HELMHOLTZ
MUNICH

Statistics for Human Genetics

Ozvan Bocher (slides from Arthur Gilly)
November 22, 2022

What can we do with Statistics ?

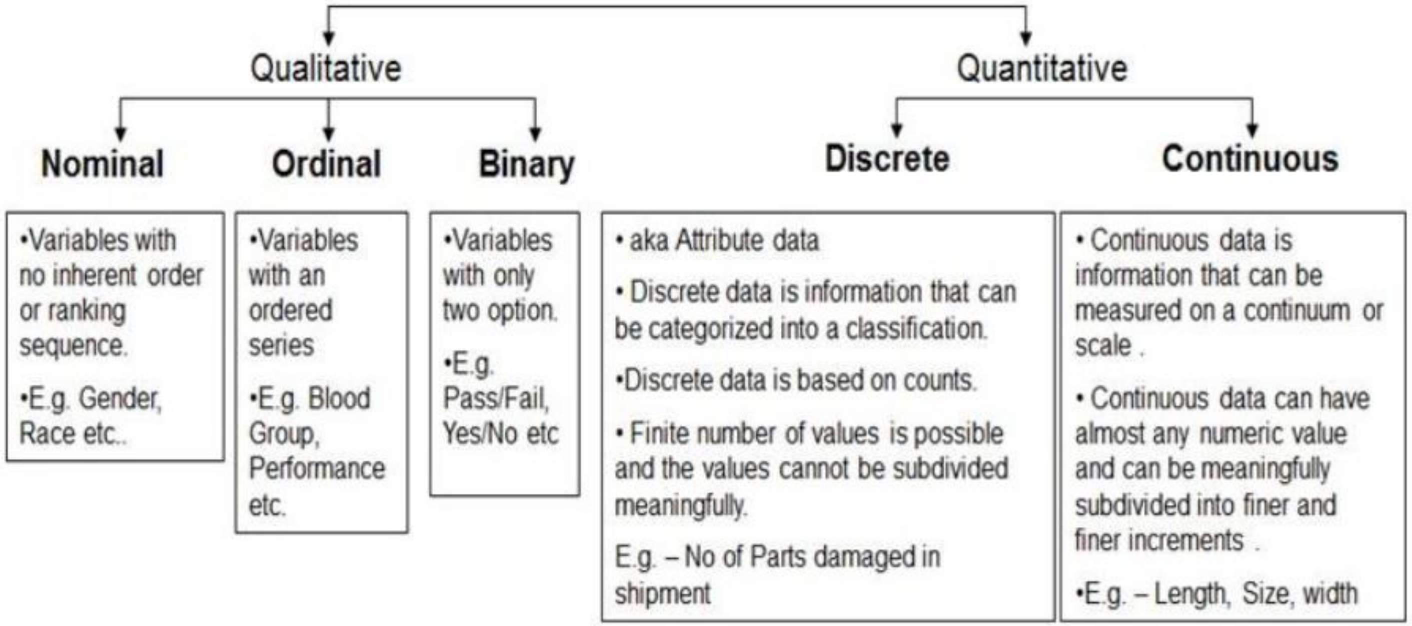
1. Estimation
2. Modelling
3. Hypothesis testing
4. Predicting

1

Random variables and estimation

Random Variables

- In statistics, we measure the realizations/observations of random variables
- Often, these random variables follow a distribution
- They can be qualitative or quantitative (continuous or discrete)



Qualitative

Quantitative

Nominal

- Variables with no inherent order or ranking sequence.
- E.g. Gender, Race etc..

Ordinal

- Variables with an ordered series
- E.g. Blood Group, Performance etc.

Binary

- Variables with only two option.
- E.g. Pass/Fail, Yes/No etc

Discrete

- aka Attribute data
- Discrete data is information that can be categorized into a classification.
- Discrete data is based on counts.
- Finite number of values is possible and the values cannot be subdivided meaningfully.

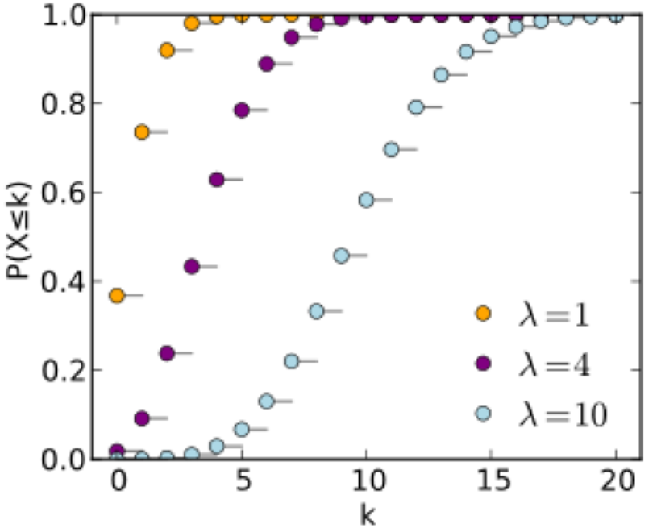
E.g. – No of Parts damaged in shipment

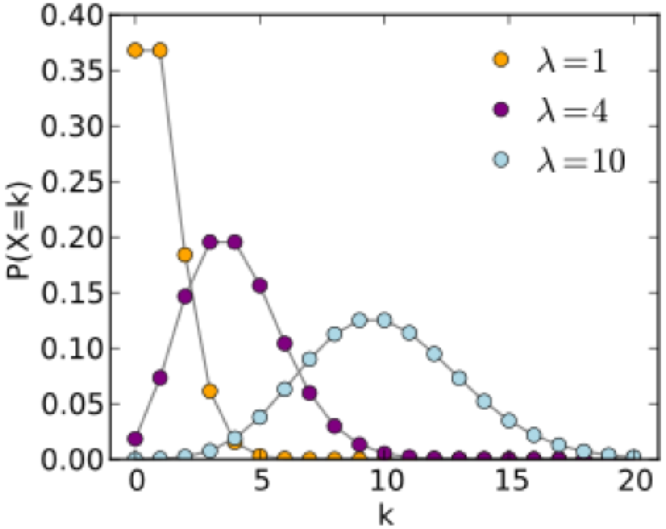
Continuous

- Continuous data is information that can be measured on a continuum or scale .
- Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
- E.g. – Length, Size, width

Distributions

- Two ways to represent them:





Cumulative distribution function (CDF)

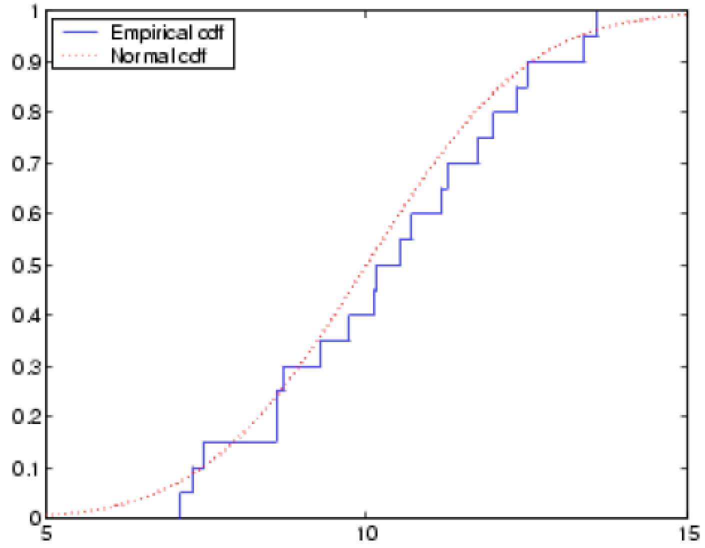
- $y = p(X \leq x)$
- Always growing
- Ideal way to represent but hard to read
- All distributions look the same

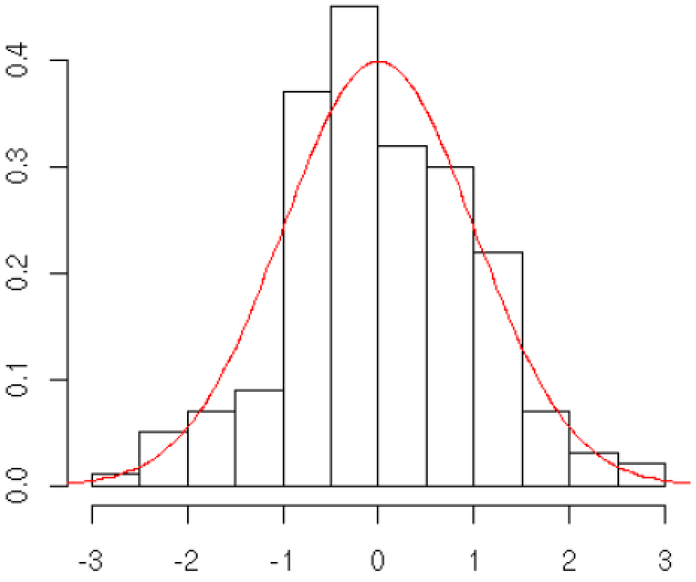
Probability density function (PDF)

- $y = p(X = x)$ for discrete
- Shows how values are distributed
- Nice visually but hard to deal with

Distributions

- How to estimate them:





Empirical CDF

- Rarely used
- Except when you want to compute empirical quantile functions

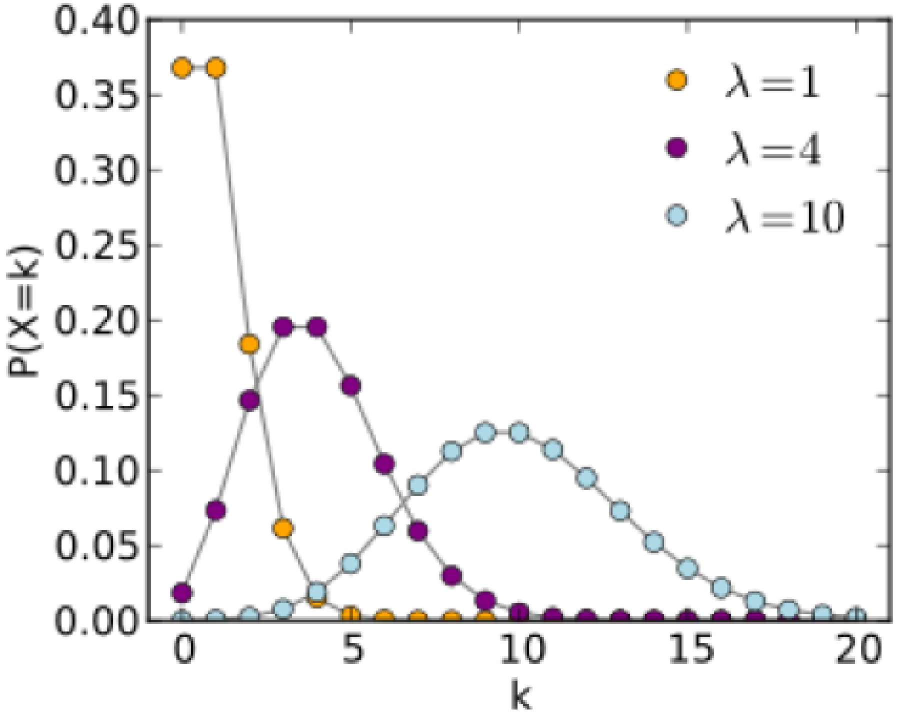
Barplot (discrete)

- For every value, count occurrences

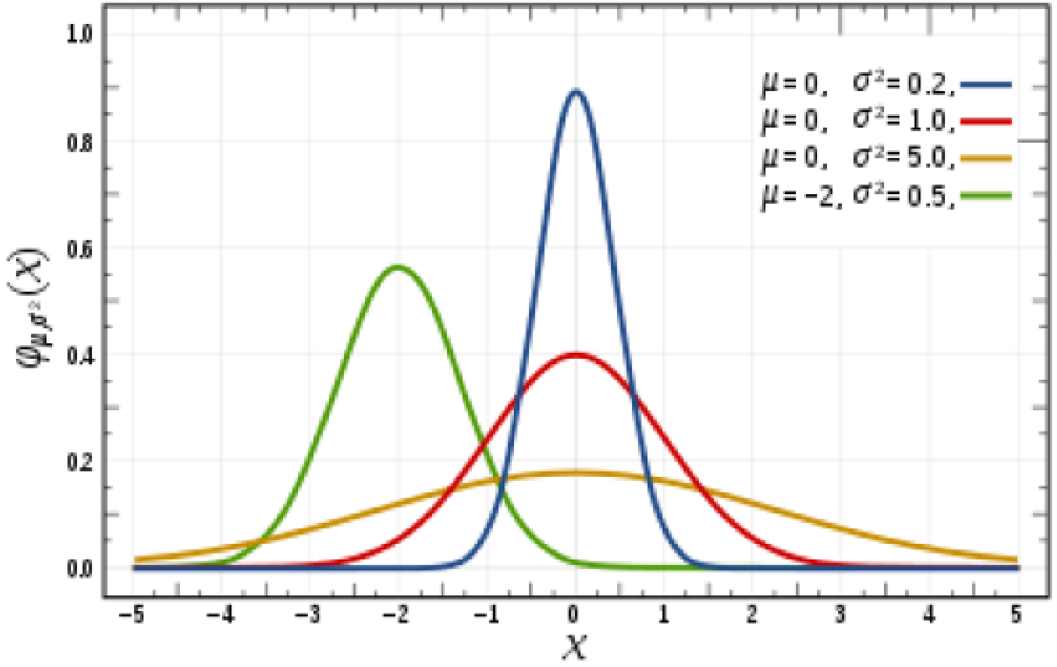
Histogram (continuous)

- Cut the interval into bins and count observations within bins

Distributions – random variables (real world data)

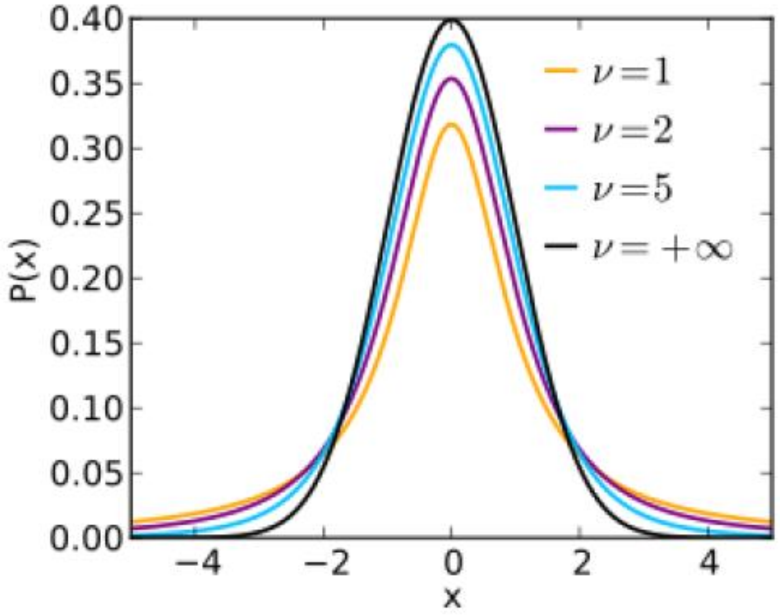


$X \sim Poisson(\lambda)$

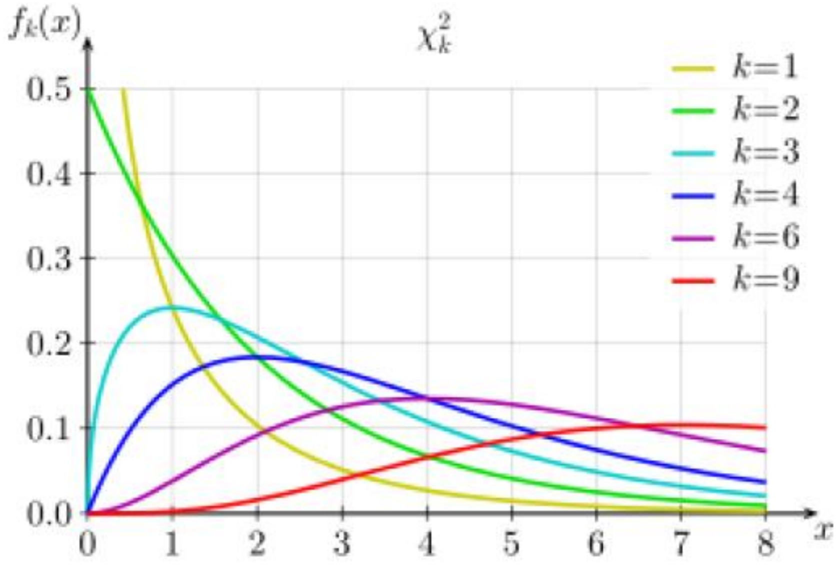


$X \sim \mathcal{N}(\mu, \sigma^2)$

Distributions – tests statistics



$X{\sim}T(\nu)$



$X{\sim}\chi^2(k)$

λ, μ, σ, ν and k are the ideal, theoretical parameters
→ Use estimations from the data to approximate them

Statistics

- A statistics is a meaningful quantity derived from the data
- Often, estimators are realizations of distribution parameters
- Examples include mean, proportion, …
- For simple distributions/parameters, there is a formula
- For more complex ones, we have to use other techniques (Monte-Carlo, Permutations, …)

$$(\hat{\mu} =)\ \bar{x} = \frac{1}{n}\sum_{k=0}^{n} x_k$$

$$\hat{p} = \frac{x}{n}$$

$$w = \frac{\left(\hat{\theta} - \theta_0\right)^2}{se(\hat{\theta})} \sim \mathcal{N}(0,1)$$

$$\left(\widehat{\sigma^2} =\right)\ s^2 = \frac{1}{N-1}$$

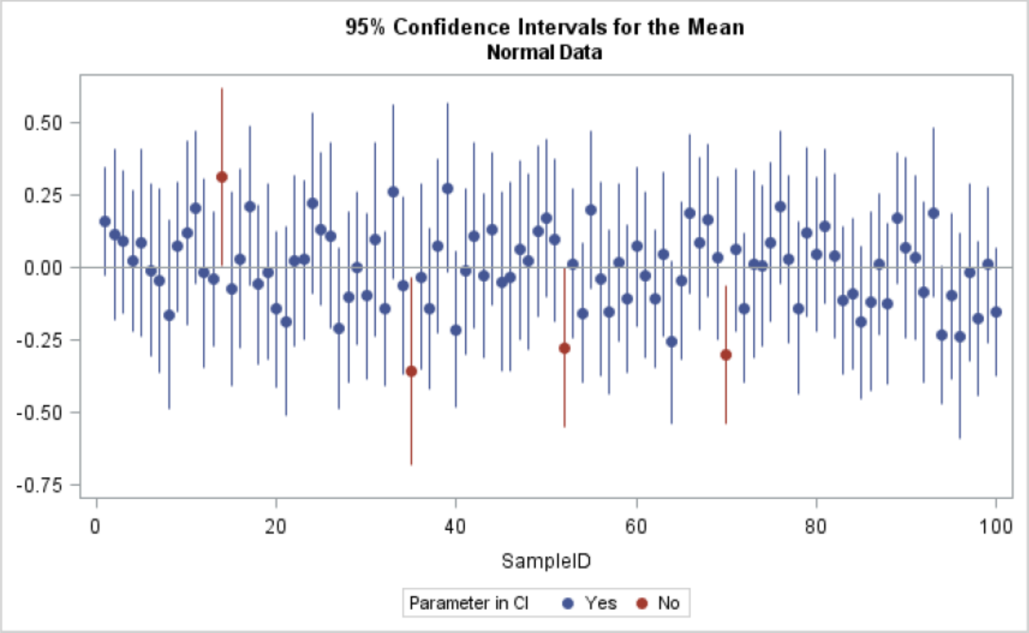
Confidence intervals

- x% confidence interval (x%C.I.):

  → x% of the time when this interval is calculated, it will contain the true value of the parameter

  → The true value of the parameter has x% chances to be in the x%C.I.

- Often 95%C.I. used corresponding to the classical α level



HELMHOLTZ MUNICH

2

Modelling

Modelling

- Estimate the effect of one variable on another variable

$$\boldsymbol{phenotype} \sim \boldsymbol{\beta} \times \boldsymbol{genotype} + \epsilon$$

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix} \qquad \begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix} \qquad \begin{bmatrix} 1 \\ \vdots \\ 2 \end{bmatrix}$$

$= \{0,1\}$ (case-control)

$\in \mathbb{R}$ (quantitative) $\sim \mathcal{N}(0,1)$

$= \{0,1,2\}$ (genotype, directly typed)

$\in [0,2]$ (dosage, imputed) $\begin{bmatrix} 0.965 \\ \vdots \\ 1.816 \end{bmatrix}$

- What is the effect of the genotype on the phenotype ?
  - ➢ β estimations for continuous phenotypes
  - ➢ Odds Ratio (OR) for binary phenotypes. Ex: Case/Control studies

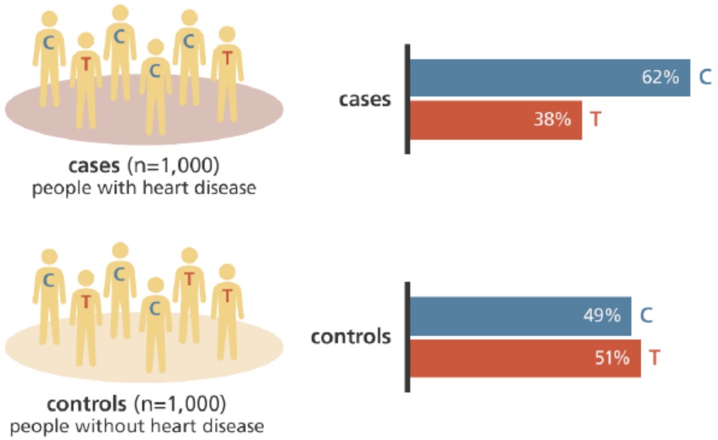
2

Modelling

3.1 – Case/Control

Case/control studies

- OR: *how much more likely are you to be a case if you carry the risk allele ?*
  - Per genotype *g* and disease *Y*, we compute the odds $O = \frac{p}{1-p} = \frac{p_{Y=1|g}}{1-p_{Y=1|g}}$



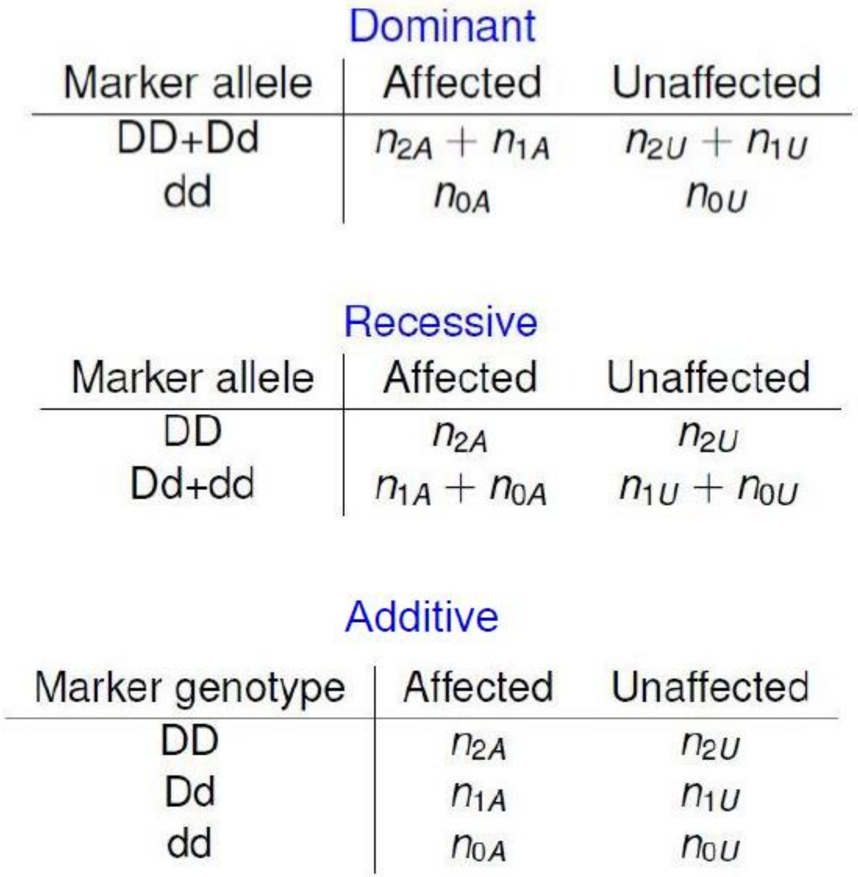
	Cases	Controls	N
T	380	510	890
C	620	490	1100

$$O_T = \frac{380/890}{510/890} \qquad O_C = \frac{620/1100}{490/1100}$$

- OR: Ratio of the odds of the two alleles
  - OR>1: the allele is ‘deleterious’
  - OR<1: the allele is ‘protective’

$$\boldsymbol{OR_{C/T} = \frac{620 * 510}{490 * 380} = 1.70}$$

Case/control studies



Dominant

Marker allele	Affected	Unaffected
DD+Dd	$n_{2A} + n_{1A}$	$n_{2U} + n_{1U}$
dd	n_{0A}	n_{0U}

Recessive

Marker allele	Affected	Unaffected
DD	n_{2A}	n_{2U}
Dd+dd	$n_{1A} + n_{0A}$	$n_{1U} + n_{0U}$

Additive

Marker genotype	Affected	Unaffected
DD	n_{2A}	n_{2U}
Dd	n_{1A}	n_{1U}
dd	n_{0A}	n_{0U}

$$\mathrm{OR} = \frac{(2 \times n_{2A} + n_{1A}) \times (2 \times n_{0U} + n_{1U})}{(2 \times n_{0A} + n_{1A}) \times (2 \times n_{2U} + n_{1U})}$$

Allelic odds-ratio

$$\boxed{OR = \frac{n_{affected\ carriers} \times n_{healthy\ non-carriers}}{n_{healthy\ carriers} \times n_{affected\ non-carriers}}}$$

	Cases	Controls
T	380	510
C	620	490

$$OR_{C/T} = \frac{620 * 510}{380 * 490}$$

Case/control studies

- Output: OR and 95% confidence interval of the OR
- Association test: is it significantly different from 1 ?
  - ➢ $H_0: OR = 1$
  - ➢ $H_1: OR \neq 1$
- Statistics: Fisher’s exact test or Chi-squared
- In case of dosages or if covariates are included: logistic regression

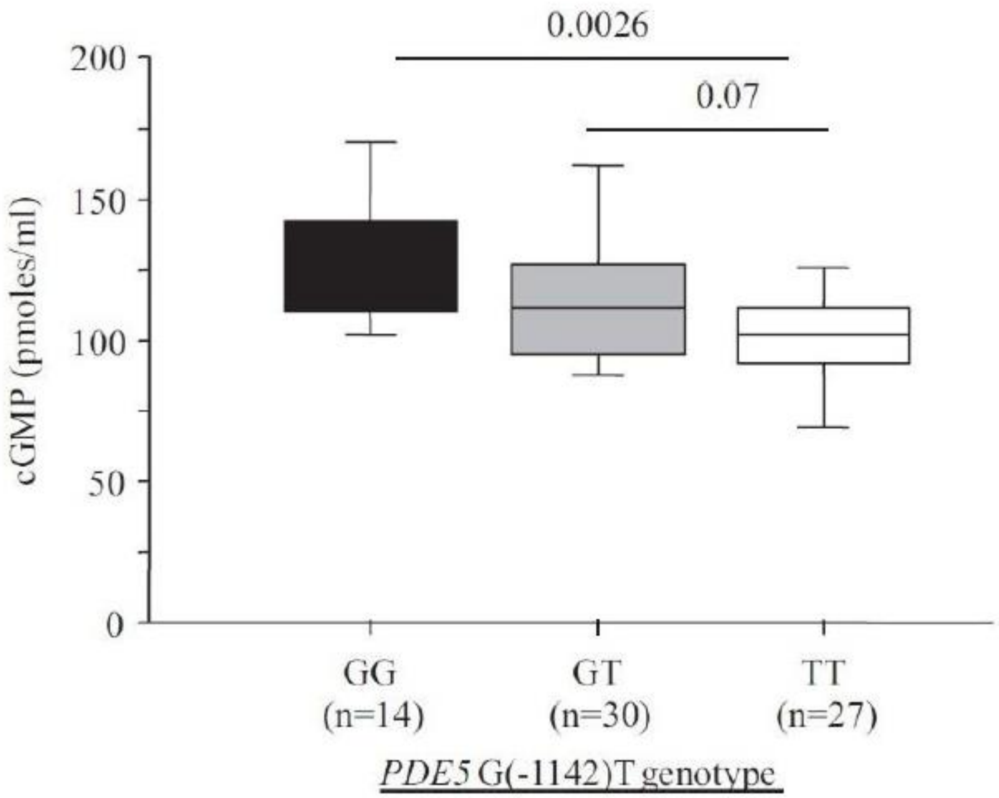
2

Modelling

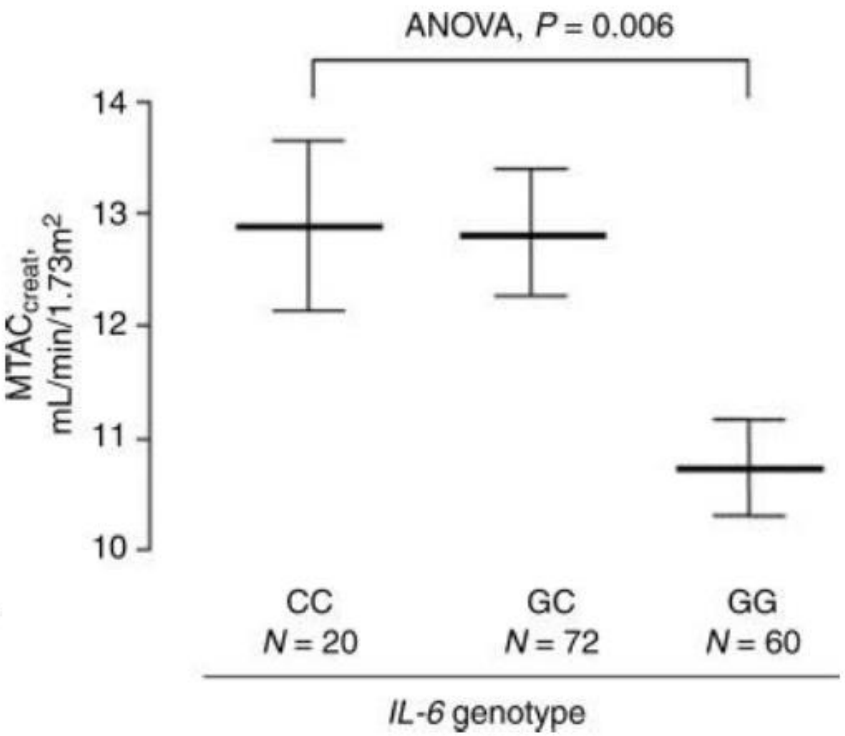
3.2 – Continuous traits

Continuous traits

- If directly typed genotypes (0, 1, 2) are analyzed: ANOVA



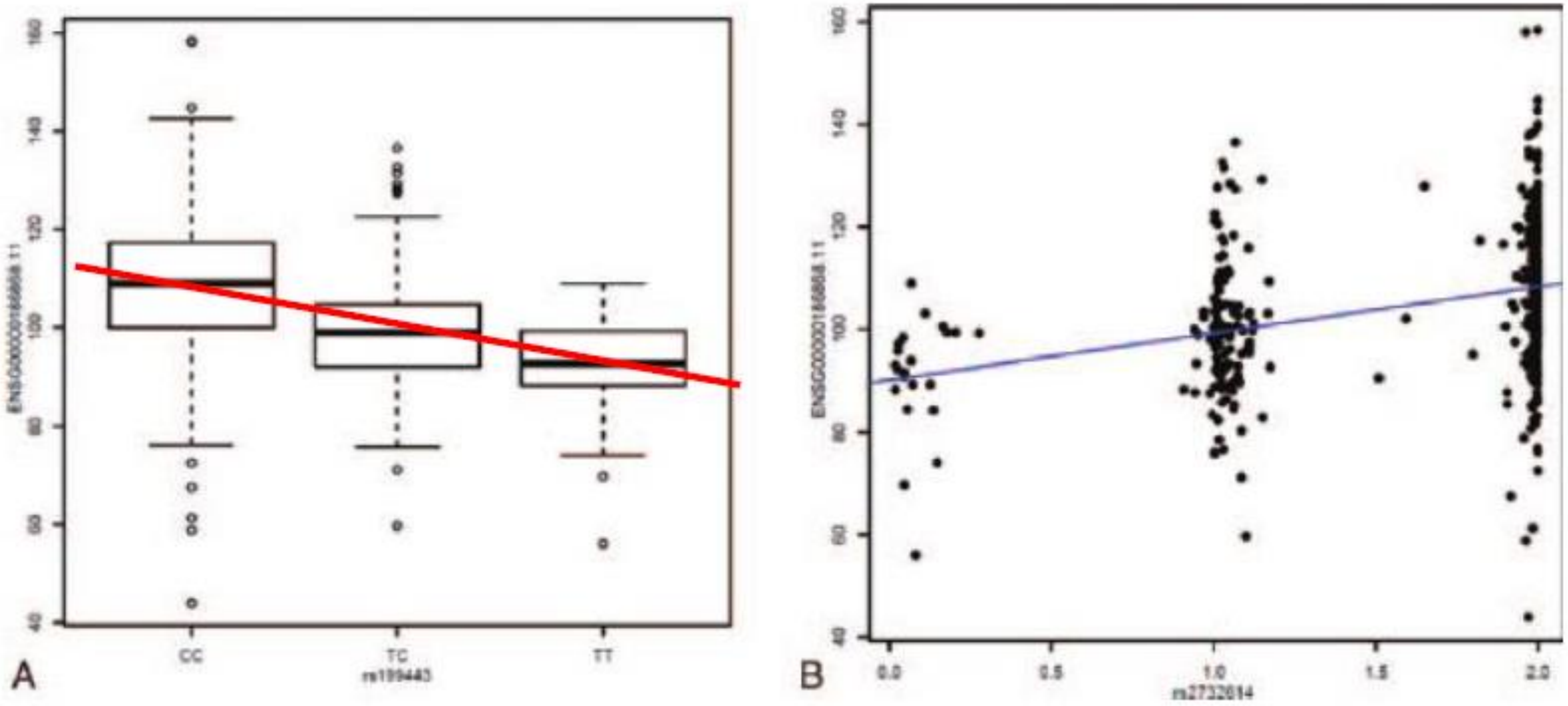
Additive



Recessive

Continuous traits

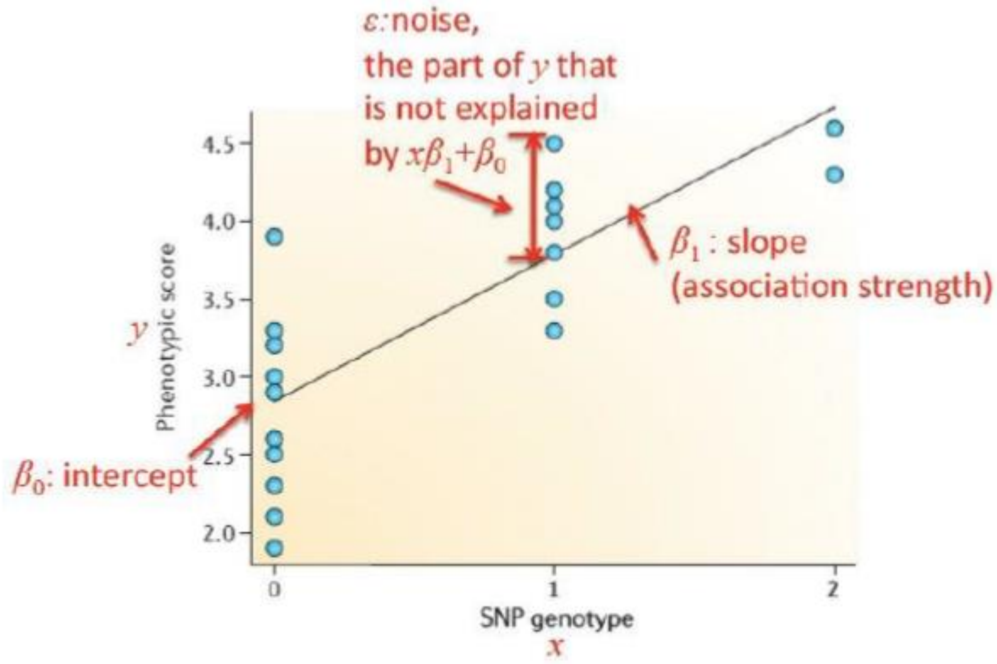
- If dosages are analyzed (imputed quantity of minor allele $d \in [0,1]$): linear regression
- In general: generalized linear models



Continuous traits

- A linear regression model is defined as:

$$y = x\beta_1 + \beta_0 + \varepsilon$$



- Data:
  - y is a continuous trait
  - x is the SNP genotype at a given locus
- Parameters:
  - $\boldsymbol{\beta_1}$ is the regression coefficient, represents the strength of association between y and x
    - ➢ $\boldsymbol{\beta_1} > 0$: for every supplementary allele, the phenotype will increase by the beta coefficient value
    - ➢ $\boldsymbol{\beta_1} < 0$: for every supplementary allele, the phenotype will decrease by the beta coefficient value
  - β_0: intercept term (is often ignored)
- Assumptions:
  - The individuals in the study are not related
  - The phenotype y has a normal distribution

3

Hypothesis testing

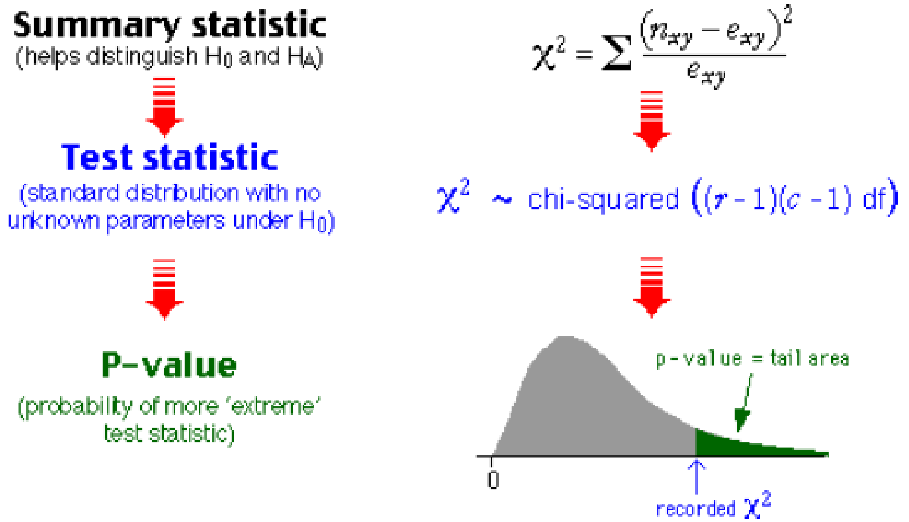
3.1 – Statistical tests

Hypothesis testing

- Measure whether the data gives sufficient evidence to reject a hypothesis

Null hypothesis H_0 vs Alternative hypothesis H_A (H_1)

- H_1 = Hypothesis of interest
- Use a statistic that follows a certain distribution under H_0
  - *Name of the test = name of the statistics*
  - *Can we reject H_0? Not rejecting H_0 is different from proving it !*



Summary statistic
(helps distinguish H_0 and H_A)

$$\chi^2 = \sum \frac{(n_{xy} - e_{xy})^2}{e_{xy}}$$

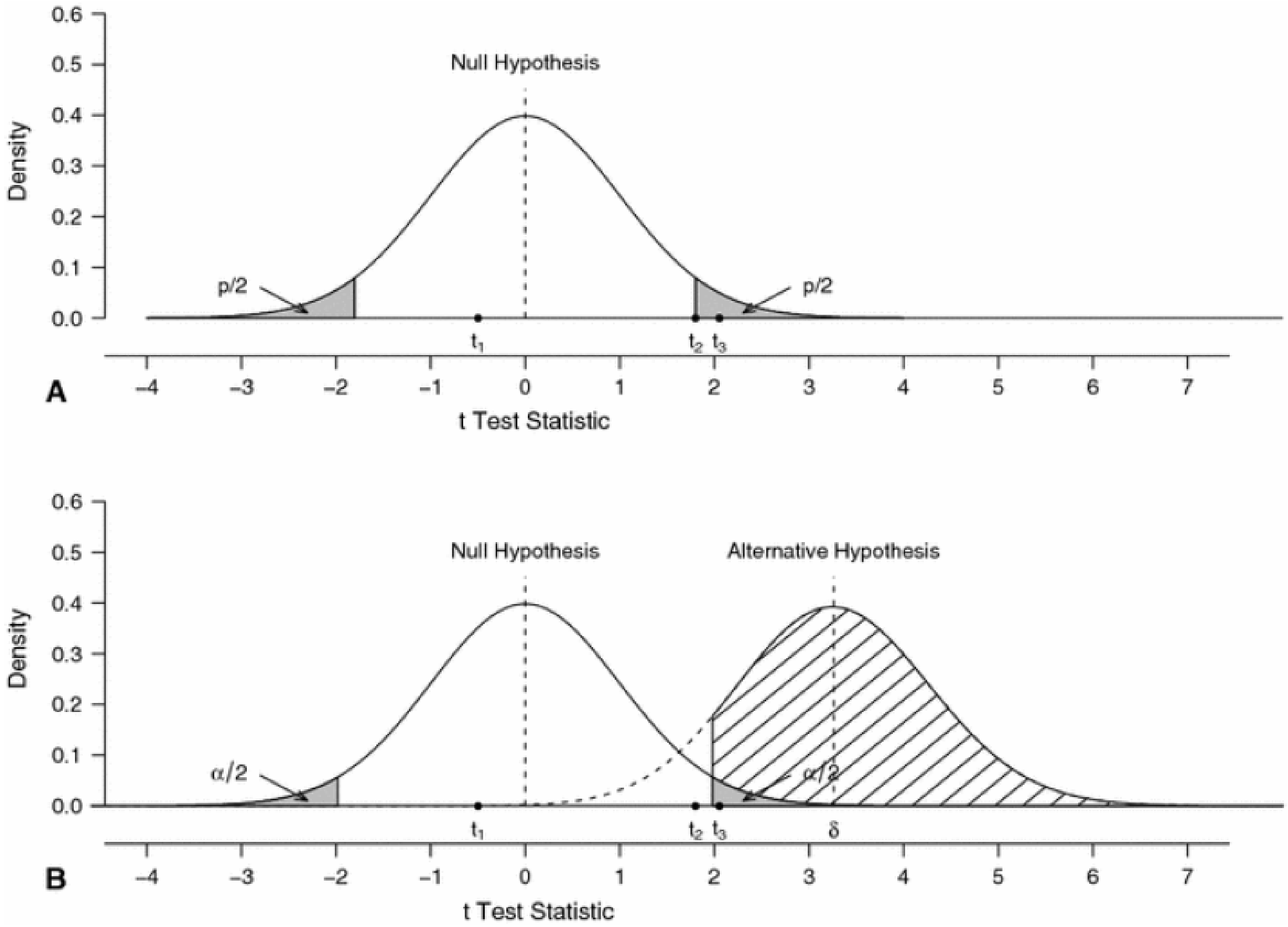
Test statistic
(standard distribution with no unknown parameters under H_0)

χ^2 ~ chi-squared $((r-1)(c-1)$ df$)$

P-value
(probability of more 'extreme' test statistic)

- We calculate the statistics based on our data
- As we know the distribution, we can compute the CDF $p(X \leq x)$
- Decision based on a significance threshold α and the p-value = how likely the measurement comes from the null
  - *If $p < \alpha$ → we reject H_0 and consider the test significant*

H_0 vs H_1

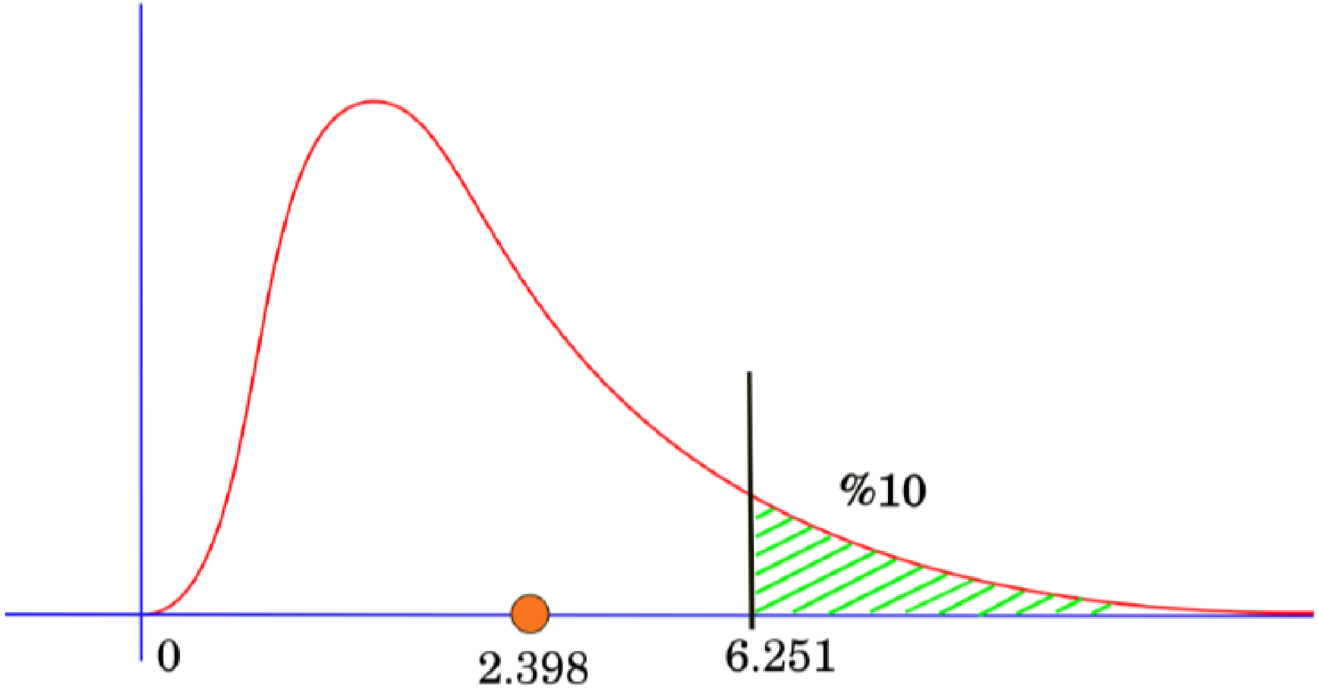


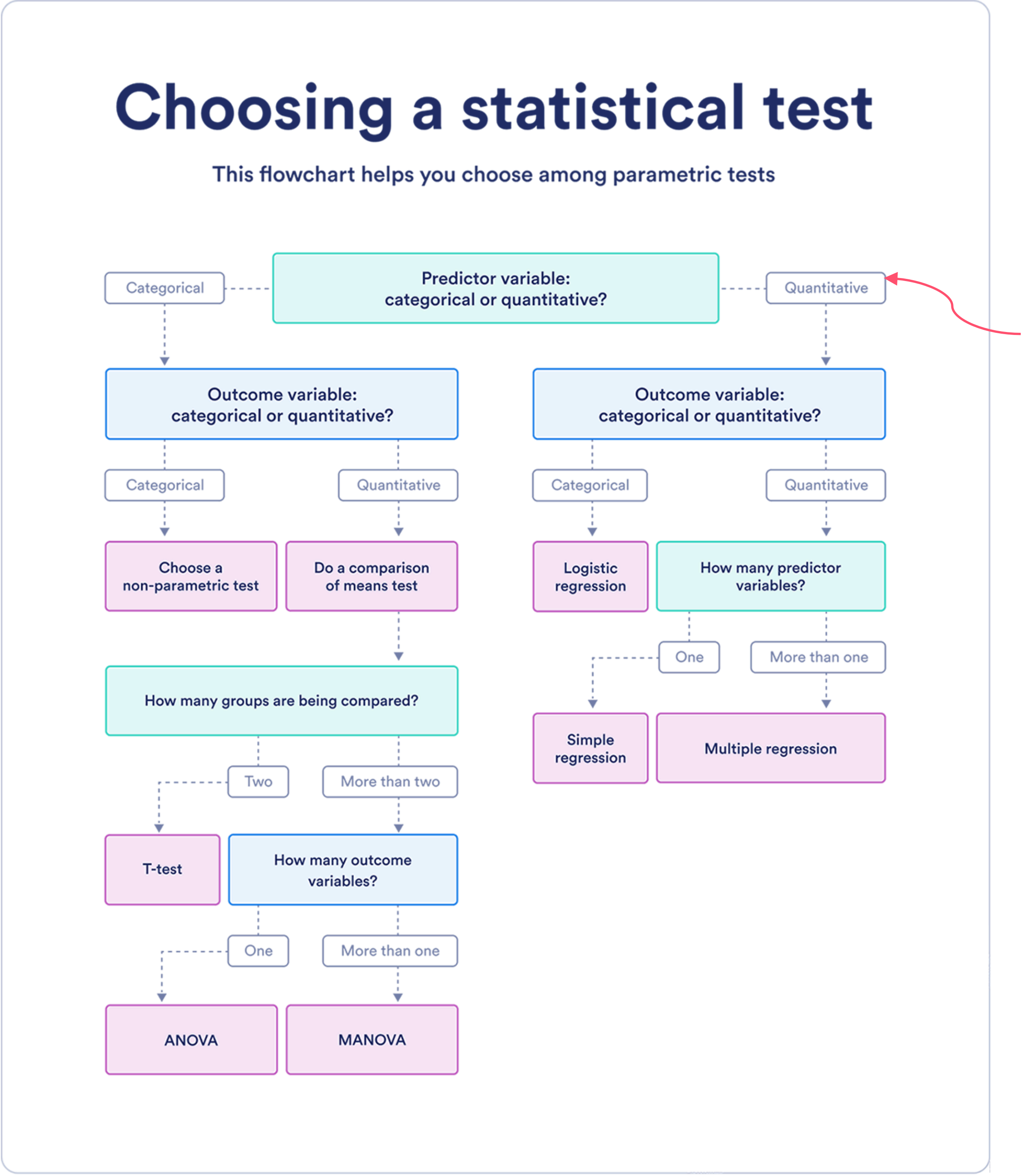
HELMHOLTZ MUNICH

One-sided vs Two-sided test

- Depends on the hypothesis

Two-sided:	$\boldsymbol{H_0}$: " = "	vs.	$\boldsymbol{H_1}$: " ≠ "
One-sided:	$\boldsymbol{H_0}$: " ≤ “	vs.	$\boldsymbol{H_1}$: " > "
	$\boldsymbol{H_0}$: " ≥ "	vs.	$\boldsymbol{H_1}$: " < "





Choosing a statistical test

This flowchart helps you choose among parametric tests

Predictor variable: categorical or quantitative?
Categorical
Quantitative
Outcome variable: categorical or quantitative?
Categorical
Quantitative
Choose a non-parametric test
Do a comparison of means test
How many groups are being compared?
Two
More than two
T-test
How many outcome variables?
One
More than one
ANOVA
MANOVA
Outcome variable: categorical or quantitative?
Categorical
Quantitative
Logistic regression
How many predictor variables?
One
More than one
Simple regression
Multiple regression

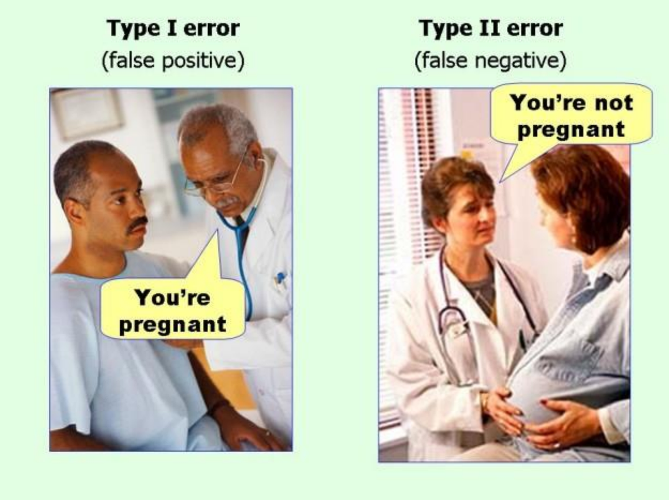
GWAS are performed under an additive model

3

Hypothesis testing

3.2 – Multiple testing

Multiple testing

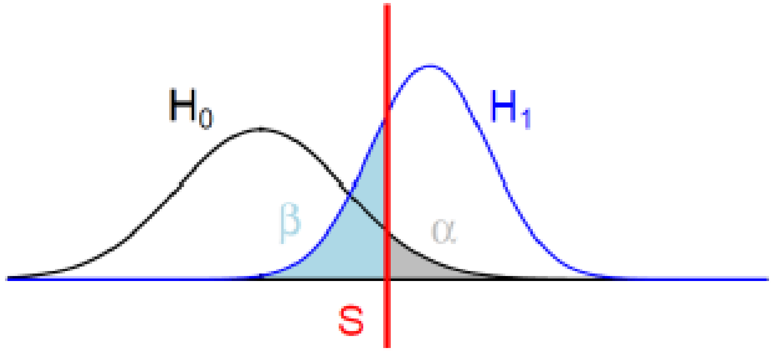


- *If $p < \alpha$ → we reject H_0 and consider the test significant*
- α corresponds to the type I error risk that we want to control

	***H_0* is true**	***H_1* is true**
***H_0* not rejected**	Correct decision	Type II eror
***H_0* rejected**	**Type I error = α**	Correct decision

- If the number of tests increases, the risk increases

→ Need to take into account the multiple tests to maintain α at the desired level across all tests



Multiple testing: m tests

Family-Wise Error Rate (FWER)

- Bonferroni correction
- Simple to implement, harder to interpret

$$p_{critical} = \frac{0.05}{m}$$

- *“If all tests are under the null, probability that* ***one or more*** *of them is a false positive.”*

False-discovery based (FDR)

- Benjamini Hochberg procedure
- Harder to implement, easy to understand

$$p_{critical} = argmax(p < \frac{i}{m}Q)$$

- *i = rank, Q = FDR*
- *“Proportion of significant tests that are false positives.”*

When to use which depends on

1. Best practices
2. Relative cost of a false negative/positive

Multiple testing: application in genomic studies

- Statistical significance:
  - 5% for one test
  - Genome-wide: one test per variant and per phenotype

$$\boldsymbol{phenotype} \sim \beta \times \boldsymbol{genotype} + \epsilon$$

  - But all the variants are not independent and in reality we account for LD = correlation between the variants
  - $5\text{x}10^{-8}$ for GWAS, 10^{-9} for sequencing based

$$p_{critical} = \frac{0.05}{m}$$

Exercise : Significance threshold

- If the adjusted genome wide significance threshold is 5×10^{-8} for GWAS, how many "effective" variants are there in a genotyped human genome? **10^{6}**
- You are writing an article about a GWAS for 16 different traits. What will be your threshold for declaring significance? **3.125×10^{-9}**

3

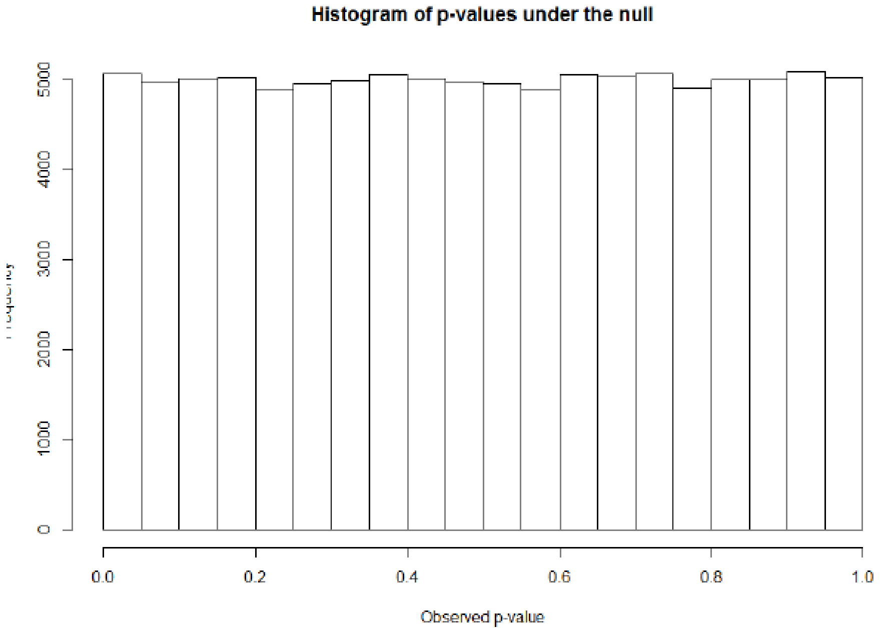
Hypothesis testing

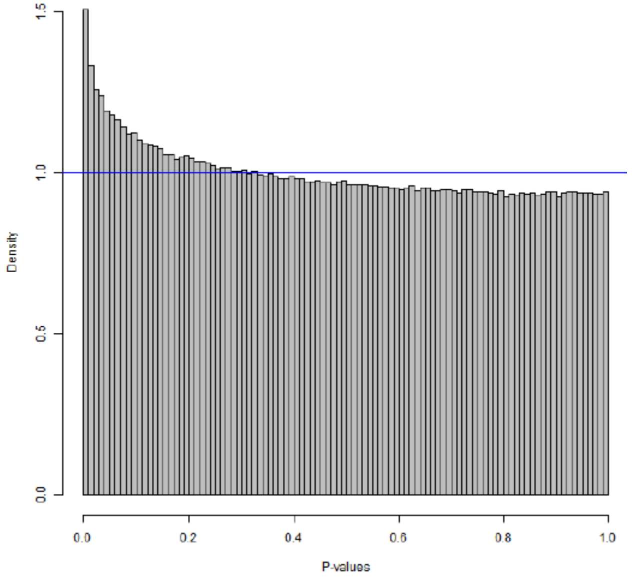
3.3 – Checking the results

QQ-plots

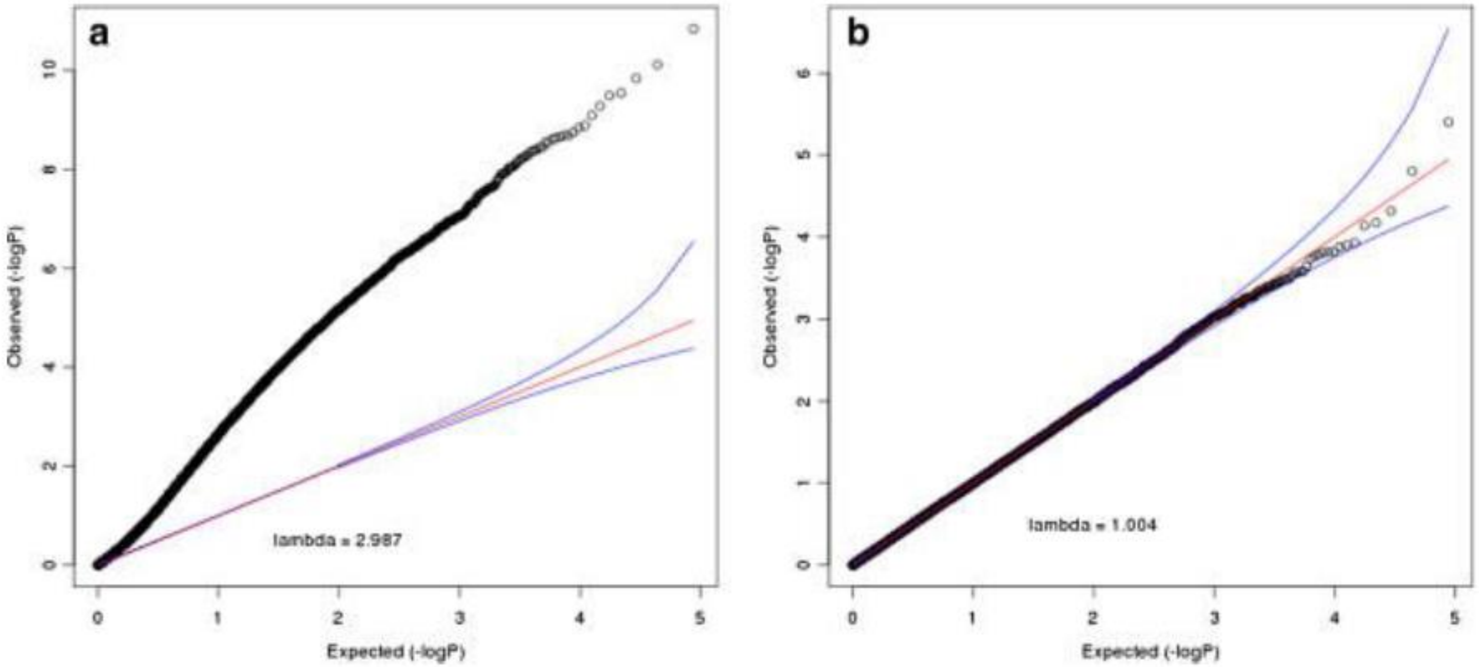
- Under H_0, p-values should be uniform [0,1]
- It is expected that most signals are around H_0

  → If we have much signal, more around 0
- Compare quantiles with expected ones: QQ-plot
- In R: *qqunif*





QQ-plots



- Inflation: too much signal
- Visual inspection but also lambda value

QQ-plots

- Appearance can be misleading:

→ A QQ-plot can seem inflated but just a lot of signal

→ And conversely…

- Calculation of the genomic inflation factor:

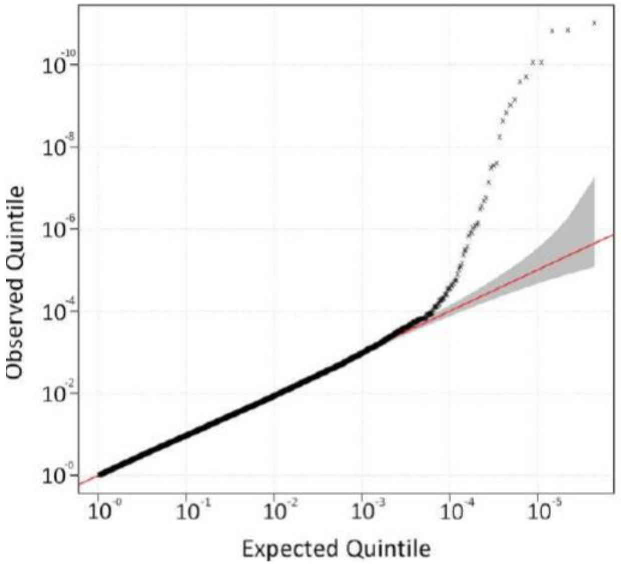
$$\lambda = \frac{median(Q_{\chi^2}(p))}{0.45}$$

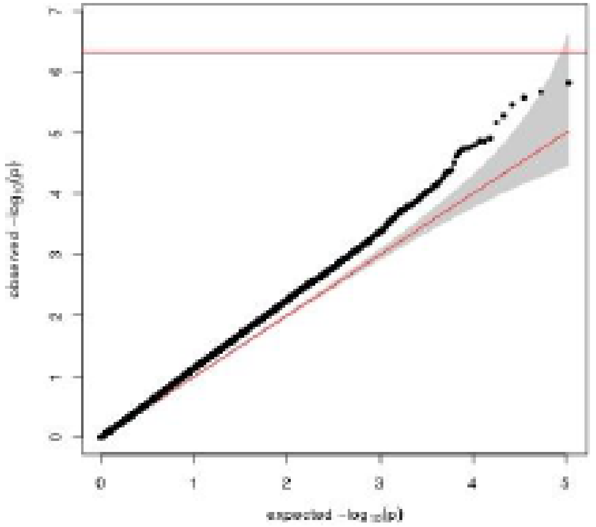
With 0.45 being the median of χ_1^2

→ $\boldsymbol{\lambda > 1}$**: inflation** (systematic bias)

→ $\lambda < 1$: deflation (potential power issue)

- Ideally, we want to correct the model
- Can also adjust: GC correction: divide by lambda

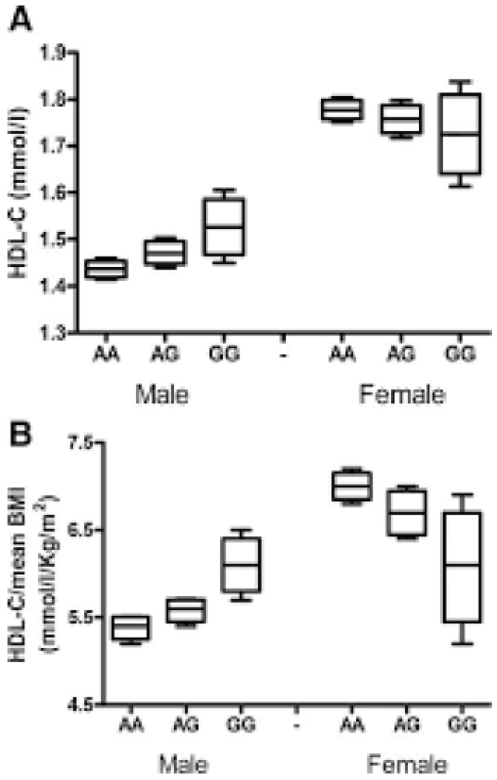




Problems in the model

- Covariates: sex, batch effects, chip effects

  → Potential bias if associated to phenotype and genotype



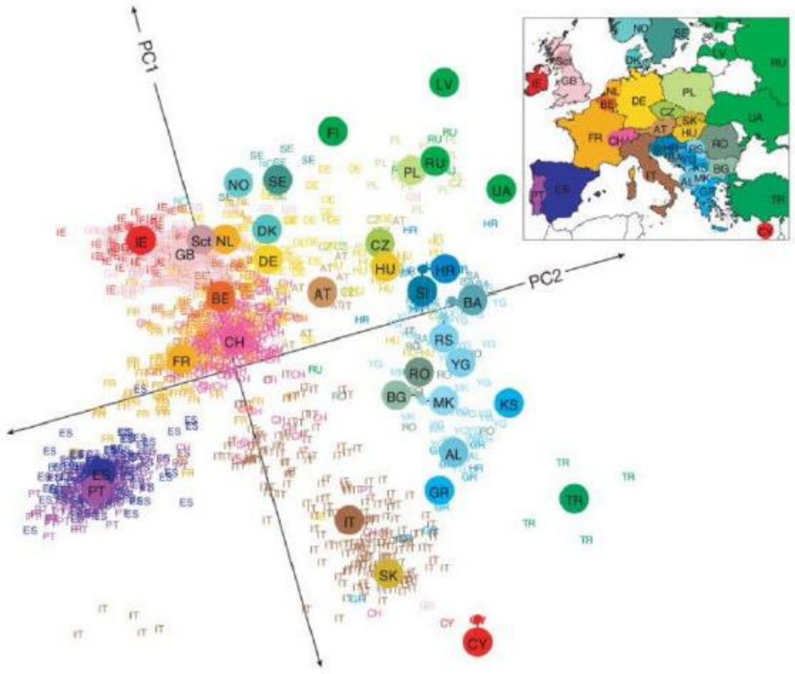
Problems in the model

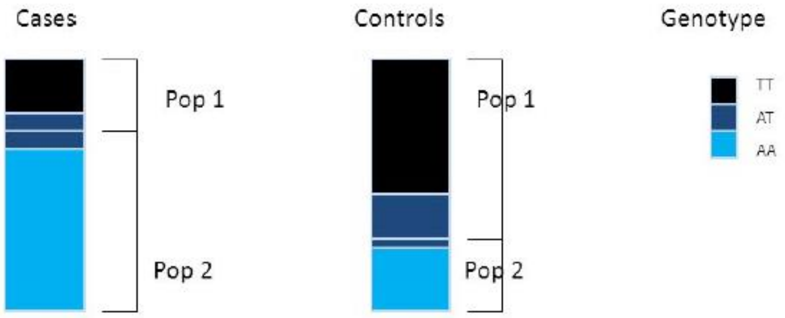
- Covariates: sex, batch effects, chip effects

  → Potential bias if associated to phenotype and genotype

- Structure or subpopulations

  → Allelic frequencies are known to be different from one population to another





Problems in the model

- Covariates: sex, batch effects, chip effects

  → Potential bias if associated to phenotype and genotype

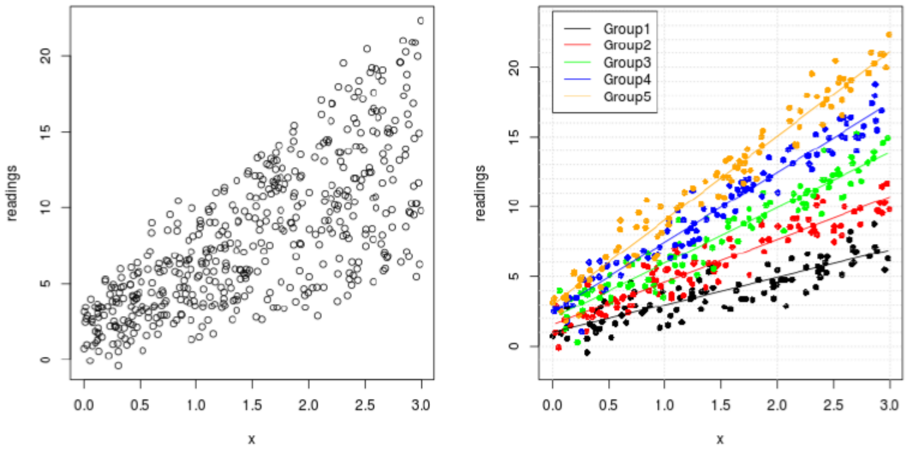
- Structure or subpopulations

  → Allelic frequencies are known to be different from one population to another

  → Linear mixed models can model the intra-group effect

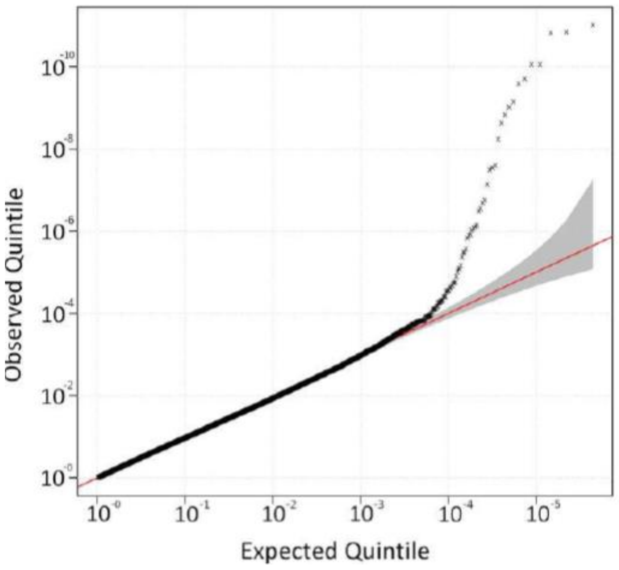
  → Adjust for principal components of a PCA

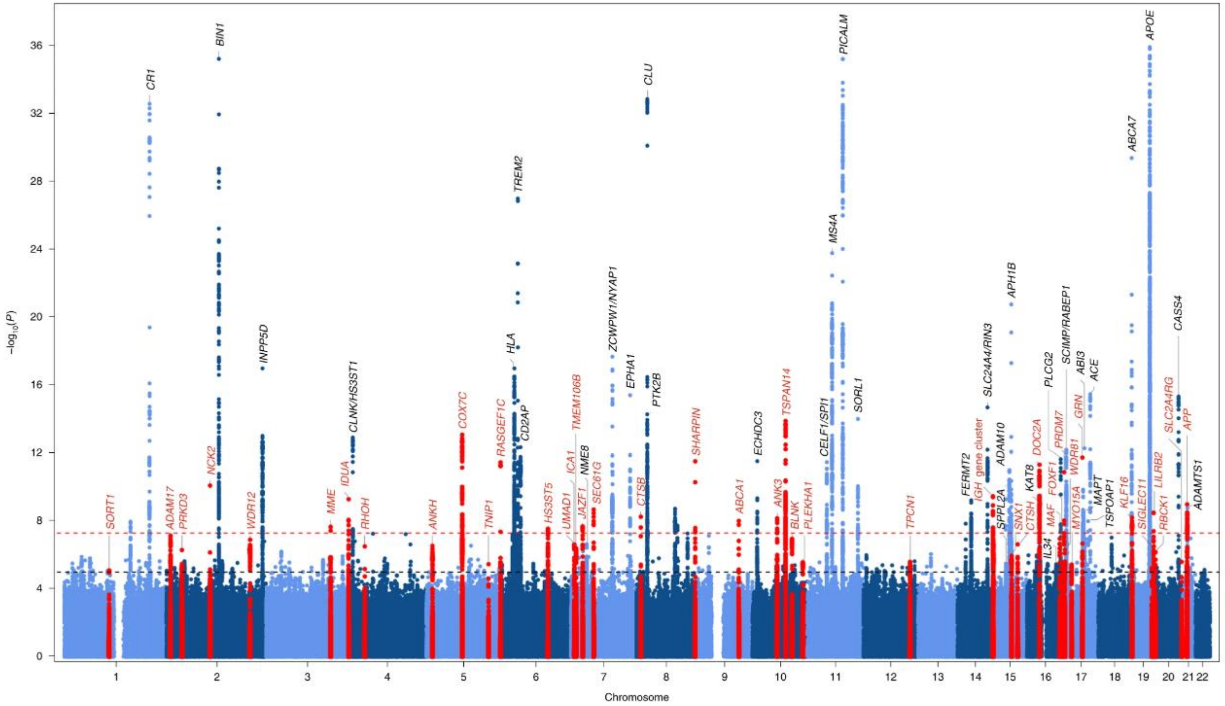
$$phenotype \sim \beta \times genotype + \beta_1 \times covariates + \beta_2 \times structure + \epsilon$$



Manhattan plot: visualization of the results

- Display $-\log_{10}(p)$ for every position in the genome
- Use a threshold ($5\text{x}10^{-8}$) to declare significance





HELMHOLTZ MUNICH

Bellenguez et al. Nat. Genet. 2022

4

Prediction

Prediction

- When we apply the estimated effects to new observations of the variable
- Suppose underlying assumptions
- Process = machine learning, predictive modelling, predictive analysis
- In human genetics, main task = **model effects of genotypes on phenotypes**
- Usually we do not predict
  - ➢ Except PRS (Lecture 5 on 6th December)

Thank you.