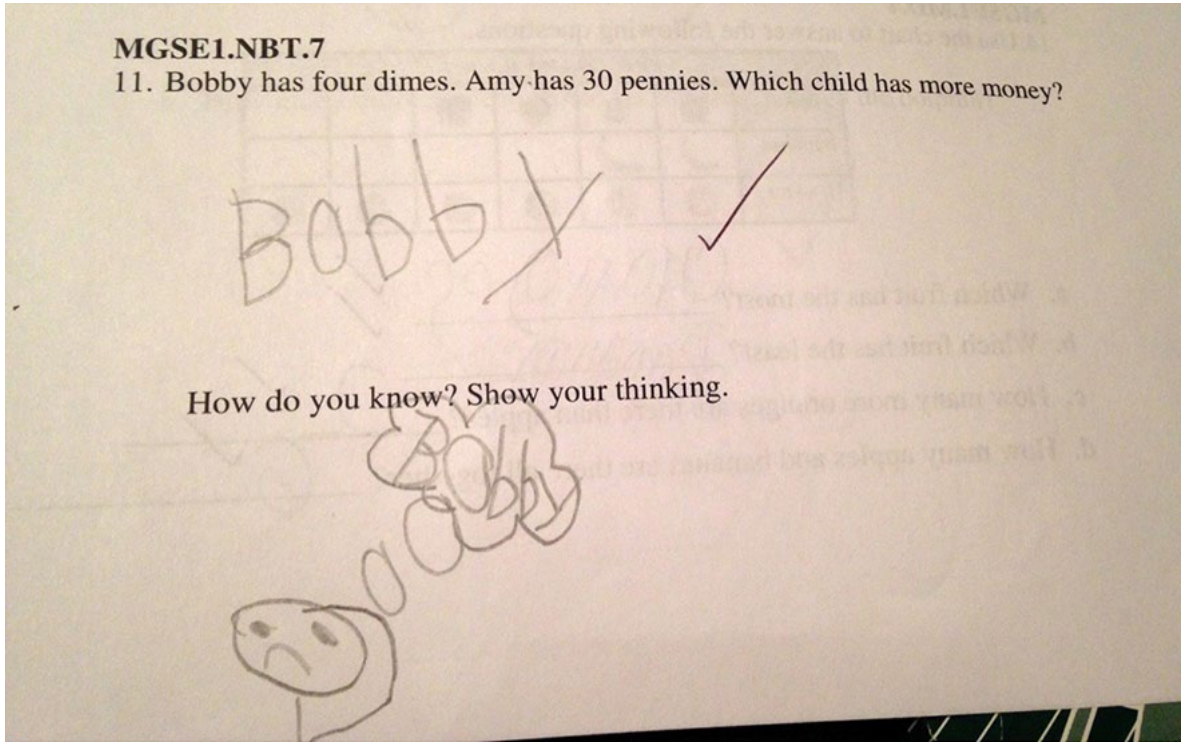
Experimental Design: Stratification

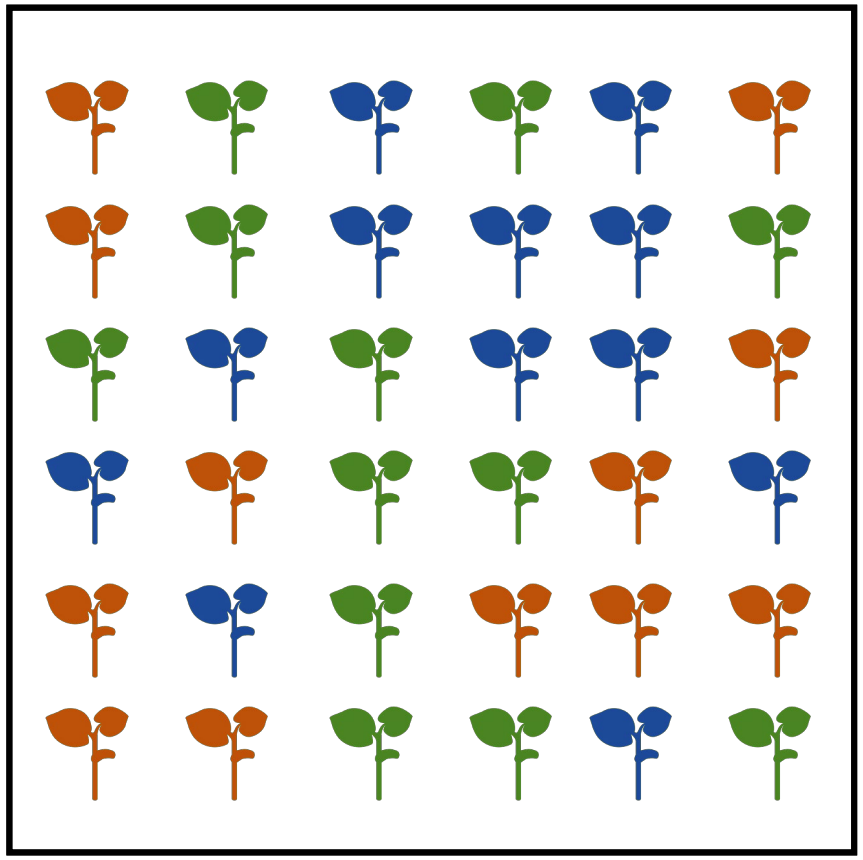
ENTMLGY 6707 Entomological Techniques and Data Analysis



Learning objectives

1. Evaluate pros and cons of stratification in experimental design
2. Compare and contrast experimental designs
3. Understand how components of an ANOVA can be influenced by design choice

Completely randomized design



In a **completely randomized design**, we assigned treatments to experimental units randomly.

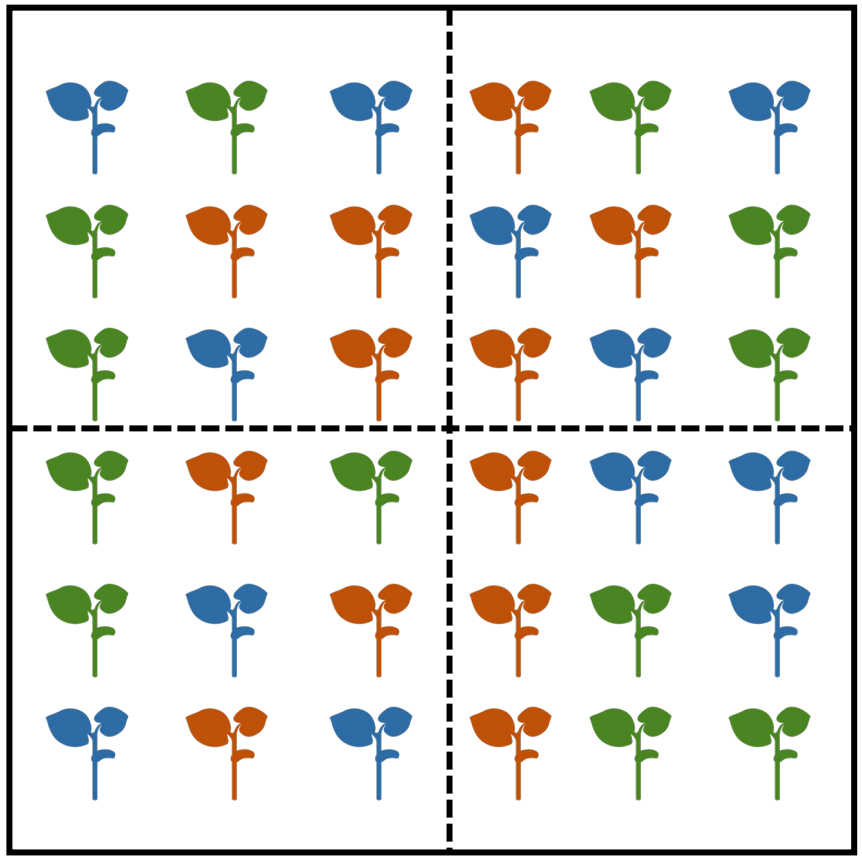
A B C

Sampling universe (e.g., farm)

single plant

Created by Alexandr Lavreniuk from Noun Project

Randomized complete block (RCB)



…but if we expect variation in our response to be influenced by location, time, shipment, etc…

…we can use **blocking** to account for that variation.

Treatments are assigned to blocks, and THEN assigned randomly to experimental units.

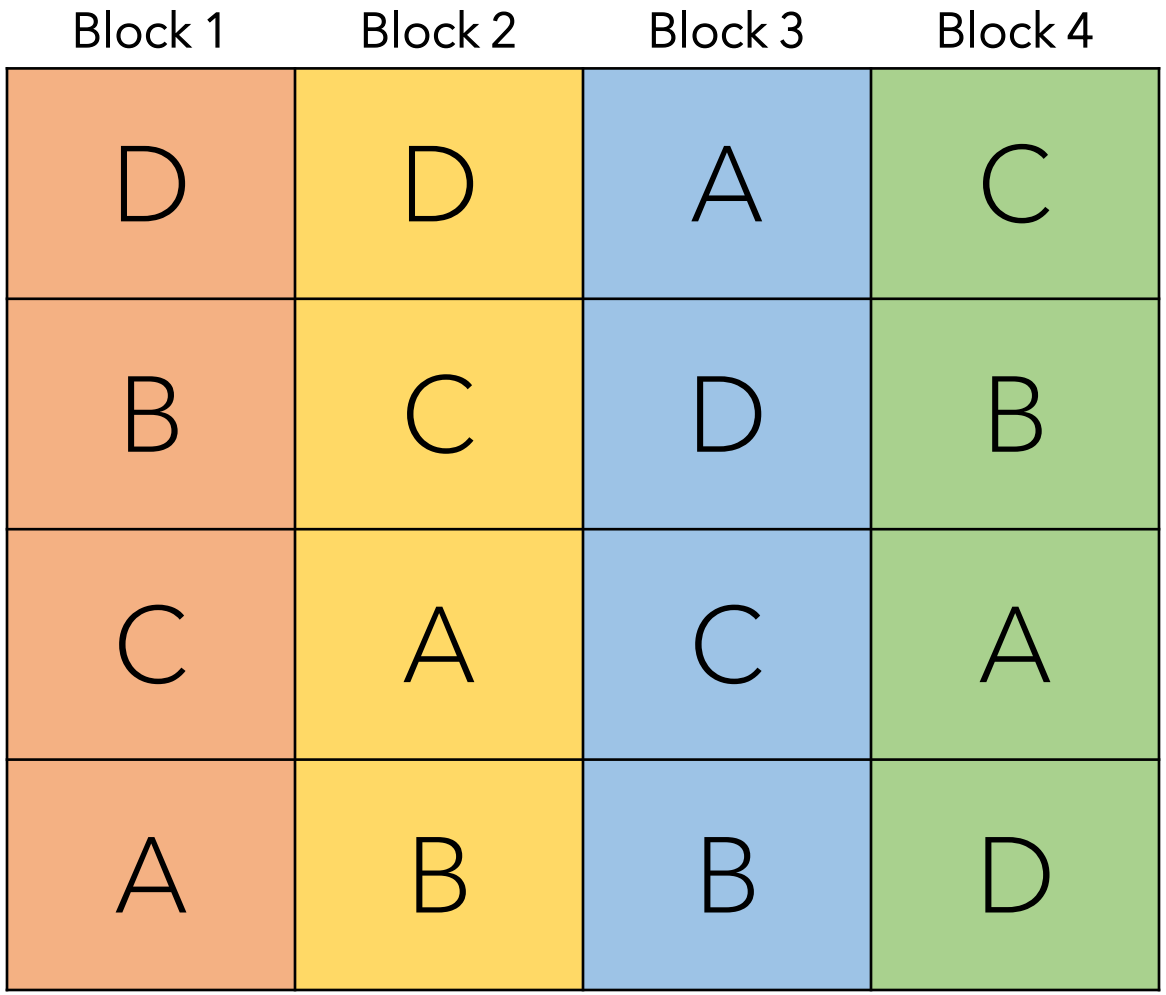
A B C

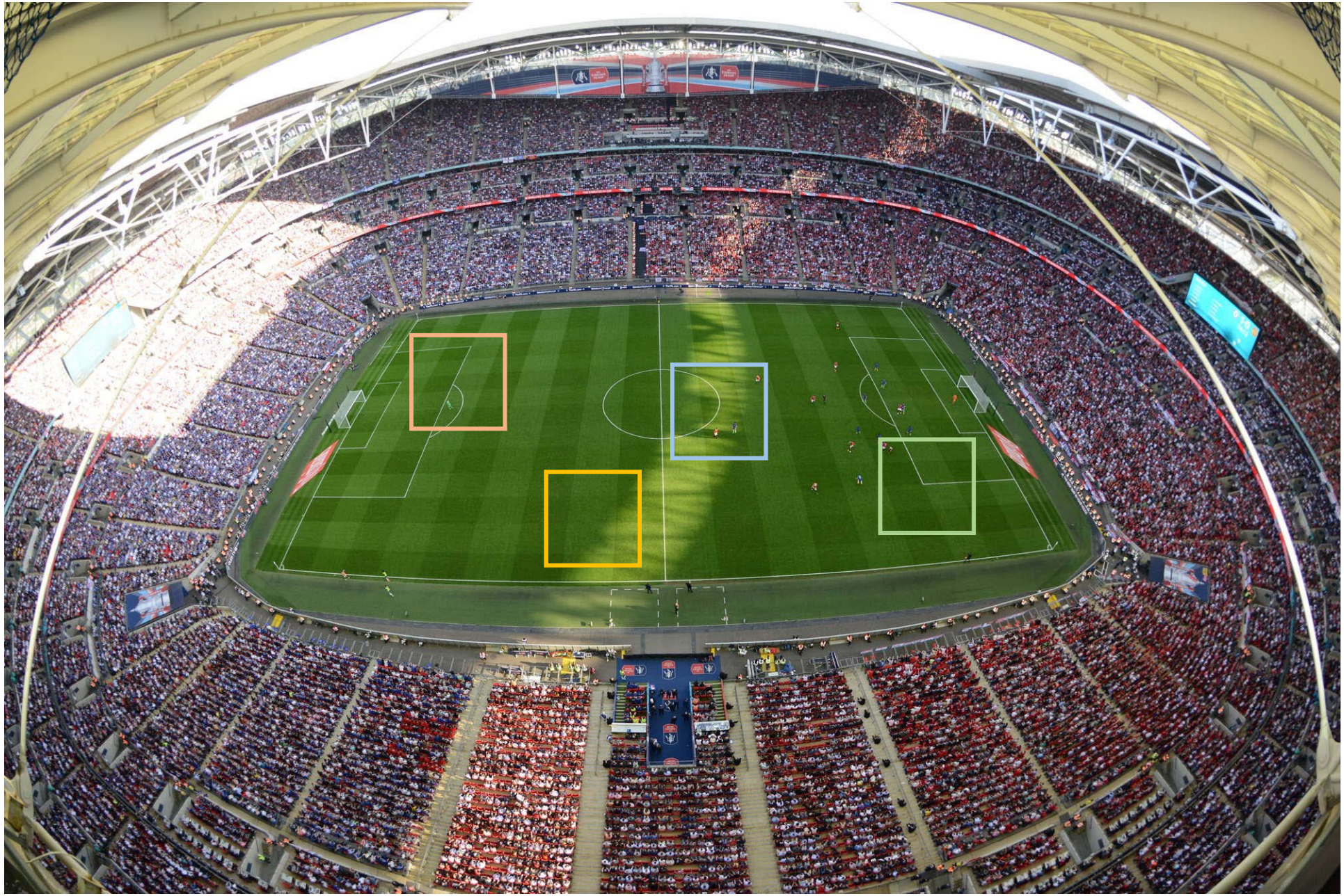
Sampling universe (e.g., farm)

single plant

Created by Alexandr Lavreniuk from Noun Project

RCB simplified





Linear model for RCB

$$Y_{ij} = \mu + b_i + \tau_j + \varepsilon_{ij}$$

Y_{ij} is the value for *j*th observation in the *i*th level of the treatment.
μ is the overall mean (average) across all observations
b_i is the blocking effect, or the difference between the *i*th block and the overall average (μ)
τ_j is the treatment effect, or the difference between the *i*th level of the treatment τ and the overall average (μ)
ε_{ijk} is the error, or the remaining difference between the *k*th observation and the mean of observations in the *i*th block and *j*th treatment level

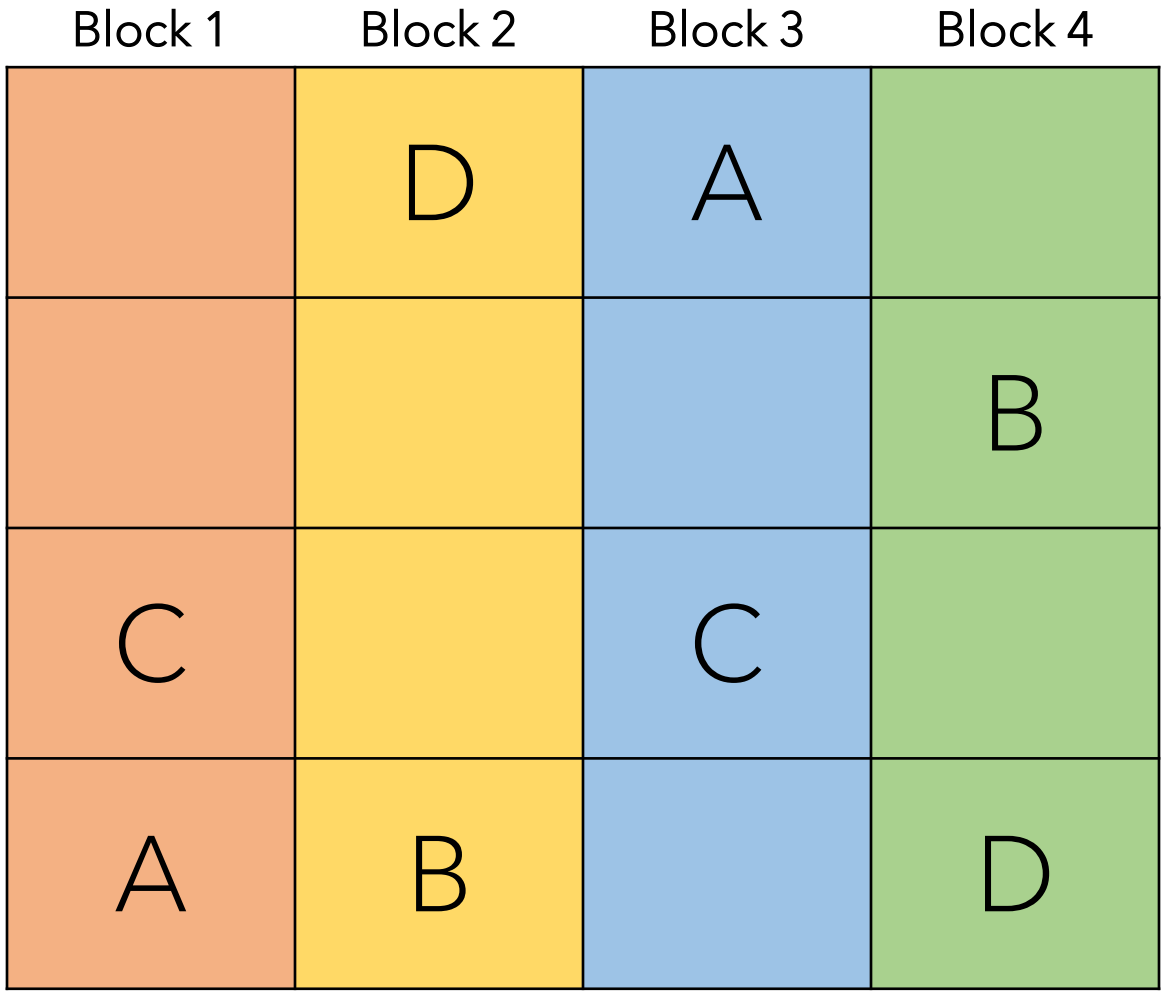
Balanced **incomplete** block design (BIBD)

Not all treatments appear in each block.

$\lambda_{ii'}$=2=number of times treatment *i* and *i'* occur together.

Block 1	Block 2	Block 3	Block 4
D		D	
	C	C	B
C	A	B	A
A	B		D

Unbalanced incomplete block design



Stratification in experimental design

The bottom line: You are trying to increase precision by experimentally and statistically partitioning out variance that is not attributable to your treatments.

Activity (discussion)

Take a moment to think about your (potential) project.

What is the response variable?
What is at least one predictor (treatment)?
Are there any nuisance variables? How will or how did you account for them?

If you don't have a study with response and predictors, try to think of and report a similar study that might fit these criteria.

Randomized complete block

Blocks can be locations in space, points or intervals of time, groups of individuals, etc.

Advantages

- Can increase precision compared with completely randomized design
- No limit to the number of treatments and blocks
- Analysis is straightforward

Disadvantages

- If blocking is not effective, you lose precision AND degrees of freedom
- More treatments → harder to include them all in a block (e.g., more physical space incorporated could lead to more heterogeneity)
- We assume no interaction between blocks and treatments (i.e., treatments are not having different effects on the response variable based on the block in which they occur). If that's happening, inference may be biased.

Activity

Source	df	SS	MS
Treatment	1	78	78
Error	4	74	18.5
Total	5	**152**	-

Pretend we added *x* blocks to the design and got the below ANOVA table.
How many blocks did we add?
Calculate a new *F*-ratio for our treatment by filling in the table.

Source	df	SS	MS
Treatment	1	78	78
Block	**1**		**59**
Error			
Total	5	152	-

ANOVA

Source	df	SS	MS
Treatment	1	78	78
Block	**1**	**59/1**	**59**
Error	**3**	**74-59=15**	**5**
Total	5	152	-

New *F*-ratio:
$F_{1,3}$=78/**5** = 15.60

We added two blocks, which (in a perfect world), accounts for some of that residual variation. The unexplained variation went down, so the denominator in our *F*-ratio went up.

If our blocks do a bad job of soaking up that variation, we are burning degrees of freedom!

Randomized complete block

Account for a source of variation

Latin squares

Account for two sources of variation (= block in two directions)

Split plots

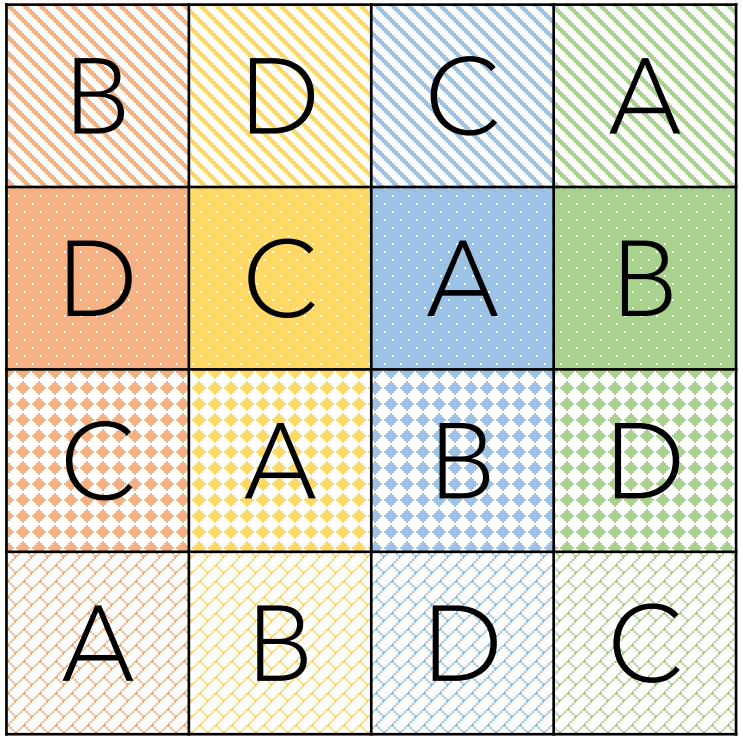
One of your treatments is difficult or expensive to vary. Establish a whole plot and subplot.

Latin square

Randomized complete block

Block 1	Block 2	Block 3	Block 4
B	C	D	A
A	D	C	B
B	D	A	C
C	B	A	D

Latin square
(block in two directions)



Latin square

Advantages:

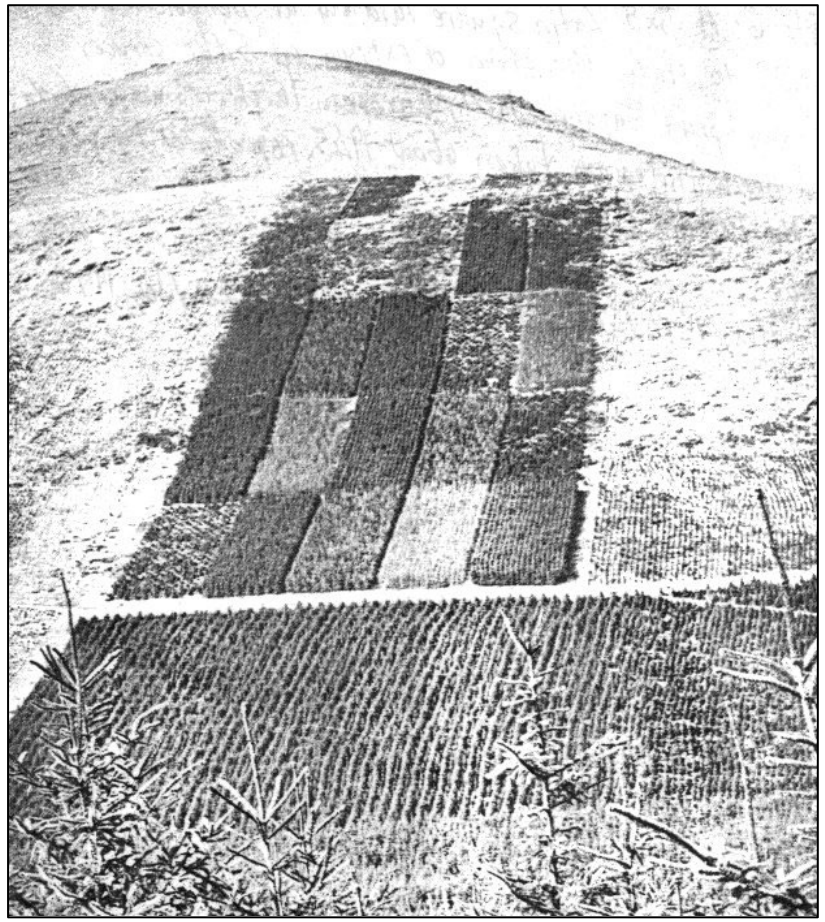
- Account for multiple (two) nuisance factors

Disadvantages:

- The number of treatments needs to equal the number of levels of blocks
- Assumes no interactions between the rows and columns (=blocks) or between the treatment and rows/columns

Latin square

Example: a 5 ×5 forestry experiment in Beddgelert in Wales, to compare varieties of a tree species; designed by Fisher, laid out in 1929, and photographed in about 1945.



1	2	3	4	5
5	3	1	2	4
2	5	4	3	1
4	1	2	5	3
3	4	5	1	2

Linear models

Randomized complete block

$$Y_{ij} = \mu + b_i + \tau_j + \varepsilon_{ij}$$

Y_{ij} is the value for the *i*th block and *j*th level of the treatment
μ is the overall mean
b_i is the effect of block *i*
τ_j is the effect of the *j*th level of treatment τ
ε_{ij} is the remaining error

Latin square

$$Y_{ijk} = \mu + r_i + c_j + \tau_k + \varepsilon_{ijk}$$

Y_{ij} is the value for the *i*th row, jth column, and *k*th level of the treatment
μ is the overall mean
r_i is the blocking effect of row *i*
c_j is the blocking effect of column *j*
τ_k is the effect of the *k*th level of treatment τ
ε_{ij} is the remaining error

Note: I've suppressed the subscript identifying individual observations

Split plot

Use when you are interested in two independent variables, but one of them is challenging to vary (e.g., you are limited by space, time, money, etc.).

1. Assign levels of the "hard-to-vary" factor to entire plots (= **whole plots**)
2. Within each whole plot, randomly assign levels of the other factor to your experimental units (= **subplots**).

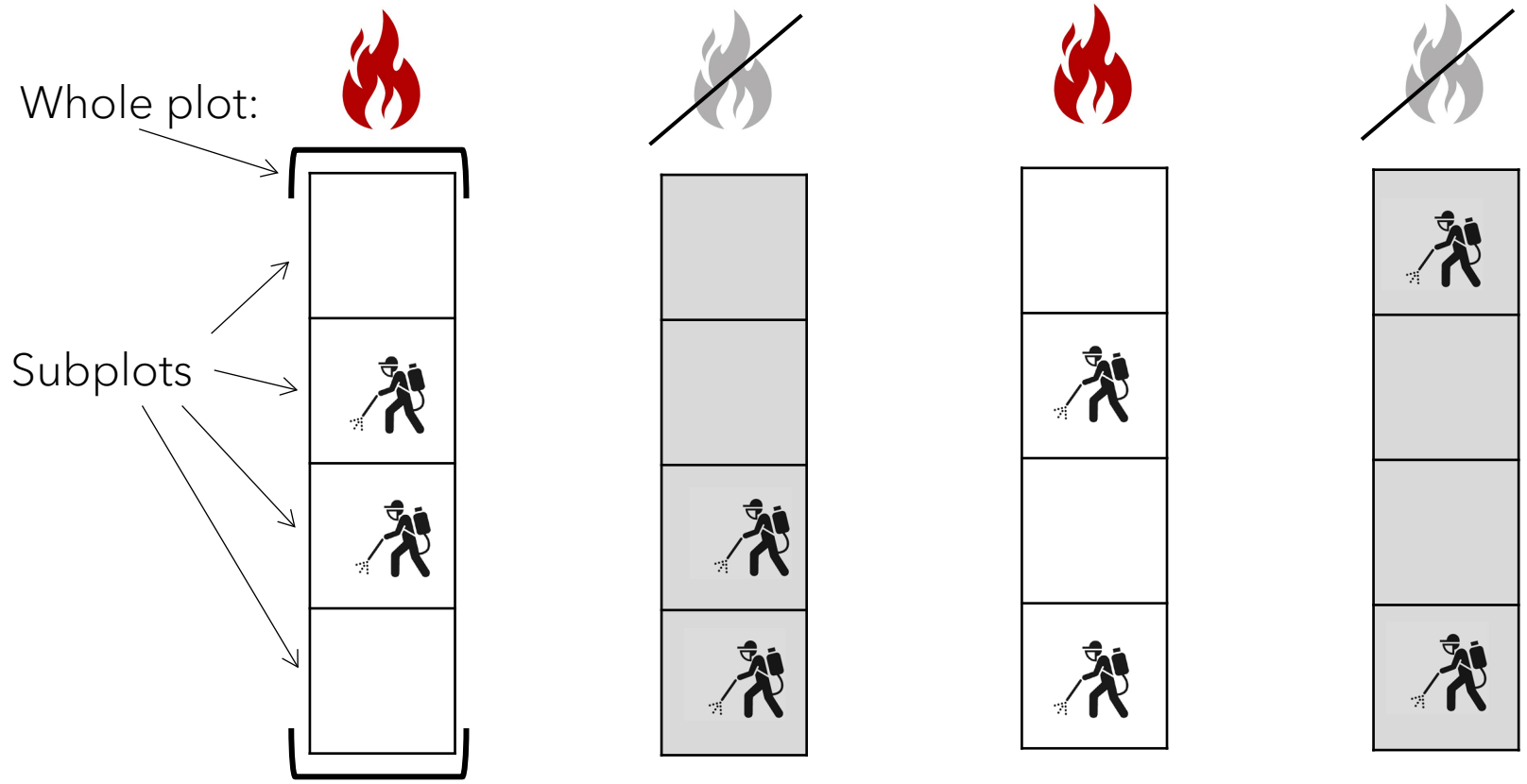
Advantages: Increase in precision on subplot factors when the experiment "works " (= the variation is decreased among subplots grouped into whole plots); cheaper

Disadvantages: Loss in precision of whole plot factor. Difficult to create/design and can be difficult to analyze.

Split plot example

We are interested in the effects of site preparation (presence vs. absence of prescribed fire) and herbicide treatment (none vs. spot treatment) on hardwood regeneration (germinated oak seeds per m^2)

Randomize at each step (including, if need be, by randomized blocking)



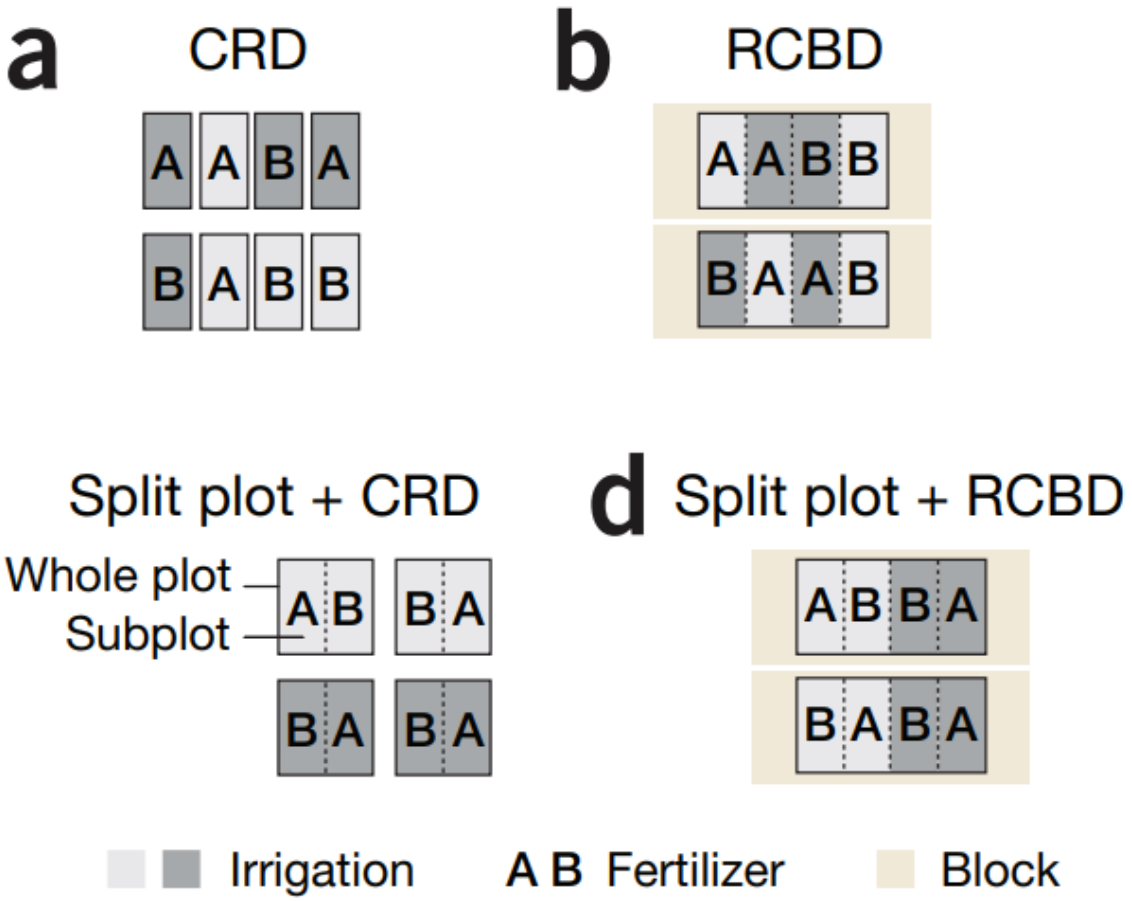
Linear model for split plot

$$Y_{ijk} = \mu + \alpha_i + w_{(i)j} + \beta_k + \alpha\beta_{ik} + \varepsilon_{ijk}$$

Y_{ijk} is the measured response of the *i*th level of the whole plot factor and *k*th level of the subplot factor within the *jth* whole plot
μ is the overall mean
α_i is the effect of the *i*th level of the whole plot factor
$w_{(i)j}$ is whole plot error
β_k is the effect of the *k*th level of the subplot factor
$\alpha\beta_{ik}$ is the whole plot by subplot interaction
ε_{ijk} is remaining error

Note: I've suppressed the subscript identifying individual observations

Split plot - comparisons



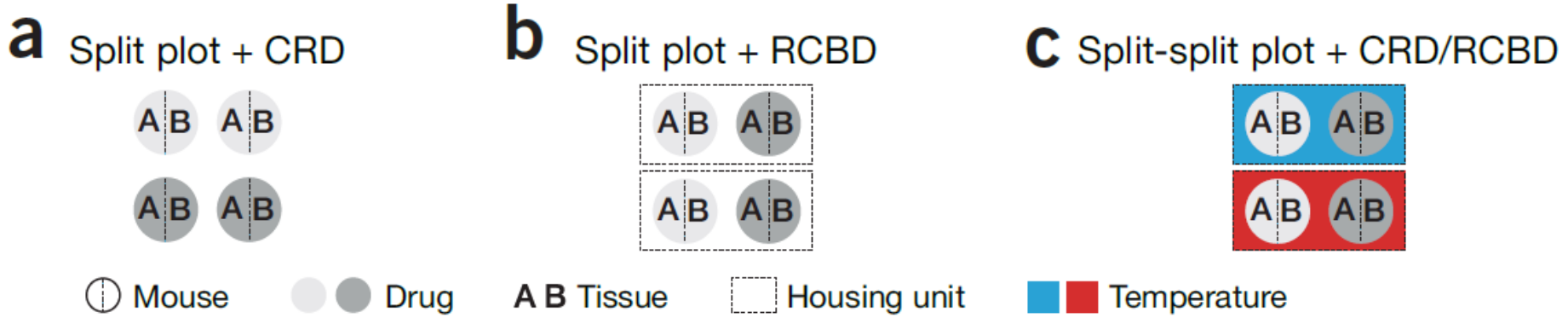
Split plot design

When some factors are harder to vary than others, a split plot design can be efficient.

Naomi Altman & Martin Krzywinski

VOL.12 NO.3 | MARCH 2015 | NATURE METHODS

Split plot



Split plot design

When some factors are harder to vary than others, a split plot design can be efficient.

Naomi Altman & Martin Krzywinski

VOL.12 NO.3 | MARCH 2015 | NATURE METHODS

F-ratios in split plot designs

Table 1 | Split plot ANOVA table for two-factor split plot designs

	CRD			RCBD		
	d.f.	MS	*F*-ratio	d.f.	MS	*F*-ratio
Block, *bl*				n'	MS_{bl}	MS_{bl}/MS_{wp}
A	a'	MS_A	MS_A/MS_{wp}	a'	MS_A	MS_A/MS_{wp}
Error wp	an'	MS_{wp}		$n'a'$	MS_{wp}	
B	b'	MS_B	MS_B/MS_{sp}	b'	MS_B	MS_B/MS_{sp}
$A \times B$	$a'b'$	$MS_{A\times B}$	MS_{AB}/MS_{sp}	$a'b'$	$MS_{A\times B}$	$MS_{A\times B}/MS_{sp}$
Error sp	$ab'n'$	MS_{sp}		$ab'n'$	MS_{sp}	
Total	$abn - 1$			$abn - 1$		

Split plot ANOVA table for two factor split plot designs using CRD (**Fig. 1c**) and RCBD (**Fig. 1d**) with *a* levels of whole plot factor *A* and *b* levels of subplot factor *B*. For CRD *n* is measurements per subplot and for RCBD *n* is number of blocks. Whole plot and subplot errors are indicated by wp and sp subscripts, respectively. For RCBD, interaction between blocking factor *bl* and *B* is usually included in the subplot error term. $a' = a - 1$, $b' = b - 1$, $n' = n - 1$. d.f., degrees of freedom; *F*-ratio, test statistic for *F* test.