So far:
* Model selection: determined the fixed effects in the mixed linear model
* Variance components estimation: genetic component of a trait showed variation, because only for traits with measureable variation, selection of parents can done
* Prediction of breeding values: ranking criterion for selection candidates, and based on this criterion, parents will be selected from the population

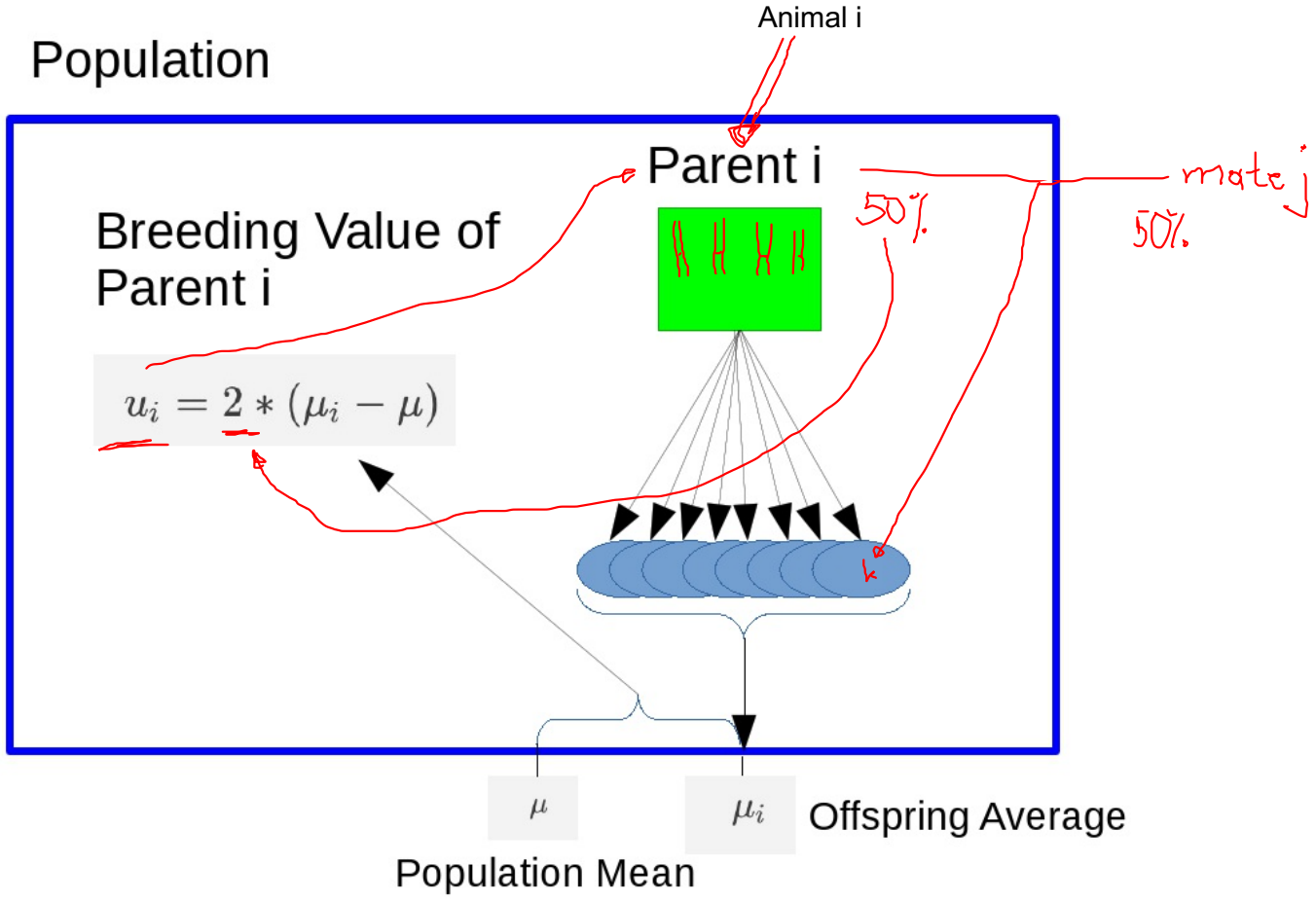
Prediction of Breeding Values

Peter von Rohr

10.05.2021

What are breeding values

Definition: two times difference between offspring of a given parent from population mean



Practical Considerations

- Definition of breeding value is based on biological fact that parent passes half of its alleles to offspring
- In practice, definition cannot be used
  - most parents do not have enough offspring
  - breeding values are needed before animals have offspring
  - different environmental factors not considered

Selection should be done as early as possible, otherwise the generation interval is increased and selection response per year is reduced

Solution

- Use genetic model to predict breeding values based on phenotypic observations
- Genetic model decomposes phenotypic observation (y_i) in different components

known non-genetic environmental factors

$$y_i = \mu + u_i + d_i + i_i + e_i$$

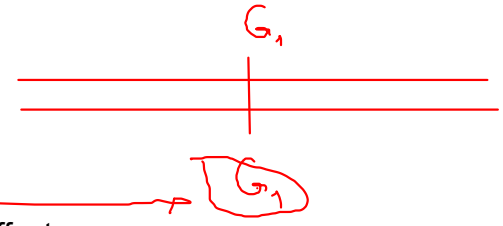
genetic factors

where μ is the general mean, u_i the breeding value, d_i the dominance deviation, i_i the epistasis effect and e_i the random error term.

Genetic Factors

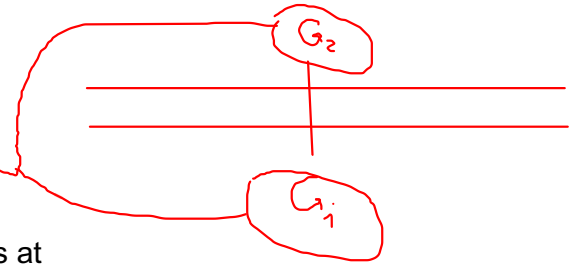
for selection, only additive effect of a single allele is important

1. Breeding value



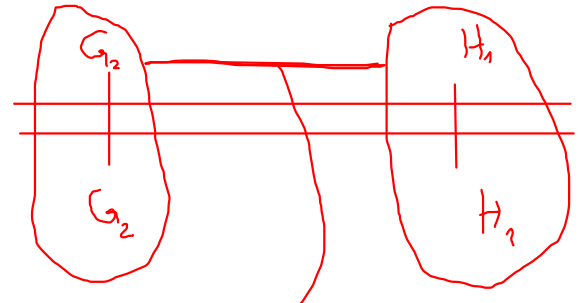
additive effect

2. Dominance



interaction between alleles at the same locus

3. Epistasis



interactions between different loci

Solution II

- For predicting breeding values d_i and i_i are often ignored, leading to a simplified version of the genetic model

Phenotype

$$y_i = \mu + u_i + e_i$$

μ: fixed; u_i: random; e_i: random

Mixed linear effect model

- Expected values and variance-covariance matrix

for random effects

$$E\begin{bmatrix} y_i \\ u_i \\ e_i \end{bmatrix} = \begin{bmatrix} \mu \\ 0 \\ 0 \end{bmatrix}$$

$E[\mu] = \mu$

$E[u_i] = 0$; $E[e_i] = 0$; $E[y_i] = E[\mu + u_i + e_i] = E[\mu] + E[u_i] + E[e_i] = \mu + 0 + 0 = \mu$

$$var\begin{bmatrix} y_i \\ u_i \\ e_i \end{bmatrix} = \begin{bmatrix} \sigma_y^2 & \sigma_u^2 & \sigma_e^2 \\ \sigma_u^2 & \sigma_u^2 & 0 \\ \sigma_e^2 & 0 & \sigma_e^2 \end{bmatrix}$$

σ_u^2: $cov(u_i, u_i)$; 0: $cov(u_i, e_i)$; σ_e^2: $cov(e_i, e_i)$

How to Predict Breeding Values

- Predicted breeding values ($\hat{u}$) are a function of the observed phenotypic data (y)

$\rightarrow \hat{u} = f(y)$

- What should $f()$ look like?
- Goal: Maximize improvement of offspring generation over parents

$\rightarrow \hat{u}$ should be conditional expected value of true breeding value u given y:

$$\hat{u} = E(u|y)$$

Henderson (1963): using conditional expected value of the true breeding value given the phenotypic observation, response to selection from a parent to an offspring generation is maximized.

true breeding values, are unknown

predicted breeding value

Derivation

- Assume: multivariate normality of u and y and $E(u) = 0$, then

$E[u] = 0$ — intercept — slope factor — corrected observations

$$\hat{u} = E(u|y) = E(u) + cov(u, y^T) * var(y)^{-1} * (y - E(y))$$
$$= E(u|y) = cov(u, y^T) * var(y)^{-1} * (y - E(y))$$

2 — 1

- $\hat{u}$ consists of two parts

μ → known non-genetic environmental factors

1. $(y - E(y))$: phenotypic observations corrected for environmental effects
2. $cov(u, y^T) * var(y)^{-1}$: weighting factor of corrected observation

→ interpreted as regression slope

So far: Two different definitions of a predicted breeding value

1: From biological facts for animal i: $\hat{u}_i = 2 \cdot (\bar{y}_i - \mu)$

2: Based on genetic model $\hat{u}_i = E[u_i | y]$

Assumptions:
* genetic model
* multivariate normality of u and y

Unbiasedness

Recall: Definition $\hat{u} = cov(u, y^T) * var(y)^{-1} * (y - E[y])$

- Expected value $(E(\hat{u}))$

$$
\begin{aligned}
E(\hat{u}) &= E(cov(u, y^T) * var(y)^{-1} * (y - E(y))) \\
&= cov(u, y^T) * var(y)^{-1} * E(y - E(y)) \\
&= cov(u, y^T) * var(y)^{-1} * (E(y) - E(y)) = 0
\end{aligned}
$$

const

- With $E(u) = 0$, it follows $E(\hat{u}) = E(u) = 0$

unbiasedness ok

Variance

variation of predicted breeding value should be as close as possible to the variance of the true breeding value: var(u)

- $var(\hat{u})$ and $cov(u, \hat{u})$ important for quality of prediction

$$
\begin{aligned}
var(\hat{u}) &= var(cov(u, y^T) * var(y)^{-1} * (y - E(y))) \\
&= cov(u, y^T) * var(y)^{-1} * var(y - E(y)) \\
&\quad * var(y)^{-1} * cov(y, u^T) \\
&= cov(u, y^T) * var(y)^{-1} * cov(y, u^T)
\end{aligned}
$$

$$
\begin{aligned}
cov(u, \hat{u}) &= cov(u, (cov(u, y^T) * var(y)^{-1} * (y - E(y)))^T) \\
&= cov(u, (y - E(y))^T) * var(y)^{-1} * cov(y, u^T) \\
&= cov(u, y^T) * var(y)^{-1} * cov(y, u^T) = var(\hat{u})
\end{aligned}
$$

$var(a \cdot y) = a^2 \cdot var(y)$

if y, a is a vector

$\Rightarrow var(a^T y)$

$a^T : var(y) \cdot a$

Accuracy

correlation between true and predicted breeding value

- Measured by $r_{u,\hat{u}}$
- Recall $cov(u, \hat{u}) = var(\hat{u})$

In general, correlation r between random variables x and y:

$$r_{x,y} = \frac{cov(x,y)}{\sqrt{var(x) \cdot var(y)}} = r_{y,x}$$

$$r_{u,\hat{u}} = \frac{cov(u, \hat{u})}{\sqrt{var(u) * var(\hat{u})}} = \frac{var(\hat{u})}{\sqrt{var(u) \cdot var(\hat{u})}}$$

$$= \sqrt{\frac{var(\hat{u})}{var(u)}}$$

- Reliability ("Bestimmtheitsmass"): $B = r^2_{u,\hat{u}} = \frac{var(\hat{u})}{var(u)}$

$0 \leq r^2_{u,\hat{u}} \leq 1$

Prediction Error Variance (PEV)

Every prediction is associated with a certain error

- Variability of prediction error: $u - \hat{u}$

$$var(u - \hat{u}) = var(u) - 2cov(u, \hat{u}) + var(\hat{u}) = var(u) - var(\hat{u})$$
$$= var(u) * \left[1 - \frac{var(\hat{u})}{var(u)}\right]$$
$$= var(u) * \left[1 - r^2_{u,\hat{u}}\right]$$

$var(u) \cdot [1 - B]$

- Obtained from coefficient matrix of mixed model equations
- Used to compute reliability

Conditional Density

Once a predicted breeding value is available, what is the distribution of the true breeding value given the predicted breeding value

- Assessment of risk when using animals with predicted breeding values with different reliabilities quantified by $f(u|\hat{u})$
- Multivariate normal density with mean $E(u|\hat{u})$ and variance $var(u|\hat{u})$

$$E(u|\hat{u}) = E(u) + cov(u, \hat{u}^T) * var(\hat{u})^{-1} * (\hat{u} - E(\hat{u})) = \hat{u}$$

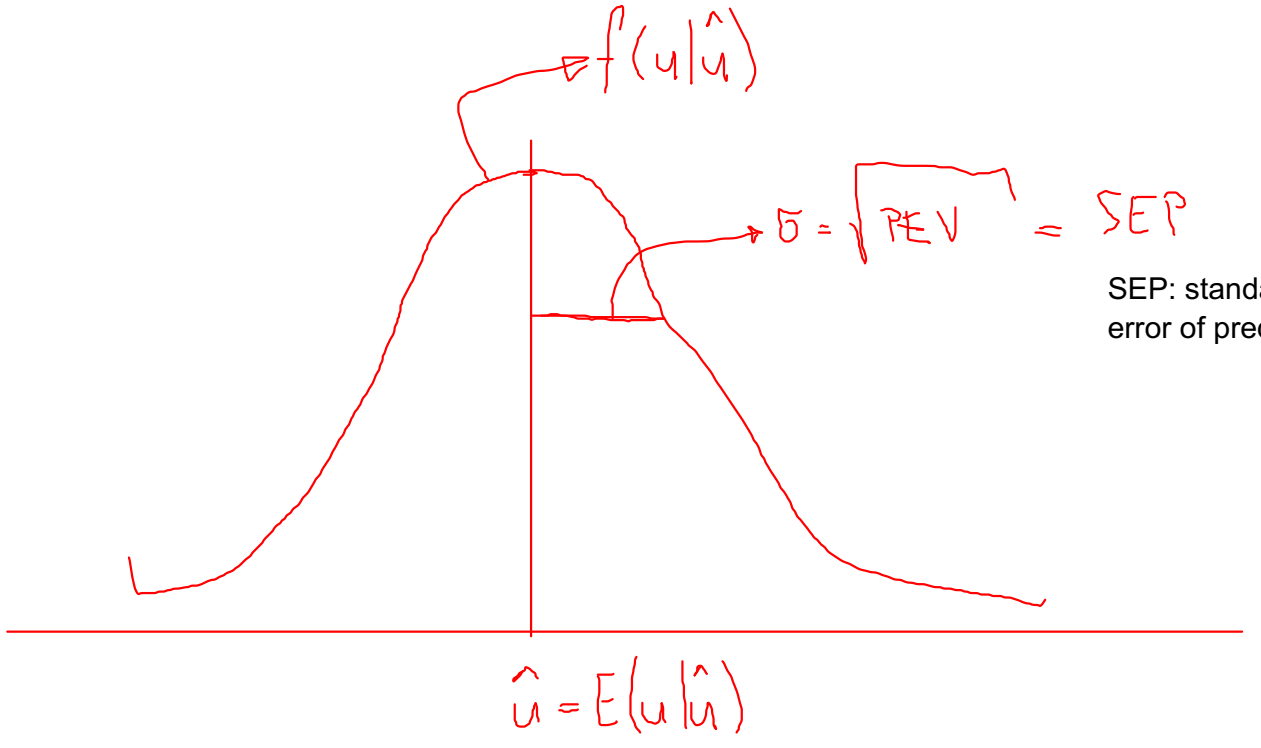
$$var(u|\hat{u}) = var(u) - cov(u, \hat{u}^T) * var(\hat{u})^{-1} * cov(\hat{u}, u^T)$$

$$= var(u) * \left[1 - \frac{cov(u, \hat{u}^T)^2}{var(u) * var(\hat{u})}\right]$$

$$= var(u) * \left[1 - r^2_{u,\hat{u}}\right] = PEV$$

Once we have a predicted breeding value: $\hat{u}$

What is the distribution of the true breeding value given the predicted breeding value



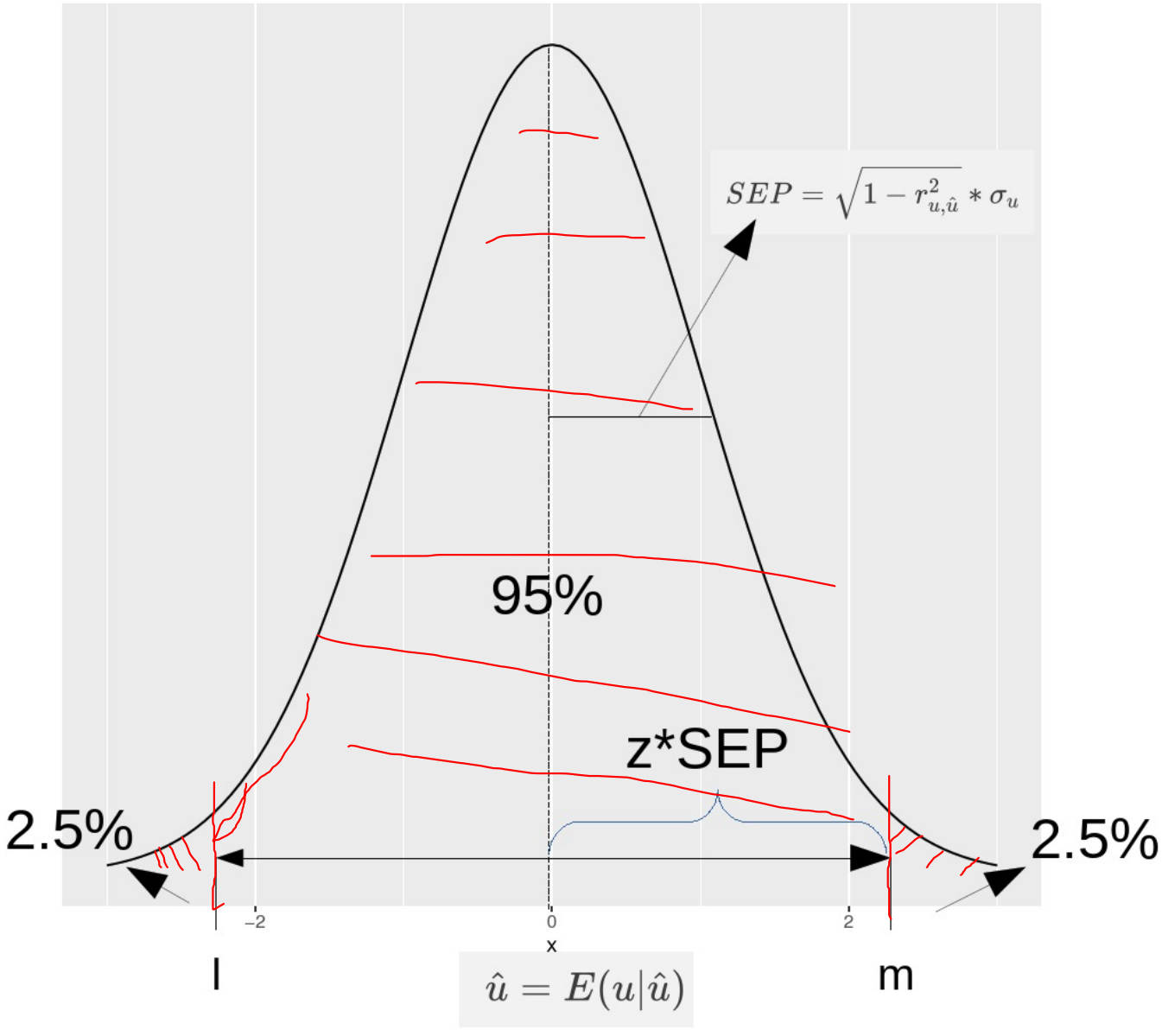
SEP: standard error of prediction

Confidence Intervals (CI)

Given a predicted breeding value, we can answer the question: What is the 95% confidence interval for the true breeding value

- Assume an error level α, this results in $100 * (1 - \alpha)$%-CI
- Typical values of α 0.05 or 0.01
- With $\alpha = 0.05$, the 95%-CI gives interval around mean which covers a surface of 0.95

CI-Plot



CI Limits

For 95% confidence interval (two-sided)

$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

$\alpha/2 = 0.025$

- lower limit l and upper limit m are given by

$$l = \hat{u} - z * SEP$$
$$m = \hat{u} + z * SEP \quad (1)$$

- z corresponds to quantile value to cover a surface of $(1 - \alpha)$
- Use R-function `qnorm()` to get value of z

qnorm $(1 - \alpha/2) \Rightarrow$ z-value

ex 95% : qnorm $(0.975) \Rightarrow z$

Linear Mixed Effects Model

Genetic model for a complete population and using matrix-vector notation

$$y_i = \mu + u_i + e_i$$

- Use more realistic model for prediction of breeding values

$$y = Xb + Zu + e$$

where

y	vector of length n with observations
b	vector of length p with fixed effects
u	vector of length q with random breeding values
e	vector of length n with random error terms
X	$n \times p$ incidence matrix
Z	$n \times q$ incidence matrix

Expected Values and Variances

$$E\begin{bmatrix} y \\ u \\ e \end{bmatrix} = \begin{bmatrix} Xb \\ 0 \\ 0 \end{bmatrix}$$

$$var\begin{bmatrix} y \\ u \\ e \end{bmatrix} = \begin{bmatrix} ZGZ^T + R & ZG & 0 \\ GZ^T & G & 0 \\ 0 & 0 & R \end{bmatrix}$$

Solutions

- Same as for simple model

$$\hat{u} = E(u|y) = GZ^T V^{-1}(y - X\hat{b})$$

with

$$\hat{b} = (X^T V^{-1} X)^- X^T V^{-1} y$$

corresponding to the general least squares solution of b

Problem

- Solution for $\hat{u}$ contains V^{-1} which is large and difficult to compute
- Use mixed model equations

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}$$

Sire Model

$$y = Xb + Zs + e$$

where s is a vector of length q_s with all sire effects.

$$var(s) = A_s * \sigma_s^2$$

where A_s: numerator relationship considering only sires

Animal Model

$$y = Xb + Za + e$$

where a is a vector of length q_a containing the breeding values

$$var(a) = A\sigma_a^2$$

where A is the numerator relationship matrix