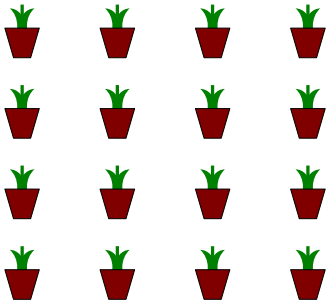
Introduction to Mixed Linear Models in Microarray Experiments

2/12/2009

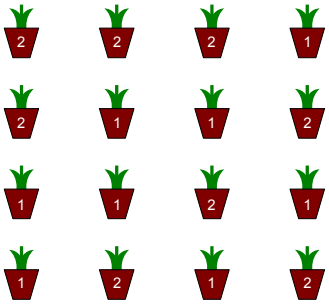
Statistical Models

A statistical model describes a formal mathematical data generation mechanism from which an observed set of data is assumed to have arisen.

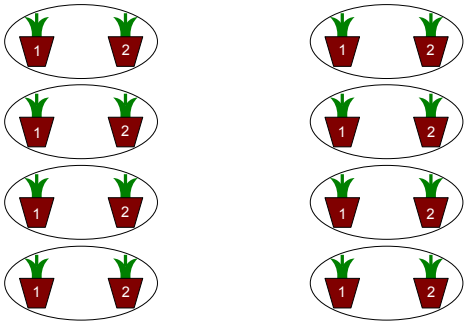
Example 1: Two-Treatment CRD



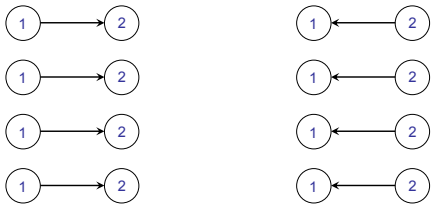
Assign 8 Plants to Each Treatment Completely at Random



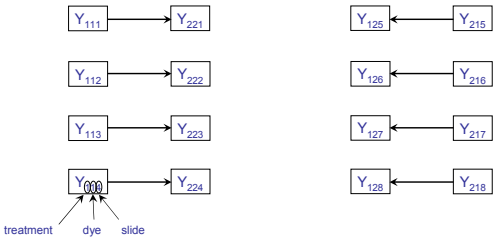
Randomly Pair Plants Receiving Different Treatments



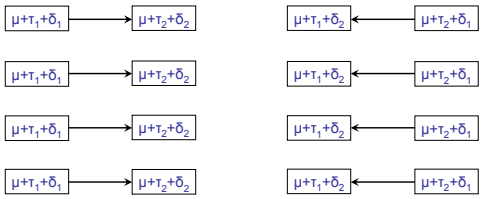
Randomly Assign Pairs to Slides Balancing the Two Dye Configurations



Observed Normalized Signal Intensities (NSI) for One Gene



Unknown Means Underlying the Observed Normalized Signal Intensities (NSI)

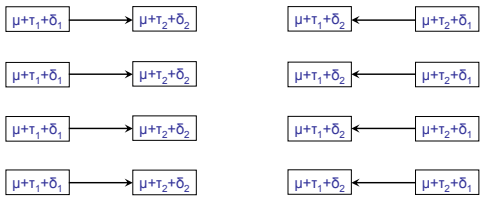


μ represents overall mean of NSI.

τ_1 and τ_2 represent the effects of treatments 1 and 2 on mean NSI.

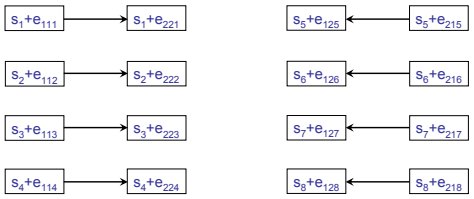
δ_1 and δ_2 represents the effects of Cy3 and Cy5 dyes on mean NSI.

The difference between τ_1 and τ_2 represents the difference in mean expression due to the effect of treatment.



A gene is differentially expressed if $\tau_1 \neq \tau_2$.

Unknown Random Effects Underlying Observed NSI



$s_1, s_2, s_3, s_4, s_5, s_6, s_7$, and s_8 represent slide effects.

$e_{111},\ldots,e_{218}$ represent residual random effects that include any sources of variation unaccounted for by other terms.

To make our model complete, we need to say more about the random effects.

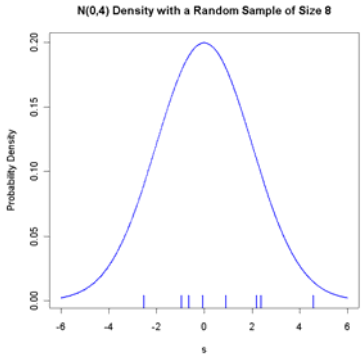
- We will almost always assume that random effects are independent and normally distributed with mean zero and a factor-specific variance.
- $s_1, s_2, \ldots, s_8 \overset{iid}{\sim} N(0,\sigma_s^2)$ and independent of

  $e_{111}, e_{112}, e_{113}, e_{114}, e_{221}, e_{222}, e_{223}, e_{224}, e_{125},$

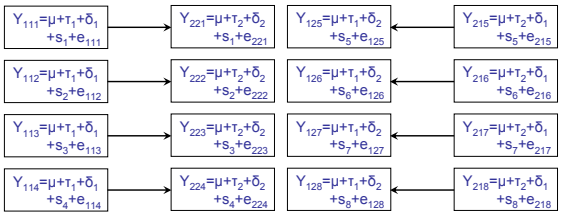
  $e_{126}, e_{127}, e_{128}, e_{215}, e_{216}, e_{217}, e_{218} \overset{iid}{\sim} N(0,\sigma_e^2)$.

  (or just $e_{ijk} \overset{iid}{\sim} N(0,\sigma_e^2)$ to save time and space.)

What does $s_1, s_2, \ldots, s_8 \overset{iid}{\sim} N(0,\sigma_s^2)$ mean?



Observed NSI are Means Plus Random Effects



$$Y_{ijk}=\mu+\tau_i+\delta_j+s_k+e_{ijk}$$

13

Our Model in Abbreviated Form

$Y_{ijk}=\mu+\tau_i+\delta_j+s_k+e_{ijk}$, $s_i \overset{iid}{\sim} N(0,\sigma_s^2)$, $e_{ijk} \overset{iid}{\sim} N(0,\sigma_e^2)$

i=1,2; j=1,2; k=1,2,...,8.

The model *parameters* are μ, τ_1, τ_2, δ_1, δ_2, σ_s^2, and σ_e^2.

The parameters σ_s^2 and σ_e^2 are special parameters called *variance components.*

We can estimate functions of model parameters from observed data.

14

Our model is a *linear model* because the mean of the response variable may be written as a *linear combination* of model parameters.

- Suppose a model has parameters θ_1, θ_2, ..., θ_m.
- A linear combination of the parameters is

$$c_1\theta_1+c_2\theta_2+ \ldots +c_m\theta_m$$

where c_1, c_2, ..., c_m are known constants.

15

Our model is a *linear model*

- $Y_{ijk}=\mu+\tau_i+\delta_j+s_k+e_{ijk}$
- $\text{mean}(Y_{ijk})=E(Y_{ijk})=\mu+\tau_i+\delta_j$.
- For example,

$E(Y_{124})=\mu+\tau_1+\delta_2$

$=1\mu+1\tau_1+0\tau_2+0\delta_1+1\delta_2$

= a linear combination of model parameters.

16

An Example of a *Nonlinear Model*

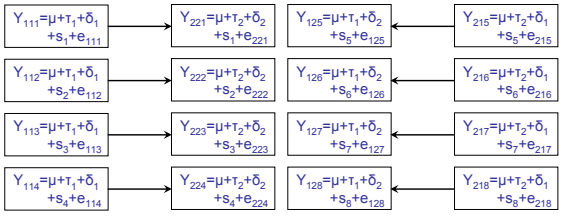
- $Y_{ijk}=\mu+\tau_i+\delta_j+\tau_i*\delta_j+s_k+e_{ijk}$
- $\text{mean}(Y_{ijk})=E(Y_{ijk})=\mu+\tau_i+\delta_j+\tau_i*\delta_j$.
- For example,

$E(Y_{124})=\mu+\tau_1+\delta_2+\tau_1*\delta_2$

which can't be written as a linear combination of the model parameters.

17

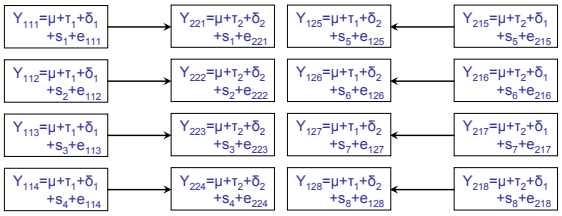
Estimating Differential Expression



Recall that a gene is differentially expressed if $\tau_1 \neq \tau_2$ or, equivalently, if $\tau_1 - \tau_2 \neq 0$.

18

Estimating Differential Expression (ctd.)



We can estimate $\tau_1 - \tau_2$ by

$$\bar{Y}_{1..} - \bar{Y}_{2..} = \tau_1 - \tau_2 + \bar{e}_{1..} - \bar{e}_{2..}$$

Estimating Differential Expression (ctd.)

$$\underbrace{\bar{Y}_{1..} - \bar{Y}_{2..}}_{\text{estimate}} = \underbrace{\tau_1 - \tau_2}_{\text{truth}} + \underbrace{\bar{e}_{1..} - \bar{e}_{2..}}_{\text{error}}$$

Using the observed data, we can estimate model parameters and determine the size of the "typical" error.

This size of the "typical" error is know as the *standard error*.

More formally, the *standard error* of an estimator is the square root of the estimated variance of the estimator.

Estimating Differential Expression (ctd.)

We can use the *standard error* to construct a 95% confidence interval for $\tau_1 - \tau_2$.

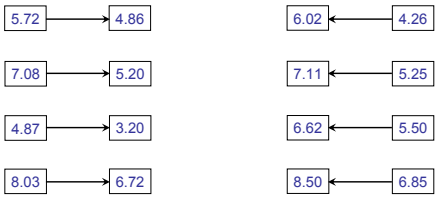
The method used to construct the interval will provide an range of values that contains the true value of $\tau_1 - \tau_2$ with probability 95%.

The form of the confidence interval is

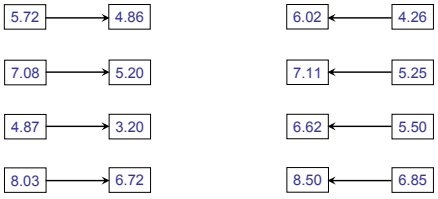
$$\textit{estimate} \pm k * \textit{standard error}$$

where *k* depends on the experimental design.

Example Observed Normalized Signal Intensities (NSI) for One Gene



Estimation of Differential Expression



Our estimate of $\tau_1 - \tau_2$ is

$$\bar{Y}_{1..} - \bar{Y}_{2..} = \tau_1 - \tau_2 + \bar{e}_{1..} - \bar{e}_{2..} = 1.514$$

Estimation of Differential Expression (ctd.)

estimate=1.514

standard error = 0.139

95% confidence interval:

$1.514 \pm 2.45 * 0.139$

i.e., 1.173 to 1.855

Estimating Fold Change

- Normalized expression measures are typically computed on the log scale.
- An estimated difference in means on the log scale can be converted to an estimated fold change on the original scale.
- For example, we estimated $\tau_1 - \tau_2$ to be 1.514.
- This translates into an estimated fold change of exp(1.514)=4.54.
- This means that treatment 1 is estimated to increase the expression of the gene by a multiplicative factor of 4.54 relative to its level under treatment 2.

25

95% Confidence Interval for Fold Change

- We determined the 95% confidence interval for $\tau_1 - \tau_2$ to be 1.173 to 1.855.
- This translates into a 95% confidence interval for the fold change of exp(1.173)=3.23 to exp(1.855)=6.39.
- Thus we can be 95% confident that the actual fold change is somewhere between 3.23 and 6.39.

26

Testing for Significant Differential Expression

- As part of our mixed linear model analysis, we can conduct a test of

  $H_0 : \tau_1 = \tau_2$ (no differential expression)

  vs.

  $H_A : \tau_1 \neq \tau_2$ (differential expression)

- The test is based on a *t*-statistic given by

  t = estimate / standard error.

- The *t*-statistic is compared to a *t*-distribution to determine a *p*-value.

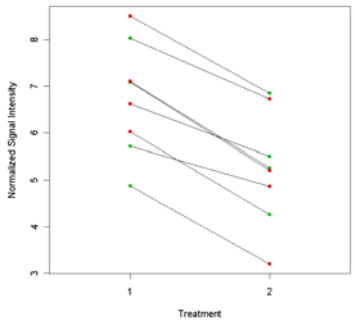
27

Testing for Significant Differential Expression (ctd.)

- The *p*-value gives the chance of seeing an estimated difference as large or larger than the one we observed ($\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot} = \tau_1 - \tau_2 + \bar{e}_{1\cdot\cdot} - \bar{e}_{2\cdot\cdot} = 1.514$) if the gene were not differentially expressed.
- A small *p*-value suggests evidence of differential expression because a small *p*-value says that the observed data would be unlikely if the gene were not differentially expressed.

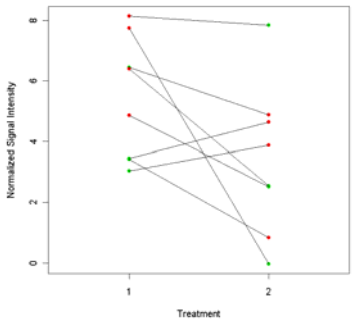
28

p-Value for Testing $\tau_1 = \tau_2$ is < 0.0001 Estimated Fold Change=4.54
95% Confidence Interval for Fold Change 3.23 to 6.39



29

p-Value for Testing $\tau_1 = \tau_2$ is 0.0660 Estimated Fold Change=7.76
95% Confidence Interval for Fold Change 0.83 to 72.49



30

Fixed Factors vs. Random Factors

- When we studied experimental design, we learned that experiments involve factors.
- Recall that a factor is an explanatory variable that takes any of two or more values in an experiment.
- Factors can be classified as fixed or random.

31

Fixed Factors vs. Random Factors

- Fixed factors are used to specify the mean of the response variable.
- Random factors are used to specify the correlation structure among the response variable observations.

32

Fixed Factors

- A factor is *fixed* if the levels of the factor were selected by the investigator with the purpose of comparing the effects of the levels to one another.
- One of the major goals of the analysis is to test for differences among the effects associated with the specifically chosen levels of the fixed factors and to describe the specific differences that exist.

33

Random Factors

- A factor is *random* if the effects associated with the levels of the factor can be viewed as being like a random sample from a population of effects.
- For random effects, we can make statements about variation in the population of random effects from which the effects at hand are considered to be like a random sample.

34

Random Factors

- Furthermore, we can generalize our conclusions about fixed factors to the populations associated with random factors.
- We are usually not interested in comparisons among the levels of random effects.
- Rather, we are interested in studying variation in the population from which the random effects are like a random sample or in controlling for that variation so that proper conclusions about fixed effects can be drawn.

35

Fixed vs. Random

- If it is reasonable to view the effects of the levels as being like a random sample from a larger population of effects, a factor may be considered random.
- If not, the factor should be considered fixed.

36

Example 2

- The factor *dye* in a two-color microarray experiment is considered to be a fixed factor.
- We cannot argue that the effects associated with the levels of *dye* are like a random sample from a larger population of effects.
- We have only two dyes that were not randomly selected from a population of dyes.

37

Example 2 (continued)

- We would use these two dyes again if we were to repeat the experiment.
- Hence *dye* will typically have only two levels whose effects are most likely not like a random sample from some larger population of effects.
- *Dye* should be considered a fixed effect even though a researcher is not interested in comparing the levels of *dye* to one another.

38

Example 3

- To understand the level of variation in expression across a population of genotypes, a researcher randomly selects 10 maize genotypes from a population of several hundred genotypes.
- Thirty seeds, three from each genotype, are randomly assigned to 30 pots and positioned in a completely randomized manner in a growth chamber.
- Six weeks after emergence, one RNA sample is taken from each plant and measured using a single Affymetrix GeneChip.

39

Example 3 (continued)

- *Genotype* is a random factor because the effects associated with the 10 genotypes can be viewed as a random sample of effects from a larger population of effects associated with the several hundred genotypes from which the 10 were randomly selected.
- Our analysis of the expression of a single gene would focus on trying to estimate the variation in expression across the whole population of genotypes from the variation observed in the 10 that we sampled.

40

Example 4

- A researcher is interested in finding genes that are expressed differently across 10 maize genotypes.
- Thirty seeds, three from each genotype, are randomly assigned to 30 pots and positioned in a completely randomized manner in a growth chamber.
- Six weeks after emergence, one RNA sample is taken from each plant and measured using a single Affymetrix GeneChip.

41

Example 4 (continued)

- *Genotype* is a fixed factor because the researcher is interested in comparisons among the 10 genotypes in this particular experiment.
- In the analysis of a single gene, we will try to characterize expression differences among the 10 genotypes considered in this experiment.
- No effort will be made to generalize the results beyond these 10 genotypes.

42

Mixed Linear Models

A model that includes effects for both fixed and random factors is a *mixed linear model* if the conditional mean of the response variable, given the random effects, is a linear combination of model parameters and random effects.

43

The model described in Example 1 is a *mixed linear model.*

- $Y_{ijk}=\mu+\tau_i+\delta_j+s_k+e_{ijk}$
- $E(Y_{ijk}|s_1,s_2,...,s_8)=\mu+\tau_i+\delta_j+s_k$ which is a linear combination of parameters and random effects.
- For example, $E(Y_{124}|s_1,s_2,...,s_8)=\mu+\tau_1+\delta_2+s_4$

$$=1\mu+1\tau_1+0\tau_2+0\delta_1+1\delta_2 +0s_1+0s_2+0s_3+1s_4+0s_5+0s_6+0s_7+0s_8$$

= a linear combination of model parameters and random effects.

44

We will often simplify the description of a linear model.

$$Y_{ijk}=\mu+\tau_i+\delta_j+s_k+e_{ijk}$$

$$Y=\mu+treatment+dye+slide+residual$$

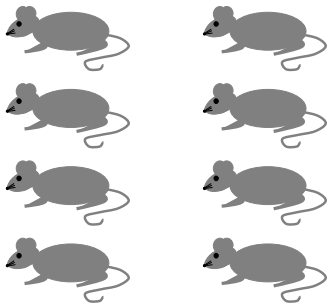
$$Y=treatment\ dye\ slide$$

The above are meant to be equivalent expressions.

Note that although μ and residual are missing from the last expression, their presence is to be assumed.

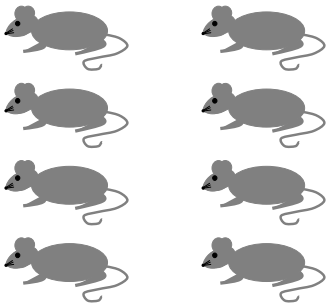
45

Example 5: CRD with Affymetrix Technology



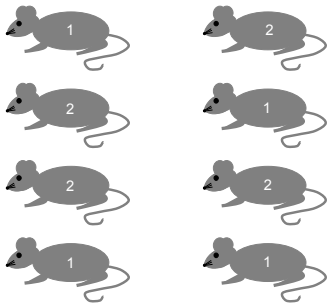
46

Randomly assign 4 mice to each treatment



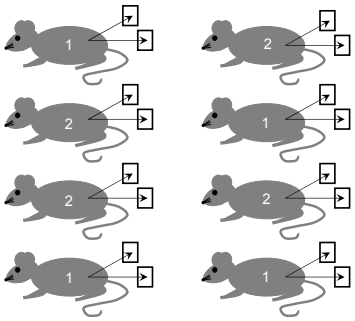
47

Randomly assign 4 mice to each treatment



48

Measure Expression using two Affymetrix GeneChips for each mouse



49

Model for One Gene

$Y_{ijk}=\mu+\tau_i+m_{ij}+e_{ijk}$ (i=1,2; j=1, 2, 3, 4; k=1,2)

Y_{ijk}=normalized signal intensity from the k^{th} GeneChip for the j^{th} mouse exposed to the i^{th} treatment.

μ=mean normalized signal intensity

τ_i=effect due to i^{th} treatment

m_{ij}=random effect due to the j^{th} mouse exposed to the i^{th} treatment

e_{ijk}=random residual effect for the the k^{th} GeneChip of the j^{th} mouse exposed to the i^{th} treatment.

50

Alternative Expressions for the Model

$Y_{ijk}=\mu+\tau_i+m_{ij}+e_{ijk}$

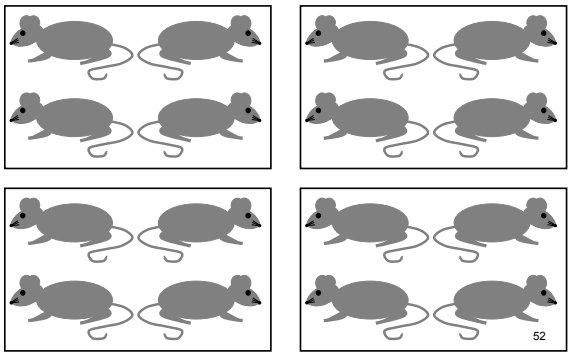
Y=μ+treatment+mouse+residual

Y=treatment mouse

A factor whose levels have a one-to-one correspondence with experimental units should be considered random

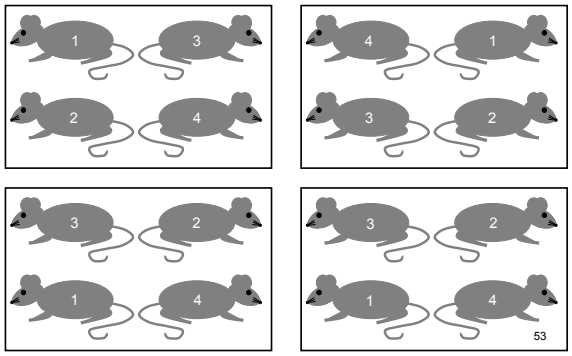
51

Example 6 : RCBD with Affymetrix Technology



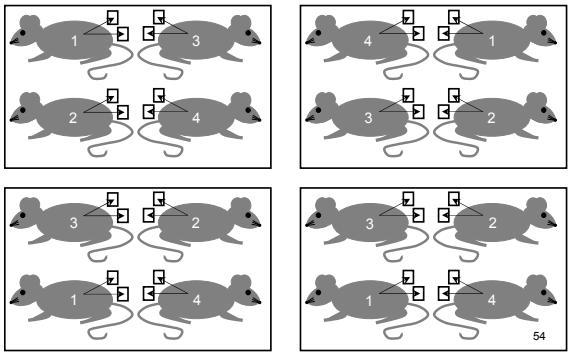
52

4 cages, 4 mice in each cage, randomly assign the 4 treatments to mice in each cage



53

Measure expression using two Affymetrix GeneChips for each mouse.



54

Model for One Gene

$Y_{ijk}=\mu+\tau_i+c_j+m_{ij}+e_{ijk}$ (i=1,2,3,4; j=1,2,3,4; k=1,2)

Y_{ijk}=normalized signal intensity from the k^{th} GeneChip for the mouse in cage j that received treatment i.

μ=mean normalized signal intensity

τ_i=effect due to i^{th} treatment

c_j=random effect for the j^{th} cage

m_{ij}=random effect for the mouse in the j^{th} cage treated with the i^{th} treatment

e_{ijk}=random residual effect for the the k^{th} GeneChip of the mouse in the j^{th} cage exposed to the i^{th} treatment.

55

Alternative Expressions for the Model

$Y_{ijk}=\mu+\tau_i+c_j+m_{ij}+e_{ijk}$

Y=μ+treatment+cage+mouse+residual

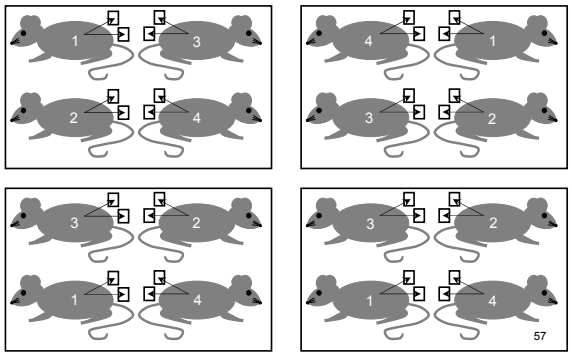
Y=treatment cage mouse

fixed random random

The factor *cage* could be considered as fixed rather than random. It is a blocking factor, and often blocking factors are considered random. The factor *mouse* must be considered random because the levels of mouse correspond to the experimental units.

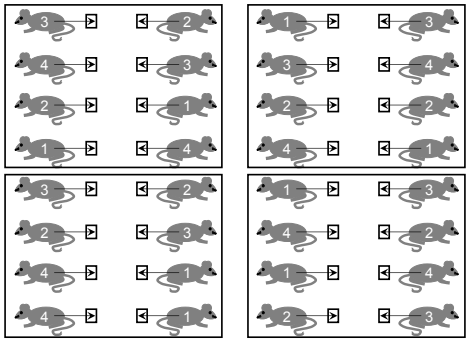
56

The random terms specify the correlation structure in the data



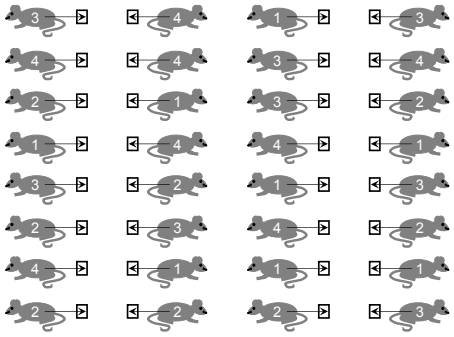
57

Without the mouse term, our model would correspond to a RCBD with 2 mice per treatment per cage.



58

With no random terms, our model would correspond to a CRD with 8 mice per treatment.



59

Design-Based Mixed Linear Modeling

- Model the mean of the response variable as a function of treatment factor main effects and interactions as well as nuisance factor effects (e.g., *dye*).
- Use random factors to specify the correlation structure among observations of the response variable that might arise due to the structure of the experimental design.
- In particular, make sure to include terms whose effects correspond to blocks and experimental units.

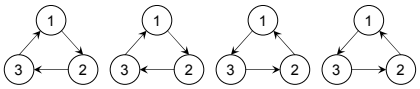
60

Specifying Effects That Correspond to Experimental Units

- When there is only one observation for each experimental unit, the residual term will contain the random effects that correspond to the experimental units.
- When an experimental unit contributes more than one observation to the data set, it is important to specify a term in the model that will have one effect for each experimental unit.

61

Example 7: CRD with Loops

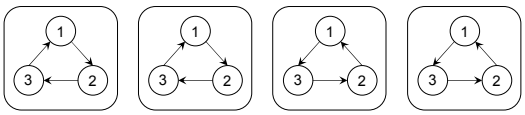


fixed

$Y = \underbrace{trt\ dye}_{fixed}\ \underbrace{slide\ xu}_{random}$

62

Example 8: RCBD with Loops

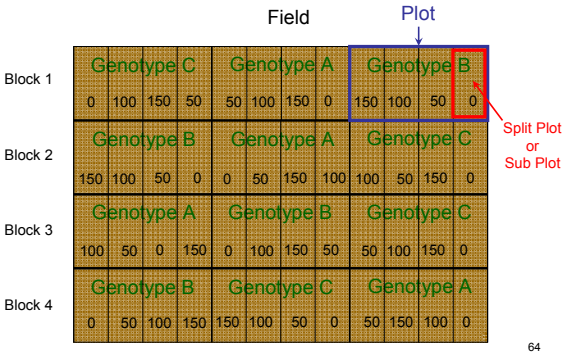


fixed

$Y = \underbrace{trt\ dye}_{fixed}\ \underbrace{block\ slide\ xu}_{random}$

63

Example 9: Split-Plot Experimental Design



64

Gene-specific mixed linear model for the analysis of the split-plot experiment if Affymetrix GeneChips will be used to measure expression

$Y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+r_k+p_{ik}+e_{ijk}$ (i=1,2,3; j=1,2,3,4; k=1,2,3,4)

Y_{ijk}=normalized signal intensity for the split-plot experimental unit associated with genotype i, fertilizer amount j, and block k

μ=mean normalized signal intensity

α_i=*main effect* of i^{th} genotype

β_j=*main effect* of j^{th} fertilizer amount

$(\alpha\beta)_{ij}$=*interaction effect* for the combination of genotype i and fertilizer amount j

65

Gene-specific mixed linear model for the analysis of the split-plot experiment if Affymetrix GeneChips will be used to measure expression

$Y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+r_k+p_{ik}+e_{ijk}$ (i=1,2,3; j=1,2,3,4; k=1,2,3,4)

r_k=random effect of the k^{th} block

p_{ik}=random effect of the plot associated with genotype i in block k

e_{ijk}=random residual effect for the split-plot experimental unit associated with genotype i, fertilizer amount j, and block k

Note that residual term corresponds to the split-plot experimental units in this case because there is one observation for each split-plot experimental unit.

66

Main Effects and Interaction

- Taken together the main effects and interaction allow for a separate mean for each combination of genotype and fertilizer amount.
- We could write μ_{ij} in place of $\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}$.
- $\alpha_i-\alpha_{i'}$ represents the difference in the mean response between genotype i and genotype i' when we average over the different fertilizer amounts.
- $\beta_j-\beta_{j'}$ represents the difference in the mean response between fertilizer amount j and j' when we average over the different genotypes.

67

Main Effects and Interaction

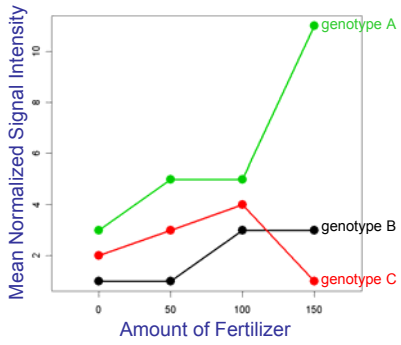
- The interaction terms $(\alpha\beta)_{ij}$ allow the difference between any two levels of genotype to depend on the level of fertilizer.
- $\mu_{ij}-\mu_{i'j} = \{ \mu+\alpha_i+\beta_j+(\alpha\beta)_{ij} \} - \{ \mu+\alpha_{i'}+\beta_j+(\alpha\beta)_{i'j} \}$

  $= \alpha_i - \alpha_{i'} + (\alpha\beta)_{ij} - (\alpha\beta)_{i'j}$
- Likewise the interaction terms $(\alpha\beta)_{ij}$ allow the difference between any two levels of fertilizer to depend on the genotype.
- $\mu_{ij}-\mu_{ij'} = \{ \mu+\alpha_i+\beta_j+(\alpha\beta)_{ij} \} - \{ \mu+\alpha_i+\beta_{j'}+(\alpha\beta)_{ij'} \}$

  $= \beta_j - \beta_{j'} + (\alpha\beta)_{ij} - (\alpha\beta)_{ij'}$

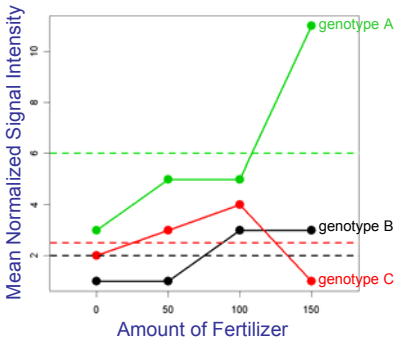
68

Interaction between genotype and fertilizer amount



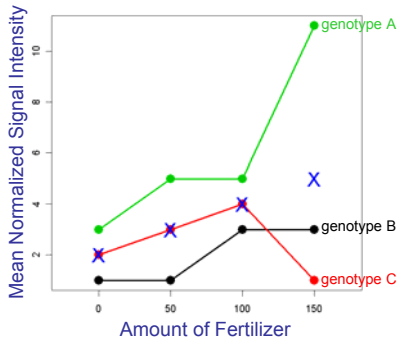
69

Differences among heights of dashed lines represent differences among genotype main effects



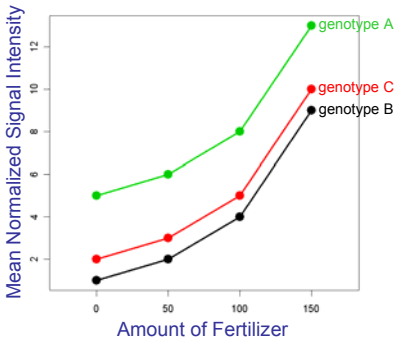
70

Differences among heights of Xs represent differences among fertilizer main effects



71

No interaction between genotype and fertilizer amount



72

Gene-specific mixed linear model for the analysis of the split-plot experiment if Affymetrix GeneChips will be used to measure expression

$Y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+r_k+p_{ik}+e_{ijk}$ (i=1,2,3; j=1,2,3,4; k=1,2,3,4)

fixed

Y=geno fert geno:fert block plot

random

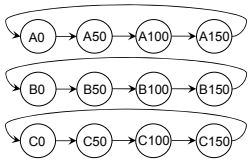
Recall that residual term corresponds to the split-plot experimental units in this case because there is one observation for each split-plot experimental unit.

73

Now suppose two-color slides will be used to measure expression.

Suppose that comparisons among fertilizer amounts for each genotype are of primary interest.

For each block, consider the following assignment of samples to slides and dyes, reversing loop direction for two of the four blocks.



74

Gene-specific mixed linear model for the analysis of the split-plot experiment if two-color slides will be used to measure expression

fixed

Y=geno fert geno:fert dye block plot splitplot slide

random

Note that the fixed factor *dye* and the random factors *splitplot* and *slide* have been added to the model. Each level of *splitplot* is associated with two observations. Similarly each level of *slide* is associated with two observations. These terms allow for additional correlation between observations that come from the same split-plot experimental unit or from the same slide.

75