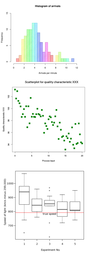
Phenotypic Data

Exploratory Data Analysis

1. Distribution of phenotypic data



Is the data representative of the total population? Are the mean and the distribution of the phenotypes as expected? (histogram, mean, variance, …)

- Check accuracy of phenotype measurements (if possible).
- Additional measurements may be necessary to represent the population.
- Any outliers (also boxplot).

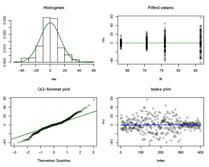
Which trend does the data follow? (scatterplot)

- Continuous data often fits a linear trend (linear regression).
- Binary data often fits a sigmoid trend (logistic regression).

Are there any cofactors we need to correct for? (boxplot by groups, linear mixed model comparison based on the residual term, log likelihood ratio, AIC, BIC, …)

- Sex.
- Herd effect, fish tank, dog breeder, field effect, … (environmental effect).
- Date of measurement.
- …

2. Distribution of the residuals (error term)



Do the residuals follow a normal distribution?

- Linear regression requires normality of the error. If not, something might be wrong with the data…
- Transformation might help, but will complicate interpretation (use carefully).

Can we assume variance homogeneity within groups?

"Real-world" data is never perfect and the models we use are robust. An approximately normally distributed data set is fine.

Genotypic Data

Quality control

Detect SNPs and samples that should be removed prior to GWAS.

1. Missing marker rate

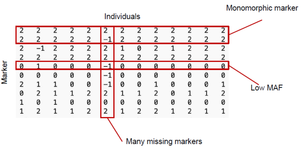
- Per-sample (2-10% missing SNPs per individual).
- Per-site (2-10% missing values per variant/SNP).
- Stricter/looser thresholds depending on data/experiment (GBS might require less stringent thresholds. Check literature for reference values).

2. Minor Allele Frequency (MAF)

- Remove monomorphic variants → non-informative.
- Remove variants at low frequency ("rare") → spurious associations.
- 1-5% MAF.
- Usually (re)done after imputation.

3. Other filtering criteria

- Sex chromosomes (might need to be removed / analyzed separately).
- Relatedness between samples (check for duplicates).



Retain approximately 75-80% (or more) of the SNPs. Use of some metrics and thresholds is species- or population-specific. Thresholds always underlie some level of subjectivity.

Genotypic Data

Imputation

Preliminary step for a wide range of genetic analyses
Most models and software for GWAS and other methods used in quantitative genetics / biostatistics methods do not handle missing data by default.

Mean substitution

- Replacement of missing values using the mean of the respective SNP across the population.
- Implemented in many GWAS and genomic selection packages.
- Simple and fast, but inaccurate.

Beagle

- Software made for phasing and imputation of genotypic data.
- LD-based approach (Hidden Markov Model; HMM).
- Very efficient and accurate using default settings.
- Other software efficient software solutions are available, but might require phasing prior to imputation using a different software package.

K-Nearest Neighbor Imputation (KNNI)

- General imputation method, applicable to any type of data (including genotypes).
- Using a similarity matrix between samples from a distance function based on available data.

SNP Marker tested for association with trait
(fixed effect)

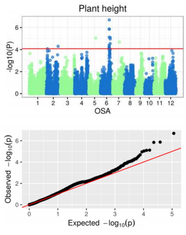
Genomic relationship matrix to correct for family structure
(random effect)

GWAS

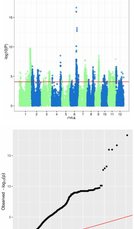
$$y = X\beta + S\alpha + Qv + Zu + e$$

Fixed effects
(other than SNP under testing and population structure)

Population structure
Subpopulation effect (fixed effect)
(Could also be calculated by PCA of the genomic relationship matrix)



- **Manattan plot shows a nice peak**
- **qq-plot looks good**
- **Publish study!**



- **Redo GWAS and correct for population structure**