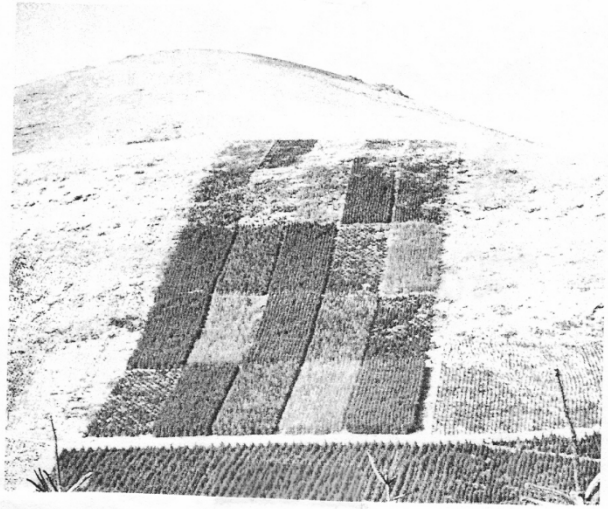
Excursional Statistics

Cramér Society Summer School

Jan-Eric Englund

June 15, 2011

Where did it all start?



BETTGELERT FOREST
LAYOUT

N

Elevation:

1730–1800′	*B*	*A*	*E*	*D*	*C*
1530–1730′	*C*	*E*	*B*	*A*	*D*
1460–1590′	*A*	*C*	*D*	*E*	*B*
1340–1460:	*D*	*B*	*A*	*C*	*E*
1250–1340′	*E*	*D*	*C*	*B*	*A*

A. Sitka spruce
B. Japanese larch
C. Sitka spruce/Japanese larch 50/50
D. Sitka spruce/Pinus contorta 50/50
E. Norway spruce/European larch 50/50

Two rows of Beech planted on each side of the series.

Plate 7. Layout of Bettgelert Experiment.

Who is who in the forest

Bertil Matérn (1917-2007)

He was from 1941 working at professor the department of Harald **Cramér** and came to the forest research institute 1945.

The PhD thesis from 1960 about spatial statistics is very famous in the area and reprinted 1986. One chapter is later on referred to as Matérn processes.

Matérn covariance

From Wikipedia:

In statistics, the **Matérn covariance** (named after the Swedish forestry statistician Bertil Matérn) is a covariance function used in spatial statistics, geostatistics, machine learning, image analysis, and other applications of multivariate statistical analysis on metric spaces. It is commonly used to define the statistical covariance between measurements made at two points that are *d* units distant from each other. Since the covariance only depends on distances between points, it is stationary. If the distance is Euclidean distance, the Matérn covariance is also isotropic.

The Matérn covariance between two points separated by *d* distance units is given by

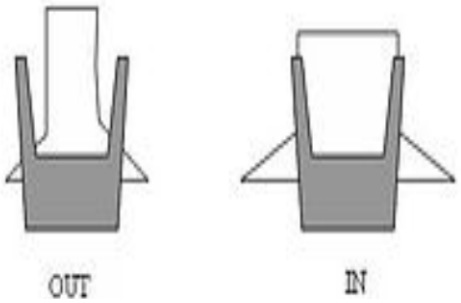
$$C(d) = \sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(2\sqrt{\nu}\frac{d}{\rho} \right)^{\nu} K_\nu \left(2\sqrt{\nu}\frac{d}{\rho} \right),$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind, and ρ and ν are non-negative parameters of the covariance.

Relascope

The Relascope is often used for point sampling. This is done by using the set spacing marked in the Relascope to gauge whether a tree is IN or OUT of the stand (Figure). If the tree is IN this means that it is counted as basal area within ones plot.

To figure out what this trees basal area is all one has to do is multiply the number of trees by the basal area factor which is based on the width of ones gauge.



Terminology: **Plot**

Fisher's agricultural experiments in Rothamsted Experimental Station are the basis for all experimental design, and therefore the terminology from agricultural experiments are transferred to other areas.

The unit you work with in the field is called a **plot**.

Here a field with 9 plots and a balanced design with three levels of treatment A.

For example, A1, A2 and A3 are three different varieties.

A1	**A1**	**A2**
A1	**A2**	**A3**
A3	**A2**	**A3**

Terminology: **Block**

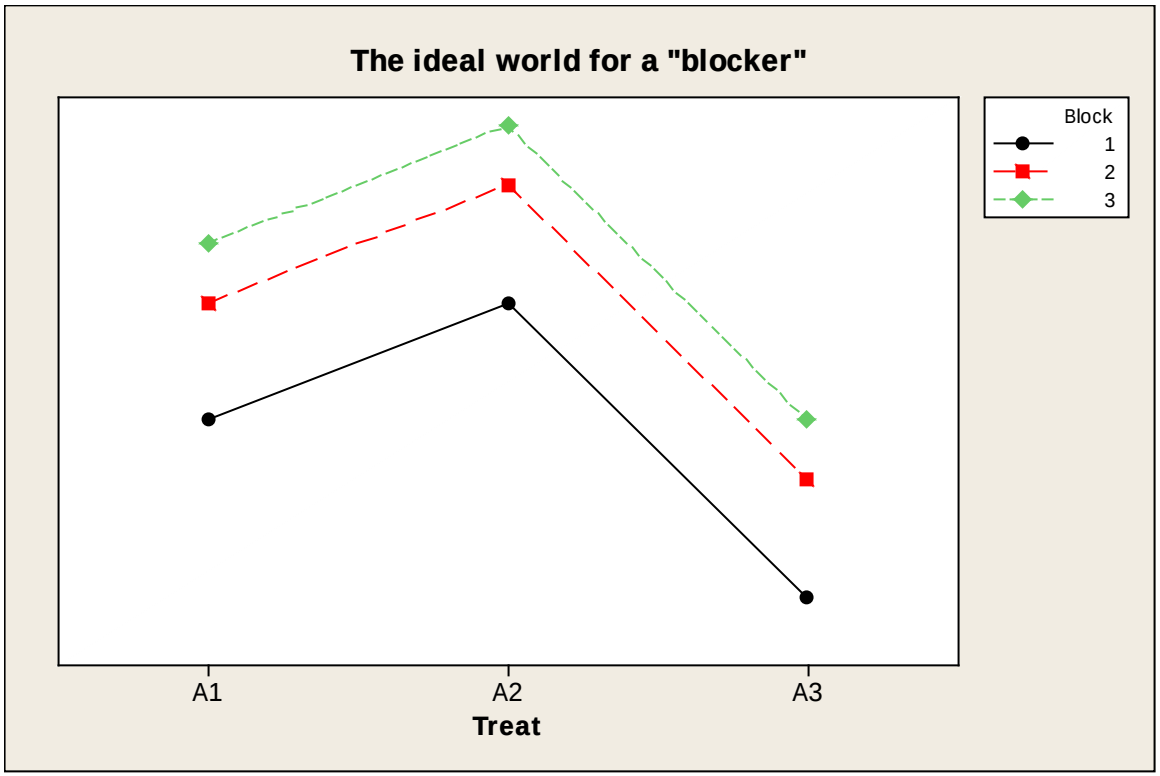
To find differences between the treatment levels you want the random error to be as small as possible. One idea to test is to put similar plots into blocks with homogeneity within the blocks.

If there is a variation (gradient) in the North-South direction but no much difference in the West-East direction it is natural to make blocks as below.

Block 1	**A1**	**A3**	**A2**
Block 2	**A1**	**A2**	**A3**
Block 3	**A3**	**A2**	**A1**

(Compare paired t-test in the basic course.)

Terminology: **Block**



In forestry the blocks sometimes are different research stations, and in this case it would be hard to fulfil the illustration of the picture because there are different soils, climate, etc.

Terminology: **Fix and Random effect**

The treatment in the example is definitely a fixed effect,
but what about the block effect?

In a simple balanced block design you can say that it doesn't matter for the comparison between the treatment levels whether the block is considered as a random or a fixed effect.

But for more complex design it matters…

Much of the problems are caused by the fact that the block was a fixed effect in older literature but now often is considered to be a random effect.

Terminology: **Split-plot designs**

Here we consider split-plot designs *with* blocks (but there are also split-plot designs without blocks).

You can consider this in two steps:

First you do a block design with one factor (the main plots).

Block 1	**A1**	**A3**	**A2**
Block 2	**A1**	**A2**	**A3**
Block 3	**A3**	**A2**	**A1**

These plots cannot be made smaller due to practical reasons.

(For example it can be soil preparation.)

Terminology: **Split-plot designs**

In the next step you split the plots (= the main-plots) into split-plots where you use your second factor.

This factor is here called C with two levels, C1 and C2.

Block 1	C1 C2 A1	C2 C1 A3	C1 C2 A2
Block 2	C2 C1 A1	C1 C2 A2	C2 C1 A3
Block 3	C2 C1 A3	C2 C1 A2	C1 C2 A1

Note: If you just get the data you cannot see whether it is a split-plot design or a block design with two crossed factors.

Model for Split-plot designs

Summarizing, the model has two error terms and is written as

$$y_{ijk} = \mu + \alpha_i + b_j + \delta_{ij} + \gamma_k + (\alpha\gamma)_{ik} + \varepsilon_{ijk}$$

(The error term δ_{ij} is actually the interaction between the main plot factor A and the block effect and could be denoted $(ab)_{ij}$.)

In general you can say that if the interaction between the two factors *not* is significant you have saved a lot of problems!

I don't present the "old" solutions considering the block to be a fixed effect and a lot of formulas for the comparisons. The problem is that this approach can give different results compared to the one from modern computer packages.

Split-plot designs; what are the main problems?

1) It is tempting to do the calculation in two steps,
   - first by the block design for the sum over the main plots,
   - second an analysis for the split-plots.
   This hand calculation does not necessarily give you the same result as the computer program if any of the variance components are estimated to be 0.

2) When the interaction is significant you have some different comparisons between means that have to be treated differently and some of these comparisons have not clearly defined degrees of freedom.

Split-plot designs; illustration of 2)

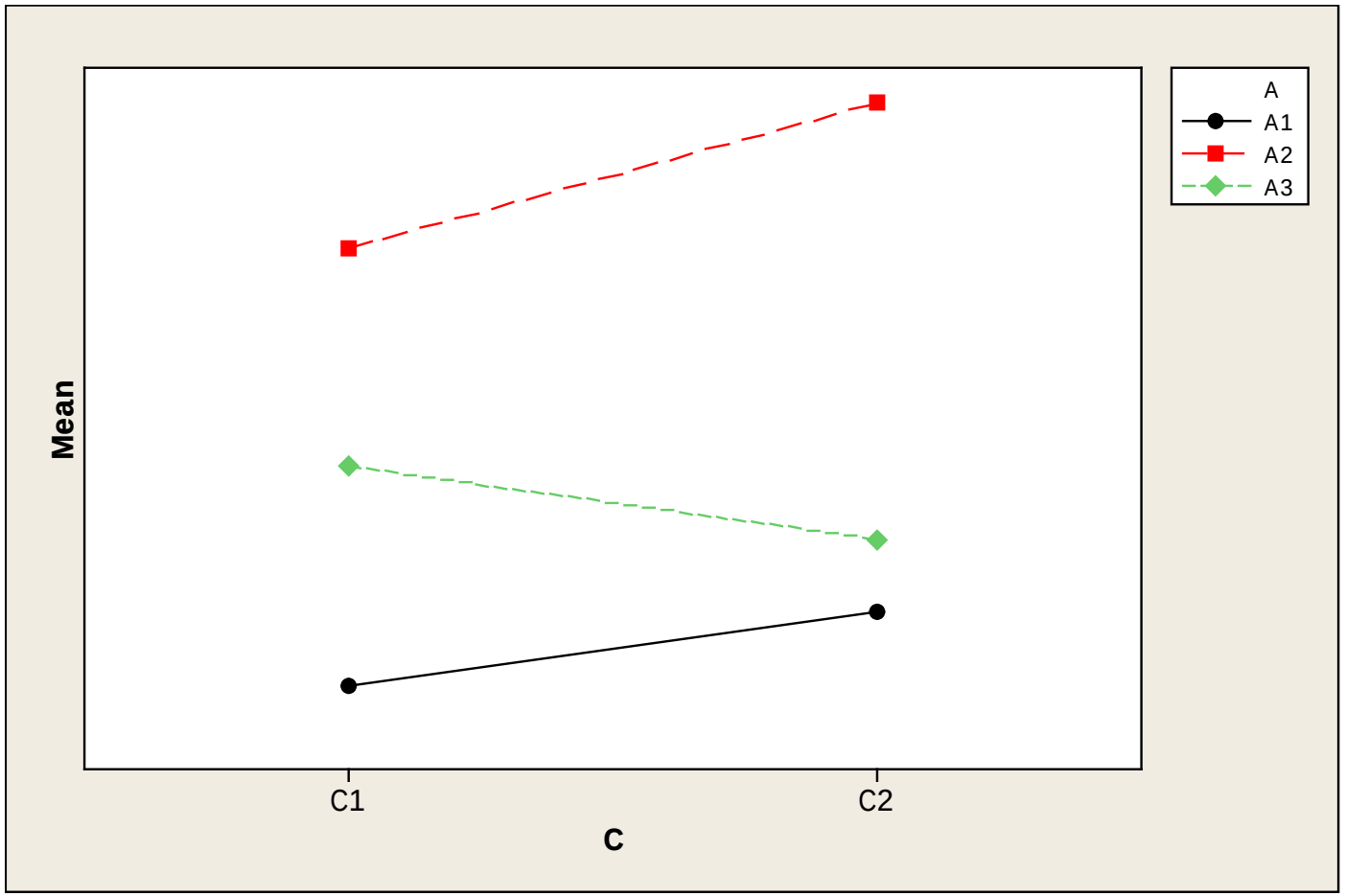
To compare two levels of the split-plot factor for the same level of the main plot factor the effect of the block and the interaction between block and main plot effect cancel out, so then you have no problem.

$$\bar{y}_{i\bullet 1} - \bar{y}_{i\bullet 2} = \gamma_1 + (\alpha\gamma)_{i1} - \gamma_2 + (\alpha\gamma)_{i2} + \bar{\varepsilon}_{i\bullet 1} - \bar{\varepsilon}_{i\bullet 2}$$

On the other hand, if you want to compare different levels of the main plot factor you don't get rid of the interaction between block and main plot effect and therefore get a combination of different error terms with no clearly defined degrees of freedom.

Split-plot designs; illustration of 2)

The comparison within lines is no problem but the problem for points that are on different lines is more complicated!



Split-plot, back to the example (and not in the forest)

Sometimes you can use time as one factor even if it sometimes not is a perfect randomization, but it give you the possibility to handle repeated measurements.

a) If the response is vitamin measured in the cultivar we could look at early and late harvest and by using different parts of the plot you have no repeated measurements.

b) If you look at yield for a perennial cultivar and look at the same plot two years you have a repeated measurement. The covariance between the two repeated measurements is

$$Cov\left(y_{ij1}, y_{ij2}\right) = Var\left(b_j\right) + Var\left(\delta_{ij}\right)$$

In conclusion, the interaction between block and main plot factor is governing the dependence between the repeated measurements.