Lecture 7

Multifactor design and analysis

Factorial Design

- A research design that includes two or more factors is called a factorial design
- In an experiment, an independent variable is often called a factor, especially in experiments that include two or more independent variables
- This kind of design is often referred to, by the number of its factors, as a two-factor design or a three-factor design
- A research study with only one independent variable is often called a single-factor design

Factorial Design

- Each factor is usually denoted by a letter (e.g. A, B, C)
- Factorial designs use a notation system that identifies both the number of factors and the number of values or levels that exist for each factor
- e.g. caffeine (3 levels) and alcohol study (2 levels) would be described as 3 x 2 two factor design

Types of factors

- FIXED - all population levels are present in the design (e.g. Gender, treatment condition, ethnicity, size of community, etc.)
- RANDOM - the levels present in the design are a sample of the population to be generalized to (e.g. Classrooms, subjects, teacher, school district, clinic, etc.)

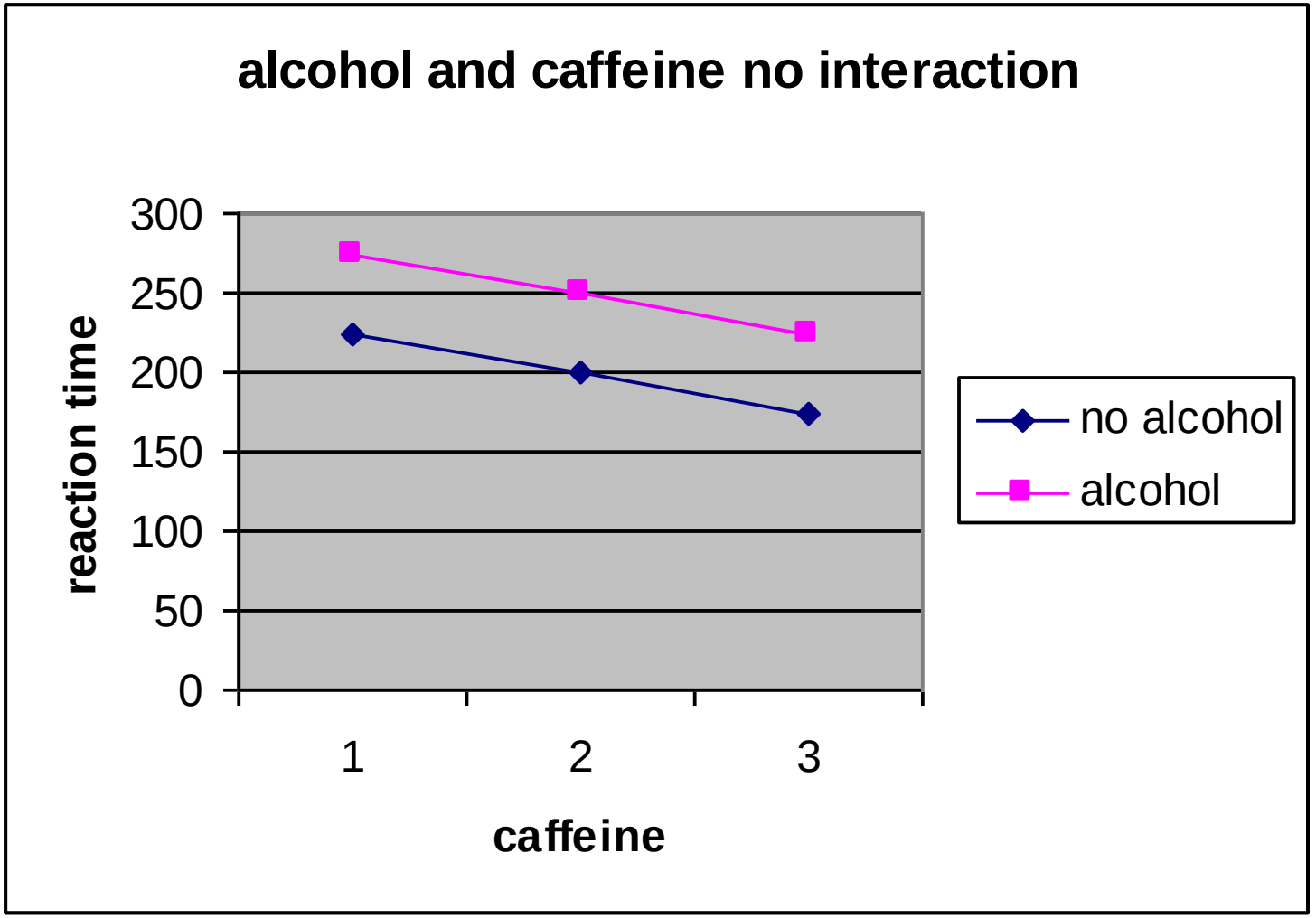
Simple effects, main effects and interactions

- The simple effects of a factor are contrasts between levels of one factor at a single level of another factor.
- The main effects of a factor are contrasts between levels of one factor averaged over all levels of another factor.
- The interaction effect measures differences between the simple effects of one factor at different levels of the other factor.

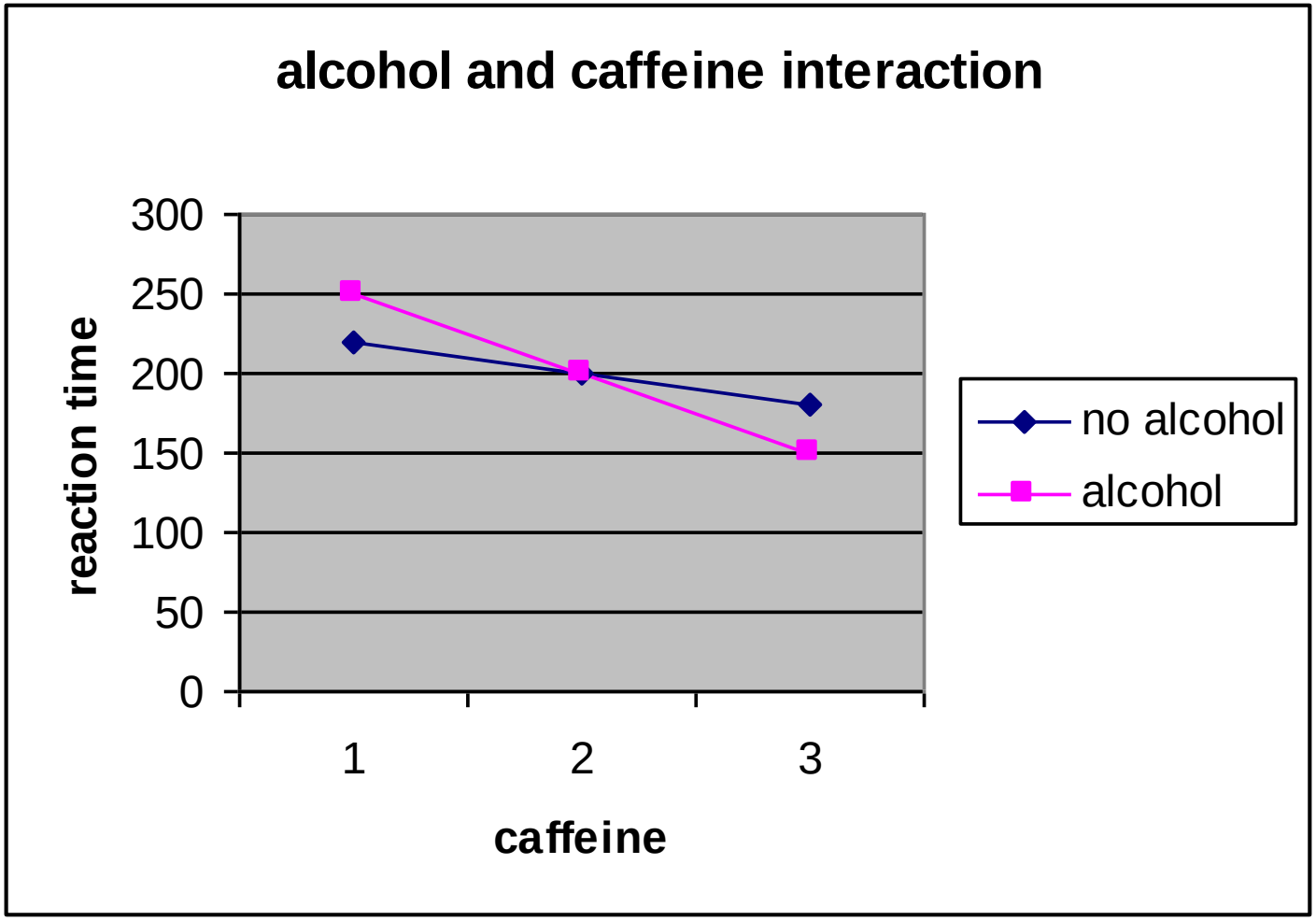
Main effects and interactions

- An interaction between factors occurs whenever the mean differences between individual treatment conditions, or cells, are different from what is predicted from the overall main effect of the factors
- When the effects of one factor depend on the levels of a second factor, then there is an interaction between the factors
- When the results of a two-factor study are graphed, the existence of nonparallel lines (lines that cross or converge) is an indication of an interaction between the two factors

Main effects of caffeine and alcohol: no interaction



Main effects of caffeine and alcohol: with interaction



A simple example

	B1	B2	Row mean	Row effect
A1	1	3	2	-1.25
A2	5	4	4.5	1.25
Column mean	3	3.5	3.25	
Column effect	-0.25	0.25		

- From major effects of factor A, we can see A2 has a higher effect than A1; From major effects of factor B, B2 has a higher effect than B1;
- But, the combination A2B2 does not have highest value. Actually, A2B1 gives the highest performance.

Factorial Designs

- Advantages of factorial designs
  - A greater precision can be obtained in estimating the overall main factor effects.
  - Interaction between different factors can be explored.
  - Additional factors can help to extend validity of conclusions derived.

Examples of models

- single factor:

  $y = \mu + gene + error$
- two factors:

  $y = \mu + treatment + gene + error$
- two factors including interaction term:

  $y = \mu + treatment + gene + treatment.gene + error$
- four factors:

  $y = \mu + treatment + gene + dye + array + error$

Two-way ANOVA

- Allows two different treatments to be examined simultaneously.
- In its simplest form, it is all but identical to 1 way, except that you calculate 2 different treatment sums of squares

Two-way ANOVA

- Get information about the main effect as well as the interaction effect
- Will be computing multiple F-ratios
- Can be both between-subjects, both within subjects, or mixed design
- Each combination of factor A and factor B creates a cell (what we are comparing is the means of each cell)

No interaction

		Factor B				Mean
		B_1	B_2	…	B_b	
Factor A	A_1	y_{11}	y_{12}	…	y_{1b}	$\bar{y}_{1\bullet}$
	A_2	y_{21}	y_{22}	…	y_{2b}	$\bar{y}_{2\bullet}$
	⋮			…		
	A_a	y_{a1}	y_{a2}	…	y_{ab}	$\bar{y}_{a\bullet}$
Mean		$\bar{y}_{\bullet 1}$	$\bar{y}_{\bullet 2}$	…	$\bar{y}_{\bullet b}$	$\bar{y} = \bar{y}_{\bullet\bullet}$

No replications

No replications

- When the data has no replications, we could not estimate interactions.
- Here

$$\bar{y}_{i\bullet} = \frac{1}{b}\sum_{j=1}^{b} y_{ij}, i = 1, 2, \cdots, a$$

$$\bar{y}_{\bullet j} = \frac{1}{a}\sum_{i=1}^{a} y_{ij},\ j = 1, 2, \cdots, b$$

$$\bar{y} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{\mathrm{b}} y_{ij}$$

The linear model

- The formal model underlying 2-Way ANOVA, with 2 treatments A and B
- $\mathbf{y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}}$
- $\mathbf{y_{ijk}}$ is the k^{th} replicate of treatment A level i and treatment B level j
- $\boldsymbol{\alpha_i}$ is the effect of the i^{th} level of treatment A (= difference between μ and mean of all data in this treatment.
- $\boldsymbol{\beta_j}$ is the effect of the j^{th} level of treatment B (= difference between μ and mean of all data in this treatment.

Hypotheses to test

- The hypothesis for rows:
  - H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_i = \ldots = \alpha_a$ (α_i is mean value of the i^{th} level for A - Overall mean value)
  - H_1: α_i (i =1,2, … , a) are not equal
- The hypothesis for columns:
  - H_0: $\beta_1 = \beta_2 = \ldots = \beta_j = \ldots = \beta_b$ (β_j is mean value of the j^{th} level for B - Overall mean value)
  - H_1: β_j (j =1,2, … , b) are not equal

Sums of squares

- Then the sums of squares are

$$SS_{\mathrm{T}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(y_{ij} - \bar{y}\right)^2,\ f_T = ab - 1$$

$$SS_{\mathrm{A}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{y}_{i.} - \bar{y}\right)^2 = b\sum_{i=1}^{a}\left(\bar{y}_{i.} - \bar{y}\right)^2 = b\sum_{i=1}^{a}\alpha_i^2,\ f_A = a - 1$$

$$SS_{\mathrm{B}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{y}_{.j} - \bar{y}\right)^2 = a\sum_{j=1}^{b}\left(\bar{y}_{.j} - \bar{y}\right)^2 = a\sum_{j=1}^{b}\beta_j^2,\ f_B = b - 1$$

$$SS_{\varepsilon} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\mathrm{y}_{ij} - \bar{\mathrm{y}}_{i.} - \bar{\mathrm{y}}_{.j} + \bar{y}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\varepsilon_{ij}^2,\ f_{\varepsilon} = (a-1)(b-1)$$

$$SS_{\mathrm{T}} = SS_{\mathrm{A}} + SS_{\mathrm{B}} + SS_{\varepsilon}$$

Mean squares and test

- For factor A $MS_A = SS_A/(a-1)$
- For factor B $MS_B = SS_B/(b-1)$
- For random errors $MS_\varepsilon = SS_\varepsilon/\left[(a-1)(b-1)\right]$
- Test for significance of A

$$\mathrm{F}_A = \frac{MS_A}{MS_\varepsilon} \sim \mathrm{F}(a-1,(a-1)(b-1))$$

- Test for significance of B

$$\mathrm{F}_\mathrm{B} = \frac{MS_B}{MS_\varepsilon} \sim \mathrm{F}(b-1,(a-1)(b-1))$$

Analysis of variance table for the two factors

Source	Degrees of freedom	Sum of squares	Mean square	F ratio
Factor A	$f_A=a-1$	SS_A	$MS_A=SS_A/(a-1)$	MS_A/MS_ε
Factor B	$f_B=b-1$	SS_B	$MS_B=SS_B/(b-1)$	MS_B/MS_ε
Error	$f_\varepsilon=(a-1)(b-1)$	SS_ε	$MS_\varepsilon=SS_\varepsilon/((a-1)(b-1))$	
Total T	$f_T=ab-1$	SS_T		

Significance test

- Compare the statistic F with the threshold value F_α under given significant level α, then give the decision on H_0
  - F test, significant level=α, corresponding to the threshold F_α
  - If $\boldsymbol{F_A > F_\alpha}$, reject H_0, i.e. the differences between the means of factor A are significant. In other words, Factor A has significant effects on observations.
  - If $\boldsymbol{F_B > F_\alpha}$, reject H_0, i.e. the differences between the means of factor B are significant. In other words, Factor B has significant effects on observations

An example

- Radioactive isotope in milk. Suppose that the concentrations of the radioactive isotope measured in picocuries per liter by three different methods in specimens of milk from four dairies are as follows:

Dairy (A)	Method (B)			$y_{i.}$
	Method 1	Method 2	Method 3	
1	6.4	3.2	6.9	5.5
2	8.5	7.8	10.1	8.8
3	9.3	6.0	9.6	8.3
4	8.8	5.6	8.4	7.6
$y_{.j}$	8.25	5.65	8.75	mean=7.55

Sums of squares

$$SS_{\mathrm{T}} = \sum_{i=1}^{4}\sum_{j=1}^{3}\left(y_{ij} - \bar{y}\right)^2 = \sum_{i=1}^{4}\sum_{j=1}^{3} y_{ij}^2 - 12 \cdot \bar{y}^2 = 43.89,\ f_T = 11$$

$$SS_{\mathrm{A}} = \sum_{i=1}^{4} 3\left(\bar{y}_{\mathrm{i.}} - \bar{y}\right)^2 = 3\sum_{i=1}^{m} \bar{y}_{i.}^2 - 12 \cdot \bar{y}^2 = 18.99,\ f_A = 3$$

$$SS_{\mathrm{B}} = 4\sum_{j=1}^{3} \bar{y}_{.j}^2 - 12 \cdot \bar{y}^2 = 22.16,\ f_B = 2$$

$$SS_{\varepsilon} = SS_T - SS_A - SS_B = 2.74,\ f_{\varepsilon} = 6$$

ANOVA

Source	Degree of freedom	Sum of squares	Mean square	F ratio	Pr>F
Dairy	3	18.99	6.33	13.86	0.004**
Method	2	22.16	11.1	24.26	0.001**
Error	6	2.74	0.46		
Total	11	43.89			

Significance test

- ANOVA tables can have many different treatments included. The skill in ANOVA is not working out the sums of squares, it is the interpretation of ANOVA tables.
- The clues to look for are always in the df column. A treatment with n levels has n-1 df - this always applies and allows you to infer the model a researcher was using to analyze data.

With replications

With replications: $a \times b \times r$

		Factor B			
		B_1	B_2	…	B_b
Factor A	A_1	y_{111}	y_{121}	…	y_{1b1}
		y_{112}	y_{122}	…	y_{1b2}
		⋮	⋮	…	⋮
		y_{11r}	y_{12r}	…	y_{1br}
	A_2	y_{211}	y_{221}	…	y_{2b1}
		y_{212}	y_{222}	…	y_{2b2}
		⋮	⋮	…	⋮
		y_{21r}	y_{22r}	…	y_{2br}
	⋮	⋮	⋮	…	⋮
	A_a	y_{a11}	y_{a21}	…	y_{ab1}
		y_{a12}	y_{a22}	…	y_{ab2}
		⋮	⋮	…	
		y_{a1r}	y_{a2r}	…	y_{abr}

Interaction

- We denote

$$\bar{y}_{ij.} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk} \qquad \bar{y}_{i..} = \frac{1}{br}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}$$

$$\bar{y}_{.j.} = \frac{1}{ar}\sum_{i=1}^{a}\sum_{k=1}^{r} y_{ijk} \qquad \bar{y} = \bar{y}... = \frac{1}{abr}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}$$

The linear model

- The formal model underlying 2-Way ANOVA, with 2 treatments A, B and their interaction
- $\mathbf{y_{ijk} = \mu + \alpha_i + \beta_j + +(\alpha\beta)_{ij} + \varepsilon_{ijk}}$
- $\mathbf{y_{ijk}}$ is the k^{th} replicate of Treatment A level i and treatment B level j
- $\boldsymbol{\alpha_i}$ is the effect of the i^{th} level of treatment A (= difference between μ and mean of all data in this treatment.
- $\boldsymbol{\beta_j}$ is the effect of the j^{th} level of treatment B (= difference between μ and mean of all data in this treatment.
- $\boldsymbol{(\alpha\beta)_{ij}}$ is the interaction effect of the i^{th} level of treatment A and the j^{th} level of treatment B.

Assumptions

- For factor A
- H_{01}: $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$
- For factor B
- H_{02}: $\beta_1 = \beta_2 = \cdots = \beta_b = 0$
- For interactions
- H_{03}: $(\alpha\beta)_{ij} = 0$; for any $i,\ j$

Sums of squares

- Then the sum of squares are

$$SS_T = \sum\sum\sum (y_{ijk} - \bar{y})^2,\ f_T = abr - 1$$

$$SS_A = br\sum (\bar{y}_{i..} - \bar{y})^2 = br\sum \alpha_i^2,\ f_A = a - 1$$

$$SS_B = ar\sum (\bar{y}_{.j.} - \bar{y})^2 = ar\sum \beta_j^2,\ f_B = b - 1$$

$$SS_{AB} = r\sum_{i=1}^{a}\sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y})^2 = r\sum_{i=1}^{a}\sum_{j=1}^{b} (\alpha\beta)_{ij}^2,\ f_{AB} = (a-1)(b-1)$$

$$SS_\varepsilon = \sum\sum\sum (y_{ijk} - \bar{y}_{ij.})^2 = \sum\sum\sum \varepsilon_{ijk}^2,\ f_e = ab(r-1)$$

- $SS_T = SS_A + SS_B + SS_{AB} + SS_\varepsilon$

Mean squares

- For factor A

$$MS_A = SS_A / (a - 1)$$

- For factor B

$$MS_B = SS_B / (b - 1)$$

- For interaction

$$MS_{AB} = SS_{AB} / \left[(a - 1)(b - 1) \right]$$

- For random errors

$$MS_\varepsilon = SS_\varepsilon / \left[ab(r - 1) \right]$$

Significance test

- Test for significance of A

$$F_A = \frac{MS_A}{MS_\varepsilon} \sim F[a-1, ab(r-1)]$$

- Test for significance of B

$$F_B = \frac{MS_B}{MS_\varepsilon} \sim F[b-1, ab(r\text{-}1)]$$

- Test for significance of interaction

$$F_{AB} = \frac{MS_{AB}}{MS_\varepsilon} \sim F[(a-1)(b-1), ab(r\text{-}1)]$$

The analysis of variance table for the two factors

Source	Sum of squares	Degrees of freedom	Mean square	F ratio
Factor A	SS_A	$f_A=a-1$	$MS_A=SS_A/(a-1)$	MS_A/MS_ε
Factor B	SS_B	$f_B=b-1$	$MS_B=SS_B/(b-1)$	MS_B/MS_ε
Interaction AB	SS_{AB}	$f_{AB}=(a-1)(b-1)$	$MS_{AB}=SS_{AB}/[(a-1)(b-1)]$	MS_{AB}/MS_ε
Error	SS_ε	$f_\varepsilon=ab(r-1)$	$MS_\varepsilon=SS_\varepsilon/[ab(r-1)]$	
Total T	SS_T	$f_T=abr-1$		

For different models

- We know that
- $\mathbf{y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}}$
- For fixed-effect model,

$$\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = 0, \sum_{i=1}^{a}(\alpha\beta)_{ij} = \sum_{j=1}^{b}(\alpha\beta)_{ij} = 0, \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$$

- For random-effect model,

$$\alpha_i \sim N(0, \sigma_{\mathrm{A}}^2), \beta_j \sim N(0, \sigma_{\mathrm{B}}^2), (\alpha\beta)_{ij} \sim N(0, \sigma_{\mathrm{AB}}^2), \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$$

Expected mean squares

Source	Degrees of freedom	Mean square	Expected MS	
			Fixed model	Mixed model
Factor A	$f_A = a-1$	$MS_A = SS_A/(a-1)$	$br\sigma_A^2 + \sigma_\varepsilon^2$	$br\sigma_A^2 + r\sigma_{AB}^2 + \sigma_\varepsilon^2$
Factor B	$f_B = b-1$	$MS_B = SS_B/(b-1)$	$ar\sigma_B^2 + \sigma_\varepsilon^2$	$ar\sigma_B^2 + r\sigma_{AB}^2 + \sigma_\varepsilon^2$
Interaction AB	$f_{AB} = (a-1)(b-1)$	$MS_{AB} = SS_{AB}/[(a-1)(b-1)]$	$r\sigma_{AB}^2 + \sigma_\varepsilon^2$	$r\sigma_{AB}^2 + \sigma_\varepsilon^2$
Error	$f_\varepsilon = ab(r-1)$	$MS_\varepsilon = Ss_\varepsilon/[ab(r-1)]$	σ_ε^2	σ_ε^2
Total T	$f_T = abr-1$			

Why do we need to know EMS?

To estimate the variance components!

For fixed-effect models

$$\sigma_{\mathrm{A}}^2 = \frac{1}{br}(\mathrm{MS}_{\mathrm{A}} - \mathrm{MS}_{\varepsilon})$$

$$\sigma_{\mathrm{B}}^2 = \frac{1}{ar}(\mathrm{MS}_{\mathrm{B}} - \mathrm{MS}_{\varepsilon})$$

$$\sigma_{\mathrm{AB}}^2 = \frac{1}{r}(\mathrm{MS}_{\mathrm{AB}} - \mathrm{MS}_{\varepsilon})$$

$$\sigma_{\varepsilon}^2 = \mathrm{MS}_{\varepsilon}$$

To estimate the variance components!

For random-effect models

$$\sigma_{\mathrm{A}}^2 = \frac{1}{br}(\mathrm{MS}_{\mathrm{A}} - \mathrm{MS}_{\mathrm{AB}})$$

$$\sigma_{\mathrm{B}}^2 = \frac{1}{ar}(\mathrm{MS}_{\mathrm{B}} - \mathrm{MS}_{\mathrm{AB}})$$

$$\sigma_{\mathrm{AB}}^2 = \frac{1}{r}(\mathrm{MS}_{\mathrm{AB}} - \mathrm{MS}_{\varepsilon})$$

$$\sigma_{\varepsilon}^2 = \mathrm{MS}_{\varepsilon}$$

Interpreting the interaction term

- The hardest part of 2 way ANOVA is trying to explain what a significant interaction term means, in terms that make sense to most people! Formally it is easy; you are testing H0: MS for interaction term is same population as MS for error.
- It means that you can't reliably predict the effect of Treatment A at level a with B at level *b*, knowing only the effect of A_a and B_b on their own.

Standard errors for means

- For factor A, $s_{\bar{y}_{i..}} = \sqrt{\frac{MS_\varepsilon}{rb}}$
- For factor B, $s_{\bar{y}_{.j.}} = \sqrt{\frac{MS_\varepsilon}{ra}}$
- For cell means ($\bar{y}_{ij.}$), $s_{\bar{y}_{ij.}} = \sqrt{\frac{MS_\varepsilon}{r}}$

Interval estimates for means

- The student t with *ab*(*r*-1) degrees of freedom is required for interval estimates of the cell means. The interval estimate for a cell mean is

$$\bar{y}_{ij.} \pm t_{0.025,\, ab(r-1)} \left(s_{\bar{y}_{ij.}} \right)$$

- When the significance level is 0.05
- Also called 95% confidence interval

An example

- Tensile strength (psi) of asphaltic concrete specimens for two aggregate types with each of four compaction methods

Aggregate type (A)	Compaction method (B)			
		Kneading		
	Static	Regular	Low	Very low
Basalt	68	126	93	56
	63	128	101	59
	65	133	98	57
Silicious	71	107	63	40
	66	110	60	41
	66	116	59	44

Cell means and means for two factors

Aggregate type	Compaction method				Aggregate means ($\bar{y}_{i..}$)
		Kneading			
	Static	Regular	Low	Very low	
Basalt	65.3	129.0	97.3	57.3	87.3
Silicious	67.7	111.0	60.7	41.7	70.3
Compaction means ($\bar{y}_{.j.}$)	66.5	120.0	79.0	49.5	$\bar{y} = 78.8$

Sums of squares

$$SS_T = \sum\sum\sum (y_{ijk} - \bar{y})^2 = \sum\sum\sum y_{ijk}^2 - 2 \cdot 4 \cdot 3 \cdot \bar{y}^2 = 19274.50,\ f_T = 23$$

$$SS_A = br \sum (\bar{y}_{i..} - \bar{y})^2 = 4 \cdot 3 \cdot \sum \bar{y}_{i..}^2 - 2 \cdot 4 \cdot 3 \cdot \bar{y}^2 = 1734.00,\ f_A = 1$$

$$SS_B = ar \sum (\bar{y}_{.j.} - \bar{y})^2 = 2 \cdot 3 \cdot \sum \bar{y}_{.j.}^2 - 2 \cdot 4 \cdot 3 \cdot \bar{y}^2 = 16243.50,\ f_B = 3$$

$$SS_{AB} = r \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y})^2 = 1145.00,\ \ f_{AB} = 3$$

$$SS_\varepsilon = SS_T - SS_A - SS_B - SS_{AB} = 152.00,\ \ f_\varepsilon = 16$$

ANOVA

Source	Degree of freedom	Sum of squares	Mean square	F ratio	Pr>F
Aggregate	1	1734.00	1734.00	182.53	0.000**
Compaction	3	16243.50	5414.50	569.95	0.000**
Interaction	3	1145.00	381.67	40.18	0.000**
Error	16	152.00	9.50		
Total	23	19274.50			

Estimation of variance components and their contribution

Source	Variance	Proportion (%)
Aggregate	143.71	12.20
Compaction	900.83	76.47
Interaction	124.06	10.53
Error	9.50	0.81
Total	1178.10	

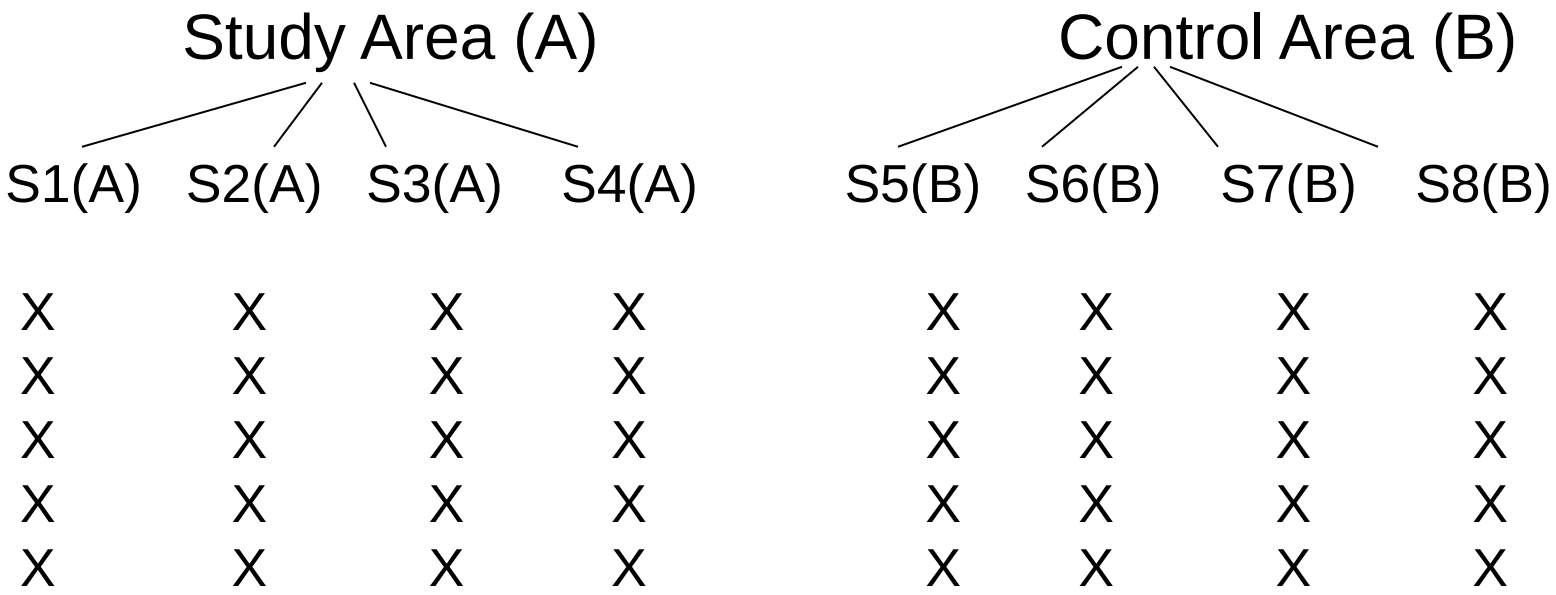
Nested Experiments

Nested Experiments

- In some two-factor experiments the level of one factor , say B, is not “cross” or “cross classified” with the other factor, say A, but is “NESTED” with it.
- The levels of B are different for different levels of A.
- For example: 2 Areas (Study vs Control)
- 4 sites per area, each with 5 replicates.
- There is no link from any sites on one area to any sites on another area.

Nested Experiments

- That is, there are 8 sites, not 2.



Number of sites (S)/replications need not be equal with each sites.
Analysis is carried out using a nested ANOVA not a two-way ANOVA.

Nested Experiments

- A Nested design is not the same as a two-way ANOVA which is represented by:
- | | A1 | A2 | A3 |
- | B1 | X X X X X | X X X X X | X X X X X |
- | B2 | X X X X X | X X X X X | X X X X X |
- | B3 | X X X X X | X X X X X | X X X X X |
- Nested, or hierarchical designs are very common in environmental effects monitoring studies. There are several “Study” and several “Control” Areas.

Objective

- The nested design allows us to test two things: (1) difference between “Study” and “Control” areas, and (2) the variability of the sites within areas.
- If we fail to find a significant variability among the sites within areas, then a significant difference between areas would suggest that there is an environmental impact.
- In other words, the variability is due to differences between areas and not to variability among the sites.

Objective

- In this kind of situation, however, it is highly likely that we will find variability among the sites.
- Even if it should be significant, however, we can still test to see whether the difference between the areas is significantly larger than the variability among the sites with areas.

Statistical Model

$$y_{ijk} = \mu + \alpha_i + \beta_{(i)j} + \varepsilon_{(ij)k}$$

- *i* indexes “A” (often called the “major factor”)
- (*i*)*j* indexes “B” within “A” (B is often called the “minor factor”)
- (*ij*)*k* indexes replication
- $i = 1, 2, \ldots, a$
- $j = 1, 2, \ldots, b$
- $k = 1, 2, \ldots, r$

Model (continued)

$$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}) + (\bar{y}_{ij.} - \bar{y}_{i..}) + (y_{ijk} - \bar{y}_{ij.})$$

- and

$$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y})^2 = \sum_i \sum_j \sum_k (y_{i..} - \bar{y})^2 + \sum_i \sum_j \sum_k (y_{ij.} - \bar{y}_{i..})^2 + \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2$$

$$SS_T = \sum\sum\sum (y_{ijk} - \bar{y})^2 = \sum\sum\sum y_{ijk}^2 - abr - \bar{y}^2$$

$$SS_A = br \sum (\bar{y}_{i..} - \bar{y})^2 = br \sum_i \bar{y}_{i..}^2 - abr\bar{y}^2$$

$$SS_{(A)B} = r \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..})^2$$

Model (continue)

- Or,
- $SS_T = SS_A + SS_{(A)B} + SS_\varepsilon$

- Degrees of freedom:
- $f_T = abr - 1$, $f_A = a - 1$, $f_{(A)B} = a(b-1)$, $f_\varepsilon = ab(r-1)$

Example

- *a*=3, *b*=4, *r*=3; 3 Areas, 4 sites within each area, 3 replications per site, total of (*abr* = 36) data points

	M_1				M_2				M_3				**Areas**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
	Sites												
Rep	10	12	8	13	11	13	9	10	13	14	7	10	
	14	8	10	12	14	11	10	9	10	13	9	7	
	9	10	12	11	8	9	8	8	16	12	5	4	
	11	10	10	12	11	11	9	9	13	13	7	7	$\bar{y}_{ij.}$
	10.75				10.0				10.0				$\bar{y}_{i..}$
	10.25												$\bar{y}$

Example (continue)

- $SS_A = 4 \times 3 [10.75^2 + 10.0^2 + 10.0^2] - 4 \times 3 \times 3 \times 10.25^2 = 4.5$
- $SS_{(A)B} = 3 [(11-10.75)^2 + (10-10.75)^2 + (10-10.75)^2 + (12-10.75)^2 + (11-10)^2 + (11-10)^2 + (9-10)^2 + (9-10)^2 + (13-10)^2 + (13-10)^2 + (7-10)^2 + (7-10)^2]$

  $= 3 \times 42.75 = 128.25$
- $SS_T = 10^2 + 14^2 + \ldots + 4^2 - 4 \times 3 \times 3 \times 10.25^2 = 240.75$
- $SS_\varepsilon = 108.0$

ANOVA table for the example

- **Nested ANOVA: Observations versus Area, Sites**

Source	DF	SS	MS	F	P
Area	2	4.50	2.25	0.158	0.856
Sites (A)B	9	128.25	14.25	3.167	0.012**
Error	24	108.00	4.50		
Total	35	240.75			

- What are the “proper” ratios?

$E(MS_A) = \sigma^2 + r\sigma^2_{(A)B} + rb\sigma^2_A$

$= MS_A/MS_{(A)B}$

$E(MS_{(A)B}) = \sigma^2 + r\sigma^2_{(A)B}$

$E(MS_\varepsilon) = \sigma^2$

$= MS_{(A)B}/MS_\varepsilon$

Summary

- Nested designs are very common in environmental monitoring
- It is a refinement of the one-way ANOVA
- All assumptions of ANOVA hold: normality of residuals, constant variance, etc.
- Need to be careful about the proper ratio of the Mean squares.
- Always use graphical methods e.g. boxplots and normal plots as visual aids to aid analysis.

Let’s work on previous data together

- Tensile strength (psi) of asphaltic concrete specimens for two aggregate types with each of four compaction methods

Aggregate type (A)	Compaction method (B)			
		Kneading		
	Static	Regular	Low	Very low
Basalt	68	126	93	56
	63	128	101	59
	65	133	98	57
Silicious	71	107	63	40
	66	110	60	41
	66	116	59	44