

Sampling for natural resources and environmental modelling and monitoring

D G Rossiter

May 1, 2017

Copyright © 2007 ITC, 2010, 2012-2013 University of Twente, 2014-2017 D G Rossiter

All rights reserved. Reproduction and dissemination of the work as a whole (not parts) freely permitted if this original copyright notice is included. Sale or placement on a web site where payment must be made to access this document is strictly prohibited. To adapt or translate please contact the author (dgr2@cornell.edu).



Cornell University
College of Agriculture and Life Sciences

Topic: Why sample?

We sample to **understand** some aspect of reality (natural or social).

With that understanding we can **take some action**.

The most basic concepts in sampling theory are:

- **individual**
- **population**
- **sample**

We first define these, and their **relation** that underlies sampling theory.



Population and sample

Individual The object about which we want to make a statement

Population All the individuals of interest, about which we want to make some statement

Actual populations:

- *Examples:* all farmers in a village / district / province / country; all 1x1 m soil bodies in a defined area

Hypothetical populations:

- *Example:* All rice crops that could be produced in a defined area.

Sample The set of individuals actually observed

- *Examples:* twenty selected farmers; fifty soil borings; twenty rice crops actually produced

So the **sample** is a **subset** of the **population** of **individuals**



Why sample?

- We want to know the attributes of some **population**
 - Mean, median, range, variance, distribution . . .
 - Predicted value of **unsampled individual**, maybe with prediction **uncertainty**
 - * Spatial: unvisited location
 - * Temporal: unsampled time point or interval
- It is usually **not possible** to observe all individuals
 - For a **hypothetical** or **infinite** population it is not possible
 - For a **finite** population, it is **inefficient** if all we need is an **estimate** from the population to a known **precision**
 - Can not sample in the past or the future
- The **cost** (field, lab.) to observe an individual may be high, budgets are limited



The sample

- So, we only observe a portion of the population
- This is called the **sample**
 - Note: One observation from the sample is sometimes also called ‘a sample’
 - But, it is more correct to call it an **observation** or **sampling unit**



How does sampling allow inferences about populations?

How can it be that we can make a statement about the **population** as a whole, if we haven't observed all of it?

This is possible because of:

1. the **law of large numbers**

- The larger the sample compared to the population, the more the sample parameters approach the population parameters;

2. the concept of **probability** sampling: each individual has some defined chance of being sampled.

- Note this is *not* necessarily **equal** chance.

3. the concept of **representativeness**: selected individuals are “typical” of the population.

- Probability sampling ensures representativeness.



Sampling contexts

- **Non-spatial, non-temporal**
 - space or time are not relevant to the analysis
 - sampling can be in space or over time but these are not considered relevant
- **Spatial**
 - location of observations is important for analysis, e.g., mapping
- **Temporal**
 - time that observations are made is important for analysis, e.g., monitoring
- **Spatio-temporal**
 - both space and time are relevant for analysis, e.g., mapping trends over time

We begin with non-spatial, non-temporal.

Topic: Sampling concepts

1. Steps in sampling
2. Designing a sampling strategy

In a later topic we will see how to compute sample sizes.



Steps in sampling

1. Define the **research questions**
2. Define the **target population**, **target variable** and **target parameter**
3. Define the **quality measure**
4. Specify the **sampling frame**
5. Specify the **sampling design**
6. Determine the **sample size**
7. Determine the **sampling plan**
8. Carry out the **sampling** in the field



Research questions

Without knowing **what you want to know** it is impossible to design a sampling scheme to **find out**!

These questions should be as precise as possible. Compare:

- Which is the most widely-grown rice variety in a district?
- What is the total area under rice in the district?
- What proportion of the rice production in a district is from this variety?
- What was the mean yield of this rice variety in a given year?
- What is the yield potential of this rice variety under optimal management?
- What is the relation between yield and soil preparation method? (etc.)

For all these: How much **certainty** (precision) is needed in the answers?



Target population

What exactly is the **population** about which we want to make inferences?

This is the (possibly hypothetical) enumeration of all the **individuals** that make up the population:

- Clear **rules** for **inclusion** or **exclusion** from the sample
- If the population is inherently **continuous** (e.g. “forest cover”), we need a **discretization** rule to divide into individuals.
 - This is the size and shape of the “individual”
 - Equivalent to the concept of geostatistical **support**



The sampling unit

Given the population, we must define it in terms of the **sampling units**.

These are the **individuals** which could be observed or measured.

We must specify:

- How to **identify** (recognize, limit) it in the field;
- How to actually **make the observation** (procedures to be followed):
 - site preparation for sampling
 - what is to be measured
 - measurement scale and resolution
 - for interviews: how to ask questions (experimental **protocol**)

and if it is a geographic individual:

- its **spatial dimensions**, called the **support**.



Target variable

This is the variable to be **measured** for each sampling unit.

Note that there may be several target variables of interest in the same sampling campaign.

Examples:

- Soil grain size fractions (gravel, coarse sand . . .) in the 0-20 cm and 30-50 cm layers;
- Whether the soil is above a regulatory threshold for some pollutant (“contaminated”) or not
- Age of each child in a household and whether s/he attends school regularly

Target parameter

This is the **statistical measure** which will summarize the target variable. It is closely related to the research question. What do we really want to know?

Examples:

- Mean proportion of each soil grain size fraction over a study area;
- Mean proportion of each soil grain size fraction of all 1 ha blocks in the study area (**mapping**)
- Minimum, maximum, percentiles . . .
- Variance (as a measure of heterogeneity)

These will be estimated by **statistical inference**.

The sampling frame

This is the technical term for the list or **enumeration** of **all possible sampling units** for the survey.

- Note that this does *not* have to be the population!
- But if it is not, the researcher must **argue** that it is **representative** of the population; this **meta-statistical** reasoning is used to make inferences about the population from the sample.



Example of sampling frame

1. The **population** is all shifting-cultivation fields in the humid tropical rainforest of Cameroon;
2. The **sampling frame** includes all shifting-cultivation fields in four “representative” villages;
3. The **sample** will be some selection of these fields.

We have to argue (with evidence) that the four villages **represent** the whole area.

We have to ensure that each individual in the sample has a known **probability** of being selected.



Sampling fraction

This is the **proportion** of individuals in the **sampling frame** that are actually selected and sampled. If N is the population size and n is the number of individuals sampled:

$$f = \frac{n}{N}$$

Example 1: In a study area of 100 ha = 1 000 000 m²; sampling individuals are defined as 10 x 10 m surface areas; so there are $10^6/10^2 = 10^4$ sampling individuals.

If we make 50 observations (e.g. biomass in the 10 x 10 m area) the sampling fraction is $50/10^4 = 0.005 = 0.5\%$.

Example 2: Sampling individuals are households; the sampling frame is the 150 households in a village.

If 20 are selected and interviewed, the sampling fraction is $20/150 = 13.\bar{3}\%$



Topic: Sampling designs

This is plan which says **which sample individuals** to **select** from the **sampling frame**. These are of several types:

1. Opportunistic
2. Purposive
3. **Probability**
 - (a) Systematic
 - (b) Random



Opportunistic sampling

Also called “convenience” or “grab”.

- wherever we happen to be, and we see something **“interesting”**
- or, wherever we are **able** to sample
- often used in reconnaissance geographical survey, going along the main roads
- also used in rapid rural appraisals

This has the obvious serious flaw that there is **no way to evaluate its representativeness**. It should **not be used for any statistical inference or even descriptive statistics**, just to get a very rough idea of the range of values in a study area.

(However, any sampling plan can be used for **model-based** spatial inference, e.g., kriging.)

Purposive sampling

We go **on purpose** to a specific site, based on its characteristics

- it's something we want to measure
- we suppose (i.e. assume based on prior evidence) that it **represents some sub-population**

This is often used with small sample sizes, where we will not make statistical inferences anyway, but want to make sure to see certain sites.

It differs from opportunistic sampling in that we have an *a priori* (before the field) reason to sample a given individual.

Example: select sites for soil profile description in locations where we expect a certain kind of soil “typical” for the landscape.

Again, **can not make any statistical statements** about the population.

Probability sampling

This is a scheme where there is a **known probability** for any sampling unit to be selected for the sample.

- This does not have to be equal probability, just as long as it is known

This is the ideal for **statistical inference**.

We now consider some **probability sampling designs**:

1. Systematic
2. Completely random; also called Simple random
3. Stratified random
4. Clustered random



Systematic sampling

Sampling units are drawn from the frame in some **regular pattern** (“system”), either in geographic or feature space

- E.g. soils sampled at intersection points of a 500x500 m grid
- E.g. interview every tenth person who walks into a community centre

The **starting point** is selected at **random** and then the other points are at fixed **offsets**; before the first point is selected, each sampling individual has an **equal probability** of being selected.

Since there is **no bias**, this can be used for statistical inference. The danger is of some **periodic effect** that matches the system.



Systematic sampling (2/2)

An advantage for **spatial sampling**: covers the whole area **evenly**; any **mapping by interpolation** from this sample will have a known minimum accuracy possible with this number of samples.

Note: in spatial sampling a small **jitter** may be added to avoid periodic effects.



Completely random sampling design

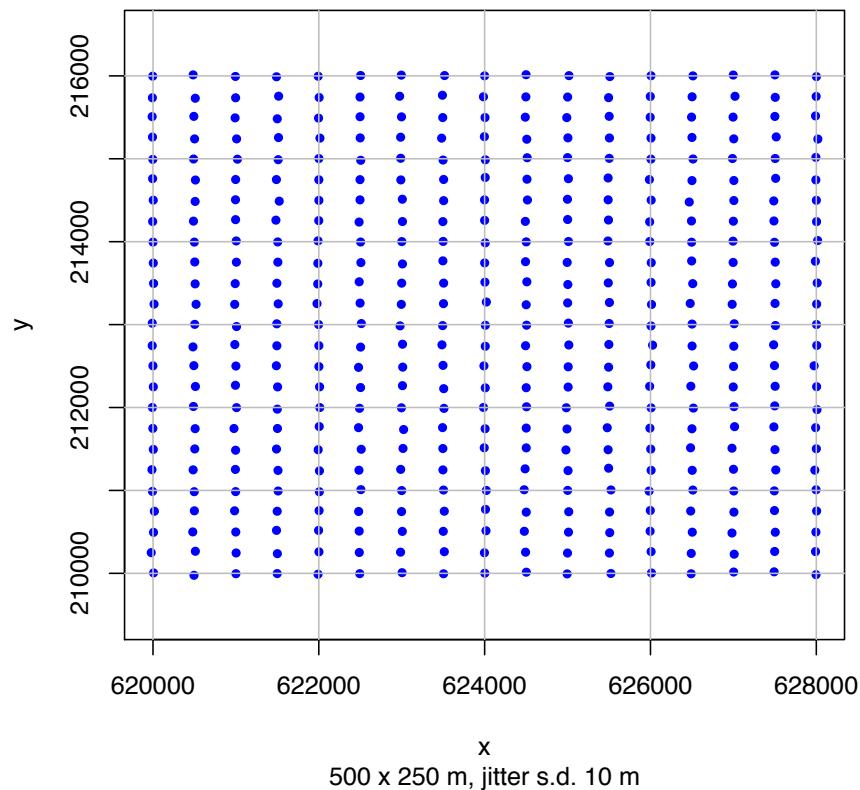
Sampling units are selected from the frame at **random**; this is also called **simple random**.

- Each sampling individual has an **equal probability** of being selected
- In practice: **number** the units, use a random number generator from the **uniform** distribution to select them
- For **geographic** samples, the **coordinates** already provide a numbering, so just select a **uniform random number** within the range of each coordinate

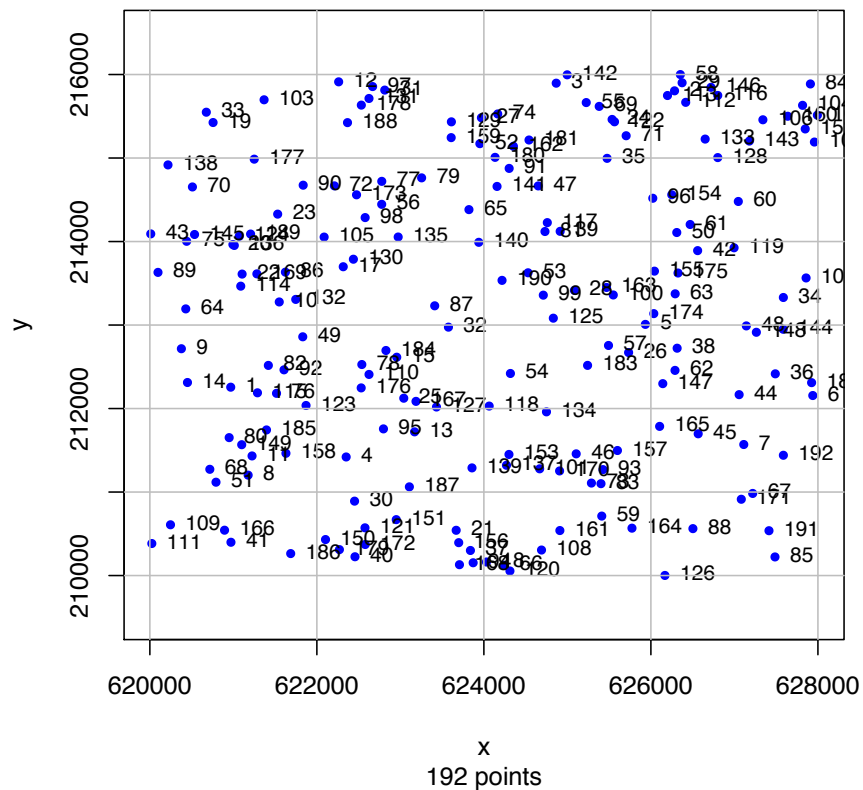


Example of two spatial sampling designs

Systematic with jitter



Simple random



Advantages of the completely random design

- Each sampling unit has the **same probability** of being selected, so there is **no need to weight the observations** in computations
- For example, the estimated population mean and its variance are calculated as:

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_i^n x_i$$
$$\hat{\sigma}_{\mu} = N^2(1 - f) \frac{s^2}{n}$$

where s^2 is the sample variance. As $f \rightarrow 1$, the variance of the mean $\hat{\sigma}_{\mu} \rightarrow 0$.

- There is **no constraint** on the **relative location** (in space) or sequence (in the sampling frame) of the observations
- In spatial sampling, provides point-pairs at many separations; good for **variogram estimation** (although not the most efficient)

Disadvantages of the completely random design

- Assumes one homogeneous population (**no sub-populations**); if these exist they may have different:
 - expected value, and/or
 - variance
- If so, this design is **inefficient**: too many samples to achieve a desired precision.
- For spatial sampling, inefficient in **transport time**



Stratified random sampling design

This addresses the situation where the population can be divided into **subpopulations** with:

- different **expected values or variances**, or
- different **importance** (required precision of results).

The sampling frame is divided into **strata** by some criterion

- *Example* in social surveys: gender, employment status, educational level . . .
- *Example* in geographical surveys: soil or geologic map units; management unit in a forestry concession . . .

Within each stratum the observations are **random**

(Note: There are also stratified **systematic** designs and **cluster** designs.)



How to distribute sample numbers in stratified designs?

1. Sample numbers in each stratum are **proportional to its size** within the full population (**proportional stratified**); analysis is then as for completely random.
 - The reason to use strata in this case is to **make sure we sample all subpopulations**; otherwise small subpopulations might be missed by chance.
2. Use larger (proportional) sample numbers where there is **more variability** (more heterogeneous); this is the most **efficient** to determining population parameters (e.g. overall mean)
3. Distribute sample numbers by **importance** of each stratum to the objectives; you will know more about the most important strata

The latter two require different formulas for computing population parameters such as inferred means, variances etc.



Clustered sampling designs

These overcome the logistical problems associated with completely random designs.

- Cluster **centres** may be chosen at random or systematically;
- From each centre, observations are distributed either:
 - **randomly 'nearby'** (within some user-defined distance)
 - **systematically** to cover some user-defined area; note starting point must be random (perhaps a random cluster centre)
- Major advantage: efficient for sample **logistics** such as transport, sample handling
- Spatial sampling: random-within-clusters also suited for estimating **variograms**

Estimates from unequal probability designs

All designs except for completely (simple) and proportional stratified random sampling have **unequal probabilities** of sampling a given individual. So any **estimate** from the sample (e.g. mean, variance, total, proportion) must take this into account.

Estimates for Stratified Random Sampling

First some notation:

- There are k strata;
- N_h is the number of **individuals** in stratum h ; total N
- n_h is the number of **observations** in stratum h ; total n
- $W_h = N_h/N$ is the proportion of the population in stratum h ; the stratum **weight**
- $w_h = n_h/n$ is the proportion of the sample in stratum h ; the sample **allocation**
- $f_h = n_h/N_h$ is the **sampling fraction** for stratum h
- \bar{y}_h, μ_h are the **sample** and **population** means for stratum h
- \bar{s}_h^2, σ_h^2 are the **sample** and **population** variances for stratum h



Estimates (continued)

Then we compute the overall mean as:

$$\bar{y} = \sum_{h=1}^k W_h \cdot \bar{y}_h$$

The variance of each stratum mean is:

$$\text{Var}(\bar{y}_h) = \frac{N_h - n_h}{N_h n_h} \bar{s}_h^2$$

and of the overall mean is:

$$\text{Var}(\bar{y}) = \sum_{h=1}^k W_h \frac{N_h - n_h}{N_h n_h} \bar{s}_h^2$$

Implications of these formulas

- As $f_h \rightarrow 1$ (more of stratum h is sampled), $v(\bar{y}) \rightarrow 0$; so if s_h is large we should allocate more samples to this stratum in order to reduce its contribution to the overall variance
- Although, as $p_h \rightarrow 1$ (the stratum represents more of the population), the contribution to the overall variance increases.



R methods for sampling design

- In the base package: `sample`; `runif` for random numbers from the uniform distribution
 - `sample` can also be used with weighted probabilities of selection
- In the `sp` package: `spsample` for spatial objects
- Package `spcosa` implements “Spatial Coverage Sampling and Random Sampling from Compact Geographical Strata”



Topic: Visualizing sampling designs

Following are screenshots from Google Earth of six spatial sampling designs covering the same study area, created with `spsample`:

1. completely random
2. stratified random (4 equal-size strata)
3. rectangular grid, random origin
4. hexagonal grid, random origin

continued . . .



...

5. non-aligned rectangular grid , random origin

6. clustered

- three cluster centres randomly located, then four points per cluster

7. stratified cluster

- one cluster of three observations in each of four strata



Google Earth

Search

Places

- NL
 - samp_area.kml
 - samp_strata.kml
 - Random sample
 - Stratified rando...
 - Grid sample
 - Hex grid sample
 - Non-aligned gr...
 - Clustered sample

Layers

Earth Gallery >>

- Primary Database
 - Borders and Labels
 - Places
 - Photos
 - Roads
 - 3D Buildings
 - Ocean
 - Weather
 - Gallery
 - Global Awareness
 - More

Image © 2014 Aerodata International Surveys

1.98 m

Google Earth

D. C. Rossiter

Another completely random sample



Google Earth

Search

Places

- NL
 - ☒ samp_area.kml
 - ☒ samp_strata.kml
 - ☐ Random sample
 - ☒ Stratified rando...
 - ☐ Grid sample
 - ☐ Hex grid sample
 - ☐ Non-aligned gr...
 - ☐ Clustered sample

Layers

Earth Gallery >>

- ☐ Primary Database
 - ☐ Borders and Labels
 - ☐ Places
 - ☐ Photos
 - ☐ Roads
 - ☐ 3D Buildings
 - ☐ Ocean
 - ☐ Weather
 - ☐ Gallery
 - ☐ Global Awareness
 - ☐ More

Image © 2014 Aerodata International Surveys

1.98 m

Google Earth

D. C. Rossiter

Google Earth

Search

Places

- NL
 - ☒ samp_area.kml
 - ☐ samp_strata.kml
 - ☐ Random sample
 - ☐ Stratified rando...
 - ☒ Grid sample
 - ☐ Hex grid sample
 - ☐ Non-aligned gr...
 - ☐ Clustered sample

Layers

Earth Gallery >>


- ☐ Primary Database
 - ☐ Borders and Labels
 - ☐ Places
 - ☐ Photos
 - ☐ Roads
 - ☐ 3D Buildings
 - ☐ Ocean
 - ☐ Weather
 - ☐ Gallery
 - ☐ Global Awareness
 - ☐ More

Image © 2014 Aerodata International Surveys

1.98 m

Google Earth

D. C. Rossiter



Google Earth

Search

Places

- NL
 - ☒ samp_area.kml
 - ☐ samp_strata.kml
 - ☐ Random sample
 - ☐ Stratified rando...
 - ☐ Grid sample
 - ☒ Hex grid sample
 - ☐ Non-aligned gr...
 - ☐ Clustered sample

Layers

Earth Gallery >>


- ☐ Primary Database
 - ☐ Borders and Labels
 - ☐ Places
 - ☐ Photos
 - ☐ Roads
 - ☒ 3D Buildings
 - ☐ Ocean
 - ☐ Weather
 - ☐ Gallery
 - ☐ Global Awareness
 - ☐ More

Image © 2014 Aerodata International Surveys

1.98 m

Google Earth

D. C. Rossiter



Google Earth

Search

Places

- NL
 - ☒ samp_area.kml
 - ☐ samp_strata.kml
 - ☐ Random sample
 - ☐ Stratified rando...
 - ☐ Grid sample
 - ☐ Hex grid sample
 - ☒ Non-aligned gr...
 - ☐ Clustered sample

Layers

Earth Gallery >>


- ☐ Primary Database
 - ☐ Borders and Labels
 - ☐ Places
 - ☐ Photos
 - ☐ Roads
 - ☐ 3D Buildings
 - ☐ Ocean
 - ☐ Weather
 - ☐ Gallery
 - ☐ Global Awareness
 - ☐ More

Image © 2014 Aerodata International Surveys

1.98 m

Google Earth

D. C. Rossiter



Google Earth

Search

Places

- NL
 - ☒ samp_area.kml
 - ☐ samp_strata.kml
 - ☐ Random sample
 - ☐ Stratified rando...
 - ☐ Grid sample
 - ☐ Hex grid sample
 - ☐ Non-aligned gr...
 - ☒ Clustered sample

Layers

Earth Gallery >>


- ☐ Primary Database
 - ☐ Borders and Labels
 - ☐ Places
 - ☐ Photos
 - ☐ Roads
 - ☒ 3D Buildings
 - ☐ Ocean
 - ☐ Weather
 - ☐ Gallery
 - ☐ Global Awareness
 - ☐ More

Image © 2014 Aerodata International Surveys

1.98 m

Google Earth

D. C. Rossiter



Another clustered sample (3 clusters of 4 each)



Stratified cluster sample (4 strata, 1 cluster of 3 per stratum)



R code (1)

```
# set up sampling area
require(sp)
max.n <- 5791250; min.n <- 5790700
max.e <- 353900; min.e <- 353100
corners <- as.matrix(rbind(c(min.e, min.n), c(min.e, max.n),
                           c(max.e, max.n), c(max.e, min.n), c(min.e, min.n)))
str(p <- Polygon(corners))
str(pp <- Polygons(list(p), "bounds"))
str(spp <- SpatialPolygons(list(pp)))
EPSG <- make_EPSG()
(EPSG[grep("WGS 84 / UTM zone 32N", fixed = T, EPSG$note), ])
ix <- which(EPSG$note=="# WGS 84 / UTM zone 32N")
(epsg.utm32.code <- EPSG[ix,"code"])
proj4string(spp) <- CRS(paste("+init=epsg",epsg.utm32.code,sep=":"))

# set up sampling plan
n.plan <- 12
samp.pts.random <- spsample(spp, n=n.plan, type="random")
samp.pts.random <- SpatialPointsDataFrame(samp.pts.random, data=data.frame(id=1:n.actual))
# convert to WGS84 long/lat as required by Google Earth and native format GPS
samp.pts.random.84 <- spTransform(samp.pts.random, CRS(paste("+init=epsg",wgs84.code,sep=":")))
```



R code (2)

```
# write out as text files for import into GPS; round to 0.00001 decimal degrees, about 1.1 m
write.table(round(coordinates(samp.pts.random.84),5),
            file="./coords/random84.txt", col.names=F, row.names=F)

# export to KML
kmlPoints(samp.pts.random.84, kmlfile="./KML/samp_random.kml",
          kmlname="Random sample", name=samp.pts.random.84$id,
          icon="http://maps.google.com/mapfiles/kml/pushpin/red-pushpin.png")
```



R code (3) – stratified random sampling

```
mid.n <- (max.n + min.n)/2; mid.e <- (max.e + min.e)/2
corners <- as.matrix(rbind(c(min.e, mid.n), c(min.e, max.n),
                           c(mid.e, max.n), c(mid.e, mid.n), c(min.e, mid.n)))
p.nw <- Polygons(list(Polygon(corners)), ID="NW")
corners <- as.matrix(rbind(c(mid.e, mid.n), c(mid.e, max.n),
                           c(max.e, max.n), c(max.e, mid.n), c(mid.e, mid.n)))
p.ne <- Polygons(list(Polygon(corners)), ID="NE")
corners <- as.matrix(rbind(c(mid.e, min.n), c(mid.e, mid.n),
                           c(max.e, mid.n), c(max.e, min.n), c(mid.e, min.n)))
p.se <- Polygons(list(Polygon(corners)), ID="SE")
corners <- as.matrix(rbind(c(min.e, min.n), c(min.e, mid.n),
                           c(mid.e, mid.n), c(mid.e, min.n), c(min.e, min.n)))
p.sw <- Polygons(list(Polygon(corners)), ID="SW")
spp.4 <- SpatialPolygons(list(p.nw, p.ne, p.se, p.sw))
proj4string(spp.4) <- CRS(paste("+init=epsg",epsg.utm32.code,sep=":"))

# random sample in each of the four strata
samp.pts.srand <- SpatialPoints(matrix(0, nrow=1, ncol=2)) # dummy first row
proj4string(samp.pts.srand) <- CRS(paste("+init=epsg",epsg.utm32.code,sep=":"))
for (i in 1:4) {
  s.tmp <- spsample(spp.4@polygons[[i]], n=n.plan/4, type="random")
  proj4string(s.tmp) <- CRS(paste("+init=epsg",epsg.utm32.code,sep=":"))
  samp.pts.srand <- rbind(samp.pts.srand, s.tmp)
}
samp.pts.srand <- samp.pts.srand[-1,] # remove dummy first row
samp.pts.srand <- SpatialPointsDataFrame(samp.pts.srand, data=data.frame(id=1:n.plan))
```

Topic: Practical issues in field sampling

1. Navigation
2. At the site
3. Inaccessible / non-population locations

In a later topic we will see how to compute sample sizes.



Navigation

- with GPS, preferably with map background (e.g., smartphone);
- map and compass from control points (e.g., road intersections);
- compare with airphoto.

In all cases compare with landscape.

Precision of location only has to match precision of support.



The sampling unit

This is the thing we observe in the field.

Recall: We must specify:

- How to **identify** (recognize, limit) it in the field;
- its **spatial dimensions**, called the **support**;
- How to actually **make the observation** (procedures to be followed):
 - site preparation for sampling;
 - what is to be measured;
 - measurement scale and resolution.



At the site

1. Record location;
2. Take photos from location in four directions to document;
3. Take photo of location.
 - This can be used later to re-find the same site, or identify the land cover etc.

Inaccessible / non-population locations

Why?

- No **permission** on site / to access site;
- Too **dangerous** or without proper equipment;
- Can access but site is not within the defined **population**
 - e.g., agricultural soil sampling, planned site is in an irrigation ditch



What to do?

Options:

1. Ignore; sample size is smaller
 - Must not be any **systematic** reason for exclusion, e.g., all sites of a particular land use or geology are inaccessible; this biases the sample
2. Pre-compute **reserve** extra sites, use one-for-one (in sequence) to substitute
 - According to the same sampling plan; so not possible for grid samples
3. A common but *incorrect* procedure: move a **random distance** (within some pre-specified bounds) and **direction** from planned site to a **substitute** site.
 - No! this destroys the probability sampling, since points near the rejected point are more likely to be chose than points further away.



Topic: Sample size

Sampling is expensive, but so is incorrect or imprecise information. These two must be balanced by determining the **sample size** that will satisfy information needs while minimizing costs.

We first illustrate the concept of sampling error, then develop theory to determine sample size, then see how to compute it:

1. Sampling error
2. Theory
3. Computation



Sampling error

Estimates from samples are almost never equal to true values, and estimates from different samples differ among themselves.

To quantify this we define the concept of **sampling error**:

- The amount by which an **estimate** of some population parameter computed from a **sample** deviates from the **true value** of that parameter for the **population**.

Example: Estimated total rice production in a district, extrapolated from a sample of fields, vs. the actual total production.

Of course we usually don't know this (since we don't know the true value).



Sampling error

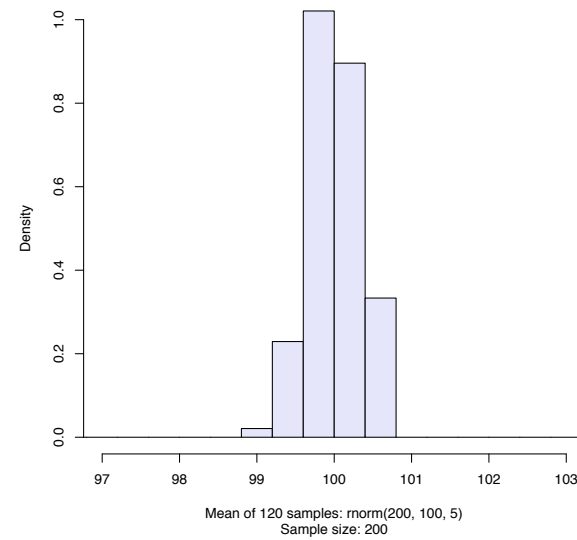
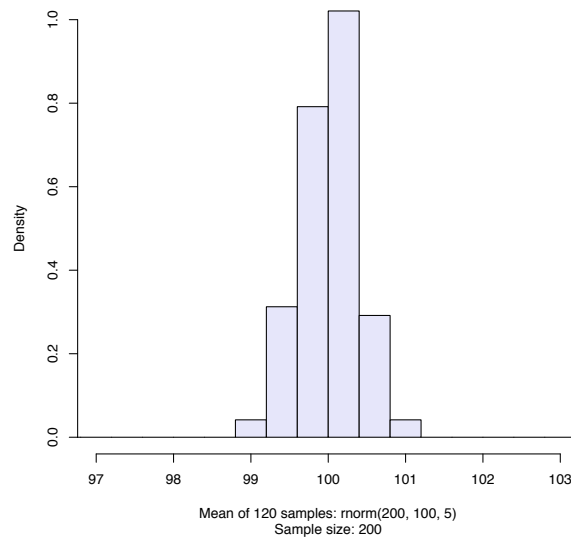
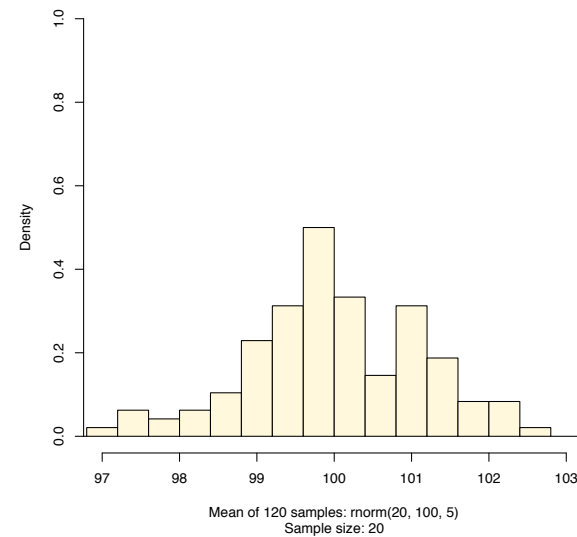
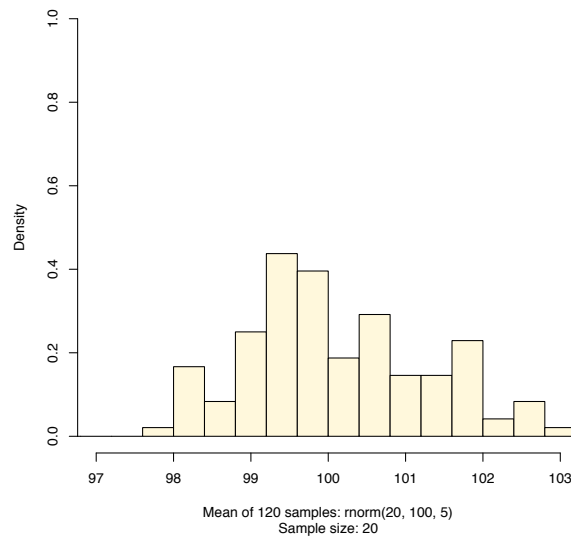
We can appreciate sampling error by **simulation** from known populations.

Example: Draw 10 different random samples from a normal distribution with true mean 100 and standard deviation 5; size of each sample is 20 observations; compute the sampling errors:

```
R> # set up a vector for the results
R> samp <- rep(0, 10)
R> # compute the means of 10 sets of 20 normal variates, true mean=100
R> for (i in 1:10) samp[i] <- mean(rnorm(20, 100, 5))
R> # compute sampling errors
R> 100 - samp
[1] -1.480606  0.256055  0.165392 -0.096576 -0.730931 -0.109797
[7]  1.118741 -0.246498  0.674641  0.887922
R> # mean sampling error
R> mean(100 - samp)
[1] 0.043834
```

Notice that the **mean** sampling error is almost zero. This is the result of the **central limit theorem**, derived from the **law of large numbers**.

Illustrating the central limit theorem: sample size 20 vs. 200



Significance testing

This is of two types, with major philosophical and practical differences:

Reject-support (RS) : the **null hypothesis** (written H_0) is set up as the **opposite** of what we would like to prove (and what we expect to prove if nature is as we think)

So, **rejecting** the null hypothesis **supports** our conclusions, which are known as the **alternate hypothesis**, written H_A .

Example: We think a new crop variety should yield at least 100 kg ha⁻¹ higher than the current one; but we set up the null hypothesis that it does not.

Accept-support (AS) : the null hypothesis is set up as what we **believe**.

So, **accepting** the null hypothesis **supports** our conclusions.

Example: The null hypothesis is that the new crop variety has average yield at least 100 kg ha⁻¹ higher than the current one.

Why Reject-support (RS) is commonly used

The idea here is that we should be **pretty sure** before rejecting a null hypothesis that is **against** what we would like to happen.

This guards against **false optimism** and **wishful thinking**.

For the remainder of the notes we are using RS testing.

The main use of AS is with very small sample sizes where it is quite difficult to achieve a low Type I error rate (see next).



Type I and Type II error

To understand how we determine sample size, we need to recall some basics of **hypothesis testing**

- n.b.: this is the so-called “frequentist” view of probability.

There are two types of inferential errors we might make:

Type I : **rejecting** the null hypothesis when it is in fact **true**; a **false positive**

Type II : **not rejecting** the null hypothesis when it is in fact **false**; a **false negative**

<i>Action taken</i>	<i>Null hypothesis H_0 is really ...</i>	
	True	False
Reject	Type I error committed	success
Don't reject	success	Type II error committed



Significance levels

There are two risk levels associated with the two types of error:

α is the risk of **Type I** error

We set α to guard against false inference. In RS testing we are inherently **conservative**.

β is the risk of **Type II** error

$1 - \beta$ is known as the **power** of the test (see below).

We get β from the form of the test and true effect (see below).



Example

Null hypothesis: A new crop variety will not yield at least 100 kg ha⁻¹ more than the current variety; that is, there is no real reason to recommend the new variety.

Note: this is an **informative** null hypothesis; not just “no difference”. It is set by the researcher. In this case, unless we can prove this much difference we won’t bother to develop the new variety. This is a management decision, not statistical.

Type I error: the new crop variety in fact **does not** have an average yield (if grown “everywhere”) at least 100 kg ha⁻¹ more than the current variety, but from our (limited) sample we say that it **does**. A “false positive”. So, we develop the variety and recommend it, but the farmer gets no significant benefit.

Type II error: the new crop variety in fact **does** have an average yield (if grown “everywhere”) at least 100 kg ha⁻¹ more than the current variety, but from our (limited) sample we say that it **does not**. A “false negative”. So, we abandon the variety, even though the farmer would have benefitted.



Comparing Type I and Type II risks

Inference from any sample smaller than the full population has two probabilities (Type I and II) of being incorrect.

Q: How to balance? A: The risk of economic loss:

Risk = **Hazard** (the **probability** of a wrong decision)
x **Vulnerability** (the cost of a wrong decision)

Solve the following minimization for α and β :

$$r = (\alpha \cdot \text{cost}_\alpha) + (\beta \cdot \text{cost}_\beta)$$

α, β are the hazards; $\text{cost}_\alpha, \text{cost}_\beta$ the vulnerabilities.



Approaches to computing sample size

We will compare two approaches:

1. Power analysis
2. Sampling to narrow a confidence interval



Topic: Power analysis

One approach to estimating the required sample size is to set the size to achieve a desired **statistical power** of detecting a true difference. This is very common in social science and medical trials.

Standard reference: Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.



Statistical power

The quantity $(1 - \beta)$ is called the **power** of a test.

It is the probability of **rejecting the null hypothesis** when it is in fact **false**; i.e. making a **correct positive decision** (to take some action).

This is what we would like to do. So, we want to **maximize the statistical power**, after setting α to guard against false inference.

Example: the probability of deciding that a new crop variety will yield at least 100 kg ha^{-1} more than the current variety, if this is in fact true. We want to know how likely we are to take the correct action, i.e. promote the new variety.



Why worry about power?

Before sampling, we would like to believe that it will be sufficient to support our beliefs, i.e. (in RS) that the **null hypothesis** is in fact **false**.

Clearly, we would like a high probability that, **if** the null hypothesis is false, our test **will detect** this.

If, before sampling, we don't feel that we can get enough power, there is **no point in sampling**! It would be wasted effort.

Note that many funding agencies require an *a priori* power analysis of proposed research.



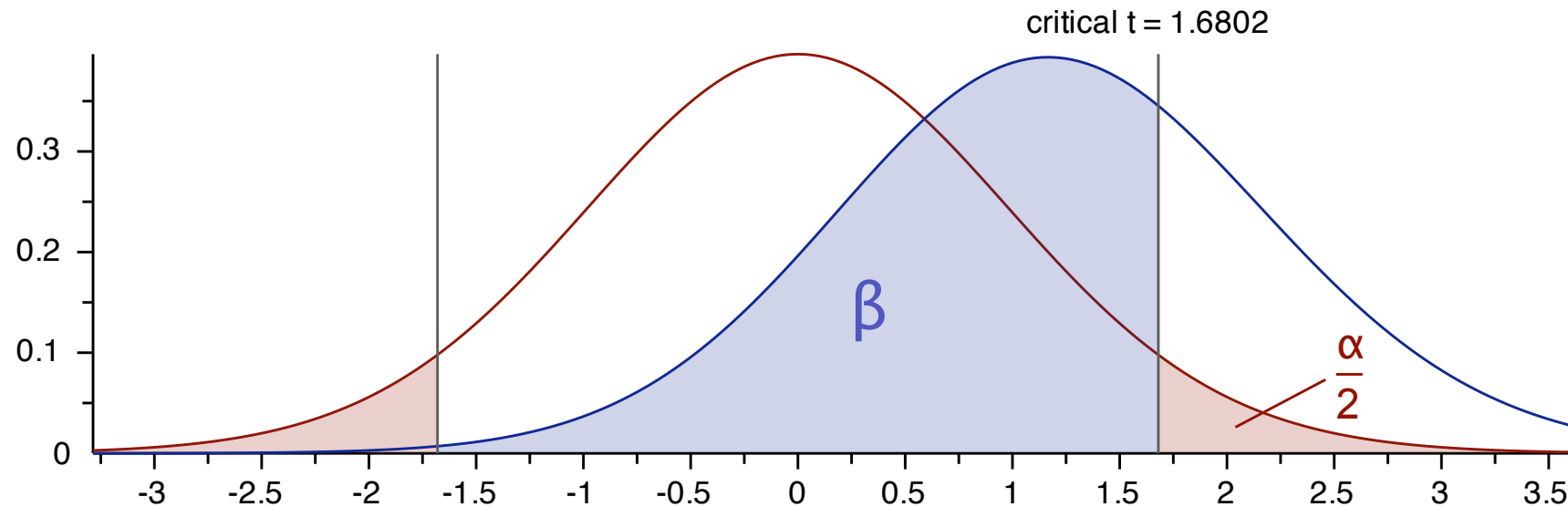
Power vs. significance

A common misconception is that power is the inverse of significance. But recall:

- α is the **risk** of a **Type I** error: rejecting the null hypothesis when it is in fact true; this was set by the analyst to guard against wishful thinking;
- $(1 - \alpha)$ is just the **lack of risk** of making this mistake, the so-called **confidence level**
- But it tells us nothing about how likely we are to **reject a false null hypothesis**; for this we need to know the **effect size**: that is, **what is the real difference?** Or, “how false?”

The power of a given test (test form, α , sample size) increases with the **effect size**. So to compute power, we need to specify the **magnitude of the effect** we would like to **detect**, if it is really true.

Graph: Power vs. significance



Red curve: t -distribution of some parameter (e.g. population mean) assuming null hypothesis is true

Red regions: **Type I error** is committed if the experimental value here

Blue curve: t -distribution of the parameter for a given real **effect**

Blue region: **Type II error** is committed if the experimental value is here



Explanation

In this figure the **power** is only 0.32. This is the **area under the blue curve, to the right of the critical t-value**.

The complement of the power, β , is $1 - 0.32 = 0.68$. This is the probability of incorrectly concluding that there is no effect, even though there is.

In the **blue region**, because the critical t-value to reach $\alpha/2$ is *not* reached, we do *not* reject H_0 when in fact it is false! So then we cannot accept any alternate hypothesis. This avoids a Type I error but **we commit a Type II error in this region**.

Conclusion: In this example, we have only a **32% chance** of a correct decision with this test and a sample of this size, to detect an effect this large.



Factors that affect power

1. The **form** of the test
 - e.g. a two-sample t-test, a paired t-test, test against a constant
 - one-sided vs. two-sided
2. The **effect size**: magnitude of the **actual differences** in the population
 - e.g. if one variety is greatly superior to the other, there will be a large effect
3. The **experimental or observational error**, or **noise**; this is variability in the data that is **not** related to the experiment; the **signal** (true effect) will be masked by this noise.
 - e.g. low-precision instruments, poor sample handling . . .
4. **Sample size**. Larger is more powerful, but too large is wasted.

Test type vs. power

Example: Detecting a true difference of $+100 \text{ kg ha}^{-1}$ (5000 kg ha^{-1} for one variety vs. 5100 kg ha^{-1} for the other) with 120 samples of each; set $\alpha = 0.1$, two-sided test (H_0 : no difference):

(See later in section for calculations)

1. Unpaired: $1 - \beta = 0.61$
2. Paired: $1 - \beta = 0.86$
3. Variety 2 vs. a constant target value of 5000: $1 - \beta = 0.70$ (only one sample set)

Note: Paired with only 60 samples of each: $1 - \beta = 0.61$, same as the unpaired test with double the samples.



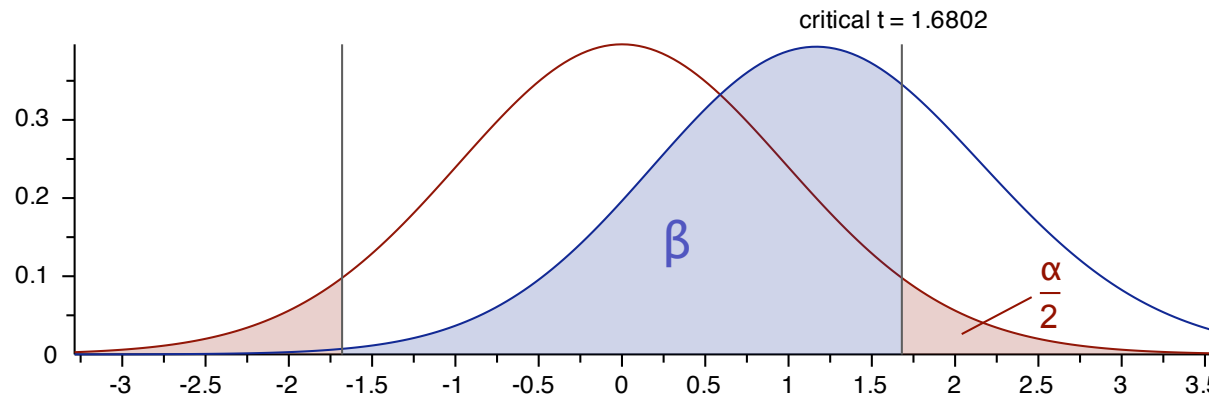
Visualizing effect of parameters on power

In the following graphs we start from a base situation:

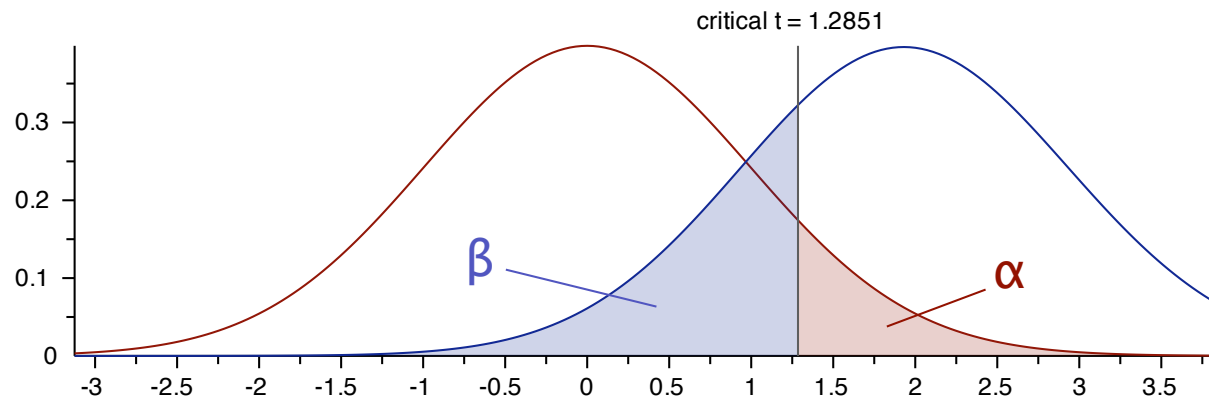
- $\alpha = 0.1$
- $H_0 = 5000$
- Effect: 100
- Paired t-test, two-sided
- 45 samples of each

Test form vs. power

One-sided tests have more power than two-sided tests:



Two-sided: Power: 0.322

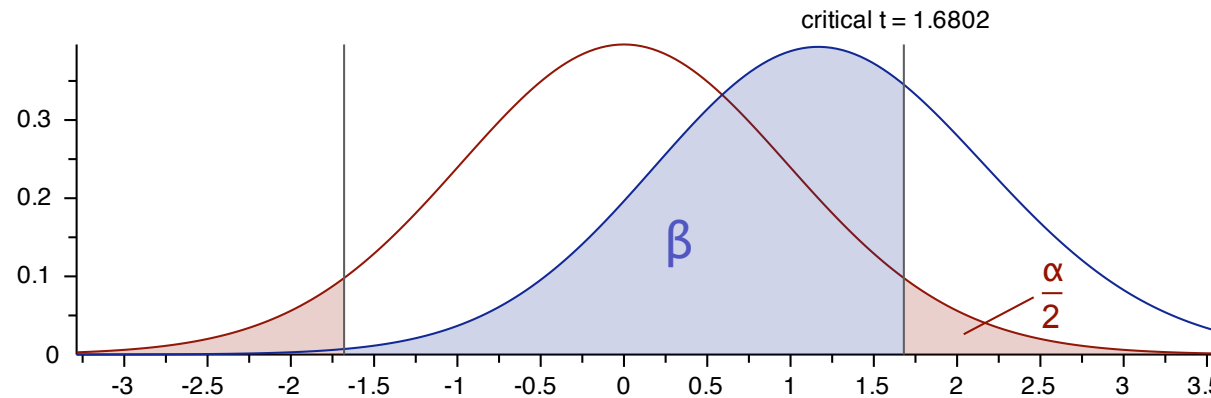


One-sided: Power: 0.460

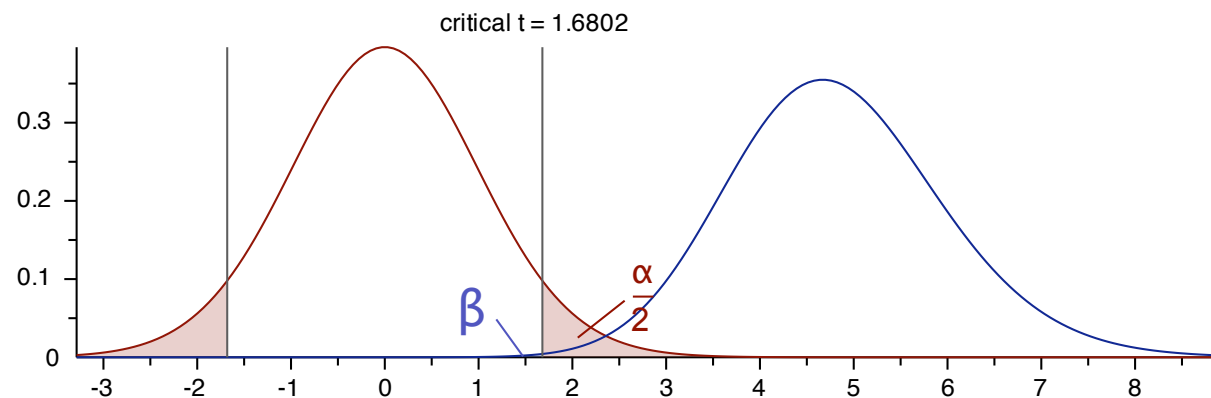


Effect size vs. power

The greater the real or assumed effect, the higher the power:



Effect: 100; Power: 0.322

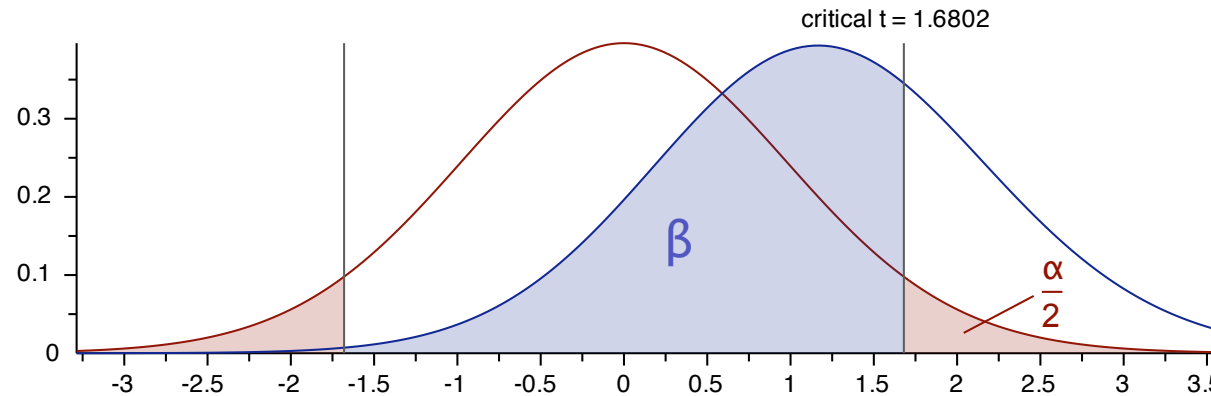


Effect: 400; Power: 0.998

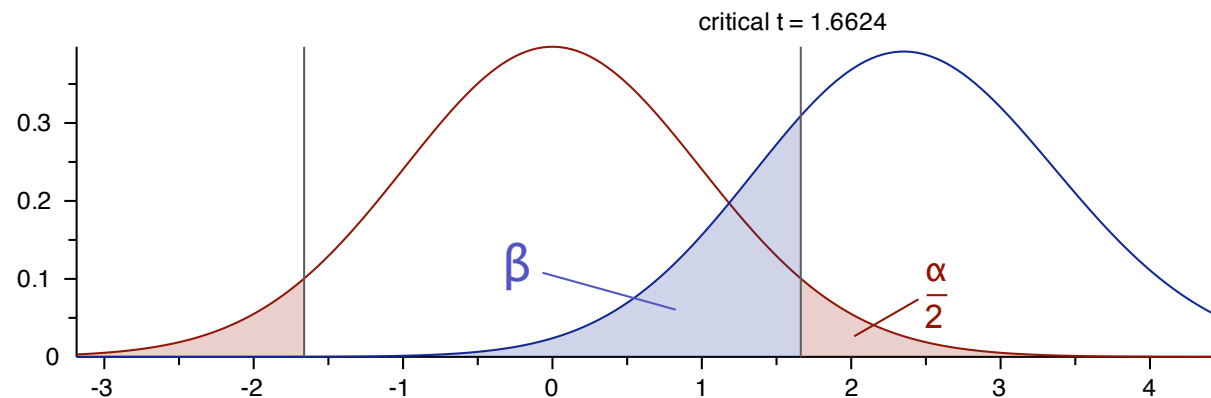


Noise vs. power

The narrower the t-distributions, the higher the power:



SD: 400; Power: 0.322



SD: 200; Power: 0.761



How do we control power?

1. **We** set the **experimental design**;
2. **We** set the **null hypothesis** (one-sided vs. two-sided, H_0 of no effect or informative);
3. **We** set the **risk level** α for Type I error;
4. **'Nature'** determines the **effect size**;
5. **We** should do our best to **control experimental error**;

The **sample size** can be used two ways:

1. We can **set the sample size** to achieve some desired power;
2. Or, if we know the sample size (e.g. experiment already done, or the maximum that can be taken), we can **compute the power** and see if it is enough



Power calculations

We can do this “forwards” or “backwards”:

1. *a priori*: compute the **required sample size** to achieve a given power, given α and effect size.
2. *post hoc*: compute the **power achieved** by a test, given α , sample size, and effect size;

Note: we don't really know the effect (only “nature” knows) but we do know the **minimum effect** that is **interesting** for our results.

For example, if the new crop variety is not at least 100 kg ha⁻¹ higher-yielding on average, it is not worth it to develop it further. This depends on the application.



R functions for power calculations

There are three functions in the standard stats library:

power.t.test : For one- and two-sample **t tests**

power.prop.test : For two-sample tests for **proportions**

power.anova.test : For balanced **one-way analysis of variance** tests

R examples (1/2)

Example: two-sample t-test for a true difference in either direction of 100, population standard deviation estimated as 400, at $\alpha = 0.1$:

1. *post-hoc*: specify sample size, compute achieved power

```
R> power.t.test(n=120, delta=100, sd=400, sig.level=0.1,  
+ type="two.sample", alternative="two.sided", strict=T)
```

```
power = 0.61279
```

With the 120 samples we achieved 61% power.

If the null hypothesis was “not less than” (one-sided), we achieve more power (74%) with the same sample size:

```
R> power.t.test(n=120, delta=100, sd=400, sig.level=0.1,  
+ type="two.sample", alternative="one.sided", strict=T)
```

```
power = 0.74267
```

R examples (2/2)

2. *a priori*: specify power, compute required sample size

```
R> power.t.test(power=0.9, delta=100, sd=400, sig.level=0.1,  
+ type="two.sample", alternative="two.sided", strict=T)
```

n = 274.72

To achieve 90% power we would need 275 samples in each group.

If we only care about detecting this difference in one direction:

```
R> power.t.test(power=0.9, delta=100, sd=400, sig.level=0.1,  
+ type="two.sample", alternative="one.sided", strict=T)
```

n = 210.64

Only 210 samples are needed to detect the difference for a one-sided test.



Graphical program for power analysis

G*Power 3 from Heinrich Heine University, Düsseldorf (D)

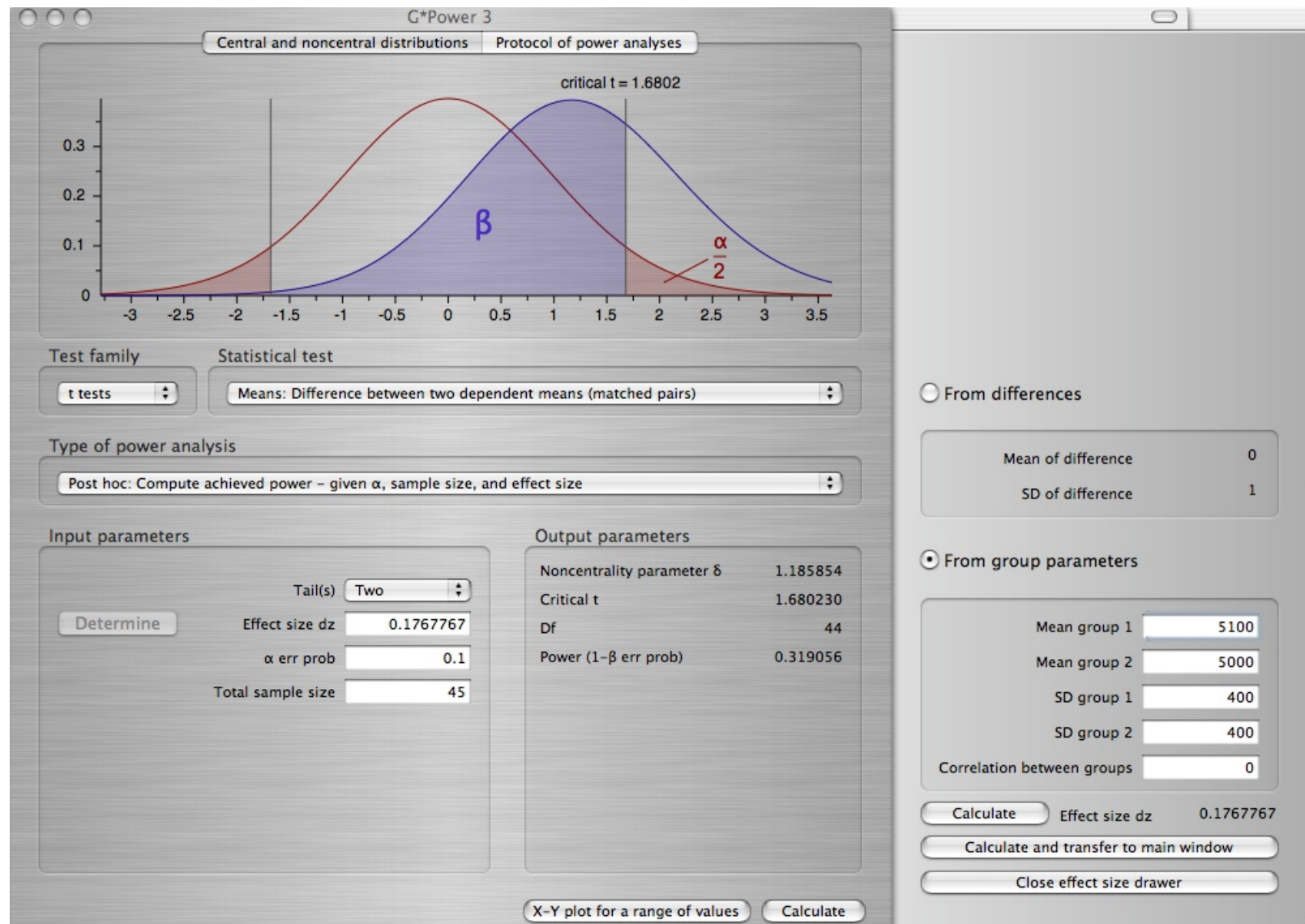
<http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>

There are Mac OS X and MS-Windows versions of this program.

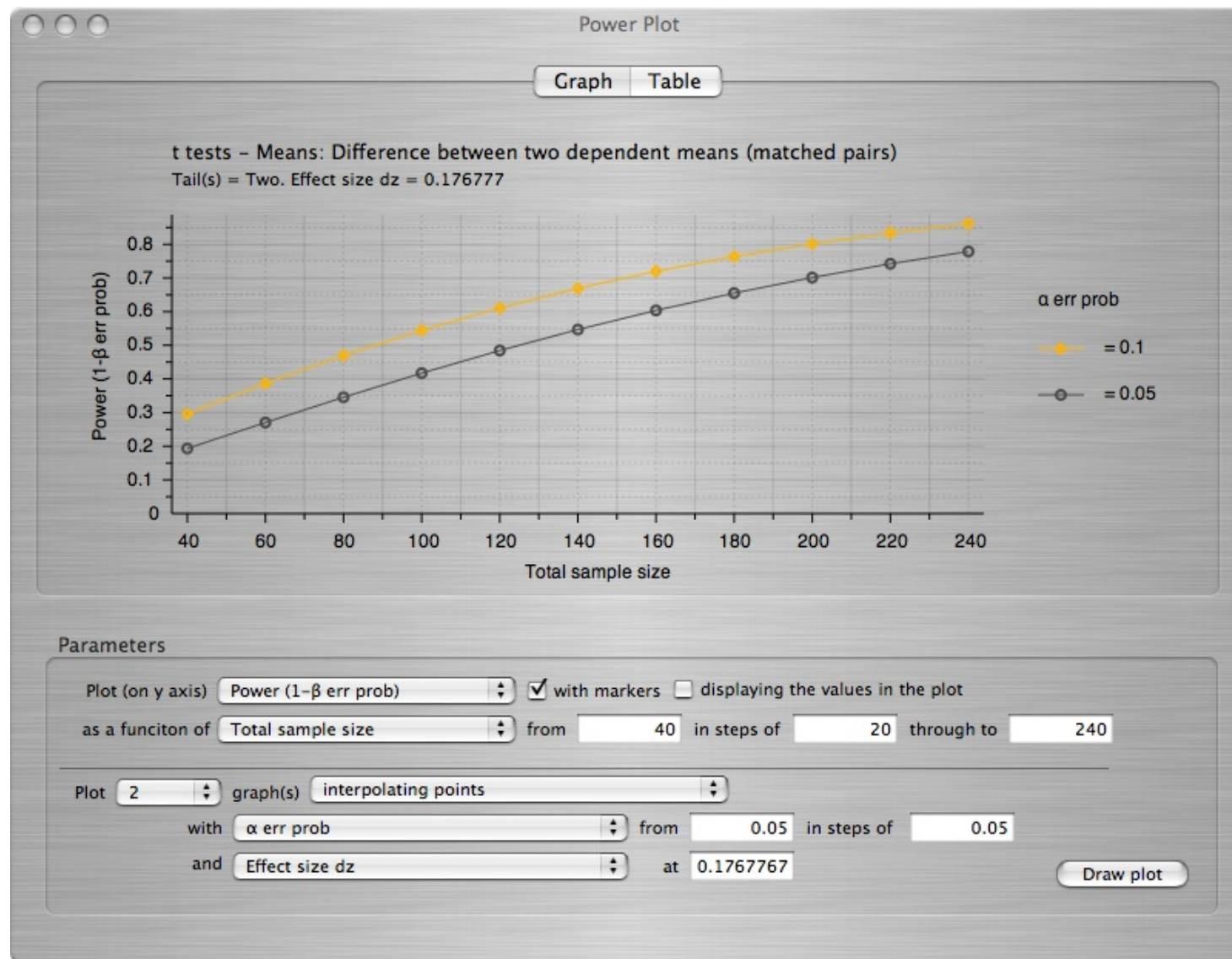
The next two slides show sample screens.



Compute *a priori* sample size and *post-hoc* power:



Trade-off sample size vs. power for various risk levels:



Topic: Sampling to narrow a confidence interval

Another approach to sample size calculation is to consider the desired **width of the confidence interval** for some parameter of interest.

This differs from power analysis because we don't specify any effect; we are just interested in one parameter.

We will use the example of a **confidence interval for a mean value**.



Confidence interval for the mean

Recall that the **confidence interval** for a mean μ is computed as:

$$(\bar{x} - t_{\alpha/2, n-1} \cdot s_{\bar{x}}) \leq \mu \leq (\bar{x} + t_{\alpha/2, n-1} \cdot s_{\bar{x}})$$

where:

- \bar{x} is the sample mean;
- $t_{\alpha/2, n-1}$ is Student's t with $n - 1$ degrees of freedom at confidence level $\alpha/2$;
- $s_{\bar{x}}$ is the standard error of the sample mean:

$$s_{\bar{x}} = \frac{1}{\sqrt{n}} s_x$$
$$s_x = \left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2}$$

Notes on these formulas

- The confidence level α , say 0.05, is halved, say to 0.025 for each side of the interval, because this is a two-sided interval.
- The ***t*-distribution** must be used because we are estimating both the mean and variance from the same sample; for reasonably-large sample sizes the **normal distribution** itself (here called the z distribution) can be used.

What affects the confidence interval?

1. The t value:

- (a) n : **sample size**: $t \rightarrow z$ as $n \rightarrow \omega$
- (b) α : **risk level** set by the experimenter that the computed interval does **not** contain the true mean; a higher risk leads to a narrow interval

2. The standard error $s_{\bar{x}}$:

- (a) n : **sample size** (again): precision increases as \sqrt{n}
- (b) the **sample standard deviation**: this is essentially the **inherent variability** of the sample



What can we control?

1. The **risk** of rejecting a true null hypothesis (α); depends on the **cost of a false positive**;
 - If there is little cost associated with making a Type I error, α can be high (lenient); this will narrow the confidence interval.
2. **Sample standard deviation**
 - We have some control by good experimental or observational procedures
 - But we can not control the **inherent variability** in the population, even with perfect technique;
3. The **half-width** w of the confidence interval, i.e. the required **precision**. We set this according to how precise the computed estimate must be; this depends on the application.



Inverting the confidence interval

With the above parameters set, we can compute the required **sample size**:

1. **Set** the required risk α that the computed mean value (or mean difference) is outside the interval;
2. **Set** the desired (half-) **width** of the confidence interval w ;
3. **Estimate** the **sample standard deviation** s_x

Then we **solve for** n :

$$\begin{aligned}w &= t_{\alpha/2, n-1} \cdot s_x / \sqrt{n} \\ \sqrt{n} &= t_{\alpha/2, n-1} \cdot s_x / w \\ n &= (t_{\alpha/2, n-1} \cdot s_x / w)^2\end{aligned}$$



No closed-form solution for n

There is a problem with this “solution” for n :

The right-hand side also contains n (which we want to compute), because the t -value depends on the degrees of freedom.

So we must somehow **approximate** t , solve, and then **iterate**. In practice either of the following two methods can be used:

1. Replacing t with z : for **larger** expected sample sizes
2. Use a conservative estimate of t : for **smaller** expected sample sizes



Solution 1: Replacing t with z

Replace $t_{\alpha/2, n-1}$ with $z_{\alpha/2}$, i.e. the normal deviate (= t with infinite d.f.).

This leads to **under-estimation** of n by a factor f of (example for $\alpha = 0.05$):

n	10	20	50	100	200	500
f	0.268	0.126	0.049	0.024	0.012	0.005

(Note: R code for this: `qt(0.975, n) - qnorm(0.975)` etc.)

So the sample size estimate will be somewhat **too low**.

Then **iterate** with this first estimate of n , using the t value this time.



Solution 2: Use a conservative estimate of t

Use a small but realistic value of n to compute $t_{\alpha/2, n-1}$; as long as the computed n is larger, this is a conservative estimate.

If a more exact n is needed, the new estimate can be used to re-compute $t_{\alpha/2, n-1}$, **iteratively**.



How to set these values?

1. We **set** the desired **risk** based on how often we are willing to be wrong (i.e. the actual value is outside the computed limits).
2. We **set** the desired **width** based on the **precision** we require.
3. We **estimate** the **sample standard deviation** from a previous study on this sampling frame, or from similar studies. This of course may not be the actual sample standard deviation we get from the new sample.



Numeric example: Problem

- **Problem**: Determine sample size for an on-farm trial to detect difference in rice yield between two on-farm treatments (e.g. early vs. normal planting date). We want to determine which is best and then recommend it.
- The **population** is all fields where rice is grown in the region.
- We use a **paired** design: sets of adjacent fields, as similar as possible in soils and management, differing only in planting date.
- We then compute the **paired difference** and, from these, the **mean difference**.
- We compute the **confidence interval** of the **mean difference** based on our sample. If this interval includes 0 we can not reject the null hypothesis H_0 of no difference between planting dates.

But we get more information here: the **interval** in which the **true difference** is expected to lie: i.e. an estimate of the **magnitude of the effect**. This can be directly used for decision-making.

Numeric example: Setup

1. We set α to 0.1 because we are willing to accept a 10% risk of falsely rejecting the null hypothesis (i.e. falsely deciding that one of the alternatives is better than the other). So for each half-width we use half of this, i.e. 0.05.
2. We set the **half-width** to 100 kg ha⁻¹ because a smaller yield difference is not important; this is 2% of the region's typical yield of 5 T ha⁻¹.
3. From a previous survey we estimate the **population standard deviation** to be 400 kg ha⁻¹; note that this will be higher with on-farm trials than in controlled experiments.



Numeric example: Solution

We begin with an estimated sample size of 20; we know this is within our budget.

$$t_{.05,19} = 1.7291 \text{ (R code: qt(.95, 19))}$$

$$n = (t_{\alpha/2, n-1} \cdot s_x / w)^2$$

$$n = (1.7291 \cdot (400/100))^2$$

$$n = 47.8 \approx 48$$

This suggests that a sample size of 48 should detect a real difference of 100 kg ha⁻¹ in either direction, with a risk of 10% of incorrectly calling a chance difference real.

Note this is 48 pairs, since it is a paired test.



Numeric solution: iteration

Now that we know $\approx n$ we can recompute t and refine the estimate; with this higher n the t value will be a bit lower and so will the required sample size.

$$t_{.05,47} = 1.6779 \text{ (R code: } \text{qt}(.95, 47)\text{)}$$

$$n = (t_{\alpha/2, n-1} \cdot s_x / w)^2$$

$$n = (1.6779 \cdot (400/100))^2$$

$$n = 45.047 \approx 45$$

The difference is small, only 3 fewer samples.



Numeric solution: normal approximation to t

$$z_{.05} = 1.6449 \text{ (R code: } \text{qnorm}(.95)\text{)}$$

$$n = (z_{\alpha/2} \cdot s_x / w)^2$$

$$n = (1.6449 \cdot (400/100))^2$$

$$n = 43.289 \approx 43$$

This is only two fewer than when using the correct t , so it is also a good approximation. We could now iterate with $t_{0.05,42}$ as above, and arrive at the final (correct) sample size, $n = 45$.



Effect of parameters

The following all **increase** the required sample size:

1. α : $\alpha = 0.10 \rightarrow 45$; $\alpha = 0.05 \rightarrow 65$; $0.01 \rightarrow 115$

Decreasing risk (i.e. a smaller α)

2. w : $w = 200 \rightarrow 11$, $w = 100 \rightarrow 45$, $w = 50 \rightarrow 180$

Detecting a **small (real) difference**

3. s : $s = 200 \rightarrow 11$, $s = 400 \rightarrow 45$, $s = 800 \rightarrow 180$

A **more variable population**



Topic: Sampling for proportions

The previous examples have been about sampling a continuous variable. Another kind of information is the **proportion** of a population that meets a given criterion.

Proportion: from 0 (none) to 1 (all) of the population.

If we are unable to observe the entire population, we take a **sample** and **infer** the population proportion from the sample proportion.

Since there will be **sampling error**, we also compute the **confidence interval** for the true proportion.

We must choose the sample size to **narrow** the confidence interval to some desired target.



Examples

- Proportion of farmers in a district who would be willing to join a new agricultural cooperative;
- Proportion of children 6-10 in a district who attend school regularly

These are examples of large populations where it is impossible to observe every possible sampling unit.

References

Cochran, W G. Sampling Techniques. John Wiley, New York, 3rd edition, 1977; §4.4

Webster, R W and M A Oliver. Statistical methods in soil and land resource survey. Oxford University Press, Oxford, 1990; Chapter 3



Terminology

- **Success** or a **positive result**: by convention, the result that is considered desirable; given the value **1**.
- **Failure** or a **negative result**: by convention, the complementary result; given the value **0**.
- p : the **true proportion** of the first result; the **probability** that a single trial will give a **positive** result
- q : defined as $(1 - p)$; the **true proportion** of the second result; the **probability** that a single trial will give a **negative** result

Note: logically, success and failure are symmetrical; the two outcomes are assigned by the researcher.



Estimating the population proportion from a sample

Given an **unbiased** sample of size n , with n_1 successes, the **population** proportion is estimated naturally as the sample proportion:

$$\hat{p} = \frac{n_1}{n}$$

Its **standard deviation** is estimated as:

$$\hat{s} = \sqrt{\frac{p \cdot q}{n}}$$

Note that \hat{s} is highest for $p = 0.5$ and decreases as either outcome becomes more probable.



Confidence interval

From the estimates of p and s , we can compute the **confidence interval** of the estimate:

$$\hat{p} \pm \left[\hat{s} \cdot Z_{1-\alpha/2} + \frac{1}{2n} \right]$$

where:

- $Z_{1-\alpha/2}$ is the one-tailed normal score for the two-tail probability of Type I error α
- $1/2n$ is the **small-sample correction**

The small-sample correction is usually ignored for sample sizes $n > 50$; it is then less than 1%.

The lower and upper limits are truncated at 0 and 1, respectively, if necessary.



Example (1/2)

1. 200 (n) representative households selected at random from a sampling frame were surveyed with a standard questionnaire;
2. 50 answered a certain question “yes” (n_1)
3. Compute: $\hat{p} = 50/200 = 0.25$
4. Compute: $\hat{s} = \sqrt{(.25 * .75)/200} = 0.030619$
5. The analyst sets the risk of Type I error, say 0.1 (true proportion outside the computed interval)



Example (2/2)

6. Then $Z_{0.95} = 1.6449$; note a two-tailed test so for one tail, half the risk

7. $\hat{p} = 0.25 \pm (0.030619 \cdot 1.6449 + 0.005) = 0.25 \pm 0.055363$

8. Confidence interval: $[0.2046 \dots 0.3054]$

We are 90% confident that the true proportion of all households in the population that would answer “yes” (if we could interview all of them) lies in the range 20.46% to 30.54%.

Is this **enough precision**? If we need a narrower interval, we must **increase sample size**. (Or, we can accept more risk of Type I error.)

Is this **too much precision**? If so, we sampled excessively and could have saved effort. (Or, we can lower the risk of Type I error.)

So, inverting this relation gives us a way to compute **required sample size**.

Required sample size

The required sample size n is estimated as:

$$n = \frac{(Z_{(1-\alpha/2)})^2}{b^2} \cdot \hat{p} \cdot (1 - \hat{p})$$

where:

- \hat{p} is a **prior estimate** of the population proportion of “successes”;
- b is the desired absolute precision as a proportion;
- $Z_{(1-\alpha/2)}$ is the one-tailed normal score for the desired two-tailed probability α of Type I error.

Note that this ignores the finite-population correction; for sample sizes < 50 an exact calculation is needed; or, the experimenter can just add a few samples.



Example

1. We want to estimate the proportion of farmers who will agree to join a new agricultural cooperative.
2. *A priori* we believe that 10% will join
3. We want to measure the true proportion $\pm 4\%$
4. We will accept only a 5% risk that our interval does not contain the true proportion; $Z_{0.975} = 1.960$.

Then:

$$n = \frac{1.96^2}{0.04^2} \cdot 0.1 \cdot 0.9 \approx 216$$



How to set these parameters?

These are all at the discretion of the analyst:

- \hat{p} estimated proportion: From prior experience in the same or similar situations; or, **wishful thinking**; the computed interval will tell us if we were right;
- b half-width of the interval; the **precision** we need; this depends on how closely we need to know the population proportion;
- α the risk of Type I error, i.e. that the interval we compute doesn't contain the true proportion; this depends on the consequences, including costs, of making a decision based on the test.

Example: if, after sampling, we estimate that between 6% and 14% of the farmers would join the cooperative, we decide to establish it. If only 3% in fact join, it will fail.



Effect of parameters on sample size

\hat{p} sample size increases from extremes (expected near 0 or 1) to the centre ($\hat{p} = 0.5$). The factor $(p \cdot (1 - p))$ is:

p	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
n	0.09	0.16	0.21	0.24	0.25	0.24	0.21	0.16	0.09

b sample size **decreases** as the **square** of the precision;

α sample size **increases** as the acceptable risk **decreases**; being more certain costs money. The factor $Z_{(1-\alpha/2)}$ is:

α	0.2000	0.1000	0.0500	0.0100
zsq	1.2816	1.6449	1.9600	2.5758

Power analysis

Testing for proportions can also be done with power analysis, as we saw for t-tests in a previous section.

Here the **effect size** is replaced by the **difference in proportions** between two samples.

The R method `power.prop.test` computes either the *post hoc* achieved power or the *a priori* required sample size.

This is a good method to determine sample size to detect a difference.

Example of power analysis for sample size

- We intend to conduct a **survey** to see **what proportion** of school-age children in a district are **attending school regularly**.
- We select **representative households** at **random** and interview or observe every child in the household.
- We **hypothesize** that there is a **difference** between the proportion of **boys and girls** that attend school, but we **don't hypothesize which is higher** (maybe boys are being kept out to work in the fields, or maybe girls are being held back to do housework, we don't know).
- **If** there is a difference we want to design a campaign **targeted** at the group that is held back.
- But there is no point in a targeted campaign unless there is **at least a 5% difference**; otherwise a general campaign for both boys and girls is sufficient.

Example (2/3)

Question 1: What sample size would be needed to detect a true 5% difference?

1. Set **risk of false detection** at $\alpha = 0.02$ (it is expensive to set up a separate campaign);
2. Estimate the **overall proportion** of children attending school regularly as 0.70 (70%) based on informal surveys and secondary information;
3. If there is a 5% difference, this 0.70 could be 0.67 vs. 0.72 (any reasonable division of the 5% will do).
4. We want the survey to have a **high probability of revealing** this difference; so we set the desired power $1 - \beta = 0.9$.



Example (3/3)

Computation:

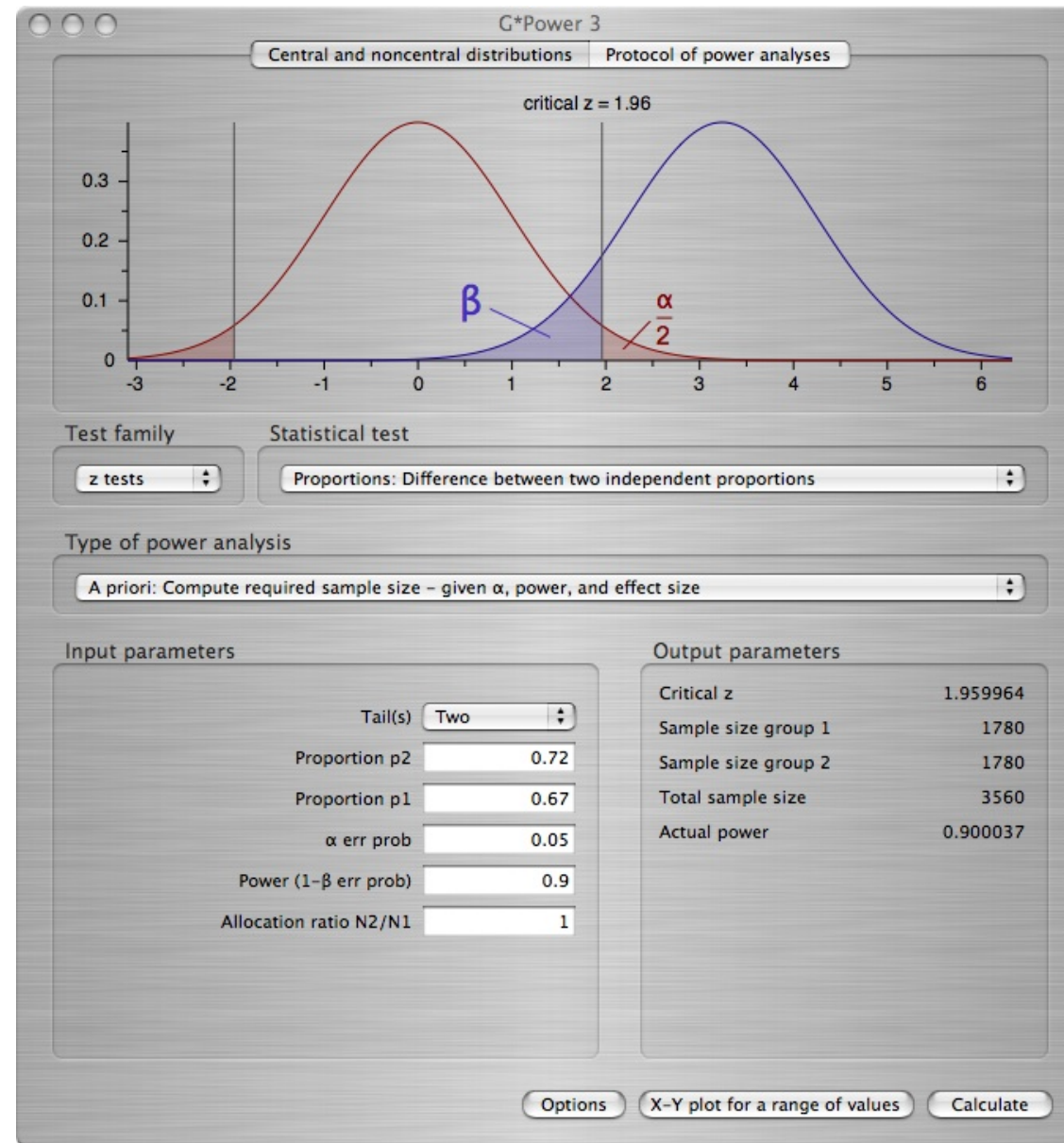
```
R> power.prop.test(power=.9, p1=.67, p2=.72,  
+ sig.level=0.05, alt="two.sided", strict=T)
```

n = 1779.8

NOTE: n is number in *each* group

Conclusion: we must survey households with a total of **at least 1780 boys and 1780 girls** to have 90% chance of detecting a true 5% difference if the overall proportion is about 70%, with a risk of 5% of falsely concluding there is a difference this large when there in fact is none.

Using G*Power 3



Example of power analysis for achieved power

Suppose we have already done the above survey with a sample size of 200 boys and 200 girls. What is the probability that this survey will reveal a true difference of 5% when the overall proportion is around 70%?

Example (2/2)

Computation:

```
R> power.prop.test(n=200, p1=.67, p2=.72,  
+ sig.level=0.05, alt="two.sided", strict=T)
```

```
power = 0.19186
```

NOTE: n is number in *each* group

Conclusion: if the observed 5% is the true difference, at this significance level and with a sample size of 200, we only have a 19% chance of detecting the difference.

More sophisticated calculations with proportions

The G*Power 3 power calculator has many variants on power calculations for proportions:

- Difference from an assumed constant
- Inequality (one or two-tailed, different methods)
- Sign test
- Generic binomial test



Topic: Sample size for multiple regression

Multiple regression is used to model a continuous variable (the **response** or **dependent** variable) as a **linear combination** of several **predictors** or **independent** variables; these can be of any type.

The aim in multiple regression is to get an accurate estimate of the **regression coefficients** β , and the **coefficient of determination**: how much of the total variance is explained by the model?

The problem is that, with many independent variables (predictors), there is a high risk of finding statistically-significant regression coefficients just by chance; there will appear to be a relation where there isn't.



Rules of thumb for multiple regression

These appear in various multiple regression texts, and are (at least heuristically) justified there.

Notation:

- m is the number of predictors (independent variables)
 - n is the required sample size (often called “cases” in the regression literature).
1. Never use $n < 5m$ observations, even for exploration; results will be too unreliable even to plan future work.
 2. To test regression coefficients β , ensure $n \geq 104 + m$ and $n > 20 m$



Confidence intervals of coefficients – why?

Reference: Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, 8(3), 305–321.

Often obtaining a significant coefficient is not so interesting, rather, we want good **accuracy** of a fitted coefficient.

That is, we want a narrow confidence interval **of the coefficient**.

Example: Predicting crop yield from fertilizer additions: what is the benefit of a single nutrient? of the interaction between two nutrients?

$$y = \beta_0 + \beta_1 N + \beta_2 P + \beta_3 K + \beta_4 N \cdot P + \beta_5 N \cdot K + \beta_6 P \cdot K$$

Compute e.g., β_1 (unit response to N), and also its confidence interval (how high/low could the response be?)



Confidence interval – calculation

$$\hat{\beta}_j \pm t_{(1-\alpha/2; N-p-1)} \sqrt{\frac{1 - R^2}{(1 - R_{XX_j}^2)(N - p - 1)}}$$

We need to estimate:

R^2 multiple correlation coefficient of the model

- better overall model → narrower interval

$R_{XX_j}^2$ multiple correlation coefficient predicting the j th predictor from the remaining $p - 1$ predictors

- lower correlation with other predictors → narrower interval for this predictor

Invert this equation with desired α to obtain N (sample size).

Topic: Geostatistical sampling

Sampling theory becomes more complicated when we need to consider the **location** in space (or time) of a sampling unit.

Geostatistical sampling is selecting sampling individuals by their **location**.

This is appropriate when the **relative geographic location** of observations is relevant to the analysis, or if there is **spatial dependence** (see below).

- The analysis depends on the relative location: e.g. to compute a trend surface, or to compute the distance of observations from focal points such as markets.

Compare:

- a random selection of crop fields from a **census list** (non-spatial), vs.
- a selection based on the **coordinates** (spatial).



Spatial sampling

Sampling in **space**: when the **location** of the observed individual is recorded and used in the analysis.

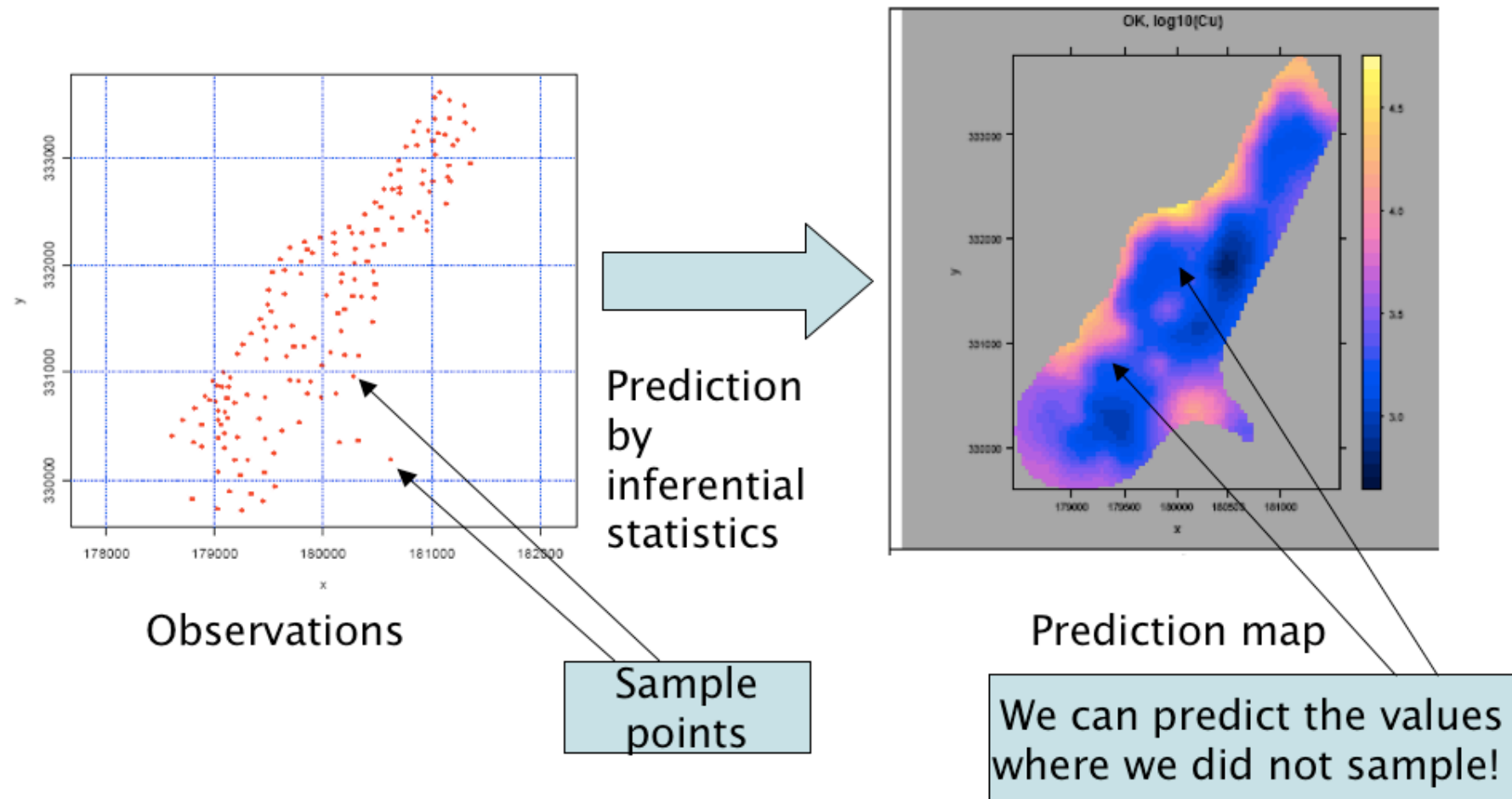
(Of course any observation is made somewhere on the Earth's surface! but it is not **spatial** sampling unless the location is recorded and used).

Extra inferences possible from spatial sampling:

- **Prediction** at unsampled locations
- Inference of **spatial dependence**: local or regional trends
- **Point-patterns**: Dispersion / clustering
- **Directional** statistics: alignment



Spatial sampling for geostatistical prediction



An example of spatial inference: Mercer-Hall wheat yields

Reference: W B Mercer and A D Hall. The experimental error of field trials. *The Journal of Agricultural Science (Cambridge)*, 4: 107–132, **1911**

A **uniformity trial**:

1. Select an apparently **homogeneous** field of 1 acre (0.40469 ha)
2. Prepare, sow **same variety** of wheat, manage **uniformly**
3. At maturity, **divide** into 500 equally-size plots (approx. 8.0937 m)
4. **Harvest** plots, measure grain and straw yield

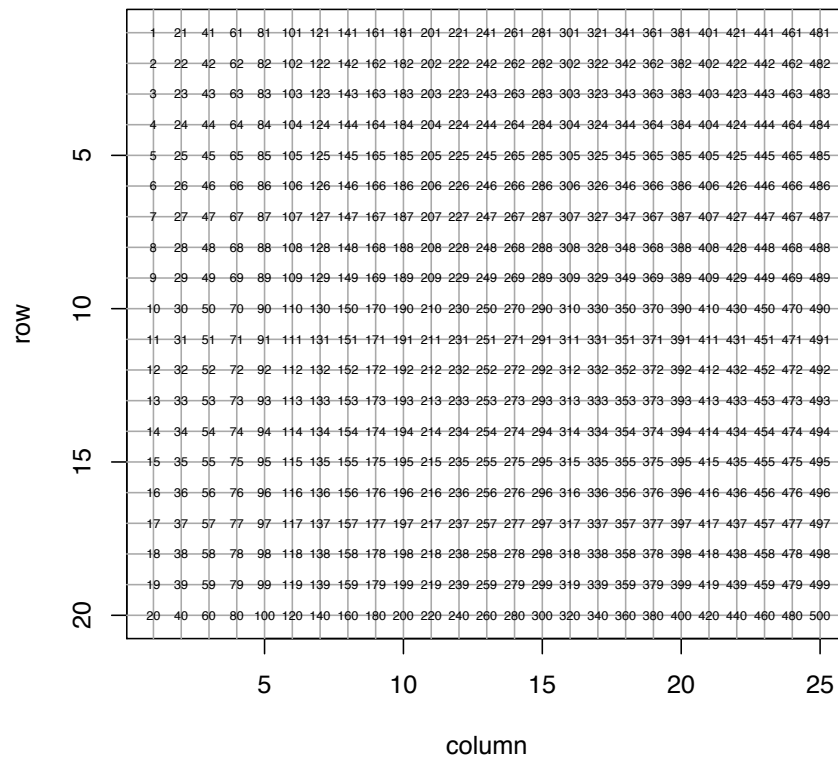
In **theory** “all plots should have the same yield” ...

Research question: How large to make plots to reduce variability between them?

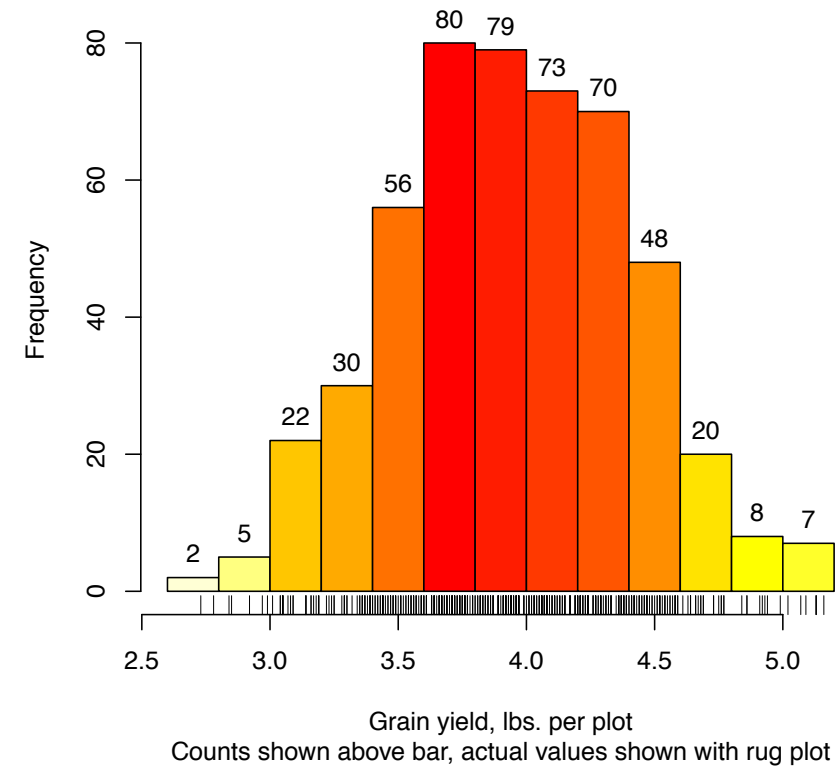


Field layout; yields

Layout of the Mercer–Hall uniformity trial



Frequency histogram, Mercer & Hall grain yield

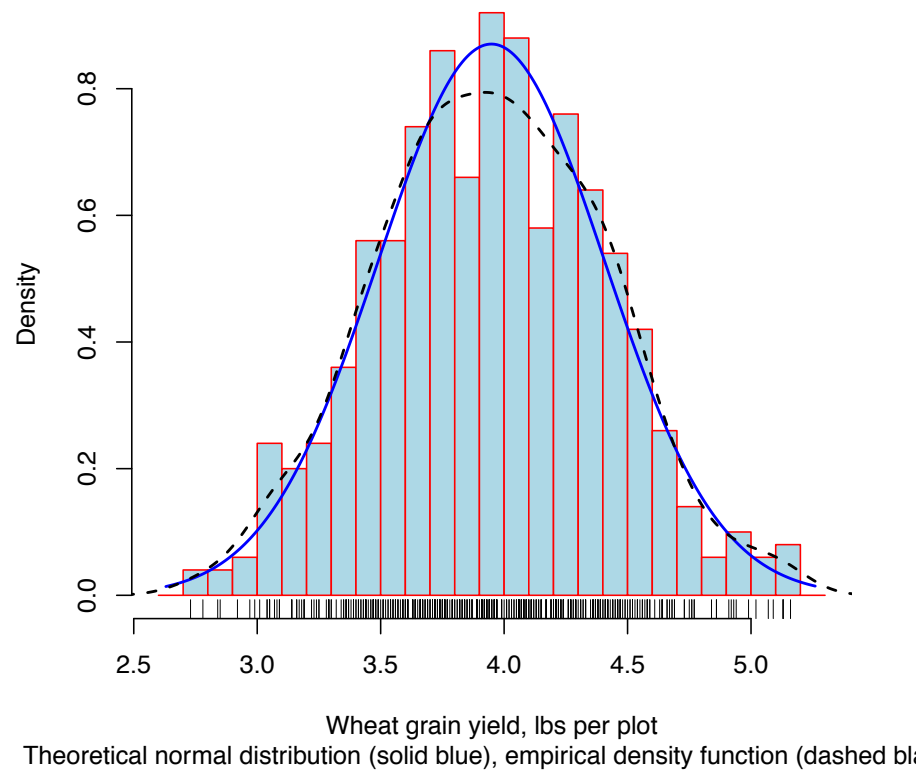


Field layout

Grain yields

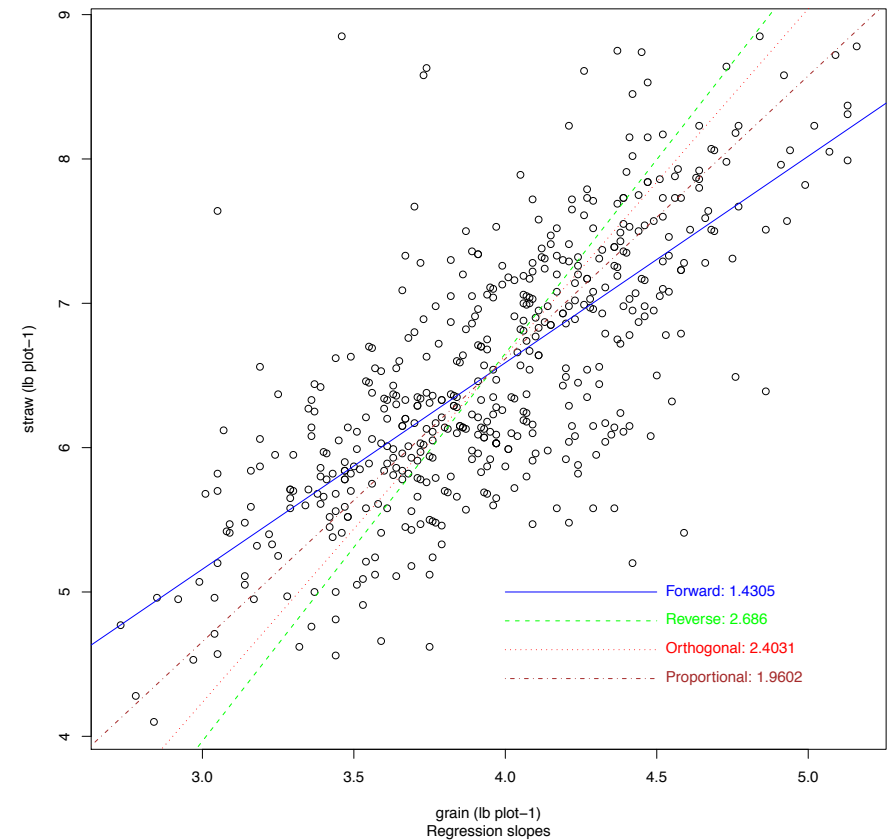
Non-spatial analysis

Mercer & Hall uniformity trial



Modelling the univariate distribution

Mercer-Hall wheat yields

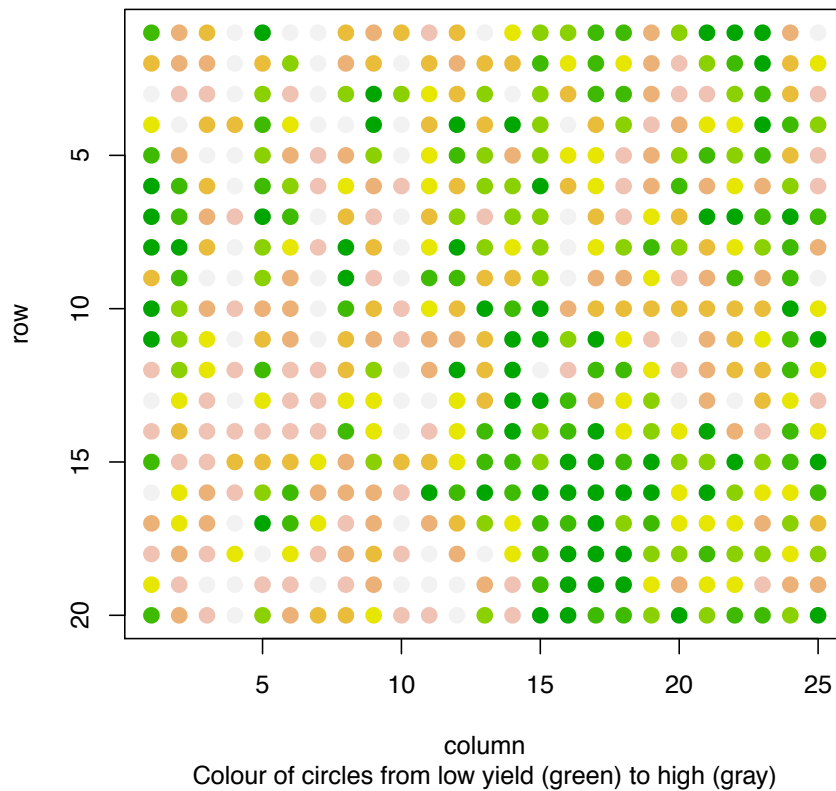


Structural relation grain/straw

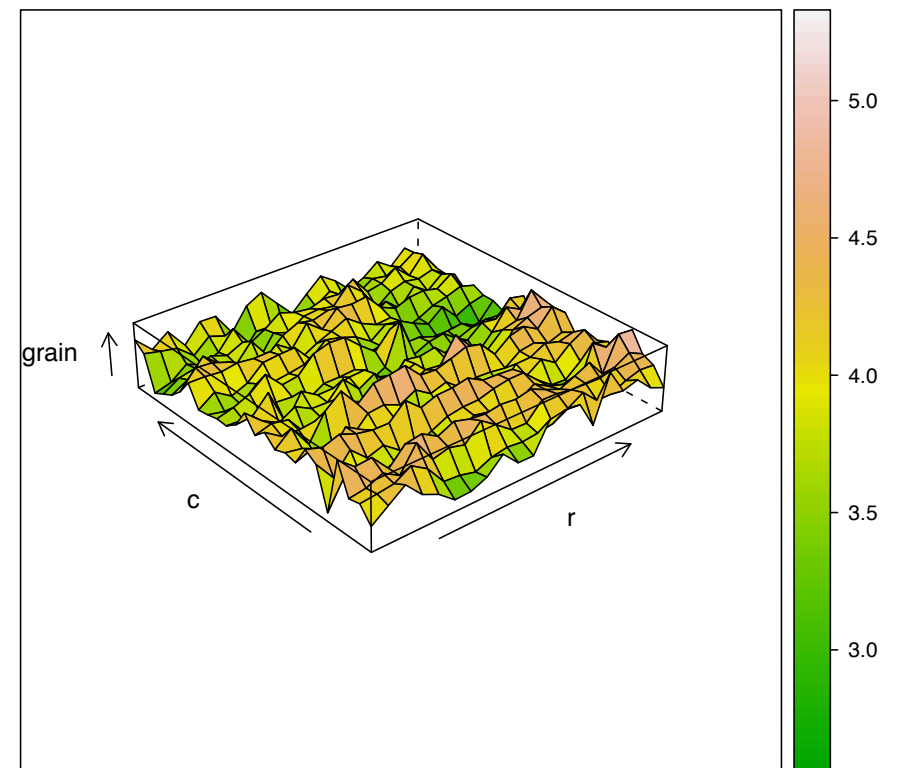


Spatial view of grain yields

Mercer–Hall uniformity trial



Grain yields, lb per plot



Looking SE from NW corner of field

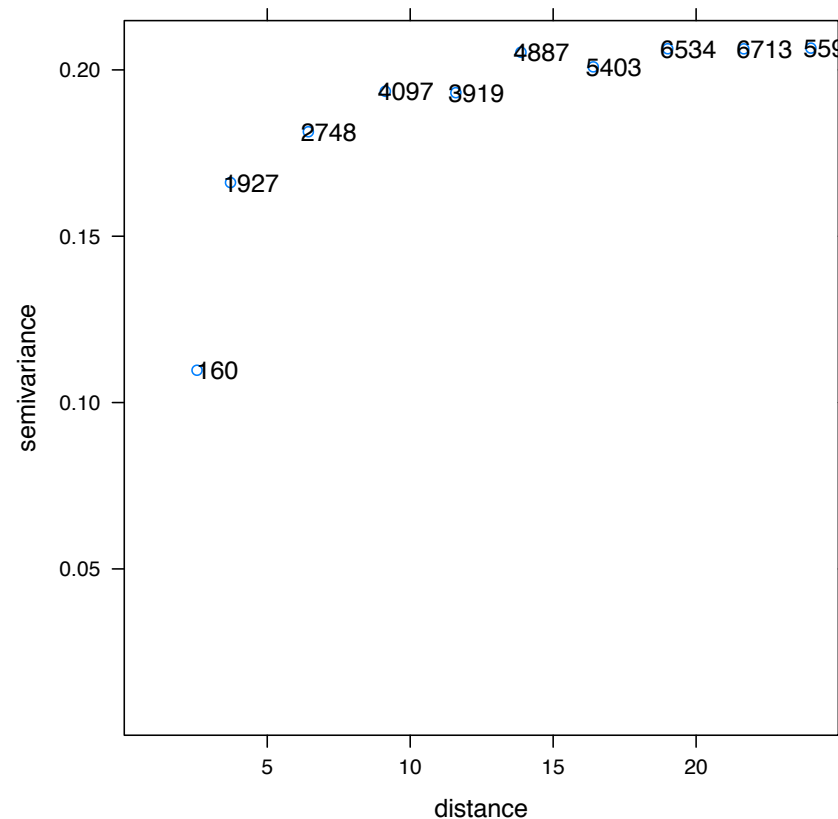
Coloured postplot

Perspective view

Obviously, there is **spatial dependence**: “hot” spots, “cold” spots

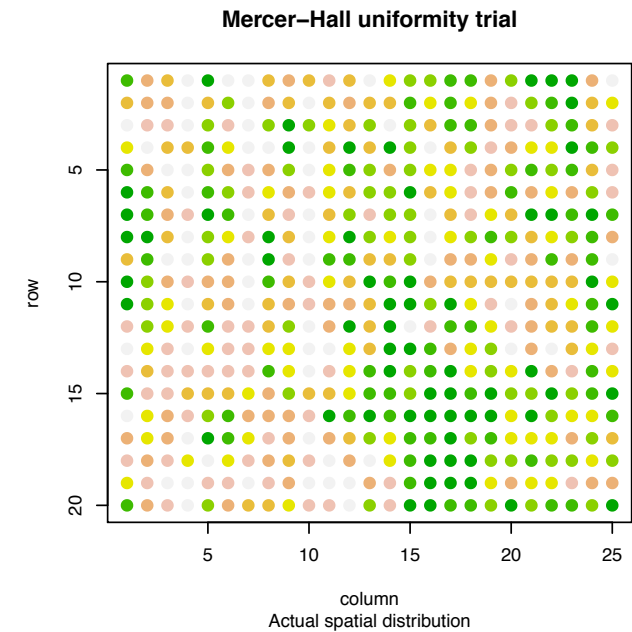
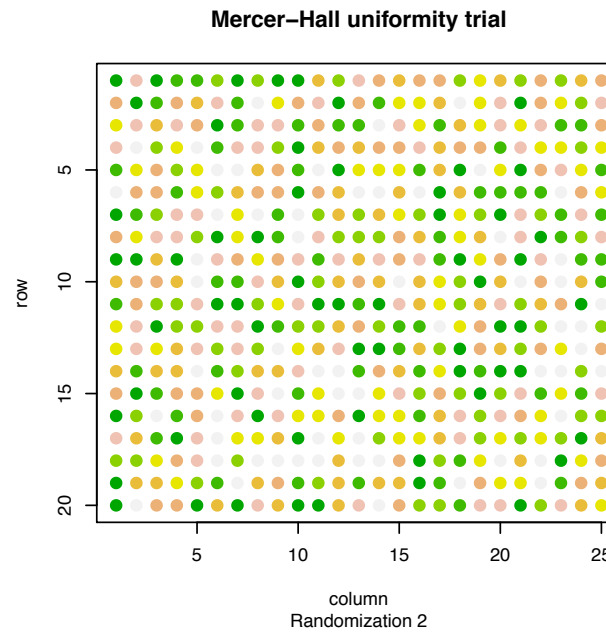
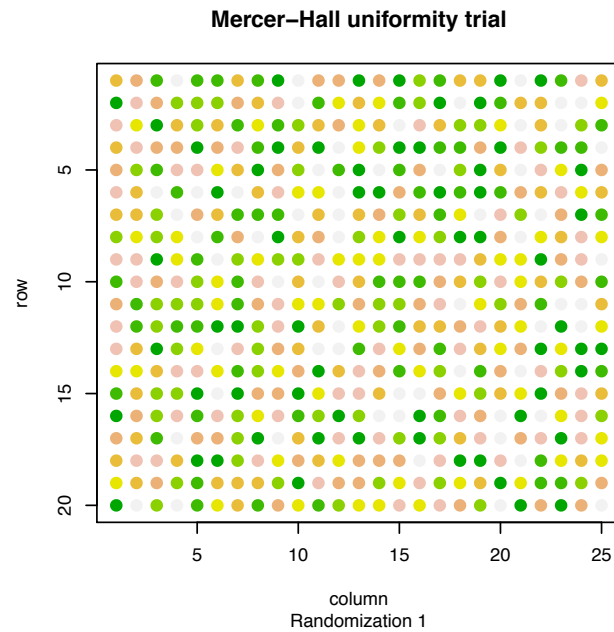


Evidence of spatial dependence



This **variogram** shows that pairs of plots are **more similar** (i.e. have **lower inter-plot variability**) as they are **closer** to each other.

Spatial independence



Two randomizations

original locations



Implications for analysis

Reference: H M van Es and C L van Es. Spatial nature of randomization and its effects on the outcome of field experiments. *Agronomy Journal*, 85:420–428, 1993.

- Nearby plots are **not independent**; observations are “repeated”
- Confidence intervals for e.g. correlations are too narrow
- Solutions:
 1. use a **spatially-balanced design** (see van Es reference)
 2. **larger block size**: include all spatial variability
 3. **replication**
 4. incorporate into analysis (mixed models, REML estimation of parameters)



Spatial dependence

All observations are **intrinsically related** by their **separation vector** (“distance”)

Fact: ‘nearby’ observations in space (also in time) are often similar

This is called **spatial** (or temporal) **dependence** or **auto-correlation** (‘auto’ = with itself)

So, **observations are not independent!**

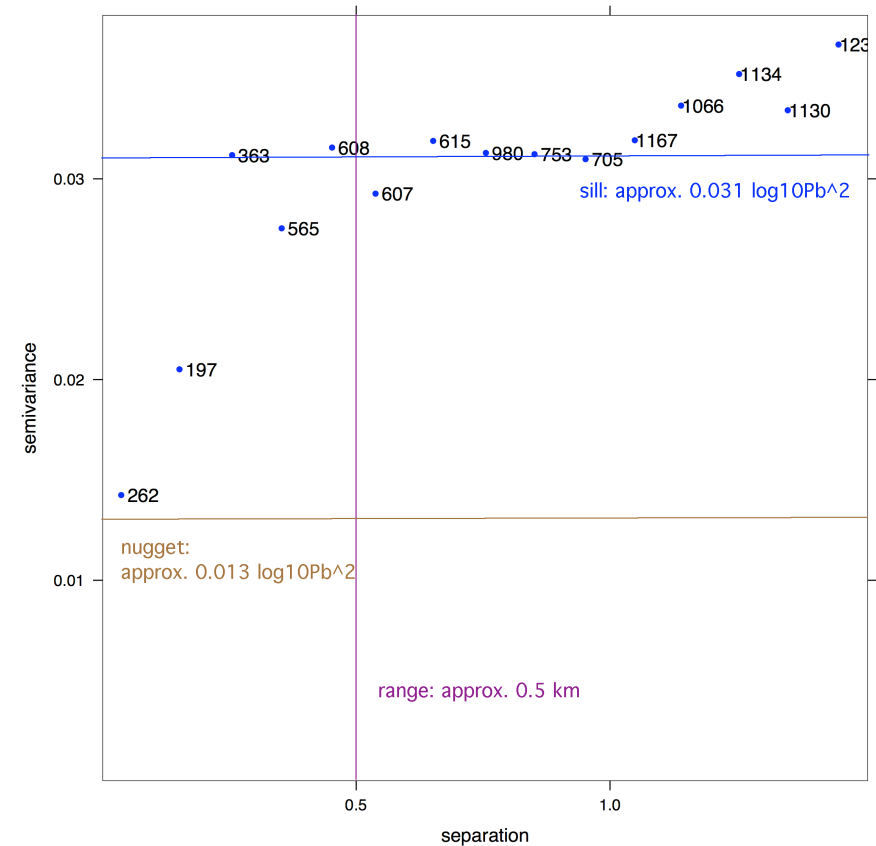
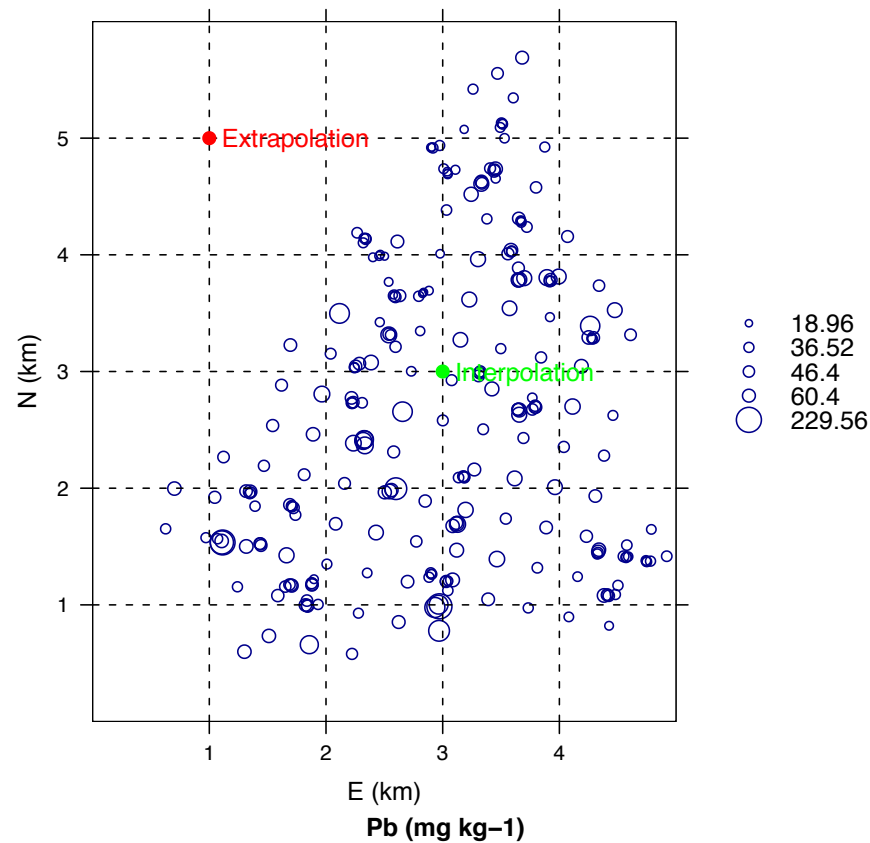
This violates a major assumption of random sampling.

(However, Brus argues convincingly that if the feature-space design was correct, so will be the inferences; although the sampling design may be inefficient.)



Evidence of spatial dependence

Soil samples, Swiss Jura



Clustering of similar values

Spatial auto-correlation (variogram)



Solution 1: avoid spatial dependence

- Do not allow any samples closer than the range of spatial dependence
- This range is known from previous studies, with correlogram or variogram analysis
- This is only valid in **systematic** (grid) designs, only in this case are probabilities of inclusion not changed by the restriction



Solution 2: use geostatistical analysis

- Use **model-based** (also called **geostatistical**) approaches to analyze the data
- Explicitly **models spatial dependence** and uses it for inference

References:

- **Goovaerts, P.** (1997). Geostatistics for natural resources evaluation. New York; Oxford: Oxford University Press.
- **Isaaks, E. H., & Srivastava, R. M.** (1989). Applied geostatistics. New York: Oxford University Press.
- **Webster, R., & Oliver, M. A.** (2001). Geostatistics for environmental scientists. Chichester: Wiley & Sons.
- **Chilés, J.-P., & Delfiner, P.** (1999). Geostatistics: modeling spatial uncertainty. New York: John Wiley & Sons.



Solution 3: account for reduced degrees of freedom

- Adjust formulas for variances etc. according to the **effective** sample size n^* .
- This is computed according to the modelled **spatial dependence** and the observation locations
- That is, the spatially-correlated part (depending on variogram model, partial sill and range) reduces the effective sample size.
- Reference: **Griffith, D.A.** (2005). Effective Geographic Sample Size in the Presence of Spatial Autocorrelation. *Annals of the Association of American Geographers* 95(4): 740–760.

Purposes of geostatistical sampling

1. To make some statement about the **area as a whole**
 - **spatial** mean, total, variance . . . ; e.g. total biomass; average biomass per ha
2. To **map** the distribution of some attribute(s) over an area;
3. To determine the **spatial structure**
 - a regional **trend**;
 - **local** spatial dependence (e.g. by **variogram analysis**);
 - **anisotropy** (direction of maximum dependence)



Models of spatial variation

- **Discrete model** (DMSV): crisp boundaries, discrete classes, no spatial dependence within polygons; variance within class estimated by all samples from the class
 - **“Design-based”** sampling, based on feature-space structure (e.g. strata, continuous feature-space predictors)
- **Continuous model** (CMSV): no boundaries, no classes, spatial dependence to some range, all spatial variability is found by the variable itself. May include a global (trend) and local component
 - **“Model-based”** sampling (‘model’ of spatial dependence)
- **Mixed model** (MMSV)
 1. Stratify by DMSV, model within by CMSV
 2. Design-based methods incorporating spatial structure

What is different about designs considering geostatistics?

The spatial structure is modelled, so ...

1. CMSV: We can place samples for maximum information or minimum cost in a **model-based** (geostatistical) sample
2. MMSV: Can consider both spatial dependence in **geographic space** and the spread of samples in **feature space**



Topic: Optimal grid sampling



Optimal point configuration for the CMSV

In a square area to be mapped, given a fixed number of points that can be sampled, in the case of bounded spatial dependence:

- Points should lie in on some **regular pattern**; otherwise some points duplicate information at others (in kriging, will “share” weights)
- Optimal (for both the “minimal maximum” and “minimal average” criteria): **equilateral triangles** (If the triangle is 1^2 , max. distance to a point $= \sqrt{7}/4 \approx 0.661$)
- Sub-optimal but close: **square grid** (max. distance $= \sqrt{2}/2 \approx 0.707$)
 - Grid should be slightly **perturbed** so samples do not line up exactly; avoids unexpected periodic effects



Computing an optimal grid size with a known variogram model

- Reference: **McBratney, A. B. & Webster, R.** (1981) “The design of optimal sampling schemes for local estimation and mapping of regionalized variables - I and II”. *Computers and Geosciences*, 7(4), 331-334 and 335-365; also in Webster & Oliver.
- In kriging, the estimation error is based **only** on the sample **configuration** and the chosen **model** of spatial dependence, not the actual data values
- So, **if** we know the spatial structure (variogram model), we can compute the maximum or average kriging variances **before** sampling, i.e. before we know any data values.
- Then we can make sampling decisions on the basis of **cost-benefit**

Error variance

- Recall: The kriging variance at a **point** is given by:

$$\begin{aligned}\hat{\sigma}^2(\mathbf{x}_0) &= \mathbf{b}^T \boldsymbol{\lambda} \\ &= 2 \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) - \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j)\end{aligned}$$

- This depends only on the **sample distribution** (what we want to optimise) and the **spatial structure** (modelled by the semivariogram)
- Note that the **values** of the target variable are nowhere in this formula!
- In a **block** this will be lowered by the **within-block** variance $\bar{\gamma}(B, B)$

Reducing kriging error

Once a regular sampling scheme is decided upon (triangles, rectangles, ...), the kriging variance is decreased in two ways:

1. **reduce the spacing** (finer grid) to reduce semivariances; or
2. **increase the block size** of the prediction

These can be traded off; but usually the largest possible block size is selected, based on the minimum decision area.

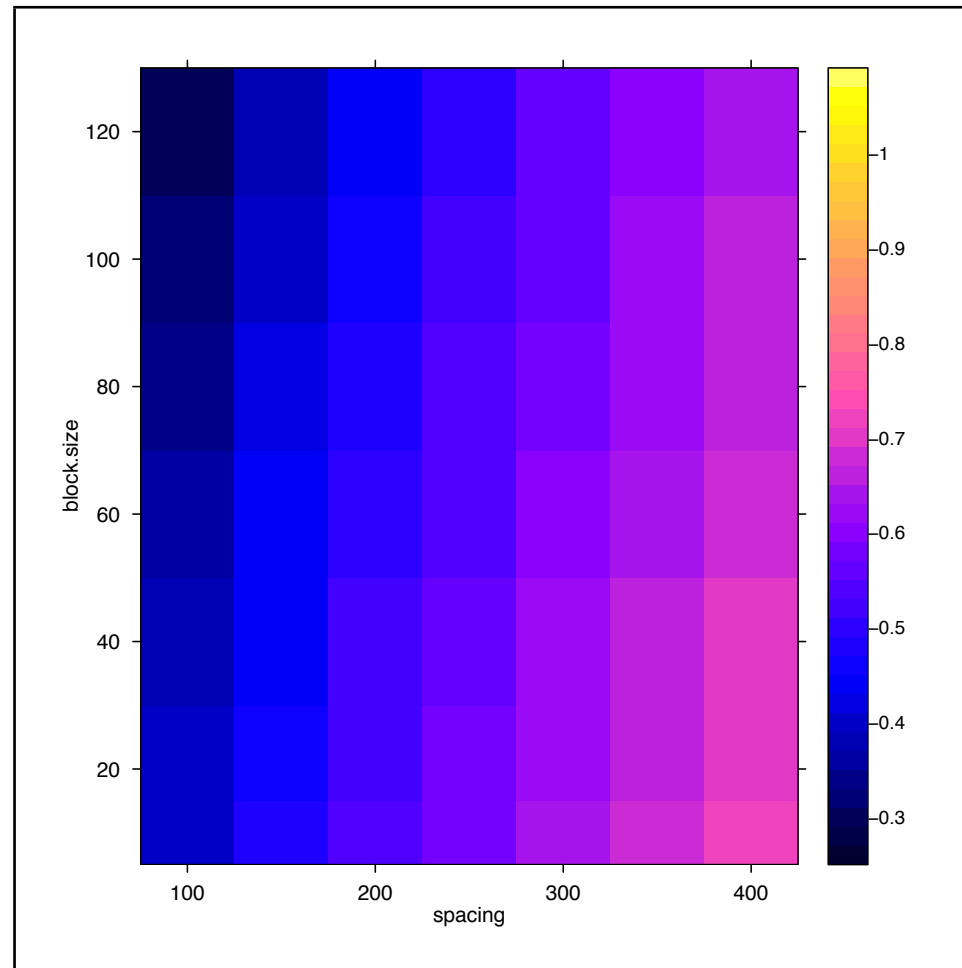


Error as a function of increasing grid resolution

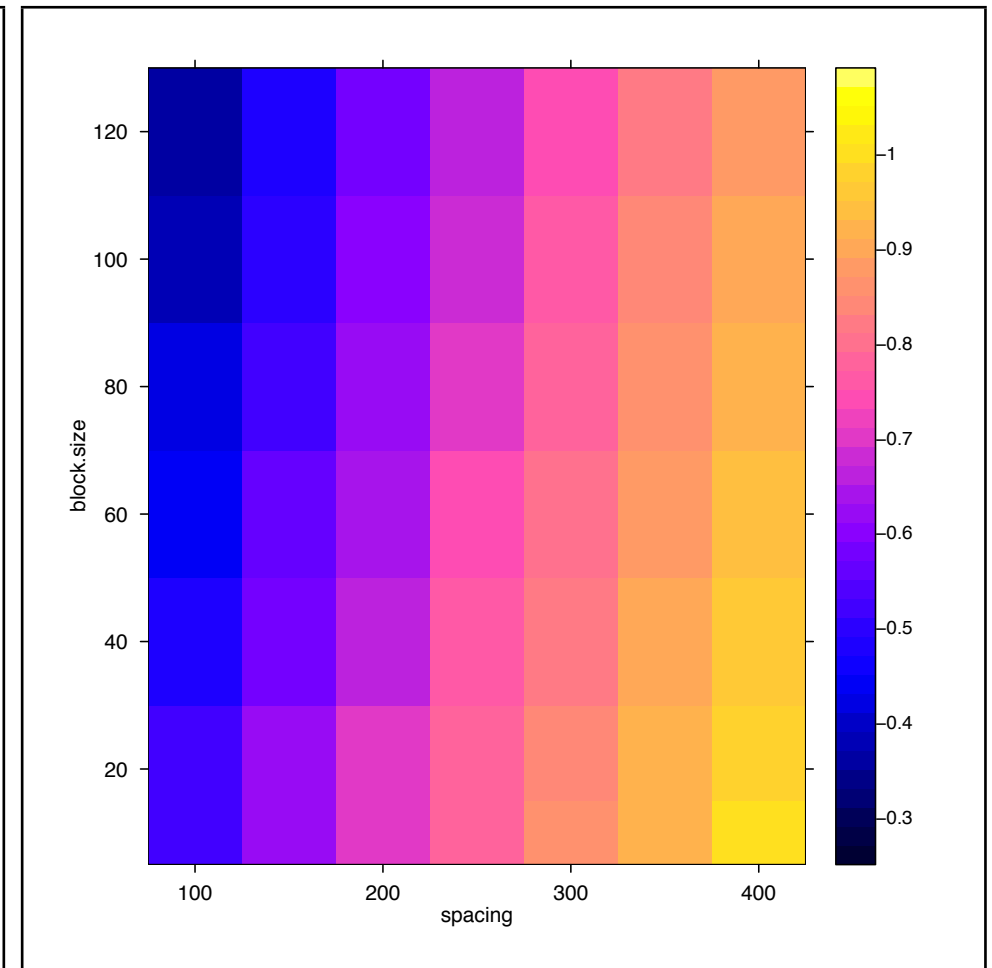
- Consider 4 sample points in a square
- To estimate is one prediction point in the middle (furthest from samples → highest kriging variance)
- Criterion is “minimize the maximum prediction error”
- If the variogram is close-range, high nugget, low sill, we need a **fine** grid to take advantage of spatial dependence; high cost
- If the variogram is long-range, low nugget, high sill, a **coarse** grid will give similar results



Kriging variances at centre point



long range variogram (1200 m)



short range variogram (600 m)

Cost of mapping an area

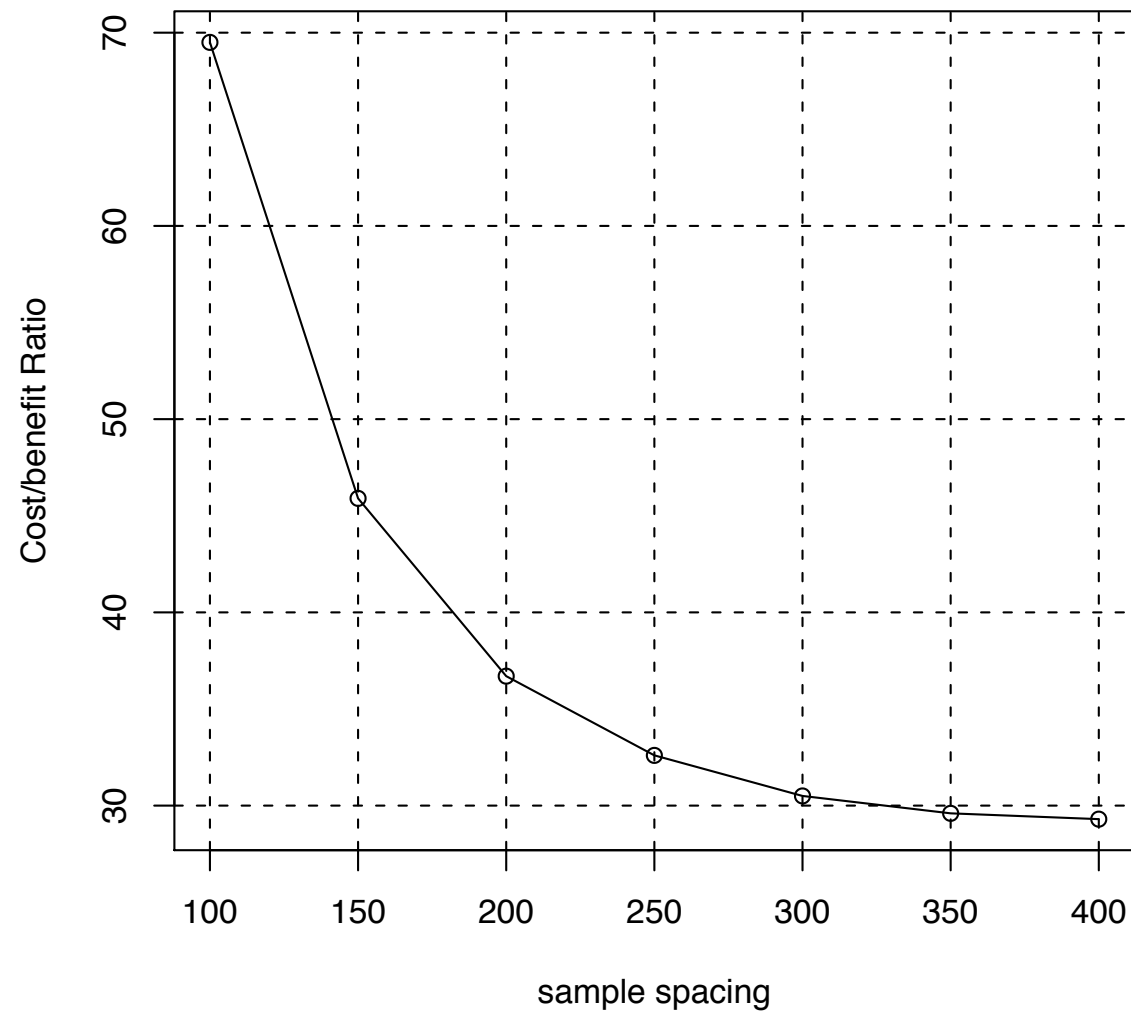
- Given sample spacing (side of grid) g and total area A , the number of sample points required to cover the area is $n = (\frac{\sqrt{A}}{g} + 1)^2$
- Example: 25 km x 25 km area ($A = 625 \text{ km}^2$)
 - $g = 5 \text{ km} \rightarrow ((25/5) + 1)^2 = 36$
 - $g = 0.5 \text{ km} \rightarrow ((25/0.5) + 1)^2 = 2601$
 - $g = 10 \text{ km} \rightarrow ((25/10) + 1)^2 = 12.25 \approx 12$
- Multiply this by the cost of each sample
 - Fixed per sample: time to acquire, equipment rental for this time, laboratory
 - Variable: travel time between samples
- In addition, there is a fixed cost to set up the sampling scheme

Cost-benefit analysis

- Compute a **cost/benefit ratio** and plot against a controllable parameter:
 - sample spacing at a given block size
 - block size for a given sample spacing



Effect of sample spacing on the Cost/Benefit Ratio



(Note: this depends on the variogram)



Topic: Simulated annealing

- Used when there are **previous samples** or **constraints** on where samples can be placed (e.g. buildings)
- **Optimizes** sampling locations by trial-and-error ...
- ... according to some **optimization criterion**, e.g. mean or maximum **kriging prediction variance**
- Must use an **“annealing”** strategy (slowly “cooling” the system) for computational efficiency

References: **D J Brus and G B M Heuvelink**. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138:86–95, 2007.

van Groenigen, J.-W. (2000). The influence of variogram parameters on optimal sampling schemes for mapping by kriging. *Geoderma*, 97(3-4), 223-236.



Problems with the “optimal” grid

The “optimal” grid presented in the previous section is optimal only in restricted circumstances. There are many reasons that approach might not apply:

- **Edge** effects: study area is not infinite
- **Irregularly-shaped** areas, e.g. a flood plain along a river
- **Off-limits** or **uninteresting** areas, e.g. in a soils study: buildings, rock outcrops, ditches . . .
- **Existing samples**, maybe from a preliminary survey; don't duplicate the effort!

Impossible to compute an optimum analytically (as for the regular grid on an infinite plane).



Annealing

Slowly cooling a molten mixture of metals into a stable crystal structure.

During annealing the **temperature** is slowly lowered.

At **high** temperatures, molecules move around rapidly and long distances

At **low** temperatures the system stabilizes.

Critical factor: speed with which temperature is lowered

- too fast: stabilize in a sub-optimal configuration
- too slow: waste of time



Simulated annealing

This is a **numerical analogy** to physical annealing:

- Some aspect of a numerical system is perturbed
- The configuration should approach an optimum
- The amount of perturbation is controlled by a “temperature”

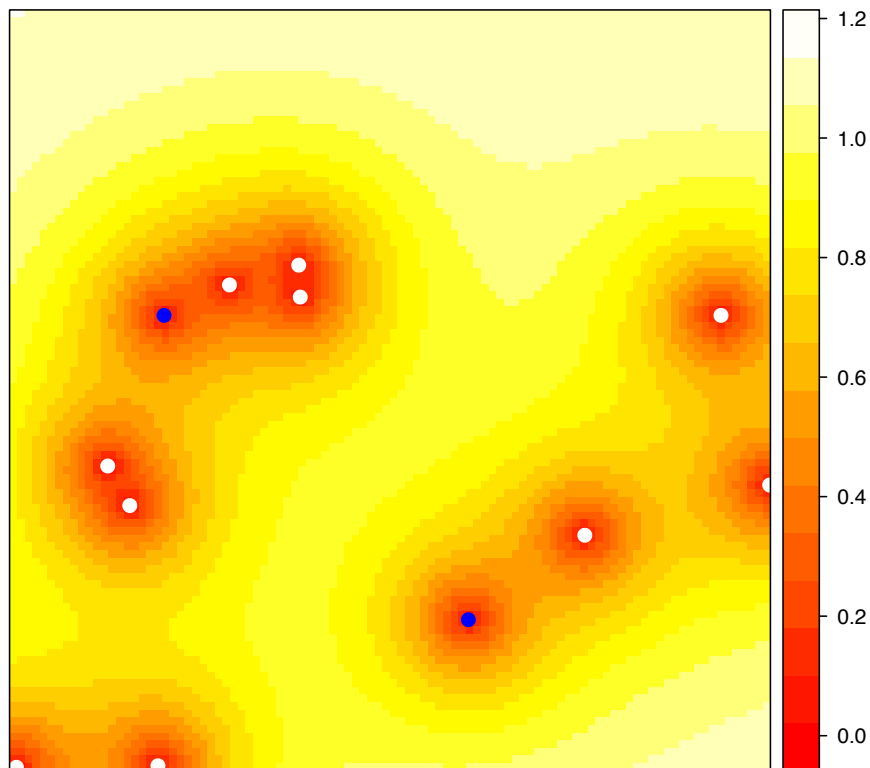
Outline of SSA

1. Decide on an **optimality** criterion
2. Place the desired number of sample points “anywhere” in the study area (grid, random ...); compute **fitness** according to optimality criterion
3. Repeat (**iterate**):
 - (a) Select a point to move; **move** it a **random distance and direction**
 - (b) If outside study area, try again
 - (c) Compute **new fitness**
 - (d) If **better**, accept new plan; if **worse** also accept with a certain **probability**
4. **Stop** according to some **stopping criterion**



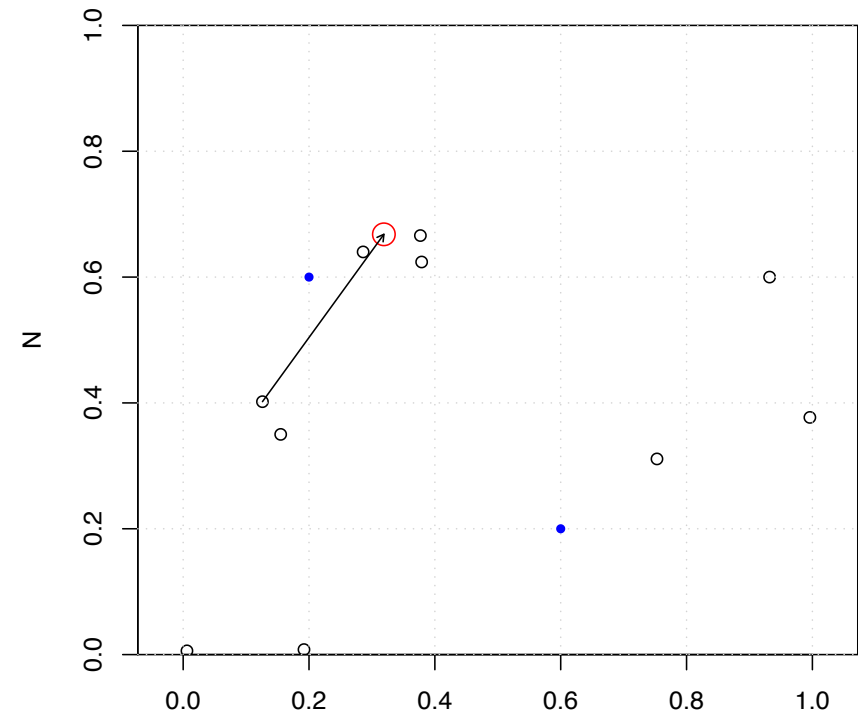
Initial configuration

Initial sampling scheme



Mean, max kriging variance: 0.8176 ; 1.1362

Spatial simulated annealing, step 1



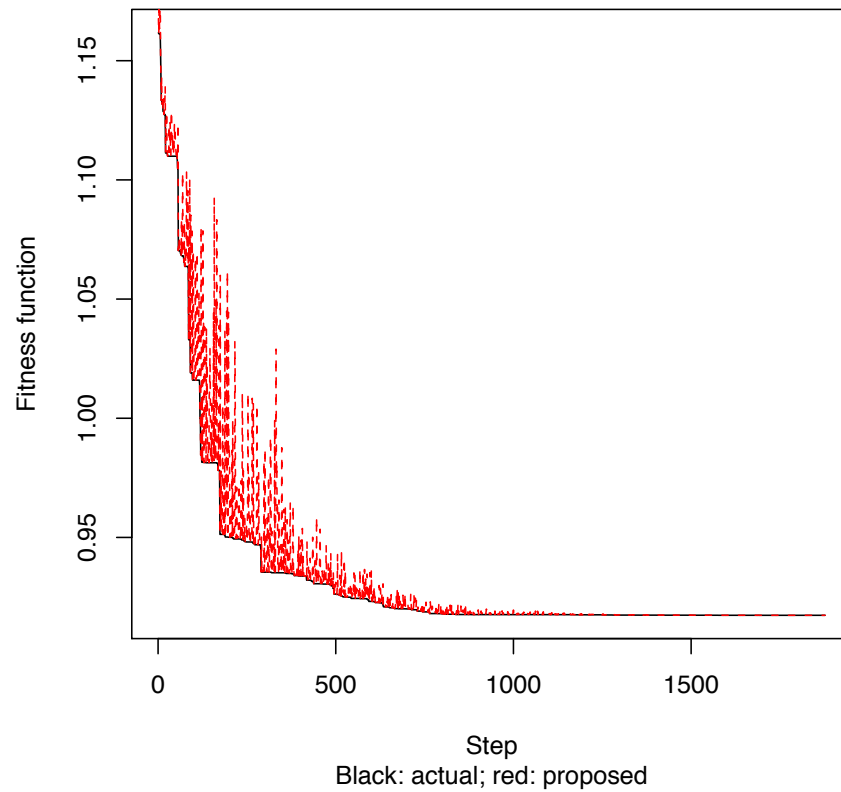
Current fitness: 0.8176 ; new fitness 0.8227

Accept the move if (1) fitness is **improved**; (2) with a certain **probability** (decreasing with time) if fitness is **worse**

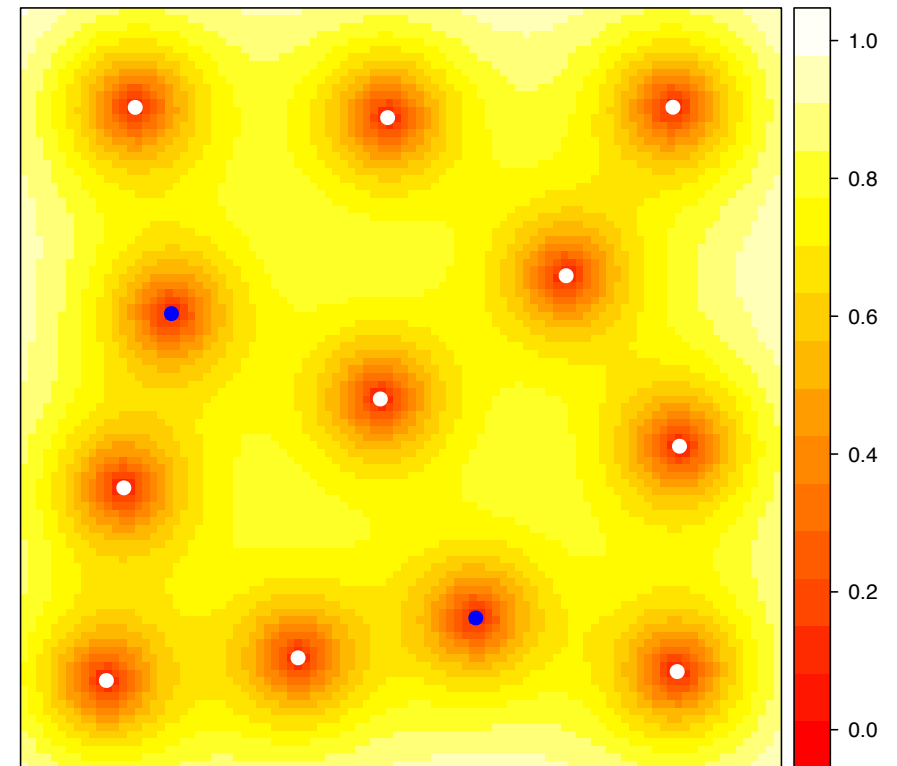


Annealing; final configurations

Fitness function vs. step



Final sampling scheme

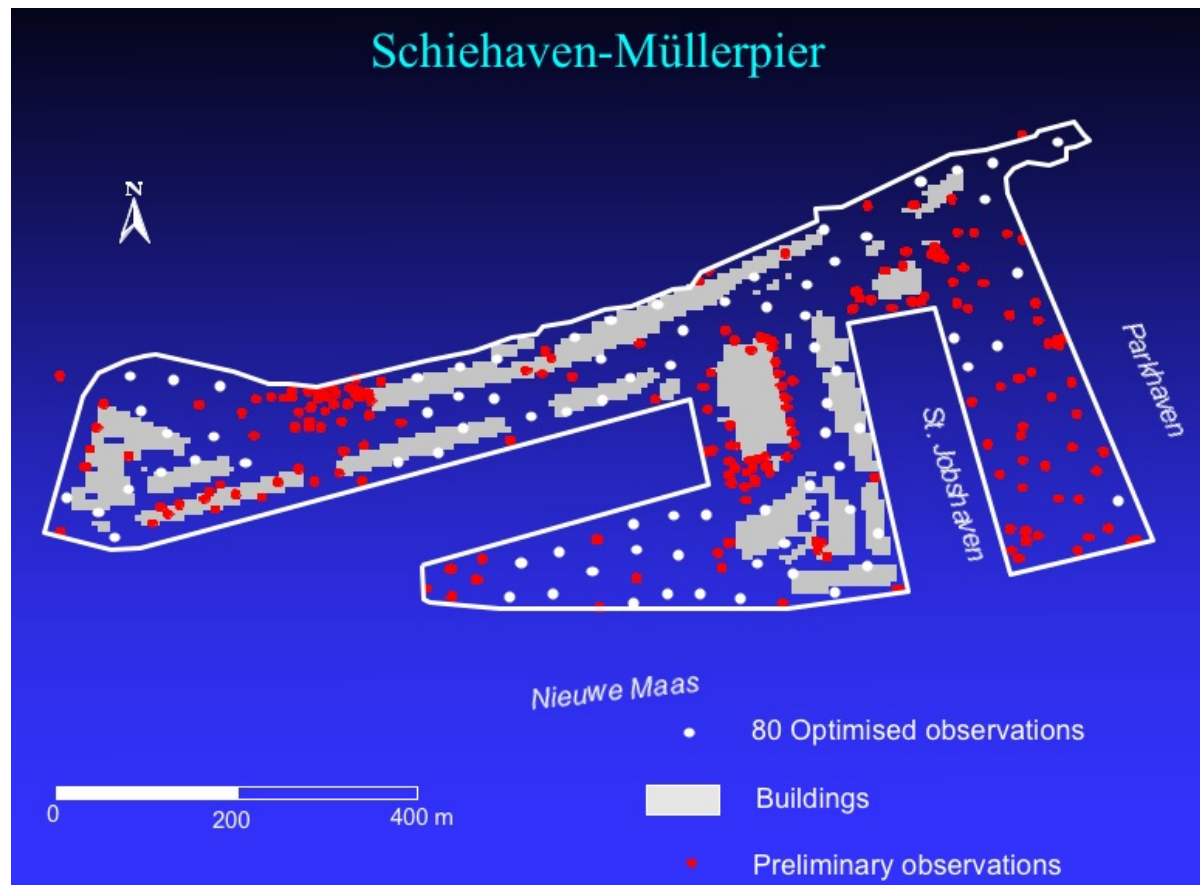


Mean, max kriging variance: 0.6886 ; 0.9796



A real example

Industrial area, existing samples; more must be taken to lower the prediction variance to a target level everywhere; **where** to place the new samples?



Reference: **van Groenigen, J. W., Stein, A., & Zuurbier, R. (1997).** Optimization of environmental sampling using interactive GIS. *Soil Technology*, 10(2), 83-97



Sampling to estimate spatial dependence

If an area has never been sampled for a variable of interest, we need to determine the **range of spatial dependence** in order to set up an efficient sampling scheme.

Note that such a sampling scheme is *not* intended to map an area; thus there may be large “holes” in the coverage; the interest is instead on determining spatial structure.

If a map is wanted, the variogram derived from this first sampling exercise can be used to design an optimum grid sample, as explained above.



Nested spatial sampling

An efficient way **estimate a variogram** is with a **nested** spatial sampling scheme. It is based on work from 1937, re-discovered and extended in 1990.

- Original work: Youden, W. J. & Mehlich A. (1937) *Selection of efficient methods for soil sampling*, Contributions of the Boyce Thompson Institute for Plant Research 9: 59-70
- **Recent paper** re-stating the method: Webster, R. Welham, S. J., Potts, J. M., & Oliver, M. A. (2006) *Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood*, Computers & Geosciences 32: 1320-1333



How to design the nested sample

- **Widest** spacing s_1 is the 'station', which are assumed so far away from each other as to be **spatially independent**
 - furthest expected dependence ...
 - ... based on the landscape ...
 - ... and expected range of process to be modelled
- **Closest spacing** s_n is the shortest distance whose dependence we want to know



Geometric series

- A **geometric series** increases terms by multiplication
- It allows us to cover a wide range of distances (possible ranges) with a few stages.
- Increases spacing in geometric series:
$$s = \sqrt{s_1 \cdot s_n}$$
- Fill in series with further geometric means

Geometric series: example

- First series: $s_1 = 600\text{m}$ (stations), $s_5 = 6\text{m}$ (closest)
- Intermediate spacing: $s_3 = \sqrt{6\text{m} \cdot 600\text{m}} = 60\text{m}$
- Series now $\{600\text{m}, 60\text{m}, 6\text{m}\}$
- Fill in with the geometric means
 - $s_2 = \sqrt{600\text{m} \cdot 60\text{m}} \approx 190\text{m}$
 - $s_4 = \sqrt{60\text{m} \cdot 6\text{m}} \approx 19\text{m}$
- **Final series** $\{600\text{m}, 190\text{m}, 60\text{m}, 19\text{m}, 6\text{m}\}$



Locating the sample points

- Objective: cover the landscape, while avoiding systematic or periodic features
- Method: **random bearings** from **centres** at each stage
- Stations can be along a transect if desired (no spatial dependence)
- From a centre at stage i (E_i, N_i), to find a point (E_{i+1}, N_{i+1}) at the next spacing s_{i+1} :
 - $\theta = \text{random_uniform}[0 \dots 2\pi]$
 - $E_{i+1} = E_i + (s_{i+1} * \sin \theta)$
 - $N_{i+1} = N_i + (s_{i+1} * \cos \theta)$

Example of nested sampling

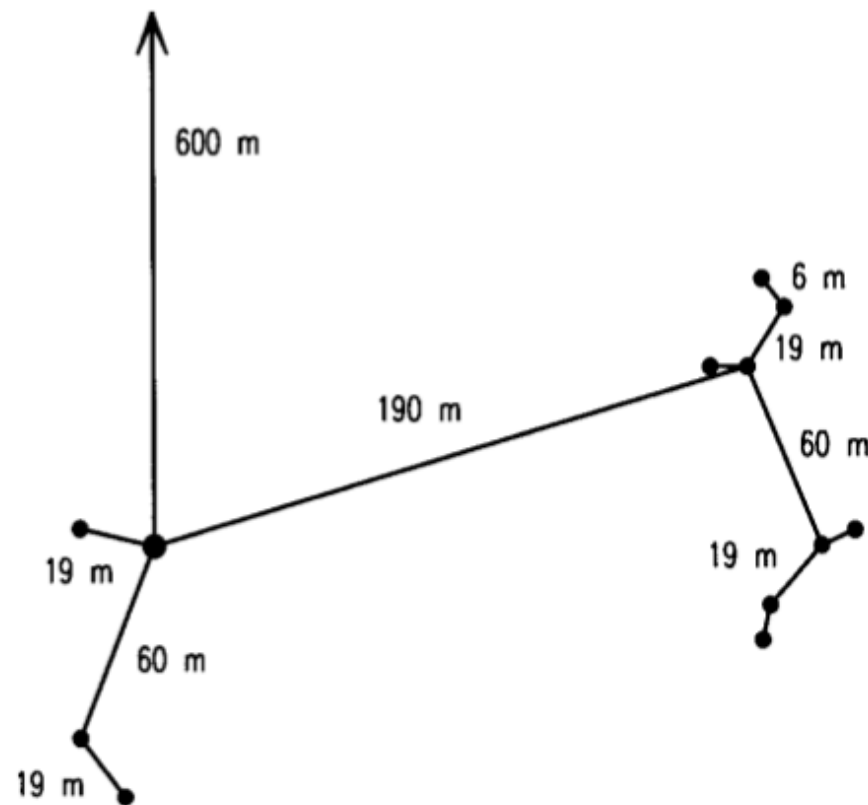


Figure 6.14 Sampling plan for one main centre of the Wyre Forest survey (not strictly to scale).

Source: **Webster, R., and M.A. Oliver** (2008). *Geostatistics for environmental scientists*. 2nd ed. John Wiley & Sons Ltd.



Number of sample points

- Number of stations selected to cover the area of interest
- At each stage S_i , the next stage S_{i+1} has in principle **double** the samples
- One is for all the previous centres from stage $S_1 \dots S_{i-1}$ and one is for the new centre from stage S_i
- So the total number doubles: half old, half new centres
- After the first 4 stages, use an unbalanced design
- This still covers the area, but only uses half the samples at the shortest ranges

Number of sample points: example

- Five stages {600m, 190m, 60m, 19m, 6m}
- Nine stations: $n_1 = 9$
- Double at stages 2 ... 4: $n_2 = 18, n_3 = 36, n_4 = 72$
- At stage 5, only use half the 72 centres, i.e. 36
- Total at stage 5: $72 + 36 = 108$ (would have been 144 with balanced sampling)



Nested ANOVA : Partition Variability by sampling level

- Linear model:

$$Z_{ijk\dots m} = \mu + A_i + B_{ij} + C_{ijk} + \dots + Q_{ijk\dots m} + \varepsilon_{ijk\dots m}$$

- Link with regional variable theory (semivariances): m stages; d_1 shortest distance at m th stage; d_m largest distance at first stage

$$\begin{aligned}\sigma_m^2 &= \gamma(d_1) \\ \sigma_{m-1}^2 + \sigma_m^2 &= \gamma(d_2) \\ &\vdots \\ \sigma_1^2 + \dots + \sigma_m^2 &= \gamma(d_m)\end{aligned}$$

- F-test from ANOVA table; for stage $m + 1$: $F = MS_m / MS_{m+1}$

Nested ANOVA : Interpretation

- There is **spatial dependence** from the closest spacing until the **F-ratio is not significant**.
- Samples from this distance are **independent**
- To take advantage of spatial interpolation, must sample closer than this
- Can estimate how much of the variation is accounted for at each spacing



Topic: Temporal sampling

Sampling in **time** allows inferences of:

- **Time series analysis**: trend, cycles, unusual events ...
- Inference of **temporal auto-correlation**
- **Prediction** of future values

It is often used for **monitoring** the state of some system.



Basics of temporal sampling

Sampling in **time**:

- **Repeat** sampling: when the **same sampled individual** is measured at several points in time;
- **Non-repeat** sampling: different individuals are sampled at the different times

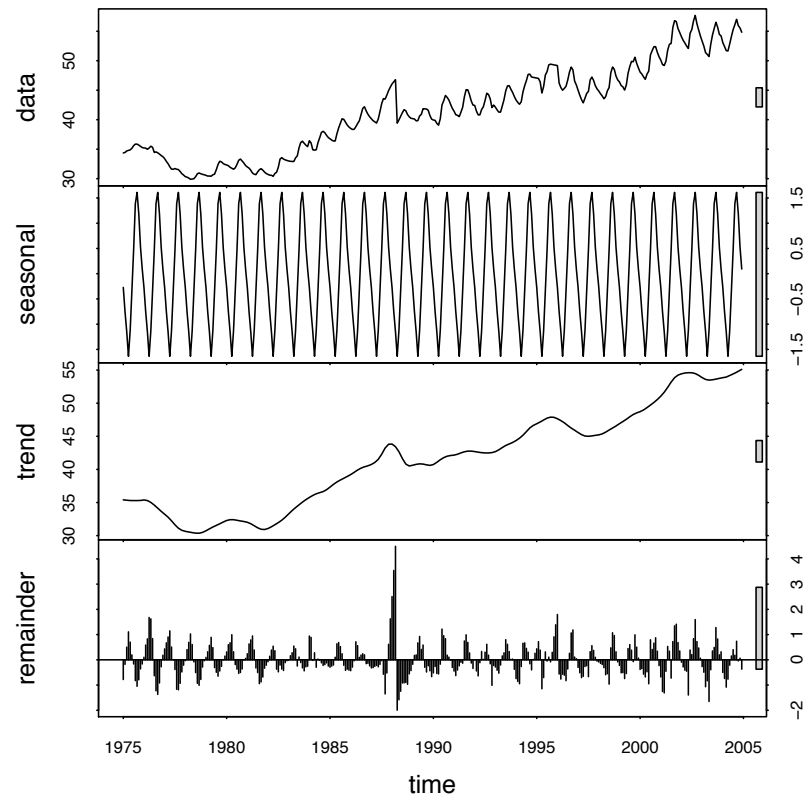
Extra inferences possible from temporal sampling:

- **Time series analysis**: trend, cycles, unusual events ...
- Inference of **temporal auto-correlation**
- **Prediction** of future values



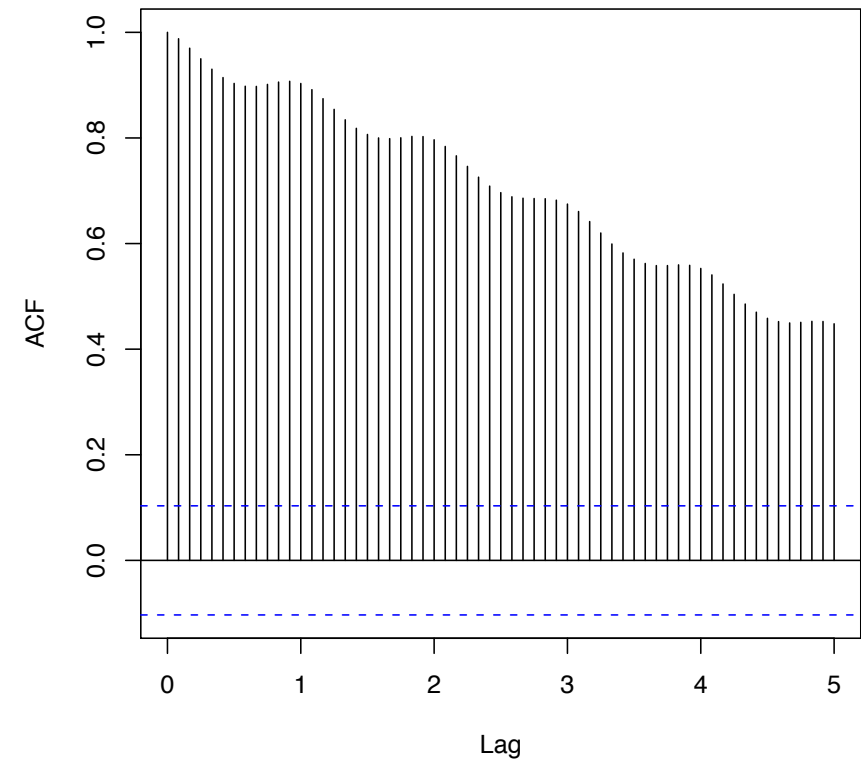
Example of time-series analysis (1)

Groundwater levels in a well, 30-year time series



Decomposition into components

Autocorrelation, groundwater levels, Anatolia well 1

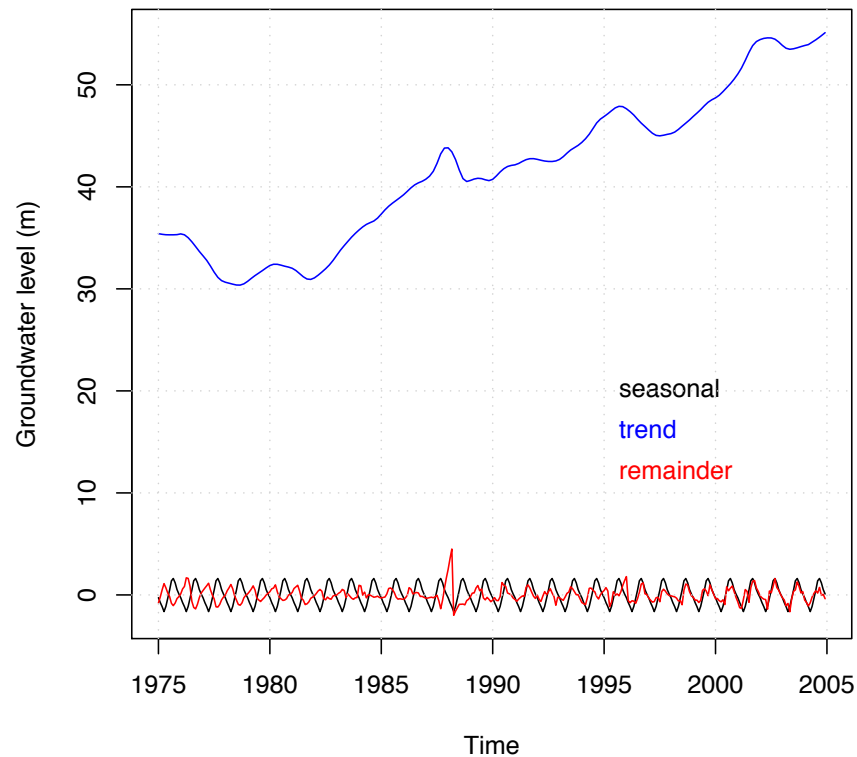


Auto-correlation

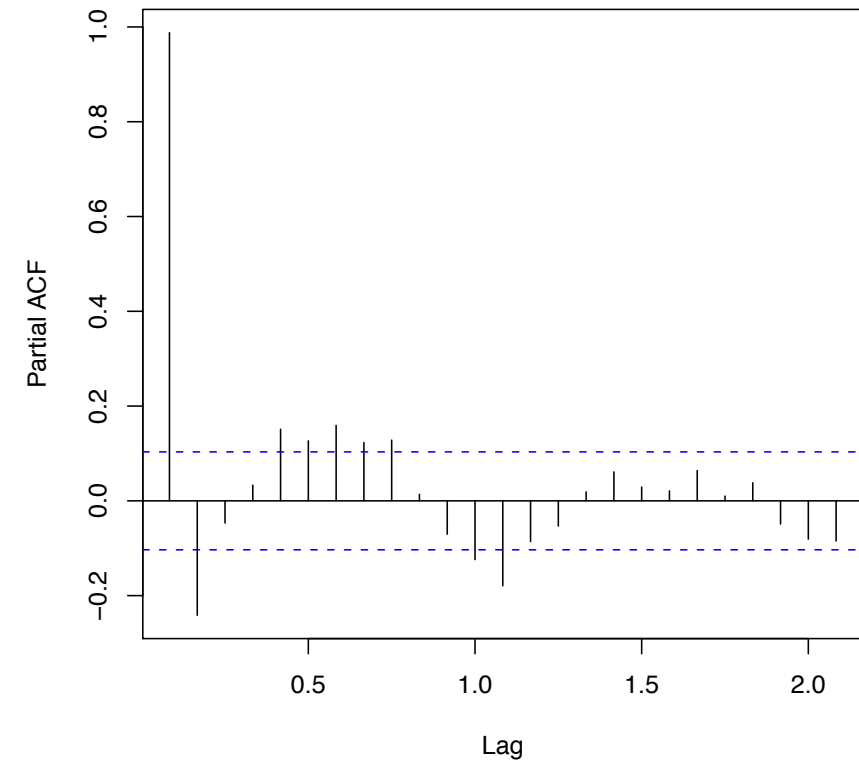


Examples of time-series analysis (2)

Anatolia well 1, decomposition



Partial autocorrelation, Anatolia well 1



Decomposition into components

Partial auto-correlation

Types of temporal sampling

- **Repeat** sampling: when the **same sampling units** are measured at several points in time;
 - Monitoring a location or individual: measuring the same **indicators**
 - *Examples*: Weather stations, stream gauges
- **Non-repeat** sampling: different units from the **same population** are sampled at the different times
 - *Examples*: Census; yearly crop field surveys



Sampling frequency

(also called the “sampling rate”)

This is the **number of samples per time period** taken from a **continuous** phenomenon.

It **discretizes** the signal.

Example: stream flow is continuous; but recording of the water level may be only once a day.

Example: crop growth is continuous, but we may measure the height, biomass etc. only at two-week intervals.

Inverse is called the **sampling interval**, i.e. the time between sampling times.



Implications of the sampling frequency

1. There is no information with which to interpolate between sampling times
2. Any **periodic behaviour** that has a **higher frequency** than $1/2$ the sampling frequency can not be identified.
 - *Example:* Daily temperature cycles can not be identified with daily measurements
 - They can be identified with twice-daily measurements



Topic: **Advanced topics**

Here we mention some advanced topics in sampling that may be applicable in certain situations.

Details may be found in the listed references.



Two-stage sampling

In many situations we would like to sample a population, but certain parts are more interesting than others.

Example: soil pollution in some region; we would like to concentrate on the pollution “hot spots”. We need detailed maps of these for action plans; but we do need to map the whole area to find these.

Two-stage sampling has:

1. A preliminary sampling design;
2. A second sampling design **based on the results of the first**

The first design is set up to be sure to find interesting subpopulations for the followup sampling. Since this second one is biased, appropriate analytical techniques must be used.



Two-stage sampling to estimate sample size

Another situation where two-stage sampling is attractive is if we **don't have any estimate of population standard deviation**.

- A **preliminary sample** is taken to estimate this; then the required **total sample size** can be computed as explained above.
- The **second sample** is used to **complete** the sampling
- Both samples can be used together for analysis.



Adaptive sampling

This is a “modify as you go” alternative to two-stage sampling. During the sampling itself, we change the design according to what we see.

These are often used in social or medical surveys. For example, if we are sampling for prevalence of tuberculosis in a city, and we find one case, we would like to sample nearby, or sample relatives and contacts of the case.

They can also be used in geographic survey. For example, if we find a polluted soil, we'd like to sample nearby to see the extent of the pollution.

Reference: Thompson, S. K., & Seber, G. A. F. (1996). Adaptive sampling. New York: Wiley.

Topic: References

Sampling is an important topic which has been treated extensively by many authors. This section lists some reference material in the following categories:

1. Textbooks
2. Technical reports
3. Papers
4. Web pages
5. Computer programs



Textbooks

- de Gruijter, J., Brus, D. J., Bierkens, M. F. P., & Kotters, M. (2006). Sampling for Natural Resource Monitoring: Springer. ISBN 978-3-540-22486-0

Especially useful for **geostatistical** sampling and **spatio-temporal** monitoring schemes. Includes useful **decision trees** for selecting sampling designs.

- Cochran, W. G. (1977). Sampling Techniques (3rd ed.). New York: John Wiley. ISBN 0-471-16240-X

The classic text, especially for **survey** sampling. Includes formulas for computing population parameters for different sampling designs.

- de Vries, P. G. (1986). Sampling theory for forest inventory: a teach-yourself course. Berlin ; New York: Springer-Verlag. ISBN 0387170669

Cornell University
Library[Home](#) • [Contact Us](#)

» Download Book (PDF, 8105 KB)



Book 2006

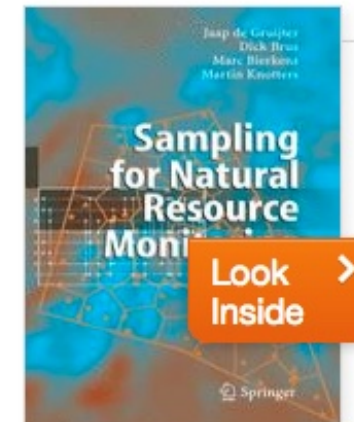
Sampling for Natural Resource Monitoring

Authors: [Jaap J. de Groot](#), [Marc F. P. Bierkens](#), [Dick J. Brus](#), [Martin Kotters](#)

ISBN: 978-3-540-22486-0 (Print) 978-3-540-33161-2 (Online)



Download Book (PDF, 8105 KB)

MyCopy Softcover Edition
24.99 EUR/USD

Technical reports

- Schreuder, H. T., Ernst, R., & Ramirez-Maldonado, H. (2004). Statistical techniques for sampling and monitoring natural resources. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. Gen. Tech. Rep. RMRS-GTR-126.

<https://www.fs.fed.us/rmrs/publications/>

statistical-techniques-sampling-and-monitoring-natural-resources

Practical information with theory, especially for **forestry applications**.

Lots of worked examples. Assumes no prior knowledge and develops the required statistical theory.

- U.S. Environmental Protection Agency. (2002). Guidance for choosing a sampling design for environmental data collection. Washington, DC: US EPA.

<https://www.epa.gov/quality/>

guidance-choosing-sampling-design-environmental-data-collection-use-developing-quality

Especially for compliance with EPA rules (typical of environmental agencies).

Papers

- Stein, A., & Ettema, C. (2003). An overview of spatial sampling procedures and experimental design of spatial studies for **ecosystem comparisons**. Agriculture, Ecosystems & Environment, 94(1), 31-47.
[http://dx.doi.org/10.1016/S0167-8809\(02\)00013-0](http://dx.doi.org/10.1016/S0167-8809(02)00013-0)
- Brus, D. J., & de Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between **design-based** and **model-based** sampling strategies for soil (with Discussion). Geoderma, 80(1-2), 1-59.
[http://dx.doi.org/10.1016/S0016-7061\(97\)00072-4](http://dx.doi.org/10.1016/S0016-7061(97)00072-4)

Web pages

- **NIST/SEMATECH e-Handbook of Statistical Methods**

<http://www.itl.nist.gov/div898/handbook/>; §3.1.3.4, 3.3.3

Emphasis on process quality control, sample plans are for proportions.

- **Electronic Statistics Textbook** (StatSoft);

<http://www.statsoft.com/Textbook/Power-Analysis/>; topic **Power Analysis**

Simple introduction to many statistics topics.



Computer programs (1/3)

R <http://www.r-project.org/>

The dominant **open-source statistical visualization and computing** language and environment; Unix, MS-Windows and Mac OS X.

- Random sampling with the `sample` method; random number generation with many distributions (Uniform, Normal, Poisson, Binomial ...), e.g. `runif`, `rnorm`, `rpois`.
- Methods for power analysis: `power.t.test`, `power.prop.test`, `power.anova.test`;
- Spatial sampling schemes with `spsample` in the `sp` library.
- Package `spscosa` “Spatial Coverage Sampling and Random Sampling from Compact Geographical Strata” from Alterra (NL); based on k-means; implements some methods described in the de Gruijter *et al.* text



Computer programs (2/3)

G*Power 3 <http://www.gpower.hhu.de/>

from Heinrich-Heine-University, Düsseldorf (D); compute the power of statistical tests to find true differences; compute required sample size for a given power. MS-Windows and Mac OS X.

Visual Sample Plan (VSP) <http://vsp.pnnl.gov/> (US Environmental Protection Agency)

From the Pacific Northwest National (USA) Laboratory; aimed at environmental modelling for remediation. MS-Windows.



Computer programs (3/3)

ArcGIS (ESRI)

Spatial analyst extension; user-contributed extensions

QGIS (open-source GIS)

Also has links to R, so can run R-based sampling methods and user-written programs

Conclusion

Sampling is the way we get information about reality.

To make **valid inferences**, the sampling scheme must be **carefully designed** according to the **research questions**.

Time, money, effort is always **limited**.

