategies
    - Simple random sampling
    - Stratified random sampling
    - **Systematic 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

Systematic Sampling

University of Glasgow

Systematic Sampling

- Assume there are N (= nk) units in the population.
- Then to sample n units, a unit is selected for sampling at random.
- Then, subsequent samples are taken at <u>every k units</u>.
- A systematic sample is thus <u>spread more evenly</u> over the population.
- Systematic sampling has a number of advantages over simple random sampling, not least of which is convenience of collection.

University of Glasgow

Systematic Sampling

- Perhaps a more practicable sampling scheme, but does require some additional thought concerning analysis of results.

- Typically only one of the units is randomly selected.

- One trick is to consider the overall sample, as comprising a series of 'transects' (or **systematic samples**) and to **estimate the mean and variance from each sub-systematic sample.**

- ***t*** is the number of sub systematic samples and ***T*** is the total number of samples

- In a spatial context such as the sediment sampling problem, this would involve laying out a regular grid of points, which are fixed distances apart in both directions within a plane surface.

University of Glasgow

Spatial Sampling

Assume that there is an attribute that is spatially continuous:

- In principle it is possible to measure the attribute at any location defined by coordinates (x, y) over the domain or area.
- in practice it is not.

Systematic sampling:

- the region is considered as being overlaid by a grid (rectangular or otherwise)
- sampling locations are at gridline intersections at fixed distance apart in each of the two directions.

The starting location is expected to be randomly selected.

Both the extent of the grid and the spacing between locations are important. The sampling grid should span the area of interest (the population).

University of Glasgow

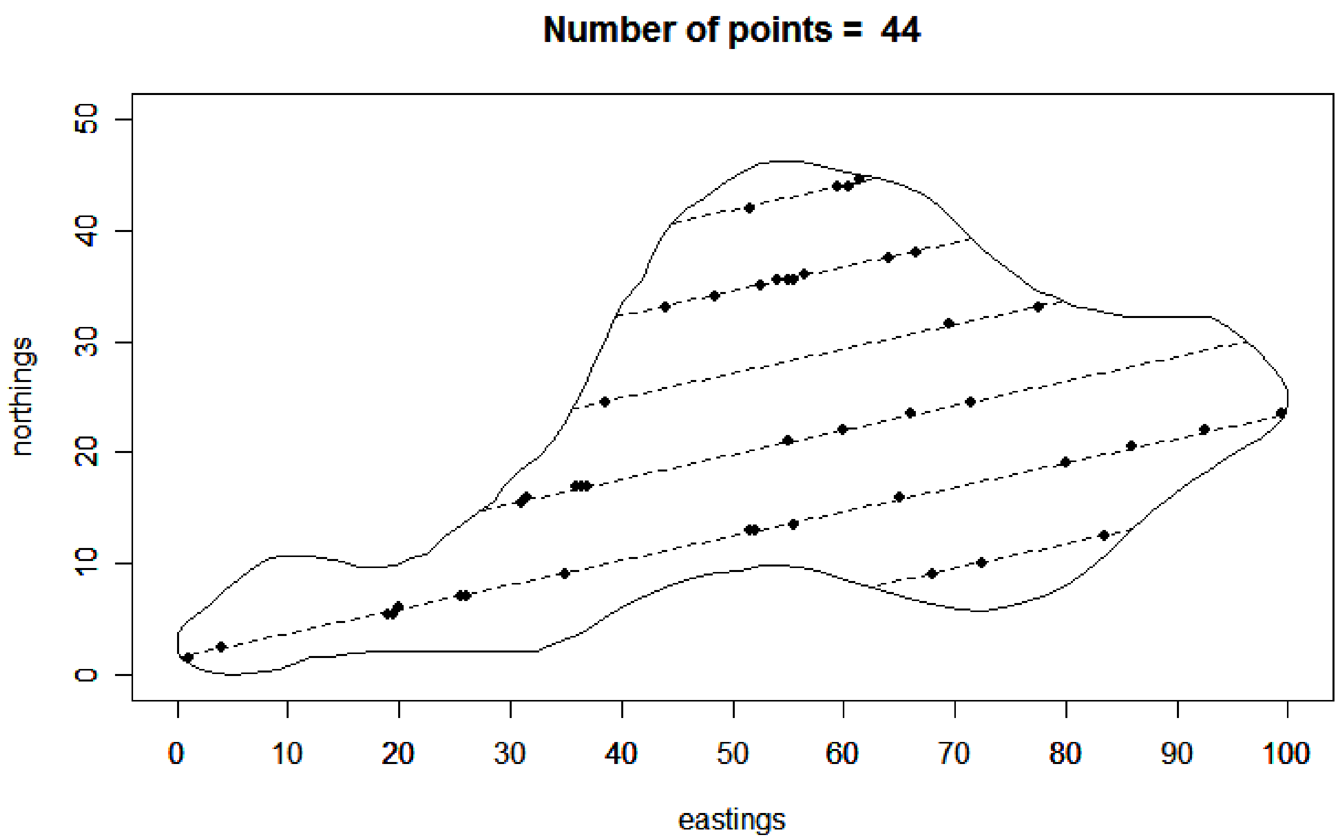
Spatial Sampling
Transects

- A line transect is a straight line along which samples are taken

- The starting point and its orientation will be chosen as part of the sampling scheme.
- In addition, the number of samples to be collected along the transect, and their spacing requires definition.

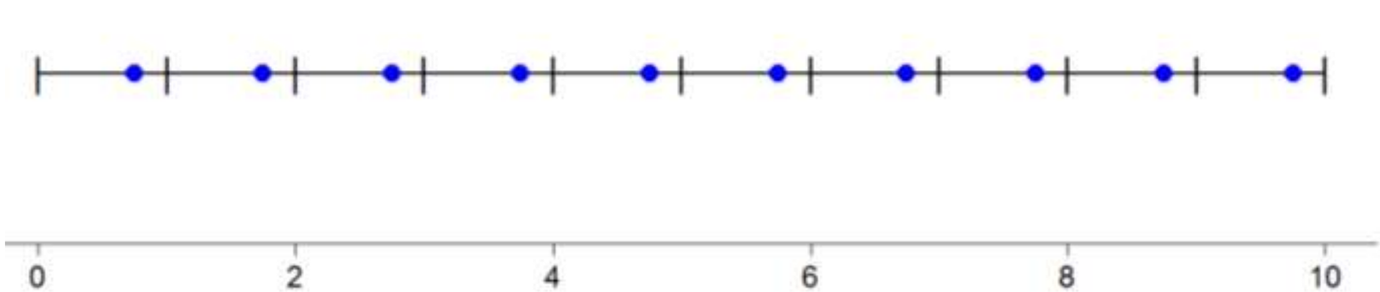
- Samples may be taken at random points along the whole length of the line (continuous sampling) or at systematically placed marked points (systematic sampling).

Spatial Sampling
Transects and Quadrats



University of Glasgow

Systematic Sampling
A: Sampling along a line

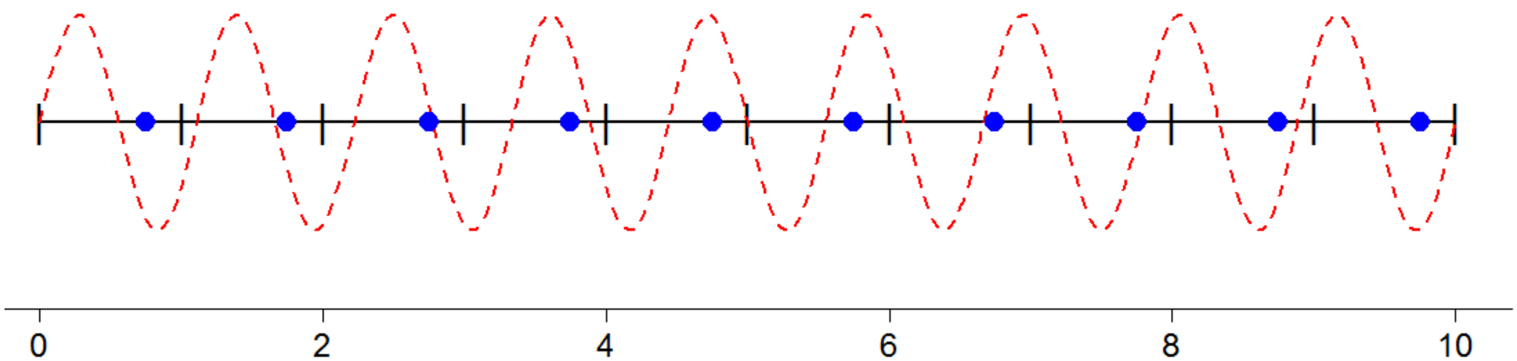


Choose an interval *k* (say 4),
Choose at random a value between 1 and *k* (say 2), then
sample 2nd, 6th, 10th....... and so on

What is the danger of the approach?

University of Glasgow

Systematic Sampling
A: Sampling along a line

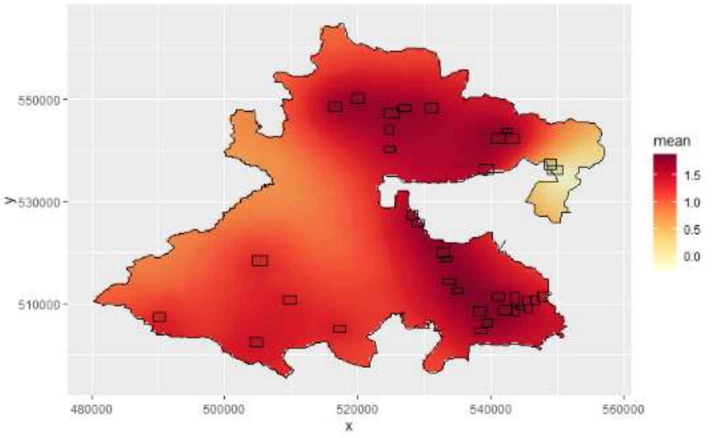


Think about sampling city air pollution on a single day every week – how would the results compare if you sampled on a Wednesday every week compared to a Sunday?

Seasonality, cyclical patterns, periodicity…

Spatial Sampling
Transects and Quadrats

- A quadrat is a well-defined area within which one or more samples are taken.
- The position and orientation of the quadrat are part of the sampling scheme.
- Quadrat sampling (or plot sampling):
- classic tool in ecology
- a series of squares (quadrats) of a set size are placed in a habitat of interest
- species within those quadrats are identified and recorded.



Locations of quadrats in an ongoing study on Orang-Utang nests in Borneo, Milne et al. in preparation

University of Glasgow

Systematic Sampling
B: Sampling over space

Start with a grid – choose the distance between the grid lines (equivalent to fixing n)

Aligned grid

- Choose co-ordinates of starting point, A, at random.
- Repeat A in each area of pre-specified grid spacing
- Choosing A to be at the centre of the square results in a **centrally aligned grid**

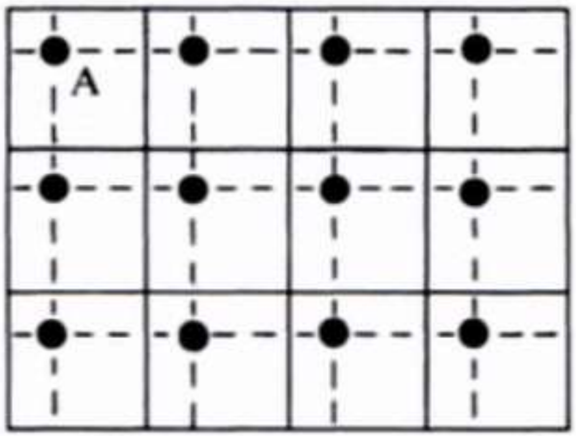
Unaligned grid – co-ordinates of points are randomly generated within each grid square

Triangular grid – modification of aligned grid where points are fixed by a triangular arrangement

Systematic Sampling
B: Sampling over space

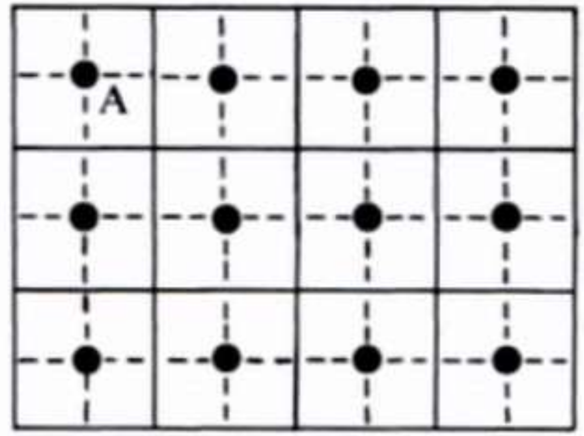
Aligned Square Grid

(a)



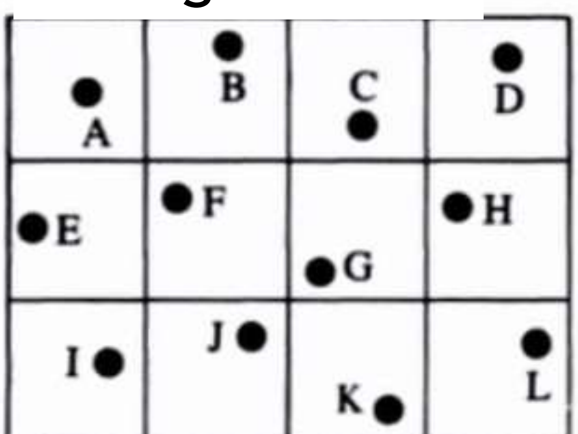
Central Aligned Square Grid

(b)



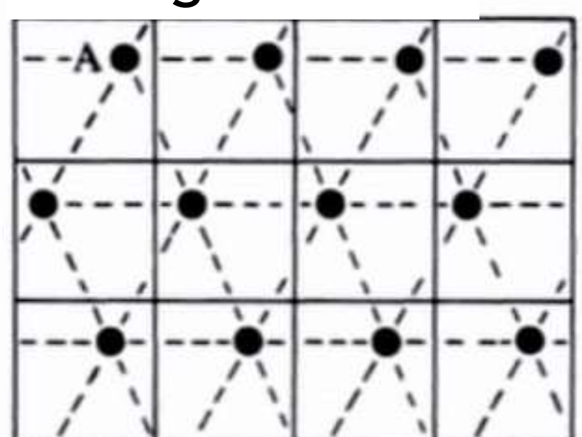
Unaligned Grid

(c)



Triangular Grid

(d)



University of Glasgow

Systematic Sampling
B: Sampling over space

a) Aligned grid and b) centrally aligned grid:
Potential issues?

c) Unaligned grid
Avoids problems with periodicities and combines aspects of both stratified and SRS

d) Triangular grid
Performs well if spatial correlation structure varies with direction

University of Glasgow

Analysis of Systematic Sampling Data

- Main advantage of systematic sampling:
  Easy to apply in practical terms.

- Main disadvantage: Systematic sampling is ordered.

- Unless the population units are in random order it is difficult to get a valid estimate of variance from a single systematic sample because the start position fixes all population units that will be included in the sample.

- **Approach**:
  Think of the systematic sample as being made of multiple systematic sub samples – each of which has a randomly determined starting point.

University of Glasgow

Analysis of Systematic Sampling Data

- The analysis of the data from a systematic sample often depends on making assumptions concerning the population.

- One approach is to consider the overall sample as being made of a series of systematic samples and to estimate the mean and variance from each sub-systematic sample.

- Let

  $t > 1$ be the number of sub systematic samples , each of sample size, n_i , where $i=1,\ldots,t$
  where
  T be the total number of possible sub systematic samples

University of Glasgow

Systematic Sampling
Population mean and variance estimates

$$\bar{y}_{sy} = \frac{\sum_{i=1}^{t} n_i \bar{y}_i}{\sum_{i=1}^{t} n_i} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{t} n_i}$$

$$Var(\bar{y}_{sy}) = \frac{1 - t/T}{t(t-1)} \sum_{i=1}^{t} \left(\bar{y}_{i.} - \bar{y}_{sy}\right)^2$$

$$= \frac{1 - t/T}{t} \frac{\sum_{i=1}^{t} \left(\bar{y}_{i.} - \bar{y}_{sy}\right)^2}{t-1}$$

NOTE: 1 – t/T is a finite population correction factor.

Examples/Summary

University of Glasgow

Sampling Examples

- Aim: Estimate the average height of trees within a forested area of $10km^2$. The distribution of species is fairly uniform.
- What sampling approach could be used – why?

University of Glasgow

Sampling Examples

- Aim: Estimate the average height of trees within a forested area of $10km^2$. The distribution of species is fairly uniform.

- What sampling approach could be used – why?

- Population: All trees within the area
- Sampling Unit; A single tree

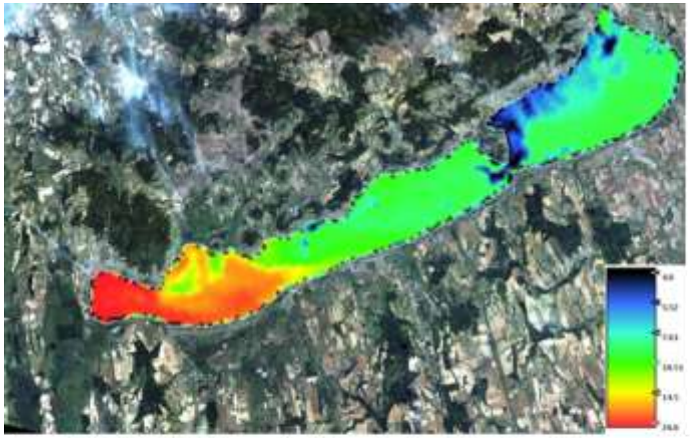
- Simple random sampling
  - uniform distribution of species so strata not necessary,
  - can cover large area at low 'cost' (financial/time),

- Alternatively systematic sampling may be used – use a grid to define quadrates and samples – easy to implement but analysis may be more complicated (in terms of estimating variance)

University of Glasgow

Sampling Examples

- Aim: Estimate average value of chlorophyll A in a lake where there is a strong trophic gradient.
- What sampling approach could be used – why?
- What problems might there be?



MERIS Chlorophyll a – Lake Balaton, Hungary

Population: All possible water samples from the lake

Sampling Unit; A single water sample

Stratified Random Sampling –
Could use transects or quadrats within areas

Underlying population is heterogeneous – 'zones'
Practical problems might be due to collection from boat – drift.
Physical constraints (i.e. depth) may prevent boat accessing some parts of lake

University of Glasgow

What have we learned?

- There are a range of sampling approaches available
- The approach selected may depend on the aim of the study and the homogeneity of the underlying population
- Practical considerations also have to be accounted for
- The spatial distribution of the population will also play a role in the sampling scheme selected
- We have not (yet) considered the effects of sampling over time (sampling frequency)

University of Glasgow

What have we learned?

- The estimates of summary statistics need to take into account the sampling scheme used to obtain the data
- Simple random sampling has the advantage of ease of collection and analysis **compared to stratified random sampling**
- In general, the variance associated with data collected using this approach is greater than that collected using stratified sampling

How many samples will I need?

University of Glasgow

How many samples do I need?

- We can think about this in terms of
    - **Precision**: with what precision do I want (need) to estimate the mean/median/proportion?
    - **Power**: How small a difference is it important to detect and with what degree of certainty?

University of Glasgow

How many samples do I need?

CI for the population mean, μ.

A general formula for a CI is given by $\bar{x} \pm t_{1-\alpha/2}\sqrt{var(\bar{x})}$

Where α is the significance level, usually 5%

The formula for the standard error, $\sqrt{var(\bar{x})}$, contains n so if we specify how precise we want our interval to be then we can solve to find n.

Note: the value from the t distribution depends on the sample size, but in practice when n>30, the value of t is close to z (standard normal) and so we often just use the z value e.g. 1.96 for a 95% CI.

University of Glasgow

How many samples do I need?

Let's say that $var(\bar{x})$ should be $\leq V$

so that $$\frac{s_x^2}{n} \leq V$$

And hence $$n \geq \frac{s_x^2}{V}$$

Where s_x^2 is the sample variance for x.

BUT…. We can't calculate s until after the sample is collected.

How do we know that in advance?

University of Glasgow

How many samples do I need?

How do we know what level of variability we will have associated with our estimate?

(a) previous experience
(b) using other published papers
(c) carrying out a pilot study

University of Glasgow

How many samples do I need? Example

PCB (Polychlorinated biphenyl):

- AIM: to estimate the mean concentration with an estimated standard error (*e.s.e.*) precision of ±0.1 mg kg^{-1}.
- The variation of PCB in salmon flesh is 3.19^2.
- How many samples would be required to obtain an estimate with this level of precision?

How many samples do I need? Example

Since the $e.s.e.$ of the sample mean is $\frac{s}{\sqrt{n}}$, then one must solve for n, for example:

$$n = \left(\frac{s}{e.s.e.}\right)^2 = \left(\frac{3.19}{0.1}\right)^2 = 1018$$

Thus this degree of improvement in precision, can only be achieved by increasing the number of samples taken to **approximately 1000**.

This may well be **impractical**; therefore the only solution may be to accept a lower precision.

University of Glasgow

How many samples do I need?

Example

We want to estimate the mean concentration of a pollutant, $\bar{x}$

We know from pilot studies that the variability of X is approximately 100.

Ideally, we want the variability of $\bar{x}$ to be less than 4

University of Glasgow

How many samples do I need?

Example

We want to estimate the mean concentration of a pollutant, $\bar{x}$

We know from pilot studies that the variability of X is approximately 100.

Ideally, we want the variability of $\bar{x}$ to be less than 4

$$var\ (\bar{x}) = \frac{s_x^2}{n} = \frac{100}{n}$$

$$4 = \frac{100}{n}$$

$$n \cong 25$$

University of Glasgow

How many samples do I need?
Stratified Random Sampling

- For a stratified sample, the problem becomes more difficult

- Not only need we consider the total sample size but also how it is allocated in the different strata.

- One approach is to specify a cost model

  - an overall cost for undertaking the survey c_0
  - and an individual cost for observations from each stratum c_i

- We could attempt to maximise the **efficiency** - minimise the variance of x_{st} for a given total cost C

University of Glasgow

How many samples do I need? Stratified Random Sampling

Cost Model: Total Cost

$$C = c_0 + \sum_{l=1}^{L} c_l n_l$$

Fixed overhead cost c_0
Cost per population unit in the l-th stratum, c_l

Then the optimum number of samples in stratum l is

$$n_l = n \frac{W_l \sigma_l / \sqrt{c_l}}{\sum_{l=1}^{L} W_l \sigma_l / \sqrt{c_l}}$$

σ_l is the population standard deviation for stratum l, n is the total number of samples in all strata. In practice, we replace σ_l with s_l

University of Glasgow

How many samples do I need? Stratified Random Sampling

a) If all stratum costs are the same,

$$n_l = n\frac{W_l\sigma_l}{\sum_{l=1}^{L} W_l\sigma_l}$$

Often called Neyman Allocation

b) else

$$n_l = n\frac{W_l\sigma_l/\sqrt{c_l}}{\sum_{l=1}^{L} W_l\sigma_l/\sqrt{c_l}}$$

A simple alternative is <u>proportional allocation</u>;

$$n_l = nW_l = \frac{nN_l}{N}$$

For prop allocation we don't need to know stratum standard deviations, BUT, if the we have a good estimate of these then (a) or (b) are more accurate

How many samples do I need?
Stratified Random Sampling

How do we know what n is?

a) pre-specify the total cost
b) pre-specify the variance
c) pre-specify the margin of error that is acceptable

How many samples do I need?

- What do we do when there are constraints which prevent us from increasing the sample size (cost etc.)

- Possibly this could be achieved by changing the design of the study.

  e.g. a paired design could be more efficient

- Within-subject differences are usually less variable than between subject differences (i.e. lower standard error) so the sample size required to detect a given difference will be lower.

This will depend on the aim of the study!