

Microarray Design and Analysis

e-sciences institute, 2-4 May 2006

Luc Janss, ETH, Zürich, Switzerland

Luc.Janss@inw.agrl.ethz.ch

Or: Luc@LucJanss.com



Outline 2nd morning

- Linear models to analyse designs with multiple treatments or times
- Further refinement of handling heterogeneity of variances using “Bayesian smoothing”

How many comparisons?

In a 2x2 design with day and treatment

- Day5 vs Day7 (within control)
- Day5 vs Day7 (within Atten)
- Contr vs Atten (within day5)
- Contr vs Atten (within day7)
-
- ... but 1 is redundant. (e.g. 4th can be known given 3rd and 2nd).

Comparison made in linear model

- For the 2x2 design the linear model is commonly used to yield:
 - An overall day effect (“main effect”): the differences between day5 and day7 pooled over treatments
 - An overall treatment “main” effect (similarly pooled over days)
 - An interaction, to specify *ONE* deviation from main effects, e.g. for treatment ATTEN within day7

1. Linear models: estimability, setting up effects

Basics of linear models

- By linear models we describe an observation as a sum of “constant” effects:
 - $M_i = \text{mean} + \text{“day 3 effect”} + \text{“challenge effect”}$
 - By “constant” effect is meant that there is a constant “challenge effect” computed irrespective of days (unless specified otherwise)

Example linear model

	Day 3	Day 5	Day 7	Day 11
Control	x	x	x	x
Challenged	x	x	x	x

Difference between
this data is the
overall effect of
challenged vs.
control, assumed the
same at all days

Linear model more formal:

	Day 3	Day 5	Day 7	Day 11
Control	μ	$\mu+d_5$	$\mu+d_7$	$\mu+d_{11}$
Challenged	$\mu+c$	$\mu+d_5+c$	$\mu+d_7+c$	$\mu+d_{11}+c$

Difference of d_5

Difference of c

Linear model

- The effects so described need not match the actual data: this is the *model* that describes the data in a “best” way under the assumptions, there will be some remaining error
- We do not have a day 3 effect or control effect: the “day 3 - control”-cell is the reference; the *model mean* refers to this reference cell.
- We can take another reference base, and therefore another model mean.

Same (“equivalent”) model but with reference cell day7-challenged

	Day 3	Day 5	Day 7	Day 11
Control	$\mu+ct+d_3$	$\mu+ct+d_5$	$\mu+ct$	$\mu+ct+d_{11}$
Challenged	$\mu+d_3$	$\mu+d_5$	μ	$\mu+d_{11}$

What now is estimated as “control effect” is the same as the -“challenged effect” in the previous model; like this all estimates can be transformed between the two equivalent models

General rules for estimable effects

- We have to select arbitrarily a reference cell which gets assigned the model mean
 - Most statistical softwares do this automatically, e.g. choosing the cell with all first level
- For every treatment there will be ($\#$ levels $- 1$) additional effects to be estimated: 3 effects for for 4 days, 1 effect for control vs. challenged.

One extension: interaction.

- Sometimes the “constant” effects model is not satisfactory, and we can also include an interaction (this remains a “linear model”)

	Day 3	Day 5
Control	μ	$\mu+d$
Challenged	$\mu+c$	$\mu+c+d+cd$

Extra term which says that something extra is happening on day 5 in challenged animals

Statistical model fitting

- You can usually say something like:
 - $\text{Obs} \sim \mu + \text{treatm} + \text{day}$
and the software will make design vectors (and chooses reference) automatically
- However, for analysis of M values from 2-colour array data this does not work
 - Dye-swaps represent a reverse effect, and we have to code design vectors manually
 - Simply reversing signs of M values does not work when we have multiple factors in our design

Common layout array data (M-values per slide)

Gene	Slide1	Slide2	Slide3	Slide4	Slide5	Slide5
A	1.3	2.4	-0.8	-1.5	0.6	-0.3
B						
....						

May need design vectors for model fit as:

Mean	1	1	-1	-1	1	1
Effect 1	1	1	-1	-1	0	0
Effect 2	1	-1	1	-1	1	-1

General approach to set 1's and -1's

Cnt D5

Slide	Cy3	Cy5		mu	D7	Chl
1	Cnt D5	Cnt D7		-1	1	0
2	Cnt D7	Cnt D5		1	-1	0
....	Chl D7	Cnt D7				

Alternative model: Woolfing model

- Goes back to cy3 and cy5 intensities (logged)
- Advantages
 - You don't have to set up design vectors yourself
 - The data doubles (sounds good, statistically)
 - Can analyse “dye effect” to detect bad dye swaps
- Disadvantage
 - Assumes good repeatability of level of expressions across slides. Analysis of M-values only assumes good repeatability of ratio of expressions across slides.

Woolfinger model data layout

Gene	Slide1	Slide2	...			
A	1.3	-1.3				
...						



Gene	S1-Ch1	S1-Ch2	S2-Ch1	S2-Ch2		
A	16.5	14.7	14.7	16.5		
....						

Design vectors

Mean	1	1	1	1		
Dye	1	2	1	2		
Effect 1	1	0	0	1		
Effect 2	...					

More handy data lay-out for Woolfinger model

Gene	Dye	Expr	Slide	Day	Treatm
A	Red	11.23	1	3	Contr
A	Green	12.64	1	5	Infect
B	Red	14.58	1	3	Contr
B	Green	16.84	1	5	Infect
...					
A	Red	12.25	2	5	Infect
A	Green	10.06	2	3	Contr

This layout allows direct use of model specification
as: $\text{Expr} \sim \mu + \text{dye} + \text{day} + \text{treatm} (+ \text{slide})$ (per gene)

Advantage of Woolfinger model: detection of “dye effect”

- Sometimes it happens that red spots are also red in a dye swap (but they should have turned to green) and vice versa -> “dye effect”
- You would fit the model (per gene)
$$\text{Expr} \sim \mu + \text{dye} + \text{day} + \text{treatm}$$
 - If everything goes well, dye should not have an effect, effects need to be taken up by day and treatment
 - If dye effects do become significant, something’s wrong....

2. Another look at variance heterogeneity: Bayesian variance smoothing

More problems of variance heterogeneity

- Variance estimates (per gene) are based on small numbers, therefore can accidentally happen to be (very) small or happen to be (very) large
 - Accidental low variance -> high t-test
 - Accidental high variance -> low t-test
- We can apply the philosophy of “random effects” to variances: we do not believe too much in extreme values, certainly not when based on small numbers, and like to pull them to a common mean

Variance smoothing

- The function `ebayes` from the `Limma` / `Bioconductor` package does this variance smoothing
- The effect is especially that accidental small variances are pulled up
 - Extreme t-test are therefore tempered
- The pulling towards the average is stronger when the variance is computed on a small number

Gene variances before and after eBayes smoothing

