

Different ANOVA designs

by Knut Helge Jensen, Dept. of Zoology, UoB

Crossed ANOVA
Nested ANOVA
ANCOVA

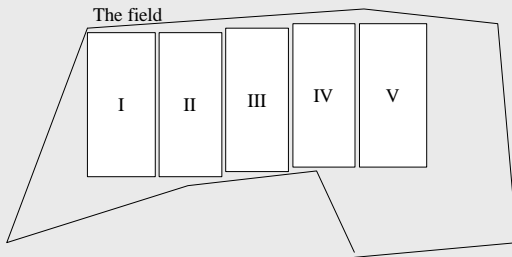
What is important in an experimental design?

- 1) **Randomization** is a warranty against confounding sources of variation (R.A. Fisher 1935, The design of experiments).
- 2) **Replication** supply an estimate of error by which the significance of the comparisons is to be judged, and increase the accuracy of the experimental comparison, (R.A. Fisher 1971, The design of experiments, 8th ed).

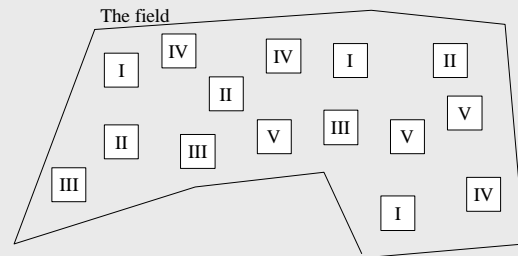
Case study "Whisky"

You want to test the yield of five different sorts of barley to be grown in the vicinity of your own whisky distillery.

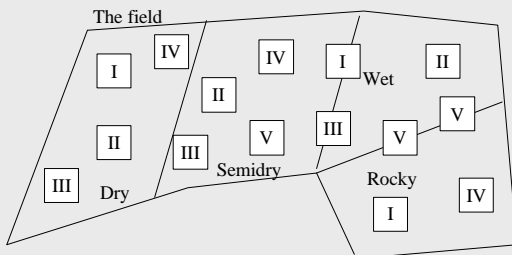
First design - what is wrong?



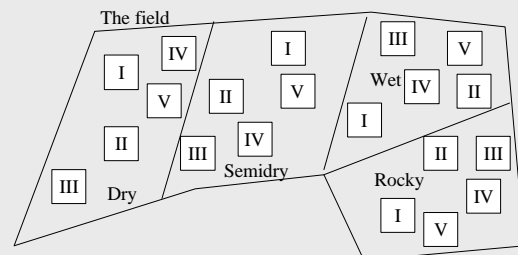
Second design - seems to be OK?



What if we have environmental heterogeneity?



Yes, this is better - environmental heterogeneity are taken into account - but is it still possible to improve the design?



$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$

And this is the final model where we have replicates of each type of barley within each block - but why is these replicates important? - It makes it possible to test for interaction between the factors "block" and "type of barley", i.e. $\alpha\beta_{ij}$.

The field

What is interaction then?

This is an example showing that the difference between varieties of Barley is different depending on Block (Block is a factor including the levels Dry and Wet). In general, if the change in the response mean resulting from a change in levels of factor A, depends on which level of factor B is applied, then there is an interaction between A and B.

It does not make sense to test main effects if you have an interaction between "Barley" and "Block" in the case study shown above. This is a general rule for interactions like this.

The option of Type III SS in many statistical packages does not make you able to test for main effects despite of an interaction. In that sense, Type III is a bug, not a feature...

What is the point of blocking?

To reduce the error term of the model

$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$

Does blocking have any costs?

Yes, the error degrees of freedom are reduced

Thus, there is a balance between reduced error term and reduced error degrees of freedom.

The final model in this case study ($Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$), is called a randomized block design with replication.

If blocking is done appropriately and if the block correspond to the various environmental conditions, the β_j term will remove from the error term variability due to environmental heterogeneity. As a consequence, the error term will be reduced, and the design will be more likely to detect a significant treatment effect than the completely randomized design.

R code:

```
lme(yield~variate+block+variate:block, random=-1|block, data=whisky, method="ML")
```

S-Plus code:

```
lme(yield~variate+block+variate:block, random=-1, cluster=block, data=whisky, method="ML")
```

yield	barley	block
0.204	i	a
0.198	i	a
0.258	i	a
0.185	i	b
0.146	i	b
0.219	i	b
0.232	i	c
0.131	i	c
0.209	i	c
0.214	i	d
0.19	i	d
0.186	i	d
0.327	ii	a
0.31	ii	a
0.353	ii	a
0.39	ii	b
etc.	etc.	etc.

$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$

This is a mixed model ANOVA

- Why?

How do we determine if a factor is fixed or random?

Two questions to be asked:

1) Does the factor represent a random sample of a population?



2) Is the factor controlled by the experimenter?



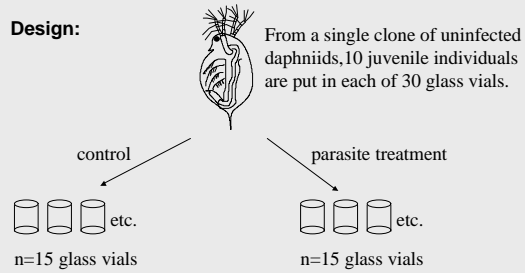
Why determination of fixed and random factor is important: It gives different calculations of the F-value:

Hypothesized effect	Model I (both factor A and B are fixed)	Model II (both factor A and B are random)	Model III (factor A is fixed and B is random)	Model III without interaction (factor A is fixed and B is random)	Model III with nesting (factor B is random and nested under the fixed factor A)
Factor A	$\frac{\text{factor A MS}}{\text{error MS}}$	$\frac{\text{factor A MS}}{A \times B \text{ MS}}$	$\frac{\text{factor A MS}}{A \times B \text{ MS}}$	$\frac{\text{factor A MS}}{\text{error MS}}$	$\frac{\text{factor A MS}}{(B \text{ nested under A}) \text{ MS}}$
Factor B	$\frac{\text{factor B MS}}{\text{error MS}}$	$\frac{\text{factor B MS}}{A \times B \text{ MS}}$	$\frac{\text{factor B MS}}{\text{error MS}}$	$\frac{\text{factor B MS}}{\text{error MS}}$	$\frac{\text{factor B MS}}{\text{error MS}}$
A × B interaction	$\frac{A \times B \text{ MS}}{\text{error MS}}$	$\frac{A \times B \text{ MS}}{\text{error MS}}$	$\frac{A \times B \text{ MS}}{\text{error MS}}$	-	-

Case study "Daphnia"

Size at first reproduction in *Daphnia magna* depending on parasitism by the intracellular gut parasite *Glugoides intestinalis*.

Design:



What kind of design is this, nested or crossed* ANOVA?

*the same as two-way ANOVA

Size	Treatment	Glass vial
2.03	parasite	1
1.74	parasite	1
2.14	parasite	1
etc.	etc.	etc.
2.23	control	2
2.32	control	2
2.49	control	2
etc.	etc.	etc.

R code:

```
lme(size~parasite, random=-1|vial/parasite, data=daphnia, method="ML")
```

S-Plus code:

```
lme(size~parasite, random=-1, cluster=vial/parasite, data=daphnia, method="ML")
```

What is the difference between a nested and a crossed ANOVA design?

Case study "Daphnia"

Size	Treatment	Glass vial
2.03	parasite	1
1.74	parasite	1
2.14	parasite	1
etc.	etc.	etc.
2.23	control	2
2.32	control	2
2.49	control	2
etc.	etc.	etc.

Nested design

All levels of factor "Treatment" does not take all levels of factor "Glass vial".

Case study "Whisky"

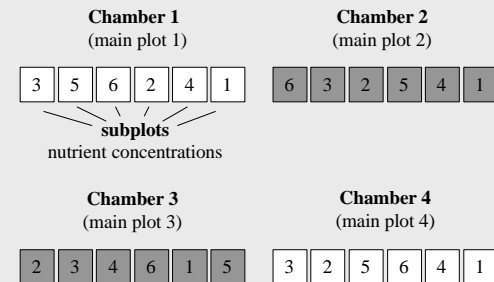
yield	barley	block
0.204	i	a
0.198	i	a
0.185	i	b
0.146	i	b
0.232	i	c
0.131	i	c
0.214	i	d
0.19	i	d
0.327	ii	a
0.31	ii	a
0.39	ii	b
etc.	etc.	etc.

Crossed design

All levels of factor "barley" take all levels of factor "block".

Case study "Tomatoes"

Growth of tomatoes depending on six nutrient levels and two CO2 concentrations.



Split-plot design

$$Y_{k(ij)} = \mu + \alpha_i + \epsilon_{k(i)} + v_j + \alpha v_{ij} + \epsilon'_{k(ij)}$$

α_i is the effect of the i th level of CO2 concentration (the between chamber factor), $\epsilon_{k(i)}$ is the main plot error term and designates the effect of chamber k within level α_i , v_j is the effect of the j th nutrient concentration (the within chamber factor), αv_{ij} is the effect of the interaction between the i th CO2 concentration and the j th nutrient level, and $\epsilon'_{k(ij)}$ is the error term associated with the subplot.

R code:

```
lme(biomass~co2+nutrients+co2:nutrients, random=~1|co2/chamber,
data=tomatoes, method="ML")
```

S-Plus code:

```
lme(biomass~co2+nutrients+co2:nutrients, random=~1,
cluster=co2/chamber, data=tomatoes, method="ML")
```

Case study "Biostat"

30 students are randomly divided in three groups of 10 students. All students are given the same lectures in biostatistics, but with a different teacher for each group. Your aim is to find the best teacher among the three. You use the test scores from the exam of the course to determine this. A confounding factor is of course the abilities of each student which may differ among the three groups of students. However, you are able to control for this factor by obtaining an aptitude test score for each student before the experiment is started. This test score is developed by the Head of Department and reflects general mathematical skills and learning abilities.



Since we have been able to measure the skills of each student before the experiment is started
- what kind of test is appropriate?



ANCOVA (Partial regression)

Exam score	Aptitude	Teacher
15	29	1
19	49	1
21	48	1
27	35	1
35	53	1
39	47	1
23	46	1
38	74	1
33	72	1
50	67	1
20	22	2
34	24	2
28	49	2
etc.	etc.	etc.

Assumptions:

- 1) the regression coefficient β_1 is the same for all treatment groups
- 2) the treatments do not influence the covariate x

Method

Full model (Full): $Y_{ij} = \mu + \alpha_i + \beta_1 x + \epsilon_{ij}$

where Y_{ij} is the dependent variable, μ = general mean effect, α_i = i th treatment effect, $\beta_1 x$ = regression coefficient of Y on x , ϵ_{ij} is unexplained variance, $\epsilon_{ij} \sim N(0, \sigma^2)$.

Reduced model without covariate (Remcov): $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

Reduced model without treatment (Remtreat): $Y_{ij} = \mu + \beta_1 x + \epsilon_{ij}$

Source	Calculation
Covariate	Full tested against Remcov
Treatment	Full tested against Remtreat
Error (within)	Full

```
S+ and R code:
full <- aov(examscore~teacher+apituede, data=biostat)
summary(full)
remcov <- aov(examscore~teacher, data=biostat)
remtreat <- aov(examscore~apituede, data=biostat)
anova(full, remcov)
anova(full, remtreat)
```

Testing the assumption of the regression coefficient (slope of the covariate) that has to be similar for all treatment groups:

```
int <- aov(examscore~apituede+teacher+apituede:teacher,
data=biostat)
noint <- aov(examscore~apituede+teacher, data=biostat)
anova(int, noint)
```