Estimation of Variance Components

Why?

- Better understanding of genetic mechanism
- Needed for prediction of breeding values
  - Selection Index / BLUP
- Needed for optimization of breeding programs and prediction of response

Variance Components	Parameters
• Add. Genetic	Heritability
• Residual	
• Maternal	Maternal Heritability
• Permanent Environment	Repeatability
• Litter,	Common full-sib comp’t (“c^2”)
• Dominance,	
• Herd	
• Covariances	Correlations
	Phenotypic/ Genetic

When to (re) estimate variance components?

- New trait
- (co)variances change over time due to environmental and/or genetic change
  - Selection
  - Upgrading
  - Trait definition

Variance and Covariance

- Variance: measure of differences (extent of)
- Covariance: measure of ‘differences in common’
    - Between individuals/ between traits

	Types of family resemblance										
	None				Moderate				Full		
sire	1	2	3		1	2	3		1	2	3
Individual Values	1	1	1		2	2	1		1	2	3
	2	2	2		3	1	3		1	2	3
	3	3	3		1	2	3		1	2	3
Var between Families	None				Moderate				Large		
Var within Fam	Large				Moderate				None		

Relating variance components to underlying effects - give it a meaning!

- Variance between groups = covariance within groups!

- Variance between HS families

  = Covariance among half sibs = $\frac{1}{4} V_A$

  They share 25% of their genes!

  Variance within HS families

  = Residual Variance = $V_P - \frac{1}{4} V_A$

  $= \frac{3}{4} V_A + V_E + V_D$

Relating variance components to underlying effects
- give it a meaning!

- Variance between groups = covariance within groups!

- Variance between FS families

  = Covariance among full sibs $= ½\ V_A + V_{ec} + ¼\ V_D$

  They share 50% of their genes!

Variance within FS families

= Residual Variance $= V_P – ½\ V_A - V_{ec} - ¼\ V_D$

$= ½\ V_A + V_{EW} + ¾\ V_D$

Analyses of Variance

Principle

- Detect the importance of different sources of effects
- Importance is determined by its contribution to variation
- Variation if derived from sums of squares and df

Analyses of Variance

Example

$y_i = \mu + e_i$ μ = mean (fixed)

e_i = residual is random (causes variation)

Var(y) = $\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$

Same as

Calculating sum of squares $\sum_{i=1}^{n} e_i^{\;2} = SSE$

Equal SS to its expectation $E(SSE) = (n\text{-}1).\sigma^2_e$

Analyses of Variance

Example Data $y = [8, 9, 11, 12]$

Model: $y_i = \mu + e_i$

Sums of squares: Total: $8^2 + 9^2 + 11^2 + 12^2 = 410$

Mean: $4 * 10^2 =$ 400

Residual SS = 10

$(= (-2)^2 + (-1)^2 + 1^2 + 2^2)$

Analyses of Variance

Example Data $y = [8, 9, 11, 12]$ a: $i = 1\ \ 1\ \ 2\ \ 2$

Model: $y_i = \mu + a_i + e_{ij}$

Estimates: $\mu = 10$ $a_i = -1.5$ $a_2 = +1.5$

					Sum of squares	
Observed:	8	9	11	12	410	SS_{Total}
Mean:	10	10	10	10	400	SS_{Mean}
a-effect	-1.5	-1.5	+1.5	+1.5	9	SSA
Residual	-0.5	+0.5	-0.5	+0.5	1	SSE

ANOVA-Table

Expected Mean Squares

	SS	df	MS	EMS
Mean	400	1		
A-effect	9	1	9	$\sigma_e^2 + 2\sigma_a^2$
Residual	1	2	0.5	σ_e^2
Total	410	4		

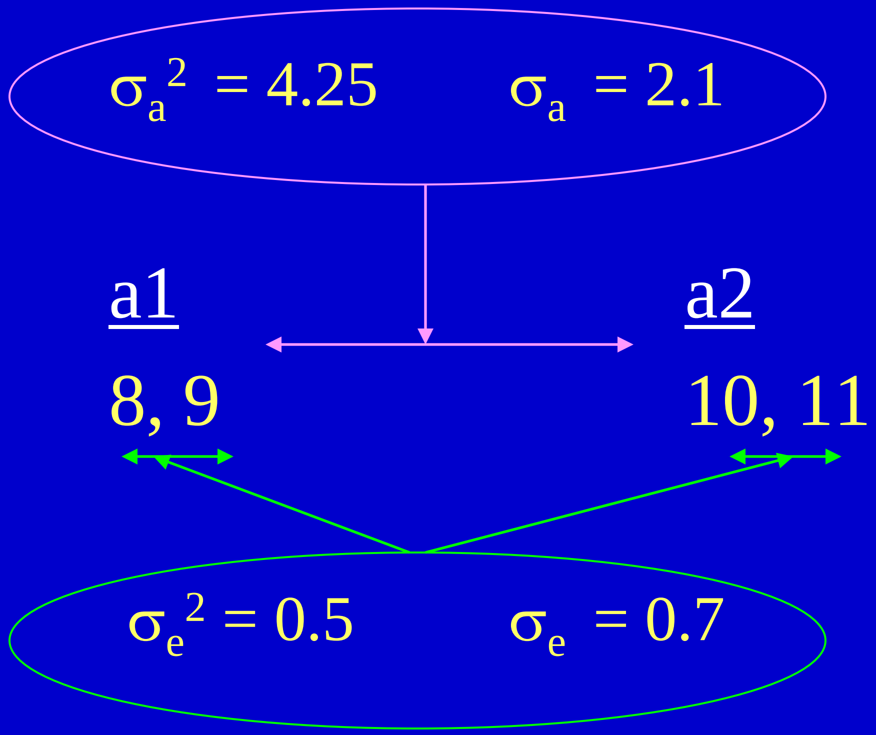
Nr. per class

Note: “a-effect” is a classification of data: e.g. according to sires (half sib groups). It relates to variance between groups

“residual” relates to variance within groups

Group (e.g. sire) differences relate to variance between groups

“residual” differences relates to variance within groups



$\sigma_a^2 = 4.25$ $\sigma_a = 2.1$

a1

8, 9

a2

10, 11

$\sigma_e^2 = 0.5$ $\sigma_e = 0.7$

Summarizing the procedure

Modeling (general)

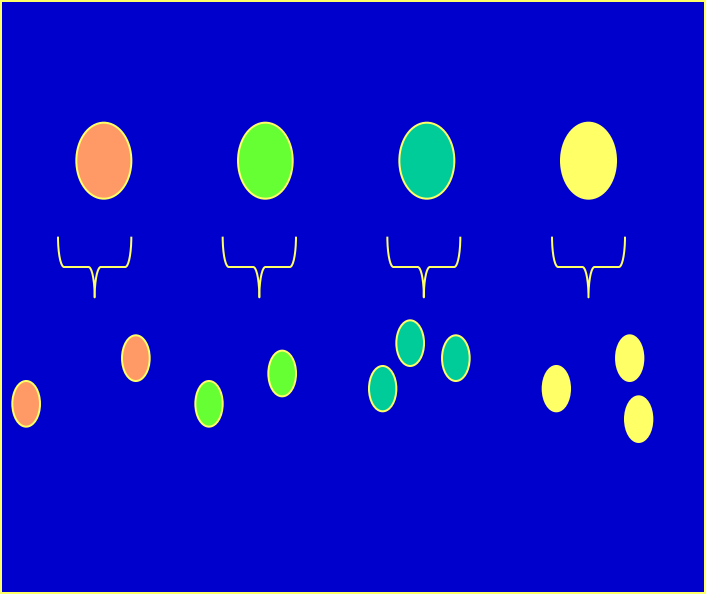
- Data = fixed effects + random effects
  - E(y) = fixed effects means
  - Var(y) = variance due to random effects

Interpretation

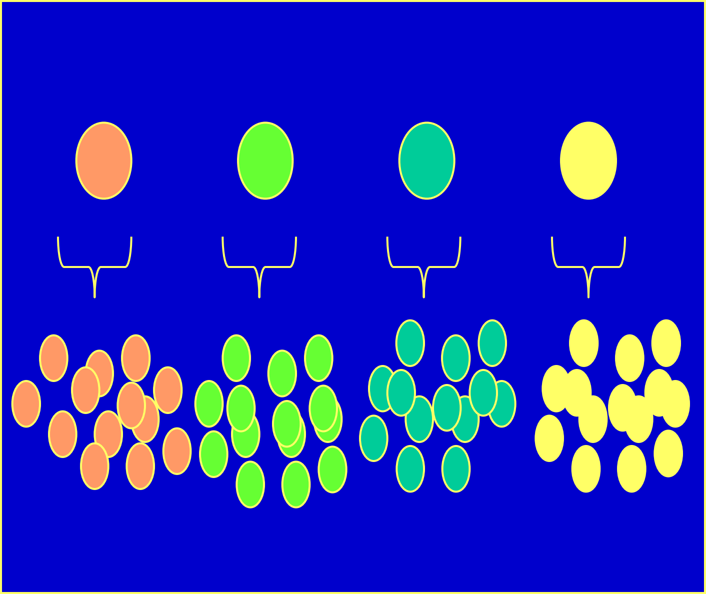
- Statistically:
  - Need sufficient data
  - Need to think about data structure
  - Sampling conditions need to be fulfilled (random?)
- Genetically
  - Translating the components into meaningful parameters
    - (e.g. sire variance = ¼ V_A)

h^2 estimates from half-sib families

Depend on number in each family (higher number → more accuracy)



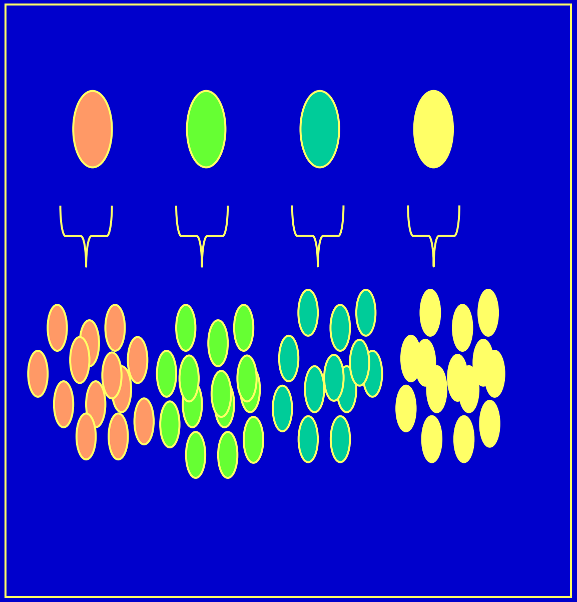
Poor estimate of family means



Good estimate of family means

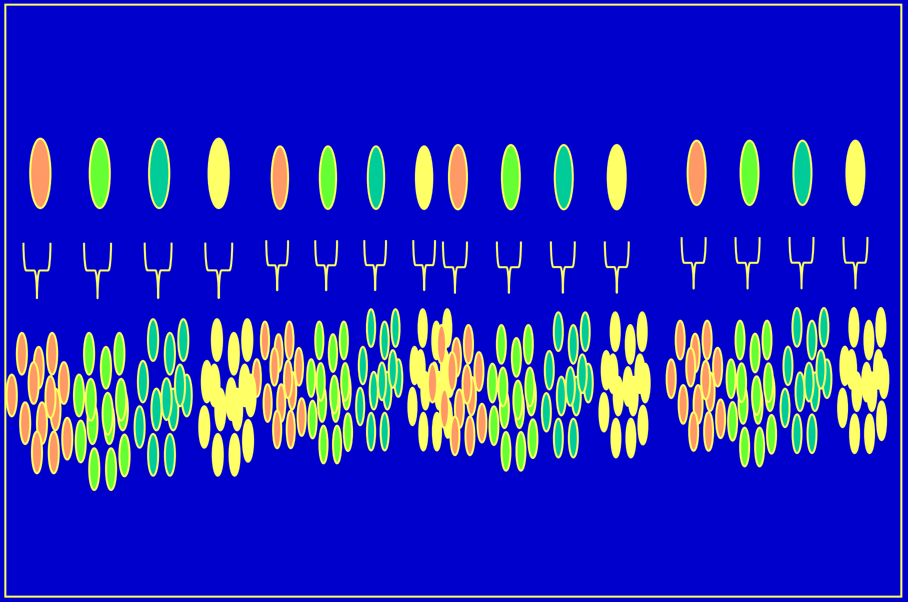
h^2 estimates from half-sib families

Depend on the number of sires (sire families) in the sample (higher number → more accuracy)



Small sample of sire,

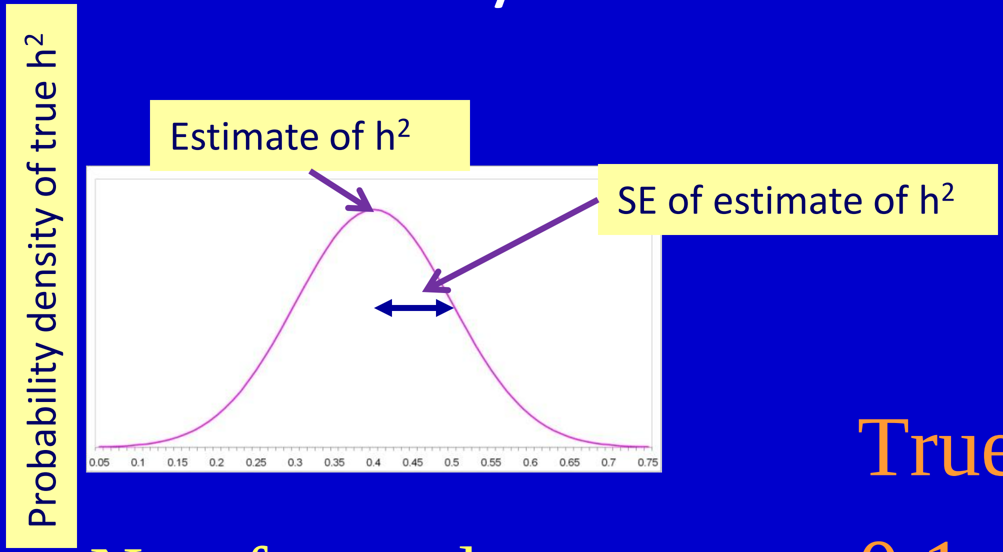
bad estimate of sire variance



Large sample of sire,

better estimate of sire variance

Accuracy: SE of heritability estimate



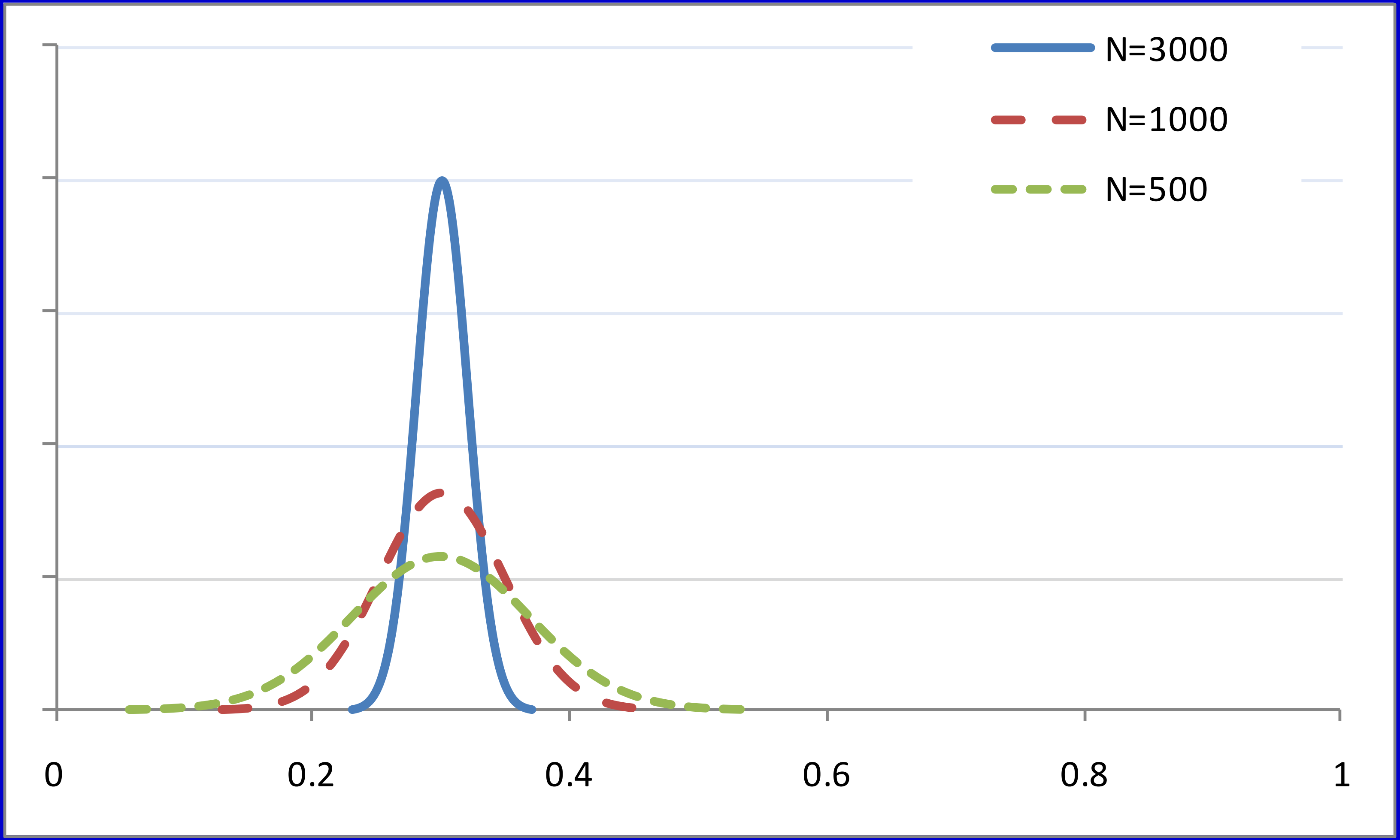
Probability density of true h^2

Estimate of h^2

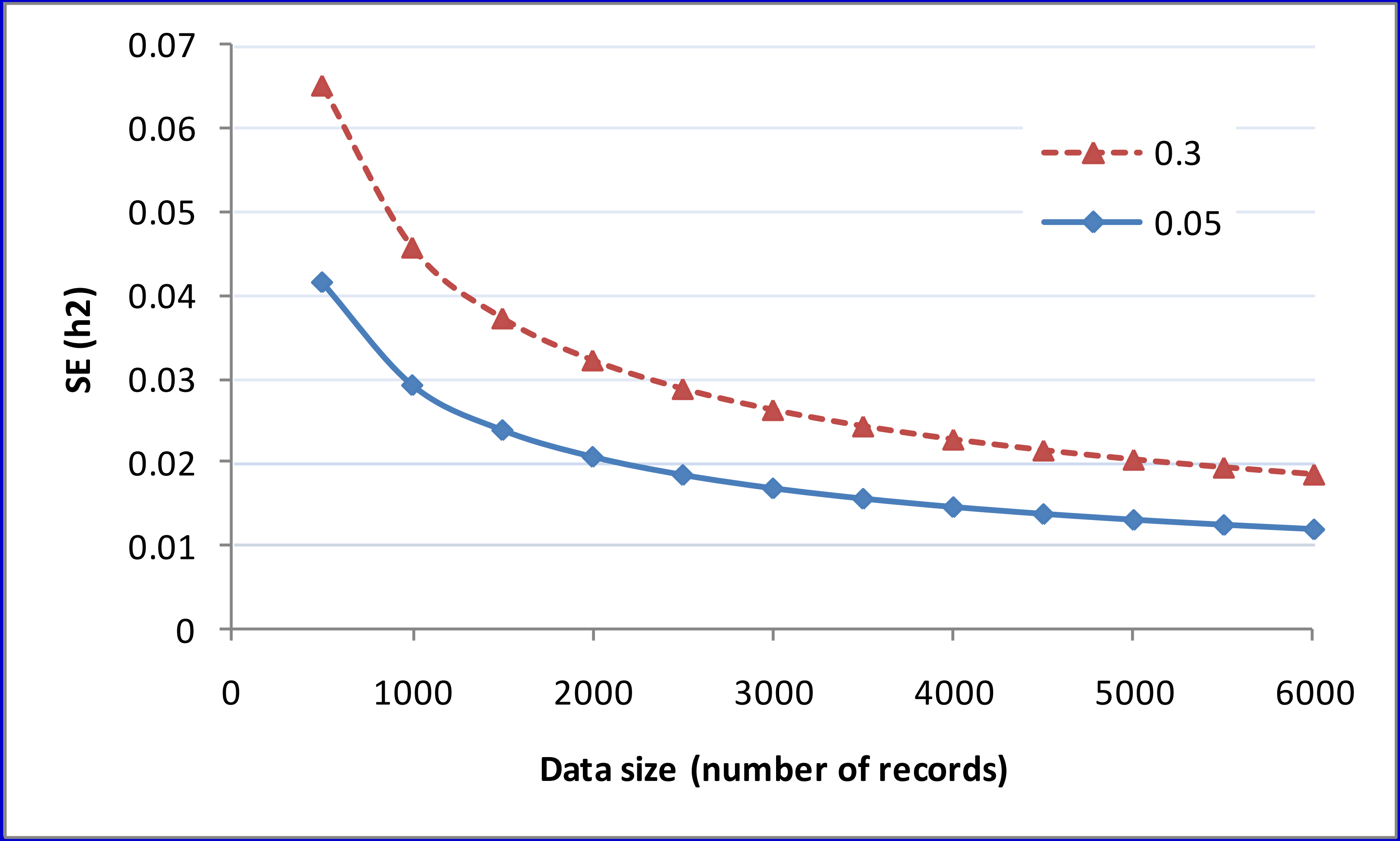
SE of estimate of h^2

Nr. of records	True heritability	
	0.1	0.3
100	0.18	0.30
500	0.08	0.14
1000	0.06	0.10
5000	0.03	0.04

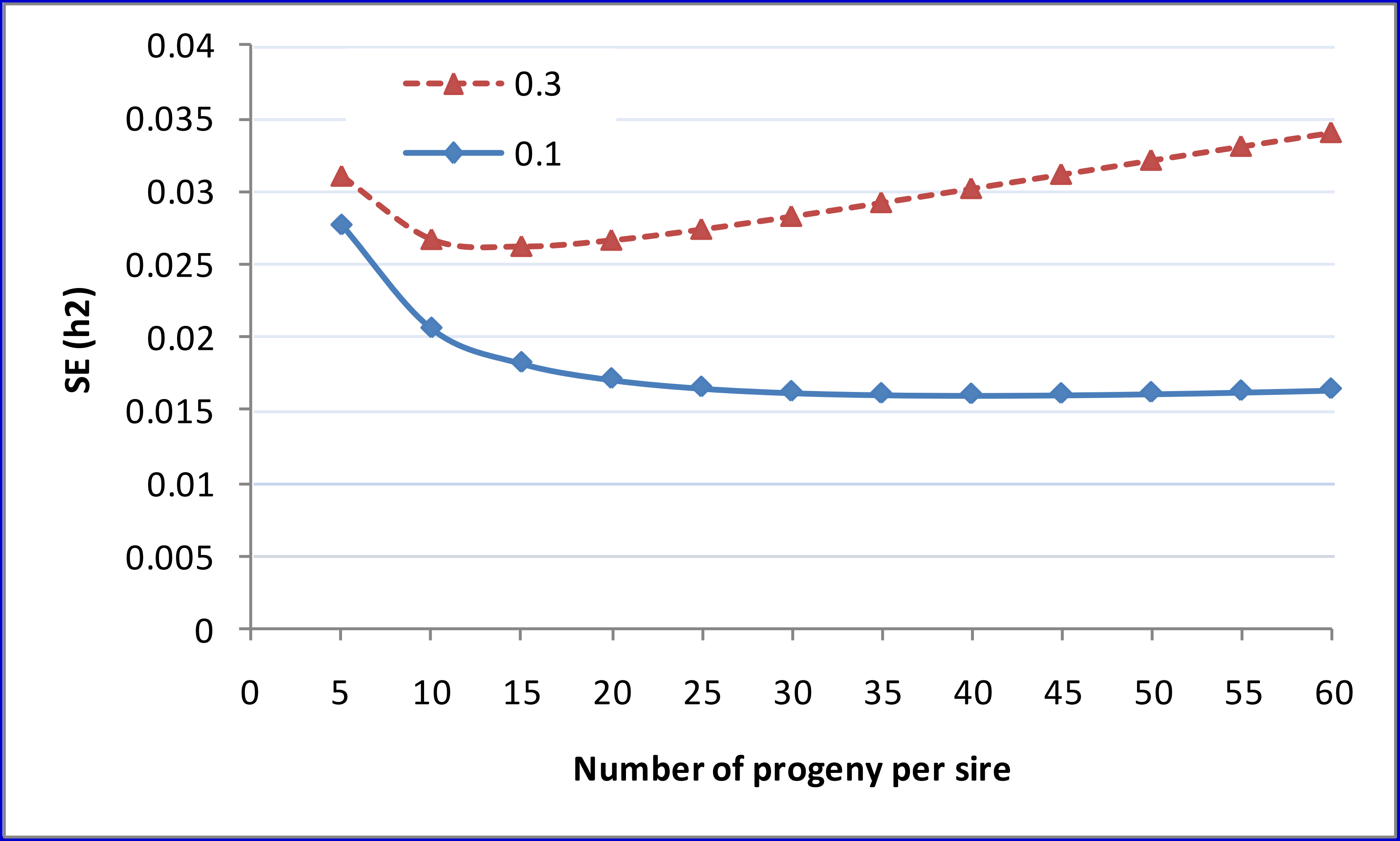
Effect of data size on SE of heritability



Effect of data size on SE of heritability



Effect of progeny group size on SE of heritability



The following slides

are not GENE422 material (reference only)

Methods for variance component estimation

- ANOVA - balanced data
- ANOVA – unbalanced data
  - Henderson’s methods (SAS etc)

- Likelihood methods
  - Maximum Likelihood
  - Restricted maximum Likelihood (REML)

- Bayesian Methods
  - Gibbs Sampling

Model: $y_i = \mu + a_i + e_{ij}$

ANOVA-Table for balanced data

Expected Mean Squares

	SS	df	MS	EMS
Mean	400	1		
A-effect	9	1	9	$\sigma_e^2 + n\sigma_a^2$
Residual	1	2	0.5	σ_e^2
Total	410	4		

Nr. per class

A-effect refers to differences ‘Between groups’

Residual refers to differences ‘Within groups’

ANOVA in Unbalanced data

Same idea as balanced (previous) but use a weighted number for “n” in: $EMS_A = \sigma_e^2 + n\sigma_a^2$

Need matrix notation to work out SS and EMS

(as in linear models)

Standard method in computer programs such as SAS, Harvey, SPSS etc.

Most general of those is called the “Henderson III method”

Likelihood methods

Each observation has a probability density, determined by its

- distribution
- expected value (e.g. mean) ‘location parameters’
- variance ‘dispersion parameters’

E.g. y with normal distribution, mean μ and variance σ^2

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

This is a Probability Density Function (PDF) for the observation

It gives the probability of the observation, given the parameters μ and σ^2

But we turn this around and get the likelihood of the parameters given y

Likelihood methods

We can multiply these probability values over the whole data, and include the fact that some of the observations may be related, i.e. we have a *joint distribution*

Data vector y with exp. means E(y) = Xb and var(y) = V

The log of the likelihood is:

$$L(b,V \mid X,y) = -\tfrac{1}{2} N \log(2\pi) - \tfrac{1}{2}\log(|V|) - \tfrac{1}{2}(y - Xb)'V^{-1}(y - Xb)$$

The expression gives the likelihood of the parameters (*b*, *V*) given data (*X*, *y*) in the right-hand side. It is a restricted (or residual) likelihood , after fitting the fixed effects.

first two terms are expectations *the last term is a (residual) sum of squares*

Restricted Maximum Likelihood

- Correct all data first for all fixed effects
- Find the maximum likelihood (solution for variance components) after these corrections

- Usually an iterative procedure is used to solve the problem
- Starting values (for the parameters) are needed to get going

An example of a REML algorithm
(EM-algorithm,for illustration only)

1. Solve mixed model equations using a prior value for the variance components (ratio)

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda \mathrm{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'Y \end{bmatrix}$$

2. Solve variance components from the MME-solutions

$$\sigma_a^2 = \left[\mathrm{\hat{a}}' \mathrm{A}^{-1} \mathrm{a} + tr(\mathrm{A}^{-1}\mathrm{C})\sigma_e^2 \right] / \ q$$

$$\sigma_e^2 = \left[\mathrm{y'y} \ - \ \mathrm{\hat{b}'X'y} \ - \mathrm{\hat{a}'Z'y} \right] / \ \left(N - r(\mathrm{X}) \right)$$

Use a new λ ($= \sigma_e^2/\sigma_a^2$) and iterate between 1 and 2

Why is REML better than ANOVA from SAS?

- It is by definition more accurate
- Uses full mixed model equations, so can utilize all animal relationships (animal model)
- Therefore, it has many properties as BLUP, e.g. it accounts for selection
- It allows more complicated mixed models (maternal effects, multiple traits etc) as with BLUP

Further notes on REML procedure

- If using an animal model, heritability is estimated from naturally combining
  - information between families (HS/FS)
  - information from parent-offspring regression

- The method and model are very flexible, but it can be hard to evaluate the estimates based on the data and the data structure
  - e.g. Is there a good family structure?

Evaluating the quality of the parameter estimates

- Accuracy
  - Look at SE of estimates (although these are approximated!)
  - Evaluate effect of number of records, and structure (nr. of groups, usually HS groups, vs nr. per group)

- Unbiasedness
  - From the data, and the possible effects, evaluate whether there was no bias from selection, or from confounding effects, e.g. sires confounded with herd or management group

Example: Analysis of weaning weight for White Suffolk

data on 9700 animals, 15,000 in pedigree

Comparison of including or not including the correlation between direct genetic (A) and maternal (M) effects and the effect of ignoring maternal effects on estimating h2

	Correlation A-M included		No correlation		No maternal effect	
PhenVar	23.45		23.26		23.94	
Heritability	0.25	0.04	0.19	0.03	0.44	0.03
Maternal Heritab.	0.28	0.04	0.18	0.02		
Correl. direct-matern.	-0.44	0.10				

Example: Analysis of weaning weight for White Suffolk

data on 9700 animals, 15,000 in pedigree

The effect of ignoring or including a permanent environmental effect (PE) of dams

	with PE		without PE	
Phenotypic Var.	23.06		23.45	
Heritability (direct)	0.25	0.04	0.25	0.04
Maternal heritability	0.13	0.04	0.28	0.04
Corr Mat-Direct	-0.50	0.12	-0.44	0.10
Permanent Env. Ewe	0.12	0.02		