BIO4300 and ENV4300

Stats fest 2007

Analysis of variance

murray.logan@sci.monash.edu.au

MONASH University

Single factor ANOVA

- **Aims**
  - Description
    - Investigate differences between population means
  - Explanation
    - How much of the variation in response variable (Y) is explained by differences in the predictor variable (X)

Single factor ANOVA

- **Data**
  - Dependent (response) variable
    - Continuous
    - Normally distributed
  - Independent (predictor) variable
    - Categorical - factor
  - *n* sampling units (replicates) per group (=level)

Single factor ANOVA

4

- **Linear model**

$y = \beta_0 + \beta_1 x + \varepsilon$ ← Regression

$y = \mu + \beta_1 G_1 + \beta_2 G_2 + \cdots + \varepsilon$ ← ANOVA

ANOVA is equivalent to a multiple regression

a)

Y	A
2	G1
3	G1
4	G1
6	G2
7	G2
8	G2
10	G3
11	G3
12	G3

b)

Y	*dummy*$_1$	*dummy*$_2$	*dummy*$_3$
2	1	0	0
3	1	0	0
4	1	0	0
6	0	1	0
7	0	1	0
8	0	1	0
10	0	0	1
11	0	0	1
12	0	0	1

$y = \mu + \alpha_1 + \alpha_2 + \alpha_3 + \varepsilon$ Where α_1 is the effect of group 1, …

$y = \mu + \alpha + \varepsilon$

Single factor ANOVA

5

- **Linear model**

$y = \mu + \alpha_1 + \alpha_2 + \alpha_3 + \varepsilon$

- Over parameterized
  - More parameters than there are groups
  - Cannot include all α_i's
  - Can only include *p*-1 terms in model, plus the overall mean
  - Must redefine the α terms

$y = \mu + \alpha + \varepsilon$

Single factor ANOVA

6

- **Contrasts**
  - redefine the α terms
    - Treatment contrasts (default behaviour of R)
    - Can define own contrasts (Planned comparisons)

$y = \mu + \alpha_2 + \alpha_3 + \varepsilon$

over-parameterized design matrix · contrast matrix · model matrix

$$\begin{array}{cccc} \text{Intercept } (\mu) & \alpha_1\,(G1) & \alpha_2\,(G2) & \alpha_3\,(G3) \end{array} \quad \begin{array}{cc} \alpha_2^* & \alpha_3^* \end{array} \quad \begin{array}{ccc} \text{Intercept} & \alpha_2^* & \alpha_3^* \end{array}$$

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} * \begin{array}{c} G1 \\ G2 \\ G3 \end{array} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Parameter	Estimates	Null hypothesis
Intercept	mean of 'control' group (μ_1)	H_0: $\mu = \mu_1 = 0$
α_2^*	mean of group 2 minus mean of 'control' group ($\mu_2 - \mu_1$)	H_0: $\alpha_2^* = \mu_2 - \mu_1 = 0$
α_3^*	mean of group 3 minus mean of 'control' group ($\mu_3 - \mu_1$)	H_0: $\alpha_3^* = \mu_3 - \mu_1 = 0$
...		

Single factor ANOVA

- **Linear model**

  $y = \beta_0 + \beta_1 x + \varepsilon$ ← Regression

  $y = \mu + \alpha + \varepsilon$ ← ANOVA

- **Reduced model (when H_0 is true, α=0)**

  $y = \beta_0 + \varepsilon$ ← Regression

  $y = \mu + \varepsilon$ ← ANOVA

- **H_0:**
  - Population group means are all equal
    - $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu$
    - $\alpha_1 = \alpha_2 = \ldots = 0$ (Where $\alpha_1 = \mu_1 - \mu$)

Null hypotheses

- $y = \mu + \alpha + \varepsilon \longrightarrow y = \mu + \beta_1 G_1 + \beta_2 G_2 + \cdots + \varepsilon$

- **Null hypotheses (H_0)**
  - Parameter based
    - Population mean = 0 (μ=0)
    - Population slope1 = 0 (β_1=0)
    - Population slope2 = 0 (β_2=0)
    - …
    - Use t-tests
    - But what do these slopes mean in ANOVA??

Null hypotheses

- $y = \beta_0 + \beta_1 x + \varepsilon$

- **Null hypotheses (H_0)**
  - Model based (variance based)

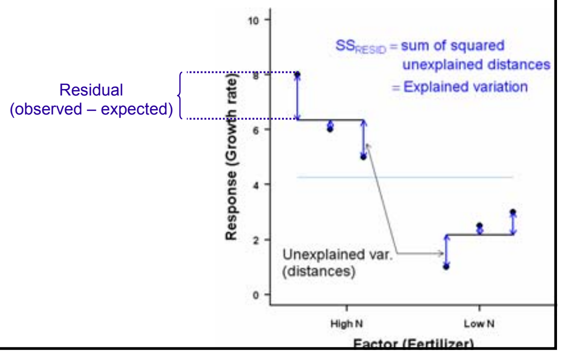
$$\text{Compare fit} \begin{cases} y = \mu + \alpha + \varepsilon \\ y = \mu + \varepsilon \end{cases}$$

- Generate a statistic based on the ratio of fit of the full and reduced models
  - F-ratio

Analysis of variance (ANOVA)

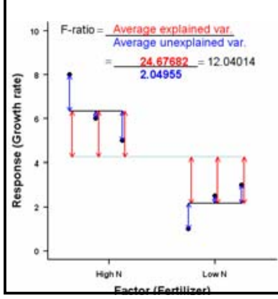
10

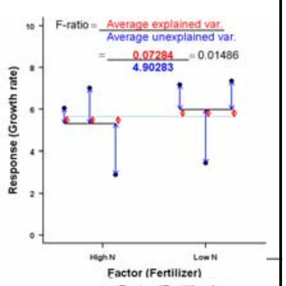
- **Partitioning total variance**



Analysis of variance (ANOVA)

11

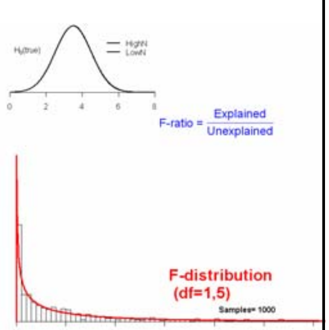




Analysis of variance (ANOVA)

12

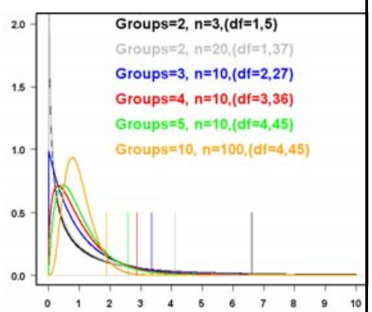
- **F-ratio**
  - Fdistribution.R



Analysis of variance (ANOVA)

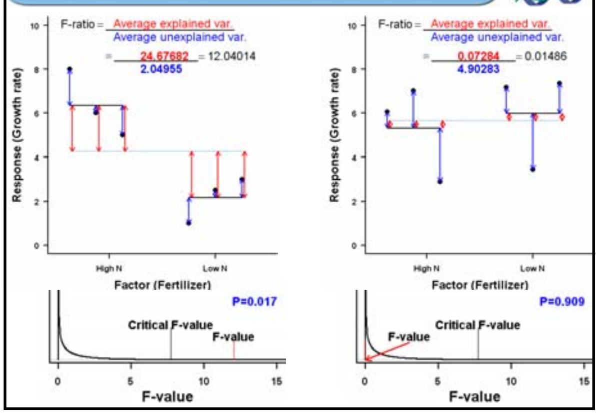
13

- **F-distributions**



Analysis of variance (ANOVA)

14



Analysis of variance (ANOVA)

15

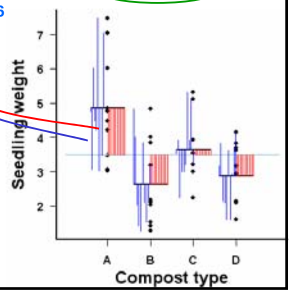
- **Weight of Lobelia seedlings grown in one of four composts (A, B, C & D)**
- **Biological hypothesis: seedlings grow differently in different composts**
  - There is an effect of compost type on the weight of Lobelia seedlings
- **H_0: $\mu_A = \mu_B = \mu_C = \mu_D = \mu$**
  - Population group means are all equal

Analysis of variance (ANOVA)

16

```
Analysis of Variance Table

Response: WEIGHT
          Df  Sum Sq  Mean Sq  F value     Pr(>F)
COMPOST    3  29.480    9.827   6.7498  0.0009943
Residuals 36  52.411    1.456
---
```



- F-ratio = $\frac{\text{Explained}}{\text{Unexplained}}$
- F-ratio = $\frac{9.827}{1.456}$ = 6.749

Single factor ANOVA

17

- There is nothing on this page!
- Well why don't you delete it then?
- Why don't you get stuffed?
- Are you talking to me?
- Don't you look at me!!
- I could have a more intelligent conversation with a thimble full of pond scum!

Single factor ANOVA

18

- Assumptions
  - Independent observations
  - Normality (residuals)
    - Boxplot of response variable
  - Homogeneity of variance (residuals)
    - Spread of observations around regression line
    - Residual plot

```
> boxplot(RESPONSE ~ FACTOR, data=DATA)
```

Single factor ANOVA

19

- Fit linear model

$y = \mu + \alpha + \varepsilon$

```
> *.lm <- lm(RESPONSE ~ FACTOR, data=DATA)
```

```
> *.aov <- aov(RESPONSE ~ FACTOR, data=DATA)
```

$y = \mu + \alpha_2 + \alpha_3 + \alpha_4 + \varepsilon$

Single factor ANOVA

20

- **Final checks (influence measures)**
  - Residual
    - How much each Y value differs from expected
    - Residual plot

```
> plot(*.lm)
OR
> plot(*.aov)
```

Single factor ANOVA

21

- **Analysis sequence**
  - Design experiment/survey
  - Collect data
  - Test assumptions
  - Fit linear model
    - Estimate parameters
    - Full vs reduced
      - Partition variability into explained & unexplained
      - r^2
        - SS_{Groups} / SS_{Total}

Single factor ANOVA

- **Analysis sequence cont.**
  - Test H_0's
    ```
    > summary(*.aov)
    ```
    - $\alpha_1 = \mu = 0$ (t-test)
    - $\alpha_2 - \alpha_1 = 0$ (t-test)
    - ….. (t-test)

    By default
    - Full vs Reduced (explained vs unexplained)
      - $\alpha_1 = \alpha_2 = \ldots = 0$
      - F-ratio statistic = $MS_{Groups}/MS_{Residual}$
      - F-distribution (df=p-1, $p(n$-1))
    ```
    > anova(*.aov)
    ```
  - Conclusions
    - Reject or not reject H_0

Planned comparisons

- **Following a significant ANOVA**
  - Some group comparisons more biologically meaningful than others
    - Control vs Treatment1, etc
- **Planned prior to data analysis!**
- ***p*-1 comparisons maximum**
- **All comparisons must be independent**

Planned comparisons

- **Consider four groups**

$$y = \mu + \alpha_2 + \alpha_3 + \alpha_4 + \varepsilon$$

  - By default, this tests
    - $\alpha_1 = \mu = 0$
    - $\alpha_2 - \alpha_1 = 0$
    - $\alpha_3 - \alpha_1 = 0$
    - $\alpha_4 - \alpha_1 = 0$

    All means are compared to group 1
- **This might not be useful**

Planned comparisons

25

- **Consider four groups**
  $y = \mu + \alpha_2 + \alpha_3 + \alpha_4 + \varepsilon$
- **We can redefine α_2, α_3, α_4**
  - *p-1* alternative comparisons
    - E.g. α_2 : group2 vs group4
    - E.g. α_2 : average of group1 & group2 vs group3
- **Define specific comparisons**
  - Need to specify which groups contribute to each effect (α_2, α_3, α_4)

Planned comparisons

26

- **Consider four groups**
  $y = \mu + \alpha_2 + \alpha_3 + \alpha_4 + \varepsilon$
- **Define specific comparisons**
  - grp1 vs grp2
  - Average gr2&grp3 vs grp4

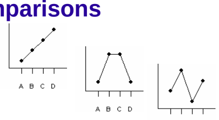
These are contrast coefficients

	Grp1	Grp2	Grp3	Grp4
α_2	1	-1	0	0
α_3	0	0.5	0.5	-1
α_4	?	?	?	?

Planned comparisons

27

- **Consider four groups**
  $y = \mu + \alpha_2 + \alpha_3 + \alpha_4 + \varepsilon$
- **Define specific comparisons**
  - Linear trend
  - Quadratic trend
  - Cubic trend



	Grp1	Grp2	Grp3	Grp4
α_2	-3	-1	1	3
α_3	-1	1	1	-1
α_4	-1	3	-3	1

Planned comparisons

- **Defining contrasts**

```
>contrasts(FACTOR) <- cbind(c(coefs), c(coefs))
```

  - Where coefs are the contrast coefficients

- **Ensure that the contrasts are independent (orthogonal)**

  - `>crossprod(contrasts(partridge$GROUP))`
  - All lower left (triangle) values must be 0

```
     [,1] [,2] [,3] [,4]
[1,]    2    0    0    0
[2,]    0   30    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

```
     [,1] [,2] [,3] [,4]
[1,]    2    0    1    0
[2,]    0   30   -4    0
[3,]    1   -5    4    0
[4,]    0    0    0    1
```

Post-hoc unplanned comparisons

- **Following a significant ANOVA**
  - None of the possible comparisons appear more or less biologically meaningful than others
  - Compare all groups pairwise
  - Lots of comparisons
    - Not independent
- **Control familywise Type I error rate at 0.05**
  - Adjust each comparison

Specific comparisons of means

- **Familywise Type I error rate**
  - Probability of at least one Type I error among all comparisons
  - Increases with increasing number of groups

No. of groups	No. of comparisons	Familywise Type I error probability
3	3	0.14
5	10	0.40
10	45	0.90

Post-hoc unplanned comparisons

- **Bonferoni corrections**
  - Divide p-value by number of comparisons
- **Tukey's test**
  - Adjust degrees of freedom according to the number of comparisons

```
>library(multcomp)
>summary(simtest(DV~FACTOR,data,type="Tukey"))
```

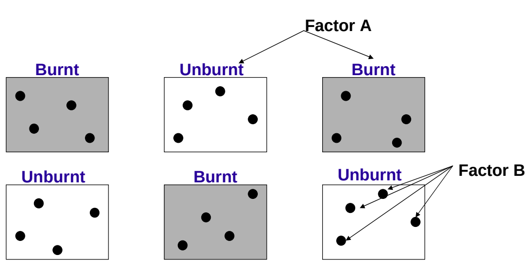
- **More comparisons – more brutal the adjustments**
  - Reduced power

Factors

- **Fixed factors**
  - Only interested in the specific groups (levels) tested
  - Can't generalize to other levels
  - E.g. control/caged
  - Levels selected to represent levels interested in
- **Random factors**
  - Want to extrapolate to all other possible levels
  - E.g. Temperature, Site, pH
  - Levels selected as random representatives

Nested ANOVA

- **Aims**
  - Investigate the effect of a main factor (A)
  - Sub-replicates collected (Factor B) within A
    - Random factor



Nested ANOVA

34

- **Data**
  - Dependent (response variable)
  - Factor A (categorical variable)
    - Main effect
    - Fixed factor
  - Factor B (categorical variable)
    - Subreplicates
    - Random factor
    - Factor B(A)

Nested ANOVA

35

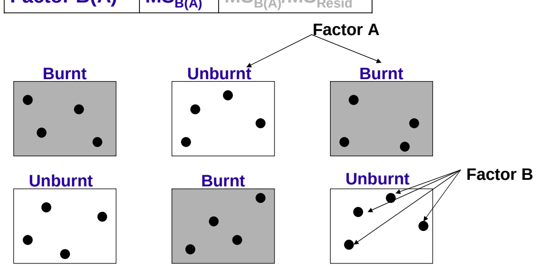
- **Linear model**
  $y = \mu + \alpha + \beta + \varepsilon$
- **Ho:**
  - Populations means of Factor A are equal
    - No effect of Factor A
    - $\mu_1 = \mu_2 = \ldots = \mu$
    - $\alpha_1 = \alpha_2 = \ldots = 0$
  - Populations means of Factor B are equal
    - No variance between levels of Factor B within Factor A
    - Not really interested in testing whether there are differences between sub-replicates!

Nested ANOVA

36

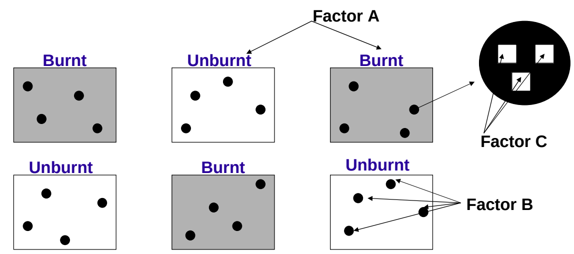
- **ANOVA table**

	MS	F-ratio
Factor A	MS_A	$MS_A/MS_{B(A)}$
Factor B(A)	$MS_{B(A)}$	$MS_{B(A)}/MS_{Resid}$



Nested ANOVA

Factor	MS	F-ratio
A	MS_A	$MS_A/MS_{B(A)}$
B(A)	$MS_{B(A)}$	$MS_{B(A)}/MS_{C(B(A))}$
C(B(A))	$MS_{C(B(A))}$	$MS_{C(B(A))}/MS_{Resid}$



Nested ANOVA

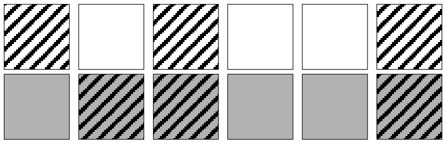
- **Assumptions**
  - As with single factor ANOVA
  - Consider what the replicates are
    - Blocks are the replicates of Factor A
    - For Factor A, the average of Factor B within A are the replicates
      - These need to be used for boxplots

Factorial ANOVA

- **Aims**
  - To investigate the effects of two or more factors (and there interactions) on a response variable
- **Data**
  - Dependent (response variable)
  - Factor A (categorical variable)
    - Fixed or random factor
  - Factor B (categorical variable)
    - Fixed or random factor
- **Every level of Factor A crossed with every level of Factor B**

Factorial ANOVA

40



Factor A – White or Gray

Factor B – Stripes or not stripes

Factorial ANOVA

41

- **Linear model**
  $y = \mu + \alpha + \beta + \alpha\beta + \varepsilon$
- **Ho:**
  - Populations means of Factor A are equal
    - No effect of Factor A
    - $\alpha_1 = \alpha_2 = \ldots = 0$
  - Populations means of Factor B are equal
    - No effect of Factor B
    - $\beta_1 = \beta_2 = \ldots = 0$
  - No interaction between Factor A and Factor B

Factorial ANOVA

42

- **ANOVA**

Factor	MS	F-ratio (both fixed)	F-ratio (A fixed, B random)	F-ratio (both random)
A	MS_A	MS_A/MS_{Resid}	$MS_A/MS_{A:B}$	$MS_A/MS_{A:B}$
B	MS_B	MS_B/MS_{Resid}	MS_B/MS_{Resid}	$MS_B/MS_{A:B}$
A:B	$MS_{A:B}$	$MS_{A:B}/MS_{Resid}$	$MS_{A:B}/MS_{Resid}$	$MS_{A:B}/MS_{Resid}$

- **Interpret interactions first!**
  - Simple main effects

Factorial ANOVA

- **Unbalanced designs (models) - non-orthogonal**
  - Unequal sample sizes
    - Source of unequal variance
    - Complicate partitioning of variance
  - When balanced
    - $SS_{Total}=SS_A+SS_B+SS_{AB}+SS_{Residual}$
  - When unbalanced
    - $SS_{Total}\neq SS_A+SS_B+SS_{AB}+SS_{Residual}$

Factorial ANOVA

- **Type I (regular, sequential) SS**
  - SS = improvement with each successive term

$$y=\mu+\alpha+\beta+\alpha\beta+\varepsilon$$

RSS_{MEAN}	$y=\mu+\varepsilon$	
RSS_A	$y=\mu+\alpha+\varepsilon$	$SS_A = RSS_{MEAN} - RSS_A$
$RSS_{A,B}$	$y=\mu+\alpha+\beta+\varepsilon$	$SS_B = RSS_A - RSS_{A,B}$
$RSS_{A,B,AB}$	$y=\mu+\alpha+\beta+\alpha\beta+\varepsilon$	$SS_{AB} = RSS_{A,B} - RSS_{A,B,AB}$

  - Order important when unbalanced

$$y=\mu+\beta+\alpha+\alpha\beta+\varepsilon$$

Factorial ANOVA

- **Type II (hierarchical) SS**
  - SS = improvement based on a hierarchical model comparison

$$y=\mu+\alpha+\beta+\alpha\beta+\varepsilon$$

Same as Type I !

RSS_A	$y=\mu+\alpha+\varepsilon$
RSS_B	$y=\mu+\beta+\varepsilon$
$RSS_{A,B}$	$y=\mu+\alpha+\beta+\varepsilon$
$RSS_{A,B,AB}$	$y=\mu+\alpha+\beta+\alpha\beta+\varepsilon$

  - Order unimportant

$SS_A = RSS_B - RSS_{A,B}$

$SS_B = RSS_A - RSS_{A,B}$

$SS_{AB} = RSS_{A,B} - RSS_{A,B,AB}$

Factorial ANOVA

46

- **Type III (marginal, orthogonal) SS**
  - SS = improvement based comparison with and without term

$$y = \mu + \alpha + \beta + \alpha\beta + \varepsilon$$

Same as Type I !

$RSS_{B,AB}$ $\quad y = \mu + \beta + \alpha\beta + \varepsilon$

$RSS_{A,AB}$ $\quad y = \mu + \alpha + \alpha\beta + \varepsilon$

$RSS_{A,B}$ $\quad y = \mu + \alpha + \beta + \varepsilon$

$RSS_{A,B,AB}$ $\quad y = \mu + \alpha + \beta + \alpha\beta + \varepsilon$

  - Order unimportant
  - Best choice??

$SS_A = RSS_{B,AB} - RSS_{A,B,AB}$

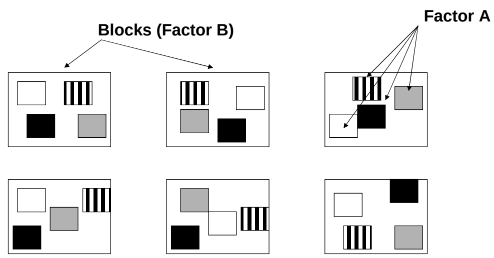
$SS_B = RSS_{A,AB} - RSS_{A,B,AB}$

$SS_{AB} = RSS_{A,B} - RSS_{A,B,AB}$

Randomized block

47

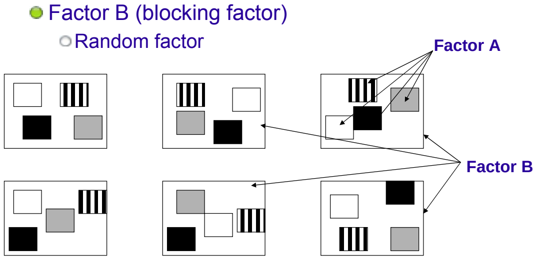
- **Aims**
  - Groups of main treatment blocked together
  - Reduce unexplained variation without increasing sample size



Randomized block

48

- **Data**
  - Dependent variable
  - Factor A – main effect
    - Each level is observed within Factor B
  - Factor B (blocking factor)
    - Random factor



Randomized block

49

- **When the levels of Factor A cannot occur simultaneously – e.g. given at different times**
  - Called repeated measures
  - Eg. Measurements taken every five minutes from the same individual

0, 5, 10, 15, 20, 30, 40	**0, 5, 10, 15, 20, 30, 40**
0, 5, 10, 15, 20, 30, 40	**0, 5, 10, 15, 20, 30, 40**

Randomized block

50

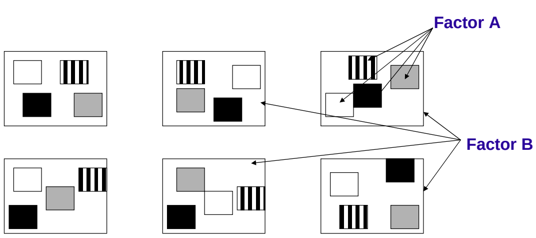
- **Linear model**
  $y = \mu + \beta + \alpha + \varepsilon$
- **Ho:**
  - Populations means of Factor B (blocks) are equal
    - No difference between blocks
    - $\beta_1 = \beta_2 = \ldots = 0$
  - Populations means of Factor A are equal
    - No effect of Factor A
    - $\alpha_1 = \alpha_2 = \ldots = 0$

Randomized block

51

- **ANOVA table**

	MS	F-ratio
Factor B	MS_B	MS_B/MS_{Resid}
Factor A	MS_A	MS_A/MS_{Resid}



Randomized block

52

- **Assumptions**
  - Normality, equal var
  - No block (B) by A interaction
    - Reduces power of test of A
    - Difficult to interpret
    - Diagnose with an interaction plot
  - Sphericity
    - Equal variances of differences between all pairs of treatments
      - var of group1 – group2 = var of group2 – group 3 etc
    - Non sphericity chances of type I errors

Randomized block

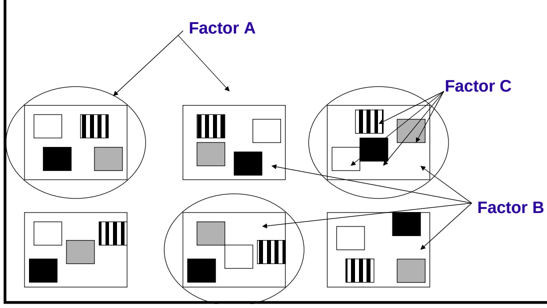
53

- **Sphericity likely to be met when treatments randomly assigned to units within blocks**
  - E.g. treatments randomized within quadrats
  - E.g. order of drugs given to rats random
- **Sphericity not likely to be met when treatments are time**
    - Since cant randomize time
    - Time 1 is likely to be more similar to Time 2 than to Time 10
  - Calculate the degree to which sphericity not met
    - Epsilon (<1)
      - Greenhouse Geisser (preferred)
      - Huyhn-Feldt
    - Adjust degrees of freedom

Split-plot designs

54

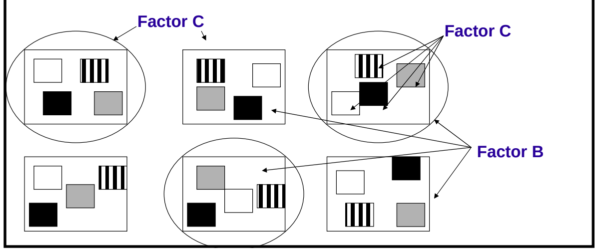
- **Aims**
  - Combination of nested and blocking designs



Split-plot designs

55

- **Data**
  - Factor A (between plot effect)
  - Factor B (plot)
  - Factor C (within plot effect)



Split-plot designs

56

- **Data**
  - Factor A (between plot effect)
    - Main effect 1
  - Factor B (plot)
    - Random factor
  - Factor C (within plot effect)
    - Main effect 2

Nested (Factor A, Factor B)

Blocking (Factor B, Factor C)

Split-plot designs

57

- **ANOVA**

Factor	MS	F-ratio
A	MS_A	MS_A/MS_B
B(A)	$MS_{B(A)}$	$MS_{B(A)}/MS_{Resid}$
C	MS_C	$MS_C/MS_{B(A):C}$
A:C	$MS_{A:C}$	$MS_{A:C}/MS_{B(A):C}$
B(A):C	$MS_{B(A):C}$	$MS_{B(A):C}/MS_{Resid}$

Split-plot designs

- **Assumptions**
  - Normality, equal variance
    - Establish correct residuals
  - Sphericity (for within plot effects)
    - As for blocking designs
  - No interaction between plot and within plot effects