Index of Lecture 7

Practical Information

Today's lecture:

- follow-up from lab session?,
- multifactorial designs:
  - 2-way and 3-way ANOVA: balanced (yes/no) & replication (yes/no),
  - main effects and interaction,
- some (new) general principles for analysis.

Guidelines for textbook reading:

- you are *not* required to know how to compute SS-values in ANOVA tables (only DF-values),
- you are *not* required to compute contrasts involving more than one factor; furthermore, polynomial contrasts may often be replaced by regression modelling,
- skip too technical parts[1] of Chapters 8–10, but focus on the models and the ANOVA tables.

Home assignments:

- second assignment for 802 students only due tomorrow,
- first assignment for everyone coming up early next week.

[1] Technical parts: 177_8-179^{15}; 179_9-180_5; $181_{11}-181_4$; $183_{10}-184^{11}$; $184_{16}-185^4$; 192_7-194_1; 205_5-208_6; Figure 8.6; Sections 9.2.3, 9.2.4 and 9.3.

Multi-Factorial Designs

Several factors in the same design?

Yes! – in good designs it is possible to separate effects of different factors from each other ⇒

- possible to study combined effect of several factors (the presence of interaction),
- if interaction is absent: cheaper (less experimental units) than in several one-at-a-time experiments,[2]
- increased scope of the experiment,

and analysing multi-factorial data by each factor separately: is generally wrong and only gives valid results if at most one factor is of importance.

Design terminology and issues:

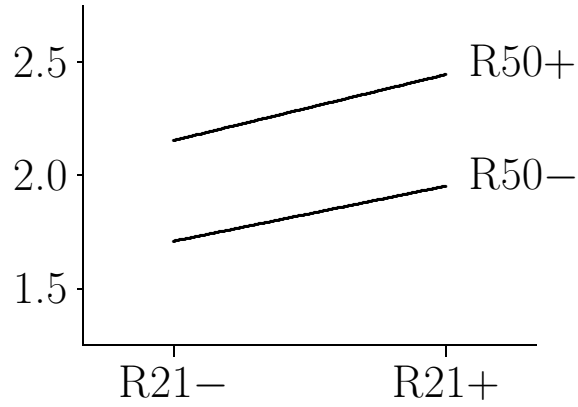
- balancedness: all (combined) groups are equally large, otherwise unbalanced,
- completeness: all (combined) groups are present (no empty cells), otherwise incomplete (*should be avoided*),
- replication: some of (combined) groups have $n > 1$, otherwise no replication (all $n = 1$),
- factorial structure can be combined with blocking structures (next lecture).

[2] Simple example: effects of alcohol (A: 0/1) and sleeping pills (B; 0/1). Two "one-at-a-time" studies (effect of A at fixed level of B; effect of B at fixed level of A), each with 20 subjects, give same precision as a combined study with 5 subjects per alcohol×pill combination (in the absence of interaction).

Decomposing a 2-Way Table of Means I

Example 8.6: Nonstarter bacteria in cheddar cheese:

Total free amino acids		Strain R21 no	yes	Mean
Strain R50	no	1.709	1.952	1.831
	yes	2.153	2.444	2.299
Mean		1.931	2.198	2.065



Different ways to look at the data:

(i) four separate groups,

(ii) two R50 groups for each R21 group,

(iii) two R21 groups for each R50 group,

(iv) (overall level), two R50 groups, two R21 groups, association between R50 and R21 groups.

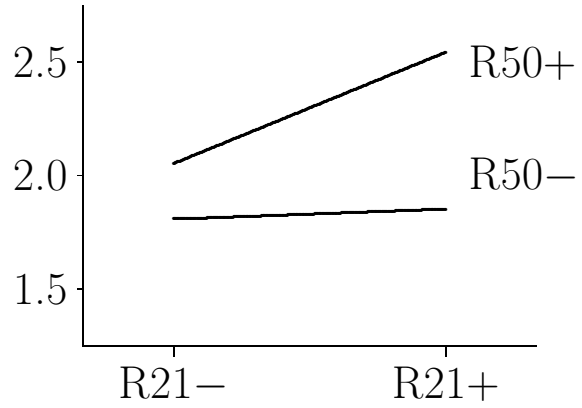
Decomposition of means corresponding to (iv):

$\bar{y}_{..}$	2.065	2.065	−0.134	0.134	$\bar{y}_{.j} - \bar{y}_{..}$
overall mean	2.065	2.065	−0.134	0.134	R21 effect
$\bar{y}_{i.} - \bar{y}_{..}$	−0.234	−0.234	0.012	−0.012	$\bar{y}_{ij} - \bar{y}_{i.}$
R50 effect	0.234	0.234	−0.012	0.012	$-\bar{y}_{.j} + \bar{y}_{..}$

Decomposing a 2-Way Table of Means II

Modified bacteria in cheese data:

Total free amino acids		Strain R21 no	yes	Mean
Strain R50	no	1.809	1.852	1.831
	yes	2.053	2.544	2.299
Mean		1.931	2.198	2.065



Decomposition of means corresponding to (iv):

$\bar{y}_{..}$	2.065	2.065	−0.134	0.134	$\bar{y}_{.j} - \bar{y}_{..}$
overall mean	2.065	2.065	−0.134	0.134	R21 effect
$\bar{y}_{i.} - \bar{y}_{..}$	−0.234	−0.234	0.112	0.112	$\bar{y}_{ij} - \bar{y}_{i.}$
R50 effect	0.234	0.234	−0.112	0.112	$-\bar{y}_{.j} + \bar{y}_{..}$

Comparison of two variants of bacteria in cheese data:

- same overall level, same overall (average) effects of both R50 and R21,
- original data: almost parallel lines ⇒ additive effects, (same effect of one factor at all levels of other factor(s)),
- modified data: non-parallel lines ⇒ non-additive effects, or interaction between the factors R50 and R21.

Interaction and Additivity

Interaction (synergism/antagonism, epistasy, covariation):

- the combined effect of two factors[3] is not predictable from isolated effects of each of them examined separately,
- the effect of one factor depends on the level of the other factor,
- non-parallel lines in plot of means versus one factor,
- no additivity between factors[4],
- note: interaction is dependent on scale (of outcome).

Dealing with interaction between two factors:

- decomposition of SS for combined factor $A \times B$:

  $$\mathrm{SS}_{A\times B} = \mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{A*B},$$

  and each SS corresponds to treatment contrast(s),
- DF formula: $\mathrm{DF}_{A*B} = \mathrm{DF}_A \cdot \mathrm{DF}_B$ (if $A \times B$ complete),
- in presence of an "important" interaction (significant and strong), the main effects are of no direct interest,[5]
- use interaction plot based on means or parameter estimates to understand character of interaction,
- possibly explore interesting contrasts in combined factor.

[3] Interaction between three factors: the interaction between two of the factors depends on the level of the third factor.

[4] i.e., interaction is the opposite of additivity, or additivity means no interaction.

[5] In GO terminology (Section 8.11), the hierarchy is retained by not removing main effects involved in an interaction, so they may not need to be tested at all.

ANOVA table: Bacteria in Cheese Data

Steps of ANOVA analysis:

- 4 treatments $\sim 2 \times 2$ factorial R50 $\times$ R21:

$$\mathrm{SS}_{\mathrm{Trt}} = 3[(1.709-2.065)^2 + \ldots + (2.444-2.065)^2] = 0.8723,$$

- Decomposition of $\mathrm{SS}_{\mathrm{Trt}}$ into orthogonal contrasts:

contrast $w(\{\mu_i\})$	w_1	w_2	w_3	w_4	$\widehat{w}$	$\mathrm{SS}(\widehat{w})$	$F = t_w^2$
main R50	1	1	-1	-1	-0.935	0.656	7.23
main R21	1	-1	1	-1	-0.535	0.214	2.36
interaction	1	-1	-1	1	0.048	0.002	0.02
						$\mathrm{SS}_{\mathrm{Trt}}$	

- ANOVA table:

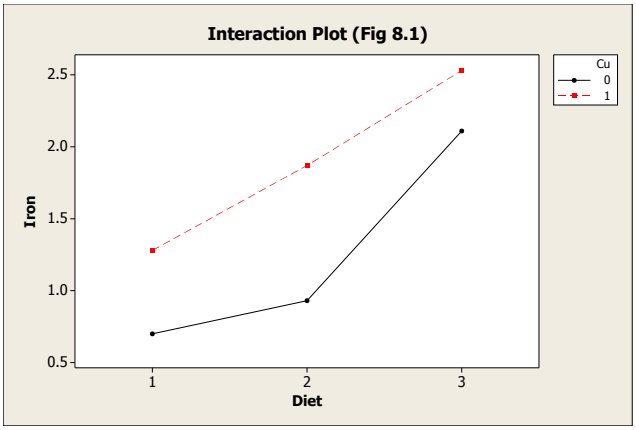
Source	DF	SS	MS	F	P-value
R50	1	0.656	0.656	7.23	0.028
R21	1	0.214	0.214	2.36	0.163
Interaction	1	0.002	0.002	0.02	0.89
Error	8	0.726	0.091		
Total	11	1.598			

- Conclusions:
  - absolutely non-significant interaction between effects of R50 and R21,
  - non-sign. main effect of R21 $\Rightarrow$ no effect of R21 at all,
  - significant, but weak main effect of R50: adding R50 increases TFAA; illustrate by estimates, CI or plots.

Interaction Plot Examples

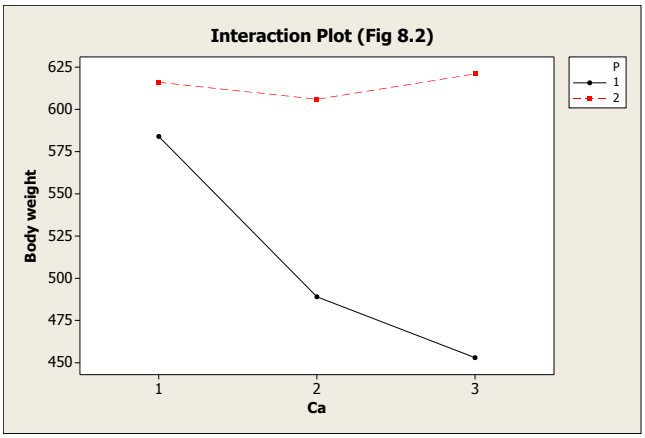
Rat data (Example 8.2):

- outcome: iron levels in liver tissue,
- factors: milk diet (3), copper deficiency (2),
- interpretation: close to additive effects.



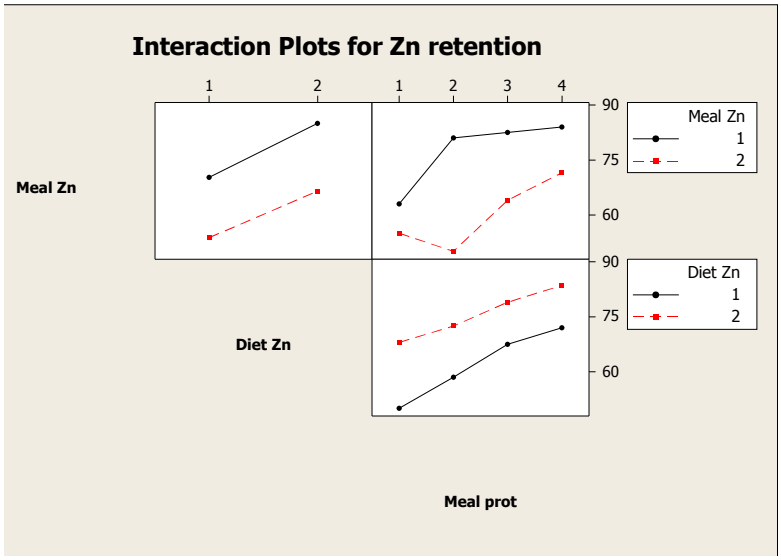
Chick data (Example 8.4):

- outcome: body weights,
- factors: Ca suppl. (3), P suppl. (2)
- interpretation: interaction: Ca effect for low P (P=1) only.



Rat data II (Example 8.5):

- outcome: Zn retention,
- factors: Diet Zn (2), final Meal Zn (2), final Meal protein (4),
- interpretation: interaction Meal Zn $*$ Meal protein; other effects additive.



Specification of Interaction

Assume a row by columns layout of two factors, and let

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

where $i \sim$ rows, $j \sim$ columns, and

- μ_{ij} = mean of (i,j)th group,
- μ = overall or baseline mean (or "intercept"),
- α_i = deviation of ith row group from overall mean,
- β_j = deviation of jth column group from overall mean,
- γ_{ij} (or $(\alpha\beta)_{ij}$, GO and commonly used notation) = *interaction* = deviation of (i,j)th group from *additivity*.

Technical note: some restrictions on α's, β's and γ's needed (otherwise too many parameters):

$$\begin{aligned}
\text{GO \& Minitab} &: \textstyle\sum_i \alpha_i = 0,\ \sum_j \beta_j = 0,\ \sum_i \gamma_{ij} = 0,\ \sum_j \gamma_{ij} = 0,\\
\text{Stata} &: \alpha_1 = 0,\ \beta_1 = 0,\ \gamma_{11}, \ldots, \gamma_{1b} = 0,\ \gamma_{11}, \ldots, \gamma_{a1} = 0,\\
\text{SAS} &: \alpha_a = 0,\ \beta_b = 0,\ \gamma_{a1}, \ldots, \gamma_{ab} = 0,\ \gamma_{1b}, \ldots, \gamma_{ab} = 0.
\end{aligned}$$

Computer software formalism for factors A and B:

- Minitab, SAS:

  $A * B$ = interaction between A and B,

  $A \,|\, B$ = A B $A * B$ (main effects and interaction).

- Stata vers. 12,13: # instead of $*$, and ## instead of $|$.
- R software: : instead of $*$, and $*$ instead of $|$.

Textbook Examples Overview

ANOVA examples of Chapters 8, 9.1-2 and 10.1-2:

- 8.6: non-starter bacteria in cheese (addition of two bacteria): 2×2 factorial with 3 replicates,
- 8.8: page faults (#) in CPU experiment (program algorithms and settings): 2×3×3×3 factorial with no replication (*models without replication*[6]),
- 8.10, 9.3: amylase activity in maize (analysis and growth temp., variety): 2×2×8 factorial with 3 replicates, both temperatures quantitative (*polynomial contrasts*),
- 9.2: unspecified outcome and factors: 2^4 $(2 \times 2 \times 2 \times 2)$ factorial with 2 replicates (*one-cell interaction*[7]),
- 9.4: seed viability (storage cond.): 3×7 factorial with 3 replicates, both factors quantitative (*polynomial contrasts*),
- 10.1-2: amylase activity (8.10) with one observation omitted (*unbalanced data*[8]),
- 10.3: unspecified outcome and factors: 2×2 factorial with highly unequal replication (*unbalanced data*[8]).

[6] Most common method is to omit highest order interaction(s) and thus obtain estimate of residual error: conservative approach if these interactions are non-zero.

[7] An interaction may be due to a single mean value deviating from additivity.

[8] Unbalanced data no longer have orthogonal factorial contrasts ⇒ effects depend on other terms in model (as in general regression); two types of sum of squares:

- partial/adjusted SS: remove term while keep all others in;
- sequential SS: remove terms sequentially (↑) so only terms above are kept in.

3-Way ANOVA with Replication

Amylase activity example: (GO Example 8.10)

- amylase specific activity of sprouted maize under 32 treatment conditions:

  y_{ijkl} = activity for maize plant batch l of type (i, j, k)

  $i = 1, \ldots, 8 \sim$ analysis temperature (10,13,15,20,25,30,35,40°C),

  $j = 1, 2 \sim$ growth temperature (13,25°C),

  $k = 1, 2 \sim$ variety (B73,O43),

  $l = 1, 2, 3 \sim$ replicate,

- completely randomized design (if full randomization),
- full statistical model = 1-way ANOVA with 32 groups,

  $$y_{ijkl} = \mu_{ijk} + \varepsilon_{ijkl},$$

  where the errors $\varepsilon_{1111}, \ldots, \varepsilon_{8223}$ are i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$.
- decomposition of combined factor levels into main effects and interactions (first order and second order),

  $\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$,

  * α_i = main effect of `atemp` (level i),
  * $(\alpha\beta)_{ij}$ = interaction `atemp*gtemp` (levels i and j),
  * $(\alpha\beta\gamma)_{ijk}$ = interaction `atemp*gtemp*var` (levels i, j and k).
- model checks based on full model $\Rightarrow$ same approach as in 1-way ANOVA: logarithmic transform appropriate.

Amylase: Summary of Results

ANOVA table (analysis on natural log scale):

Source	DF	Seq SS	Adj SS	Adj MS	F	P
atemp	7	3.01613	3.01613	0.43088	78.86	0.000
gtemp	1	0.00438	0.00438	0.00438	0.80	0.374
var	1	0.58957	0.58957	0.58957	107.91	0.000
atemp*gtemp	7	0.08106	0.08106	0.01158	2.12	0.054
atemp*var	7	0.02758	0.02758	0.00394	0.72	0.654
gtemp*var	1	0.08599	0.08599	0.08599	15.74	0.000
atemp*gtemp*var	7	0.04764	0.04764	0.00681	1.25	0.292
Error	64	0.34967	0.34967	0.00546		
Total	95	4.20202				

- high $R^2 = 1 - 0.3497/4.2020 = 91.7\%$,
- non-significant terms: `atemp*gtemp*var` and `atemp*var` (but not `gtemp` because involved in a significant interactions), and `atemp*gtemp` is close to signficant,
- final model: `atemp*gtemp gtemp*var`, but no need to refit model (pool variance terms) because DF_E is large,
- additive effects of `atemp` and `var` for given `gtemp`,
- interaction plots: strong, parabolic-type effect of `atemp` (no obvious interpretation of `gtemp` interaction), and different effects of `gtemp` for the varieties,
- presentation of `gtemp*var` results by means with SE:

var=B73		var=O43		
gtemp=13	gtemp=25	gtemp=13	gtemp=25	SE
5.85	5.92	5.75	5.70	0.015

- `atemp` effects: multiple comp. or polynomial modelling.

Amylase: Further Analyses

Polynomial modelling (GO Example 9.3):

- `atemp` contrasts not so attractive due to non-equidistant temperatures,
- for simplicty, refit with (clearly) non-significant `atemp` terms omitted before polynomial modelling,
- quadratic model in `atemp` has lack-of-fit test:

$$\begin{aligned} F &= [(0.52148 - 0.42489)/(88 - 78)]/0.00545 = 1.71 \\ &\sim F(10, 78) \text{ under } \mathrm{H}_0,\ P = 0.094, \end{aligned}$$

  — no formal evidence against quadratic model, but higher order terms may improve fit,
- cubic model in `atemp` has lack-of-fit test:

$$\begin{aligned} F &= [(0.45037 - 0.42489)/(86 - 78)]/0.00545 = 0.58 \\ &\sim F(8, 78) \text{ under } \mathrm{H}_0,\ P > 0.5, \end{aligned}$$

  — cubic polynomial model seems appropriate.
- interpretation of fitted model: plots of predicted curves.

Illustration: effect of unbalancedness (GO Examples 10.1-2): dropping one observation (first in dataset):

- sequential and partial/adjusted (SAS type I and III) sum of squares don't coincide: SS values for `gtemp` range within 0.00140—0.00330 across different models,
- model building must be sequential (as in regression),
- simple means are no longer best estimates (next slide).

Tools for Factor Effects

Least squares means = tool for interpret. factor effects:

- in balanced or other "nice" designs: just simple means,
- generally: factor level estimates, while keeping all other predictors *at their average*:
  - continuous predictors at their sample mean,
  - categorical predictors by averaging across all levels,
- interpretation: estimated level for "average subject" wrt. all other predictors (but often not a "real" subject),
- usually better than simple means because factor levels are compared "all other things being equal"; exception is with strongly correlated predictors (difficult case),
- directly available in Minitab and SAS (in Stata via `margins` command).

The parameter estimates themselves:

- continuous predictors: always look at regression coef.,
- factors: recall the restrictions (depending on your software program); usually, `lsmeans`/`margins` are sufficient,
- interaction between continuous and factor: separate regression coef. for different levels of factor (beware of parameter restrictions).

Model Specification

Why specify models (maybe not a silly question...)

- documentation of analysis,
- aid for development of good models, by making clear what is taken into account and what is not.

Types of model specifications – illustrated by Amylase example (`atemp,gtemp,var`):

- fully specified model, multi-index notation:
  $$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + \varepsilon_{ijkl},$$
  where $i = 1, \ldots, 8$; $j, k = 1, 2$; $l = 1, 2, 3$,
- fully specified model, single-index notation:
  $y_i = \mu + \alpha_{\texttt{atemp}(i)} + \beta_{\texttt{gtemp}(i)} + \gamma_{\texttt{var}(i)} + (\beta\gamma)_{\texttt{gtemp*var}(i)} + \varepsilon_i$,
  where $i = 1, \ldots, 96 \sim$ observation number,
- fully specified model, no-index notation:
  $$y = \mu + \alpha_{\texttt{atemp}} + \beta_{\texttt{gtemp}} + \gamma_{\texttt{var}} + (\beta\gamma)_{\texttt{gtemp*var}} + \varepsilon,$$
- model formula (software: without +'s and error term):
  $y =$ atemp + gtemp + var + gtemp*var + error.

Regression terms / Covariates:

- in specified models: terms like $\beta_1 \cdot \texttt{atemp}_i + \beta_2 \cdot \texttt{atemp}_i^2$,
- in model formulae: special notation (options or boxes in software).

Stata Do-file (selection)

```
insheet using ch08ta6.csv, clear /* Example 8.6 */
anova tfaa r50 r21 r50#r21 /* same with r50##r21 only */
regress  /* Stata parametrization; note different P-values! */
* interaction plot
margins r50#r21
marginsplot, noci /* CIs often look messy */
* contrasts for SS decomposition computed manually
generate tx=10*r50+r21 /* combined tx variable */
anova tfaa tx
lincom 0.tx+1.tx-10.tx-11.tx /* r50 */
lincom 0.tx+10.tx-1.tx-11.tx /* r21 */
lincom 0.tx-1.tx-10.tx+11.tx /* r50*r21 */
insheet using ch08ta7.csv, clear /* Example 8.8 */
generate lnfault=ln(fault)
anova lnfault alg##seq##size##alloc /* DFE=0 */
anova lnfault alg##seq##size alg##seq##alloc alg##size##alloc
      seq##size##alloc
insheet using ch10ta1.csv, clear /* Example 10.3 */
anova y a##b /* partial SS */
anova y a##b, sequential /* SS for b without interaction */
anova y b##a, sequential /* SS for a without interaction */
insheet using ch08ta9.csv, clear /* Example 8.10 */
egen tx=group(at gt v)
xi: boxcox amylase i.tx
generate lnam=ln(amylase)
anova lnam atemp##gtemp##v
anova lnam atemp##gtemp gtemp##v
* atemp modelled as continuous
anova lnam c.atemp##gtemp gtemp##v
anova lnam c.atemp##c.atemp##gtemp gtemp##v
anova lnam c.atemp##c.atemp##c.atemp##gtemp gtemp##v
* Example 10.1: analysis with missing value
anova lnam atemp##gtemp gtemp##v if _n>1
margins gtemp#v, asbalanced /* least squares means */
margins gtemp#v /* all estimates different */
```