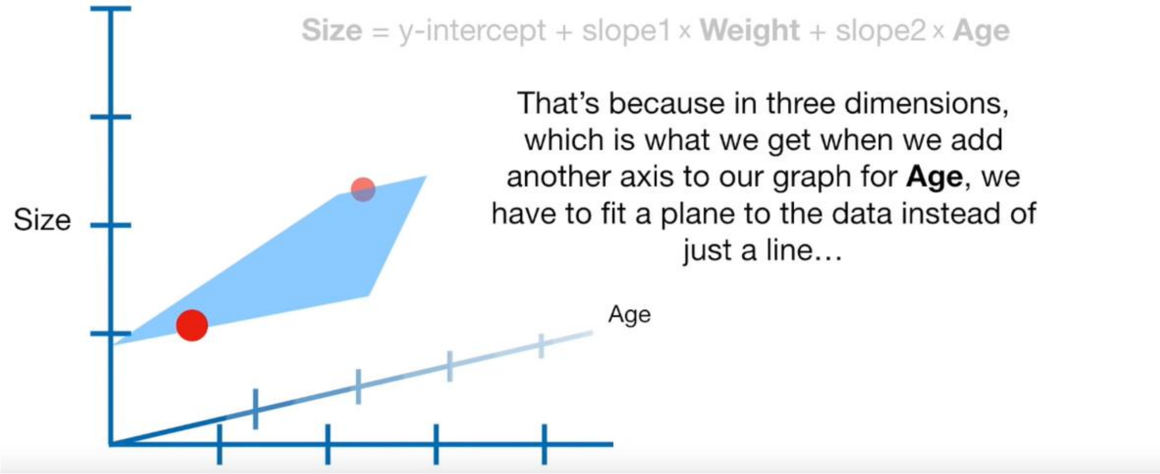
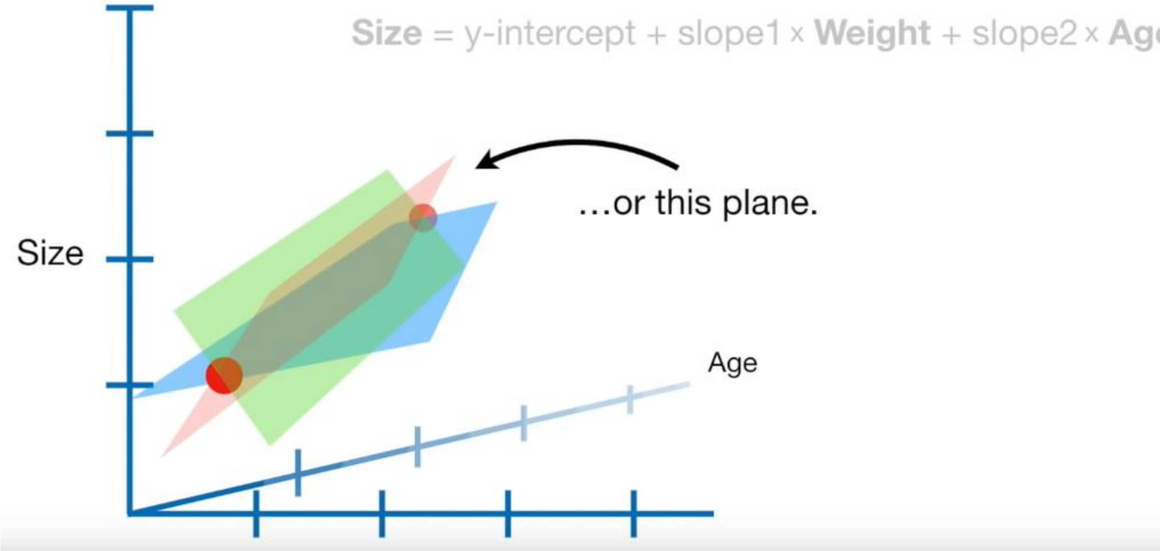
Introduction to Genomic Selection

1. Although there were a few applications in cattle breeding, MAS based on a few markers was not contributing appreciably to livestock improvement simply because **most of the traits of interest are quantitative and complex**, meaning *phenotypes are determined by thousands of genes with small effects* and influenced by environmental factors.
2. **only a few genes that contribute more than 1% of the genetic variation** for any given polygenic trait
3. (using genotypes, phenotypes and pedigree information), introduced by Fernando and Grossman (1989), Meuwissen et al. (2001) proposed some methods for what is now termed genome-wide selection or genomic selection (GS).
    a. 1) using SNP information can help to increase genetic gain and to reduce the generation interval;
    b. 2) *the biggest advantage of genomic selection would be for traits with low heritability*;
    c. 3) animals can be selected early in life prior to performance or progeny testing
4. *More genotyped individuals with accommodating pheno data lead to better genomic based predictions*
    a. more animals should be genotyped to reap the full benefits of GS. VanRaden et al. (2009) showed an increase in accuracy of 20 points when using 3,576 genotyped bulls, opposed to 6 points when using 1,151 bulls.
5. Distinct *training and validation populations* were needed to develop molecular breeding values (MBV) or direct genomic values (DGV),
6. The *effectiveness of genomic selection* can be predicted based on the **proportion of variance on the trait the SNP can explain**. There are mainly two classes of methods for genomic selection:
    a. 1) SNP effect-based method (SNP-BLUP or RR-BLUP – random/ridge regression-BLUP)
    b. 2) Genomic relationship-based method (GBLUP)
7. As the number of parameters is greater than the data points used for estimation, a solution is to assume SNP effects are random (or to have a prior distribution); in this way, all effects can be jointly estimated (**SNP effect-based method**)
    a. Why not simply a least square regression?
        i. More predictors (markers) than y (phenotypes) results in a multidimensional plane in a graph to fit the data points. As there are y<x, we can indicate the position of each phenotype score in the multi-dimensional plot, but we cannot fit a regression line/plane which fits all – we just have too few data points to determine the right position and slope

b. 

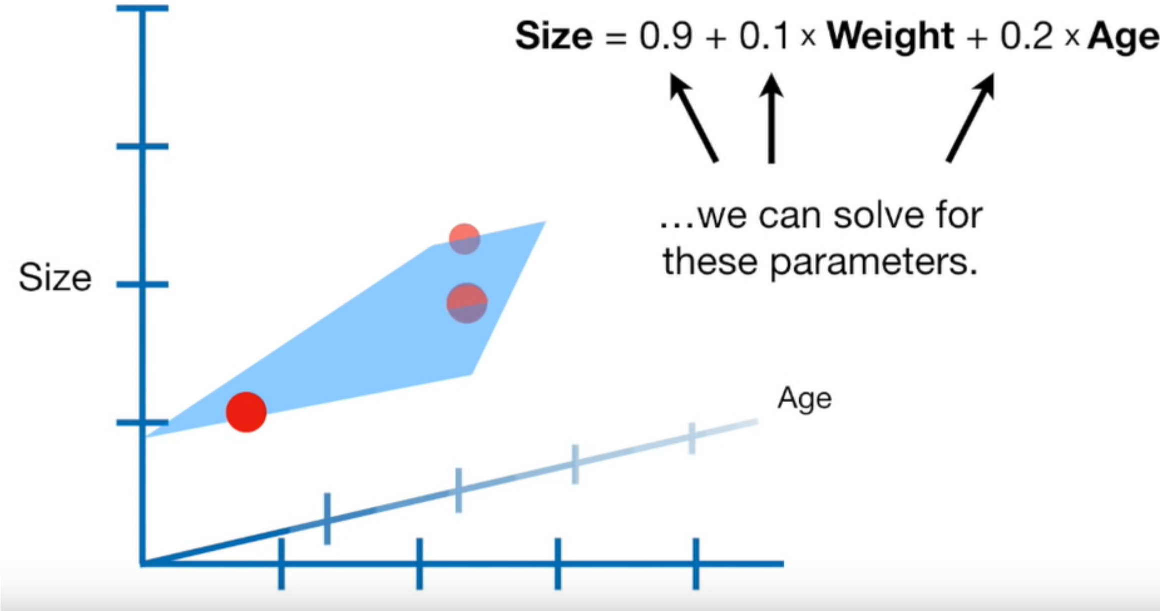
Size = y-intercept + slope1 × **Weight** + slope2 × **Age**

That's because in three dimensions, which is what we get when we add another axis to our graph for **Age**, we have to fit a plane to the data instead of just a line…

c. 

Size = y-intercept + slope1 × **Weight** + slope2 × **Ag**

…or this plane.

d. 

Size = 0.9 + 0.1 × **Weight** + 0.2 × **Age**

…we can solve for these parameters.

e. https://www.youtube.com/watch?v=Q81RR3yKn30

f. An other approach is the introduction of a bias by λ, which biases the regression

i. Based on this approach, the explained variance of the regression curve will not end up with a value of 1 (which typically happens when using a lot of predictors to estimate a trait y – which is what we do using thousands to millions of markers), but remain below

ii. 3 approaches exist to shrink the regression – shrinkage = reduction of errors

1. *Ridge regression*

a. Sum of squared residuals + λ * slope^2

i. λ * slope^2 = penalty

ii. will lead to the regression becoming less steep as λ becomes bigger

iii. How to determine the best λ? By cross validation

iv. The slope attribute contains all parameters despite the intercept (so, all markers)

v. Not all parameters are shrunk equally

vi. Ridge Regression assumes that effects are a priori normally distributed

2. *Lasso* (**least absolute shrinkage and selection operator**)

a. Sum of squared residuals + λ * |slope|

b. Many similar attributes to ridge regression

c. Differences:

i. Lasso can shrink the slope all way to 0

ii. => Meaningless variables can be eliminated as terms in the equations.

d. Lasso can exclude useless variables from the equation

i. Ridge can only minimize their effect

e. Lasso assumes that (marker) effects are a priori distributed following a Laplace (double exponential) distribution

3. *Elastic net*

a. Is a combination of both lasso and ridge regression

iii. λ & slope are both estimated using prior information, so that the marker information is utilized as a conditional probability, which depends on priory gained or assumed knowledge, e.g. a marker should not have an effect of, one phenotypic standard deviation of the trait

iv. Even though prior information is available, cross-validation will verify the best fit

8. RR-BLUP or SNP-BLUP provides SNP effects, but genomic estimated breeding values (u) can be derived as linear combinations of the SNP effects: u = Za (Z marker matrix, a allele effect)

a. marker effect follows a priori a normal distribution with a variance σ^2_{a0} (variance of marker effects)

$$p\left(a_i\right) = N\left(0, \sigma^2_{a0}\right)$$

b.

c. “0” implies that this variance is constant across markers.

d. markers are independent one from each other

i. => the prior assumption of normality precludes few markers of having very large effect

9. Bayesian approaches to genomic prediction:

a. Bayes's theorem $p(A\&B|B) = \frac{p(A\&B|A)p(A)}{p(B)}$

b. Where A is unknown (can be a parameter) and B is known (can be a trait

c. phenotype). Therefore, we want to infer values A by knowing B.

i. $p(A|B)$: posterior probability of unknown A given B is known.

ii. $p(B|A)$: likelihood function, determined by both A and B.

iii. $p(A)$: prior probability of unknown A.

iv. $p(B)$: probability to observe B without having any knowledge of A.

d. Bayesian methods are non-linear and likely to be affected by shrinkage, meaning small effects became even smaller and big effects even bigger.

e. Bayesian regressions are affected by the prior distribution that we assign to marker effects

i. each marker has a priori a different variance from each other

$$p\left(a_i \middle| \sigma^2_{\text{ai}}\right) = N\left(0, \sigma^2_{\text{ai}}\right)$$

ii.

f. BayesA

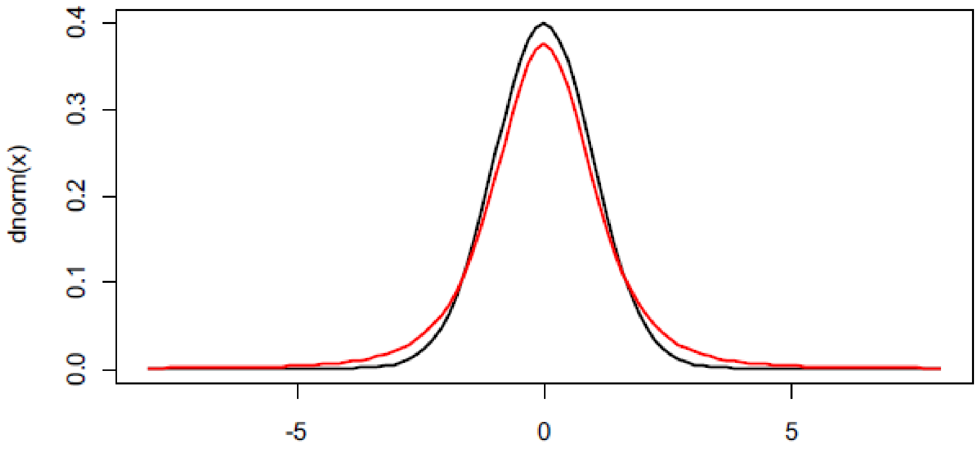
i. All SNPS have an effect on the trait

1. Few have a large effect, most have a small effect

2. => different variances are assumed

3. Red is BaysA, back is rr-BLUP

a. Larger effects are much more likely in BaysA than rrBLUP



g. BayesB

        i. Not many QTL were effecting the traits => many loci have zero variance
        ii. Π = proportion of SNP have no effect
        iii. 1- Π = have a non-zero effect
            1. When Π = 0, BayesB becomes BayesA
    h. BayesC
        i. A combination of SNP-BLUP and BaysB
        ii. Combines a distribution with constant variance (SNP-BLUP) and assumes some fraction Π of SNP have no effect (BayesB)
        iii. If Π = 0, BayesC = SNP-BLUP
    i. BayseanLasso
        i. Sets marker values a prioir to small values, instead setting some to 0
        ii. This is very similar to BayesA, in that a prior distribution is postulated for marker variances. The difference is the nature of this prior distribution (exponential in Bayesian Lasso and inverted chi-squared in BayesA),
    j. Important to realize:
        i. There is no need for multiple testing, as all SNP were introduced at the same time, and the prior already penalizes their estimates.
        ii. So the outcome of marker effects if Bayes is comparably to a common, p adusted GWAS.

10. GBLUP method
    a. Based on Genomic relationships - identical by state (IBS)
    b. quantifying the number of alleles shared between two individuals
    c. genomic relation ship can be calculated for
        i. additive ..
        ii. dominance ..
        iii. and epistasis effects
    d. the genomic relationship is given in a nxn matrix, where n denotes the genotypes objected to study – this matrix is also known as **kinship**
        i. multiple ways exist to calculate this matrix
    e. instead of SNP effects (u), genomic breeding values (Za) are estimated
    f. GBLUP is a BLUP where the pedigree relationship matrix is replaced by the genomic relationship matrix **G**.
        i. G contains ony information from genotypes individuals.

11. Many priors for marker effects have been proposed in the last years. These priors come more from practical (ease of computation) than from biological reasons. Each prior originates a method or family of methods, and we will describe them next, as well as their implications.
    a. 1. Normal distribution: Random regression BLUP (RR-BLUP), SNP-BLUP, GBLUP
    b. 2. Normal distribution with unknown variances: BayesC, GREML, GGibbs
    c. 3. Student (t) distribution : BayesA
    d. 4. Mixture of Student (t) distribution and spike at 0: BayesB
    e. 5. Mixture of Normal distribution and spike at 0: BayesCPi

f. 6. Double exponential: Bayesian Lasso

g. 7. Mixture of a large and small normal distribution: Stochastic Search Variable Selection (SSVS)

12. Recommendations for method selection

a. The trend to similar accuracy with a high number of QTL has been observed before in several studies

b. variable selection methods (e.g., BayesB) performed better than shrinkage methods (e.g., GBLUP, Lasso) with few (major) QTLs

i. setting some variances to zero seems to be of advantage when the number of QTL is low

ii. We suggest that two methods, one where loci are weighted equally (e.g., GBLUP) and one where some loci are given greater emphasis (e.g., Bayes B), be used when comparing new approaches.

13. Steps running a genomic prediction

a. Estimation of phenotype scores across multiple environments / years / experiments by simple BLUP

b. Separating the genotypes in 10 or more groups for cross-fold validations

c. Perform the genomic prediction with any of the presented models to estimate the marker effect (a) or the breeding value (Za)

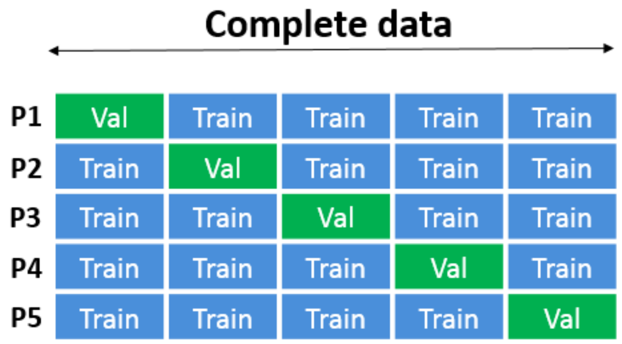
d. Using a with marker alleles or Za to predict genotype's pheno scores

14. How is validation performed?

a. We split the population of individuals in two groups – **training and testing**

b. K-fold cross-validation

i. In this method, the genotyped population is randomly divided into k subsets, and phenotypes are removed from one subset a time



Complete data

P1	Val	Train	Train	Train	Train
P2	Train	Val	Train	Train	Train
P3	Train	Train	Val	Train	Train
P4	Train	Train	Train	Val	Train
P5	Train	Train	Train	Train	Val

ii.

iii. Predictivity for each fold is calculated by a Pearson correlation

iv. The final prediction ability across all k-folds is calculated by summing up the correlation scores and dividing them by the number of k-folds.

v. The partitioning of training and testing populations will affect the accuracy attained

vi. large testing sets will reduce the reference population size and reduce accuracy

vii. When the testing set is too small, assessing differences in accuracy between methods for a particular data set may not be possible.

c. Testing and training data – problems and pitfalls:

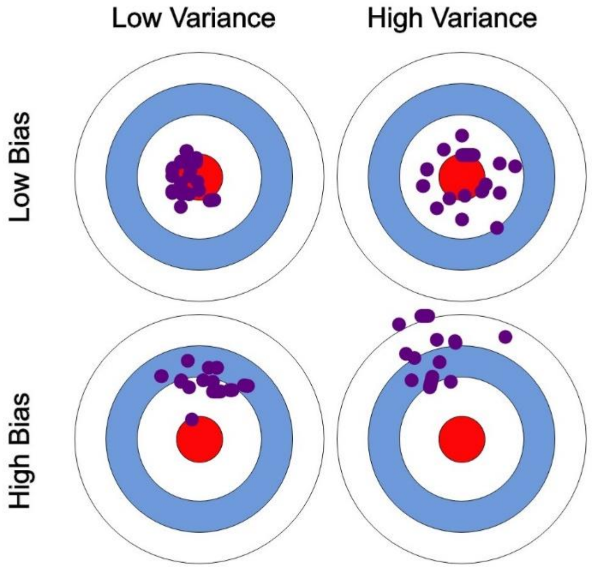
i. Relatedness is an important component of prediction accuracy

1. If the testing population is more related to the training population than the selection candidates, then the estimate of prediction accuracy will be inflated
2. Achieved accuracy may be significantly lower than within-family accuracy if individuals in selection candidates do not share pedegrees

ii. Overfitting the data

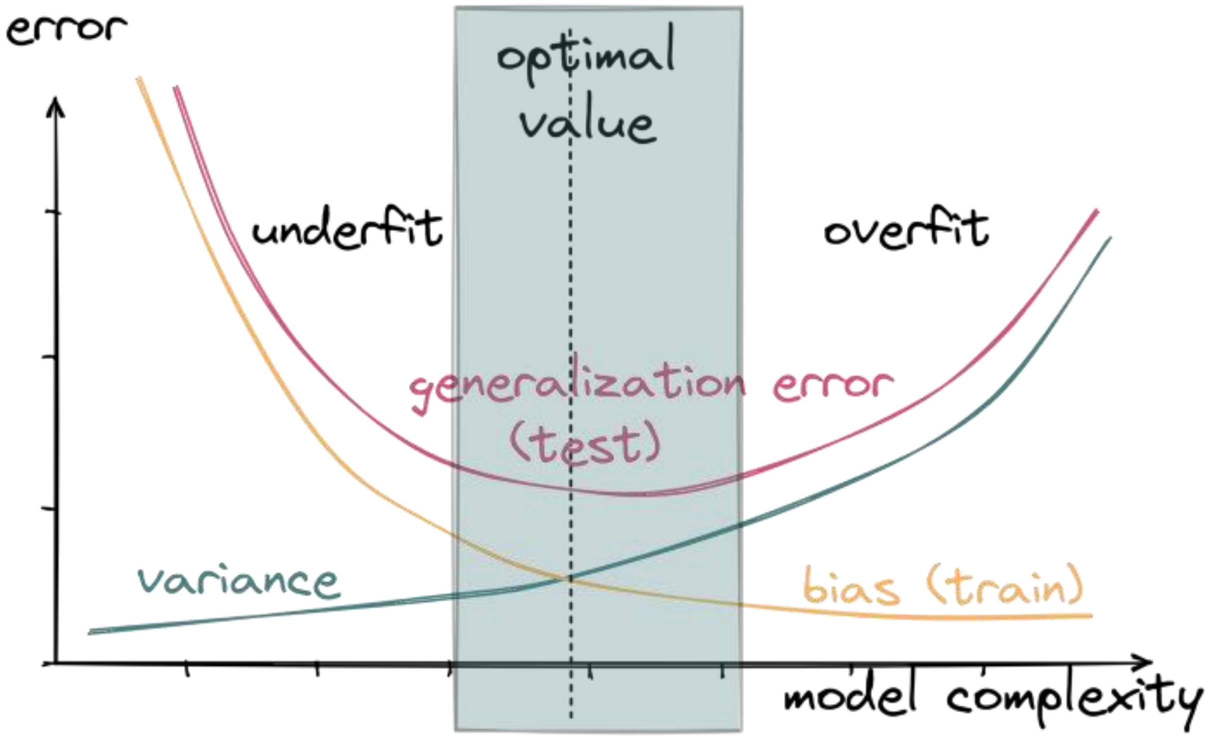
iii. Bias-Variance Trade-off

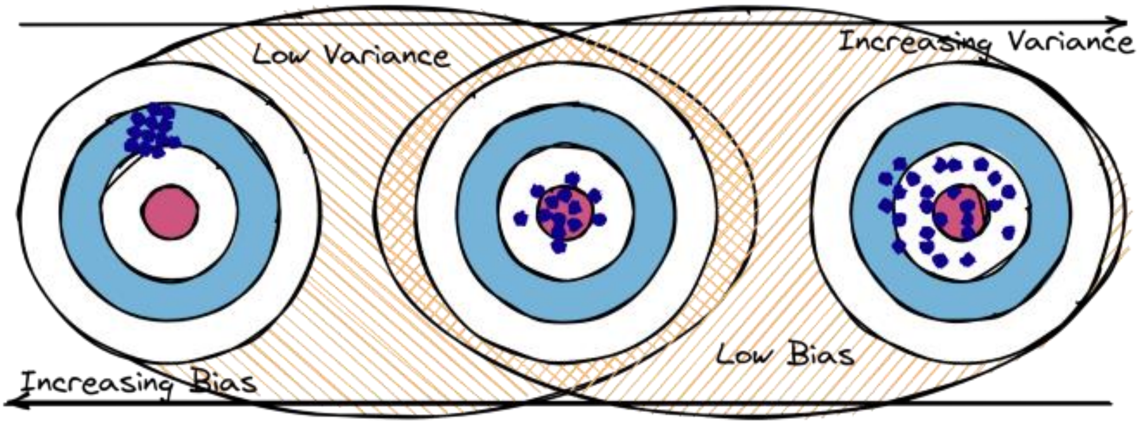
1. "The two variables to measure the effectiveness of your model are bias and variance."
2. Bias is the error or difference between points given and points plotted on the line in your training set. (Sums of squares of the points to the regression lines in the training data set)
3. Variance is the error that occurs due to sensitivity to small changes in the training set (Sums of squares of the trained regression to the test-set points)
4. https://nvsyashwanth.github.io/machinelearningmaster/bias-variance/



- High variance high bias – chose a different model to predict
- High bias low variance – the overall variation in the training set must have been much smaller than in the testing set, run cross-folds

- Low bias high variance – the variance in the testing set might be higher than in the training set – change composition of test/train sets.
- Low bias low variance – this is what we ideally want





https://medium.com/@ivanreznikov/stop-using-the-same-image-in-bias-variance-trade-off-explanation-691997a94a54

Fix vs random effects

A fixed factor assumes that the levels are separate, independent, and not similar. A random effect assumes the levels come from a distribution of levels and while they each have their own independent estimates, they are assumed to be related and exchangeable.

Random effects

- In a random effect each level can be thought of as a random variable from an underlying process or distribution.

- given levels are representative levels from in larger collection of levels (which might not even exist)
- The way to think about random effects is that each level of the effect could be considered a draw from a random variable.
- extend beyond the data and can even be used predictively
- Random effects models are computational highly expansive, so not commonly chosen
    - Iterative solving, where new solutions are based on old and improve the overall fit, can overcome this problem partly
- Fixed effects – different mean of levels, but the variance is equal across all levels
    - Random effects have an unique mean and variance for each level

- Marker effects as random:
    - the marker has different effects across populations because it is on feeble LD with some QTL
    - the "true" effect of the marker may change all the time, because at each generation LD will be different
- Marker effects as fixed
    - High variance in the test population will result in high deviations from the expected mean for some individuals in the population, which goes along with a high error in predictions
        - Using shrinkage methods like ridge regression or lasso, the (marker) effects are slightly underestimated but large exaggerations never happen and so, the mean square error is minimized.

- Maximum likelihood
    - Find the optimal way to fit a distribution to the data
    - Maximum likelihood of the mean = maximized likelihood of observing the measurements distribution (in a distribution with mean x and variance y)

- **Transforming marker variance into genetic variance**

$$\sigma_u^2 = 2\sigma_{a0}^2 \sum_i^{nsnp} p_i q_i$$

    -
    - We can utilize this fact to predict the (genomic) heritability
        - According to literature, estimates of genetic variance with pedigree or markers usually agree up to, say, 10% of difference to phenotyping variance (multiple locations, replications, years, etc.).