Lecture 3 : Statistics for Human Geneticists

Volos Summer School

21 / 05 / 2018

Arthur Gilly

Data input

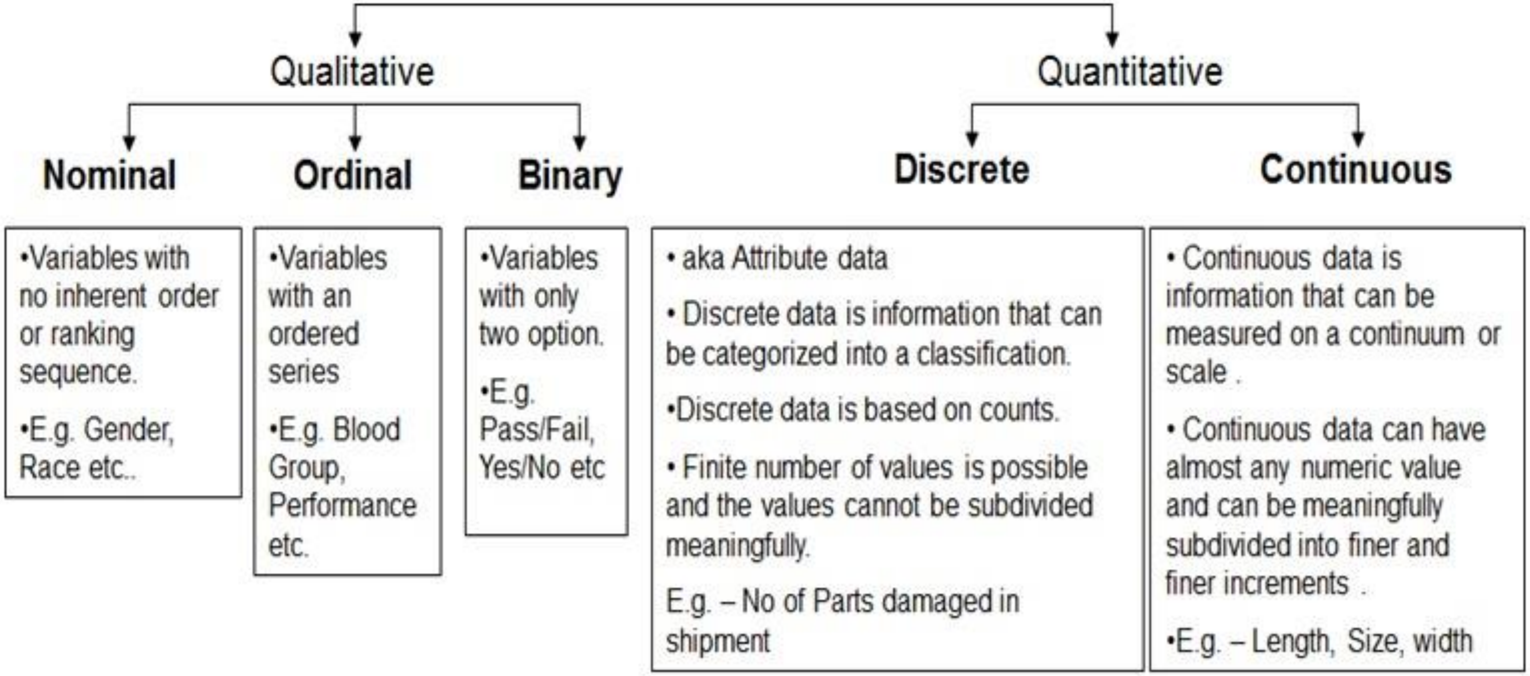
What can we do with statistics?

- Estimation
- Hypothesis testing
- Modelling
- Predicting

Data input

Random Variables

- In statistics, we measure realizations of random variables
- Often, 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

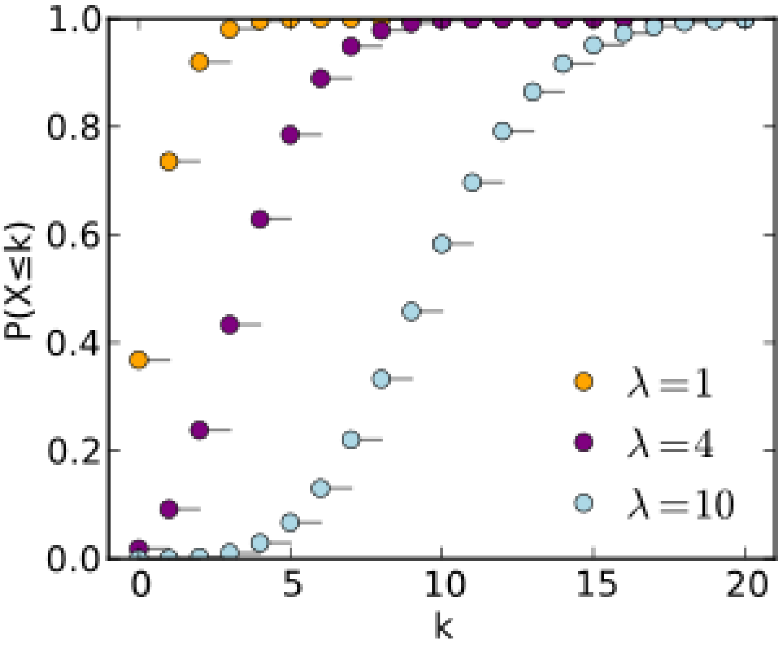
• Continuous data is information that can be measured on a continuum or scale .

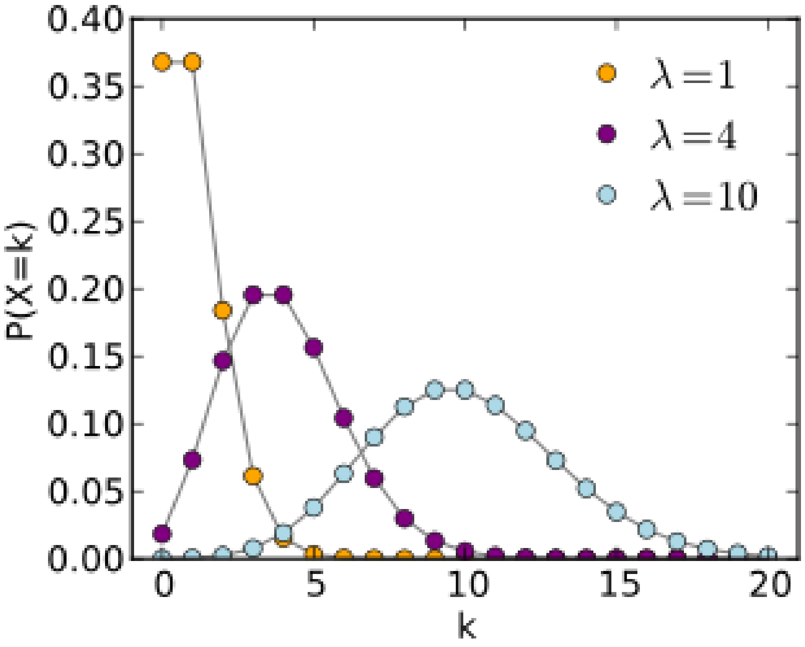
• Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .

•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

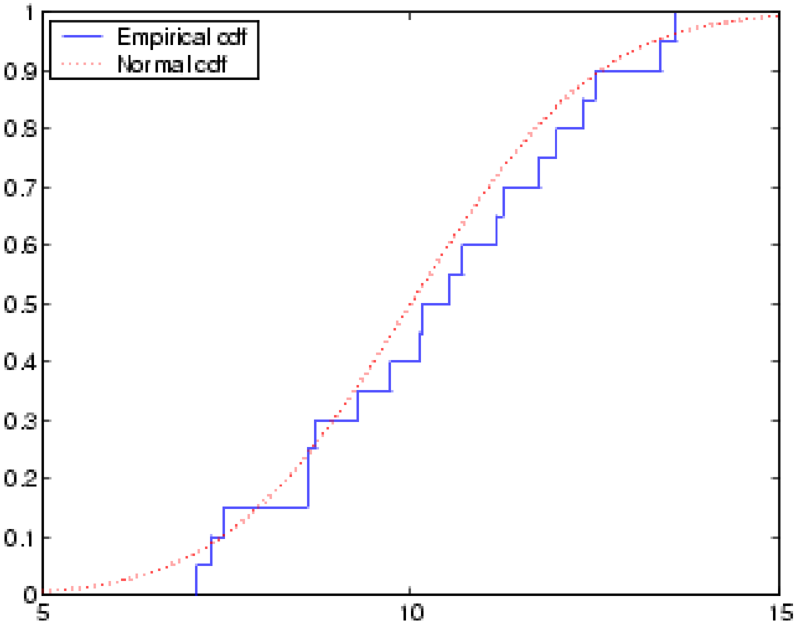
Probability density function (PDF)

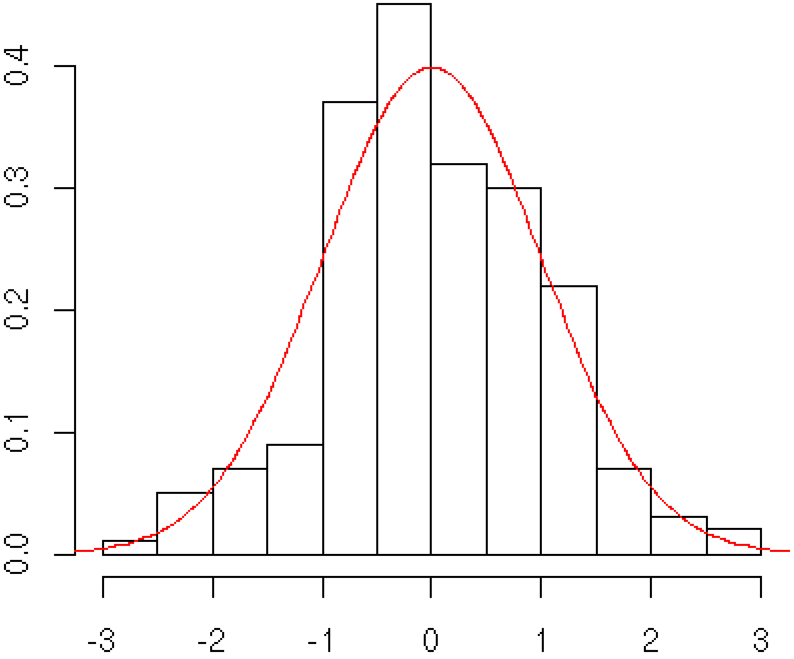
- $y = p(X = x)$ for discrete
- Shows how values are distributed
- Nice visually, but mathematically hard to deal with

Data input

Distributions

- How to estimate them:





Empirical CDF

- Rarely used
- Except when you want to compute empirical quantiles

Barplot (discrete)

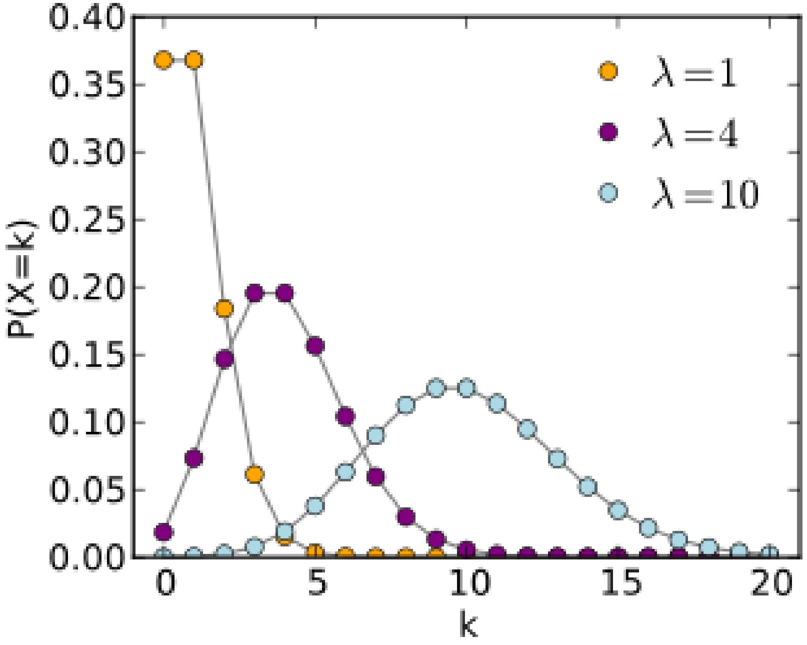
- For every value, count occurrences

Histogram (continuous)

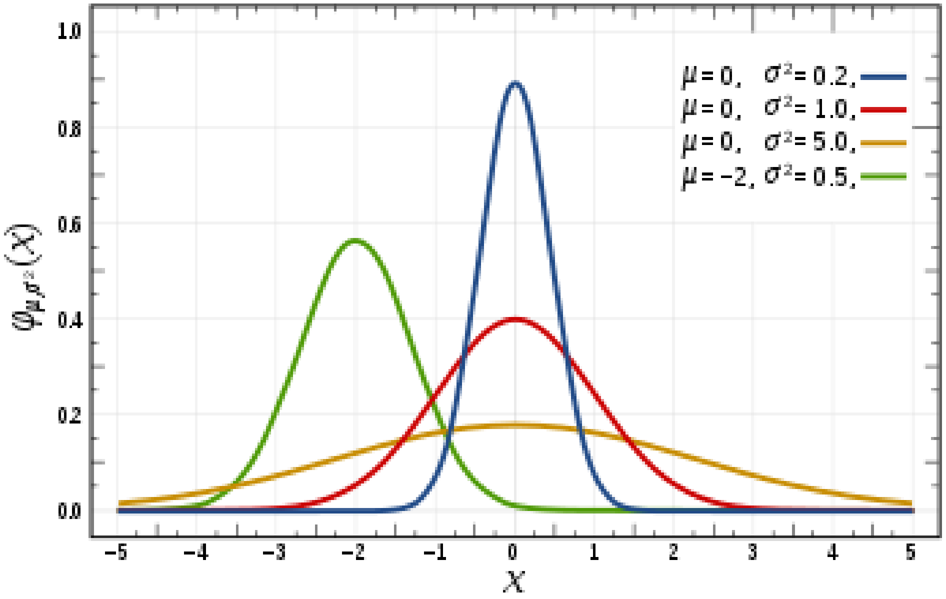
- Cut the interval into bins, count observations within bin

Distributions

- Two broad types:
  - those followed by random variables (real world data)



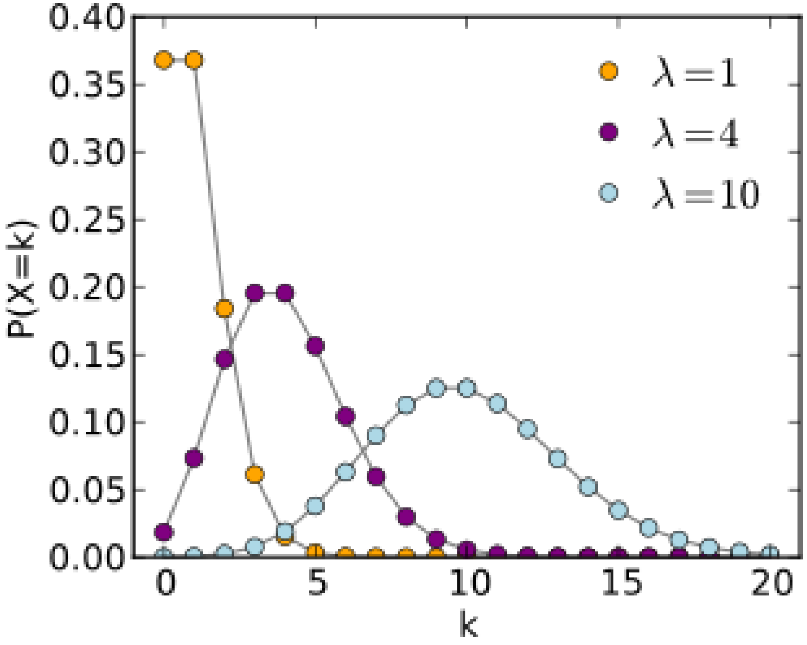
$X \sim Poisson(\lambda)$



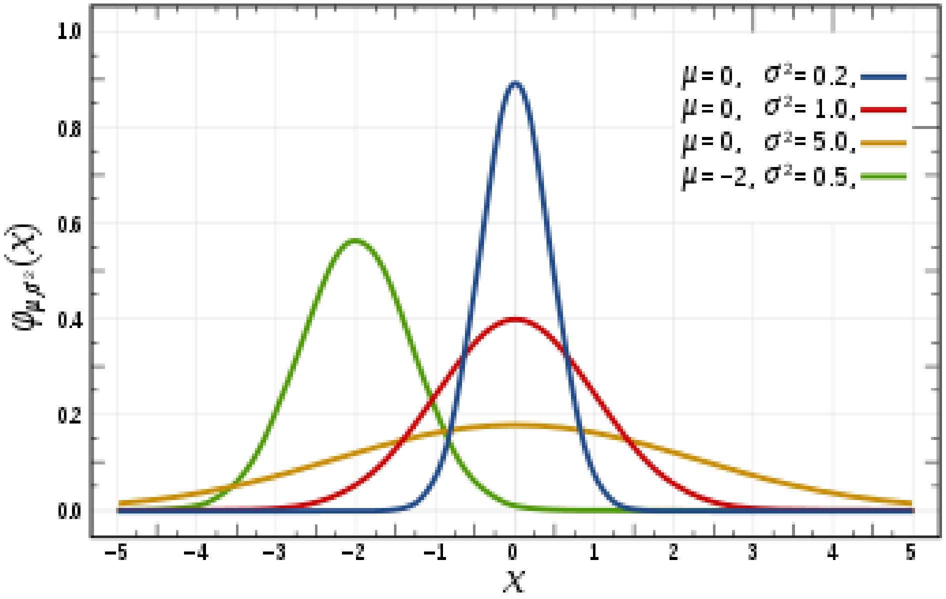
$X \sim \mathcal{N}(\mu, \sigma^2)$

Distributions

- Two broad types:
  - those followed by random variables (real world data)



$X{\sim}Poisson(\lambda)$

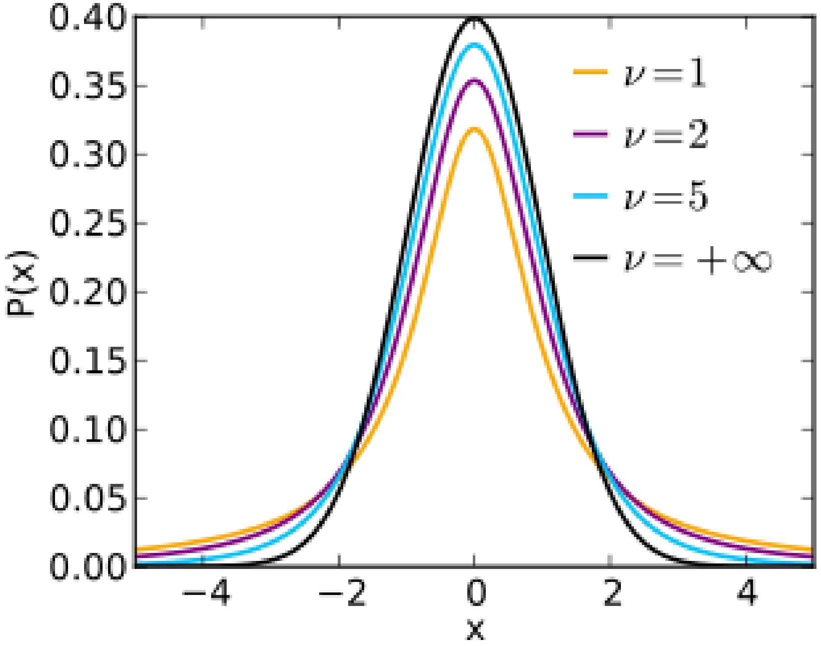


$X{\sim}\mathcal{N}(\mu,\sigma^2)$

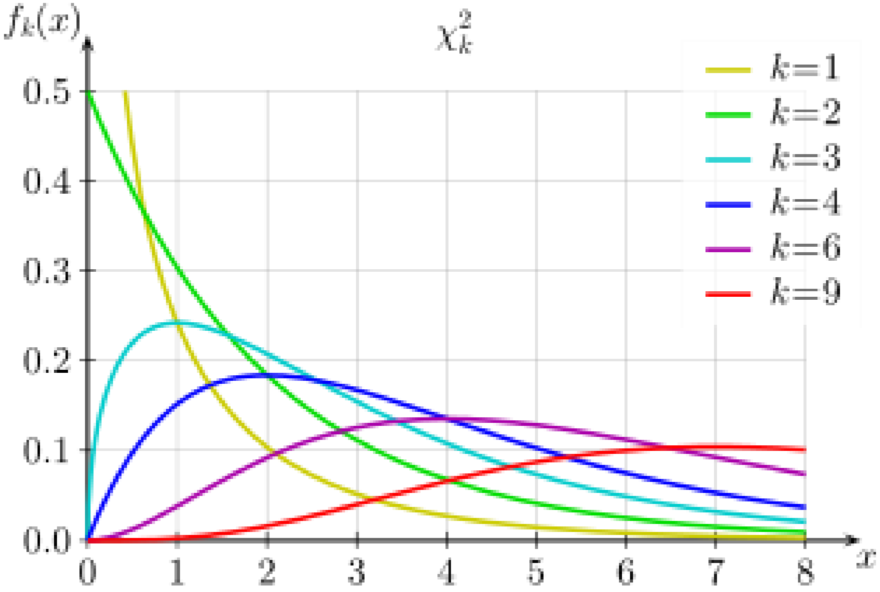
Data input

Distributions

- Two broad types:
  - those followed by test statistics



$X \sim T(\nu)$



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

λ, μ, σ, ν and k are ideal parameters. How to estimate them?

Statistics

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

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

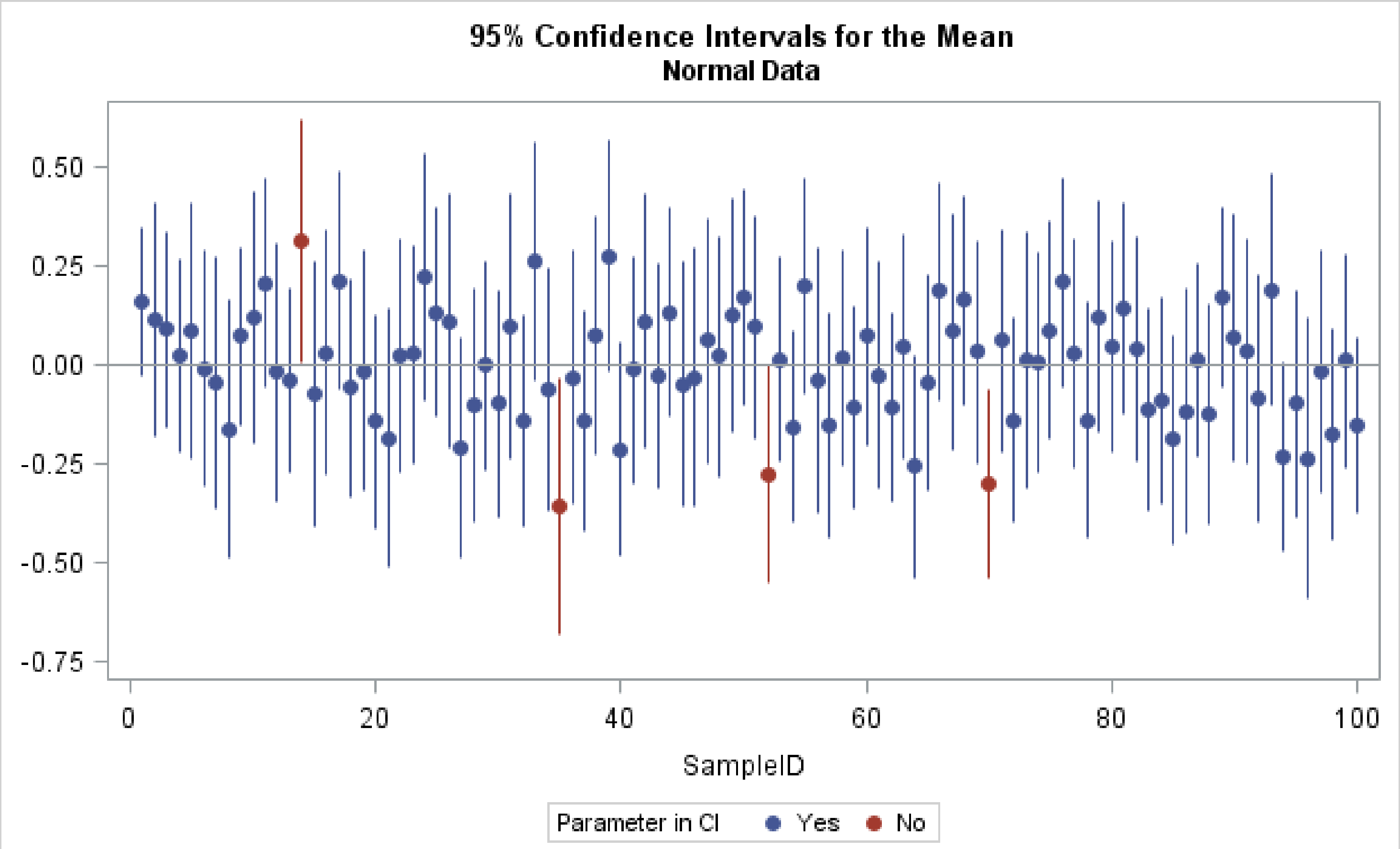
$$\hat{p} = \frac{x}{n}$$

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

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

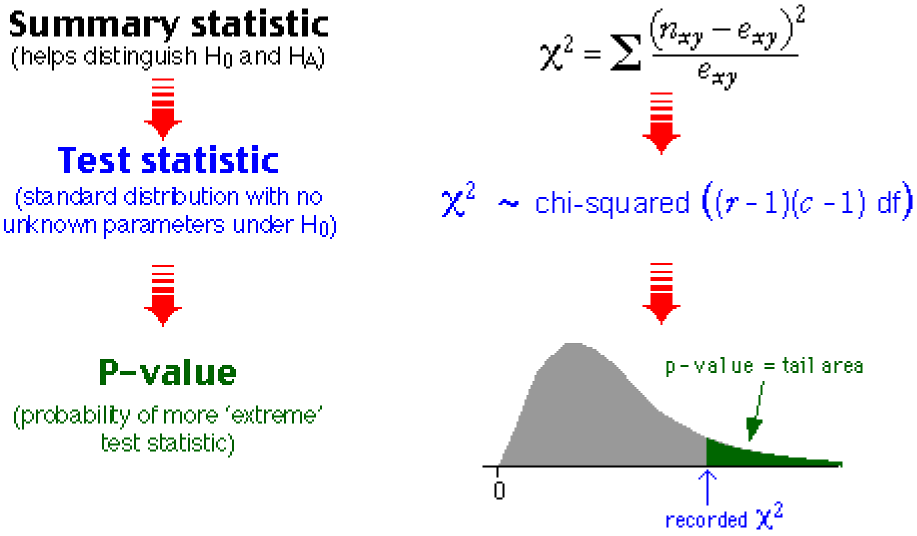
One particular statistic: Confidence intervals

- x% confidence interval (x%C.I.) : x% of the time when this interval is calculated, it will contain the true value of the parameter



Hypothesis testing

- We want to measure whether the data gives sufficient evidence to reject a hypothesis
- Null/Alternative hypothesis ($\mathcal{H}_0/\mathcal{H}_A$)
- We prove that we can produce a statistic that follows a certain distribution if the null hypothesis is true = name of the test
- We calculate the statistic based on our data
- Because we know the distribution, we can calculate the CDF $p(X \leq x)$
- = how unlikely it is that our measurement comes from the null : p-value
- Example: proportion test, t-test, chi-squared test…



Summary statistic
(helps distinguish H_0 and H_A)

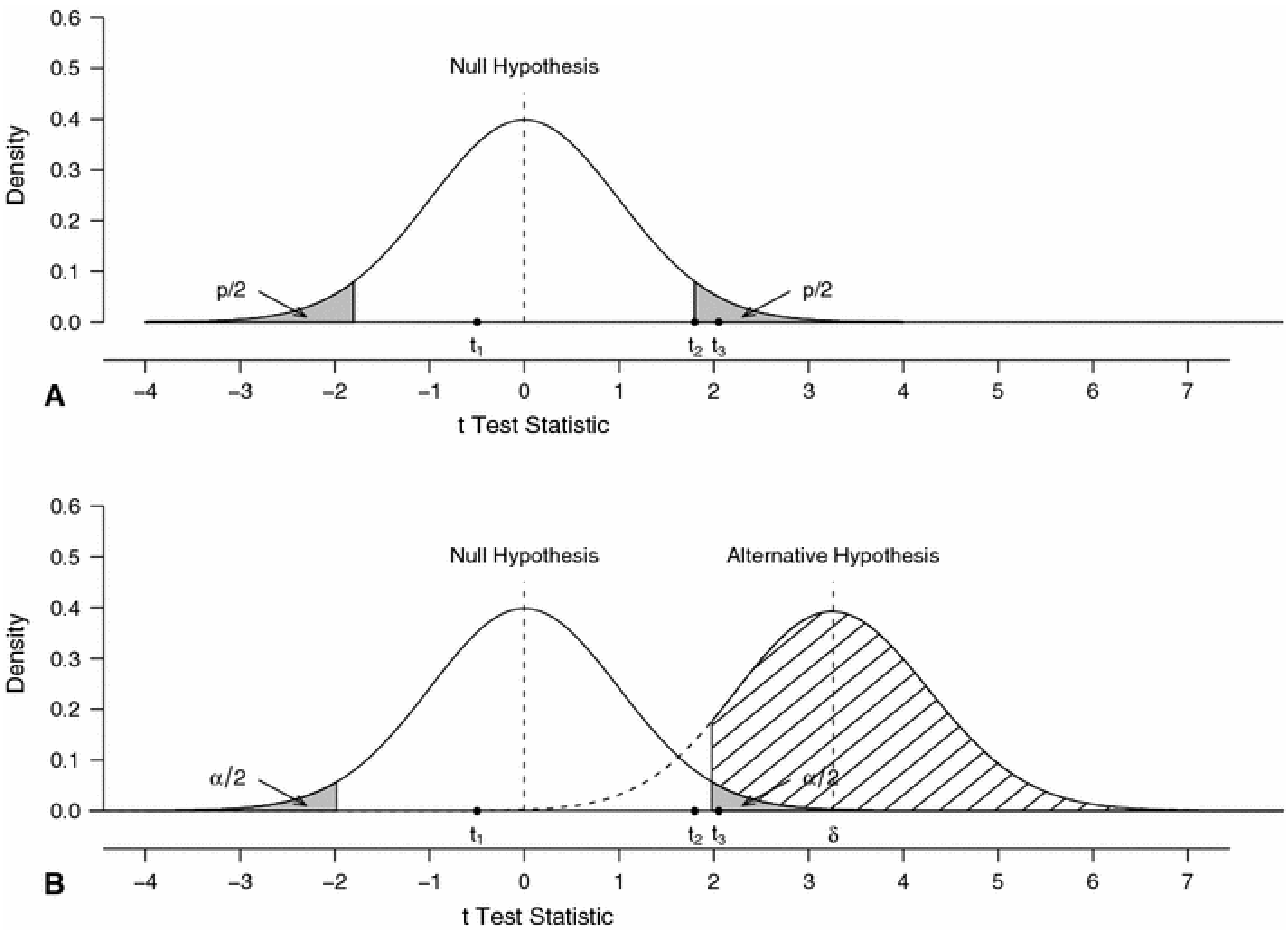
$$\chi^2 = \sum \frac{(n_{xy} - e_{xy})^2}{e_{xy}}$$

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

χ^2 ~ chi-squared ($(r - 1)(c - 1)$ df)

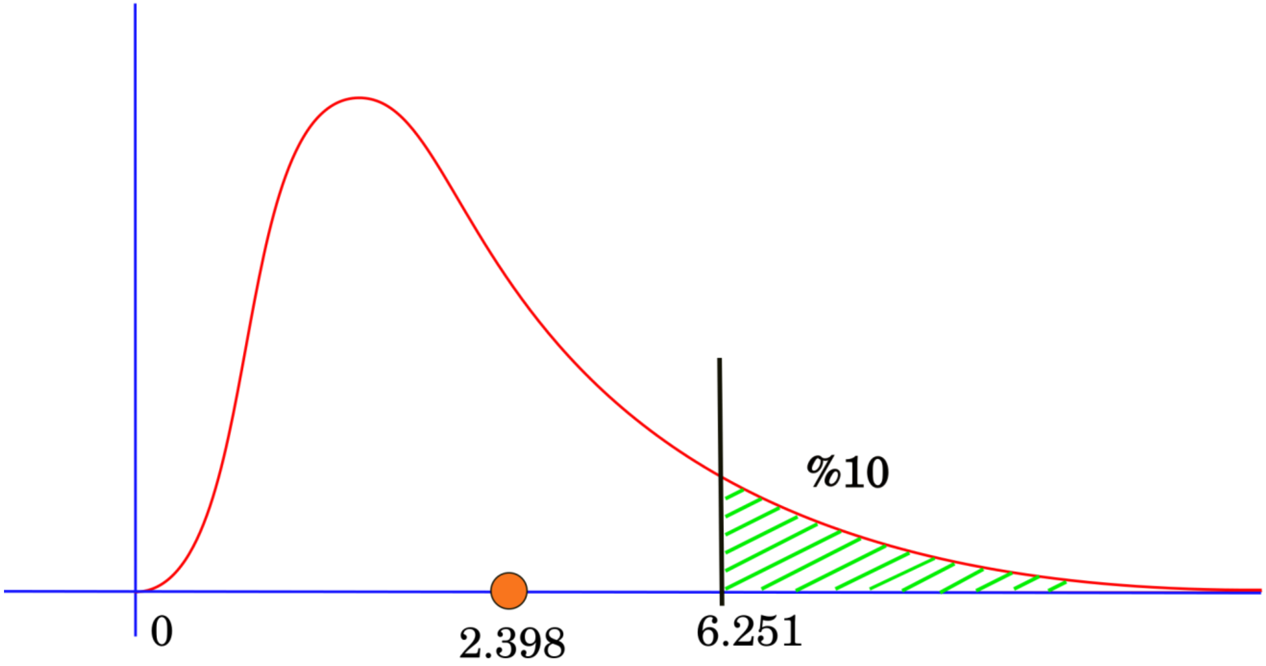
P-value
(probability of more 'extreme' test statistic)

Hypothesis testing



Data input

Two tails or one



Exercise 1 : Proportion test

- In a population, we observe 41,009 potentially damaging variants among 14,281,180 variants
- What is the proportion? In a very large reference population, we observe a proportion of 1.52×10^{-3}. Is it significantly different? (prop.test, binom.test)

Multiple testing

Family-wise (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.”*

Multiple testing

Family-wise (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.”*

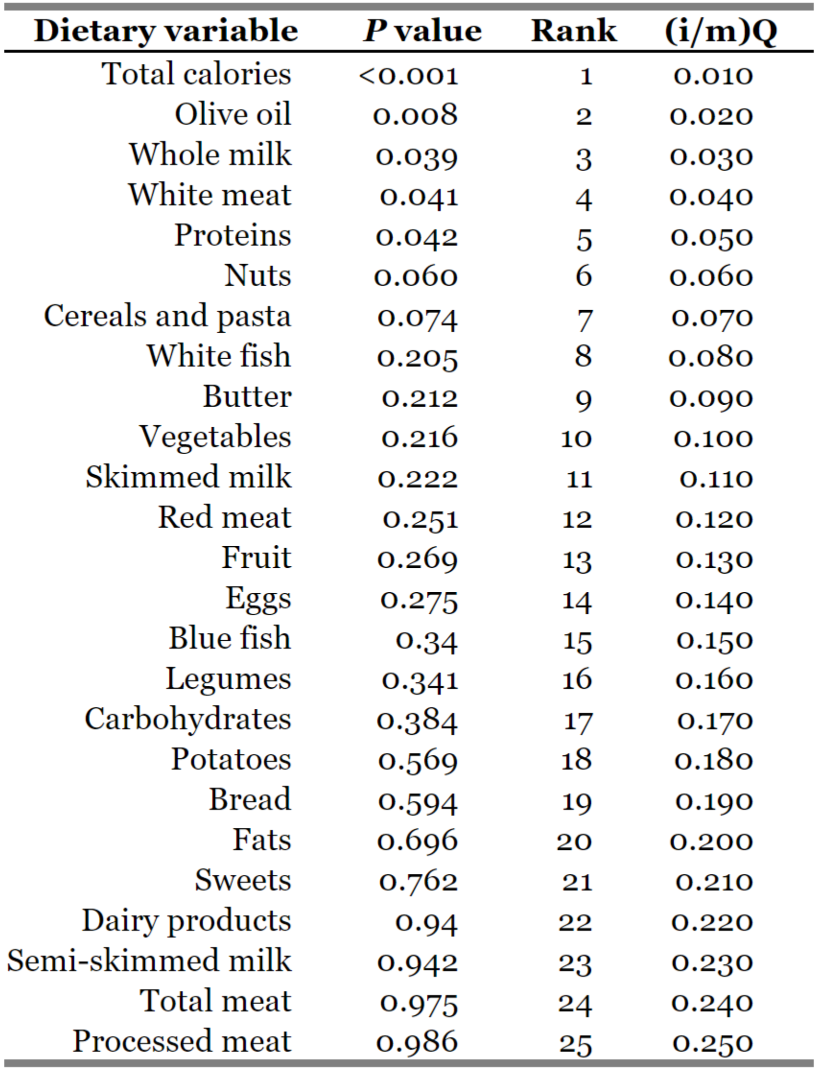
When to use which depends on

- **1) best practices**
- **2) relative price of false negative/positive**

Checking results

- Statistical significance:
  - One test: p<0.05
  - Genome-wide: one test per variant and per phenotype
  - But all variants are not independent, in reality, we account for LD
  - 5x10-8 for GWAS, 10-9 for sequencing-based

Dietary variable	***P* value**
Total calories	<0.001
Olive oil	0.008
Whole milk	0.039
White meat	0.041
Proteins	0.042
Nuts	0.060
Cereals and pasta	0.074
White fish	0.205
Butter	0.212
Vegetables	0.216
Skimmed milk	0.222
Red meat	0.251
Fruit	0.269
Eggs	0.275
Blue fish	0.34
Legumes	0.341
Carbohydrates	0.384
Potatoes	0.569
Bread	0.594
Fats	0.696
Sweets	0.762
Dairy products	0.94
Semi-skimmed milk	0.942
Total meat	0.975
Processed meat	0.986



Dietary variable	*P* value	Rank	(i/m)Q
Total calories	<0.001	1	0.010
Olive oil	0.008	2	0.020
Whole milk	0.039	3	0.030
White meat	0.041	4	0.040
Proteins	0.042	5	0.050
Nuts	0.060	6	0.060
Cereals and pasta	0.074	7	0.070
White fish	0.205	8	0.080
Butter	0.212	9	0.090
Vegetables	0.216	10	0.100
Skimmed milk	0.222	11	0.110
Red meat	0.251	12	0.120
Fruit	0.269	13	0.130
Eggs	0.275	14	0.140
Blue fish	0.34	15	0.150
Legumes	0.341	16	0.160
Carbohydrates	0.384	17	0.170
Potatoes	0.569	18	0.180
Bread	0.594	19	0.190
Fats	0.696	20	0.200
Sweets	0.762	21	0.210
Dairy products	0.94	22	0.220
Semi-skimmed milk	0.942	23	0.230
Total meat	0.975	24	0.240
Processed meat	0.986	25	0.250

Modelling and predicting

- If we estimate the effect of one variable on another variable, we do modelling
- When we apply this effect to new observations of the variable, we do prediction
- Process is called machine learning, predictive modelling or predictive analysis
- In human genetics, main task is to model effect of genotypes on phenotypes

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

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

Modelling and predicting

- If we estimate the effect of one variable on another variable, we do modelling
- When we apply this effect to new observations of the variable, we do prediction
- Process is called machine learning, predictive modelling or predictive analysis
- In human genetics, main task is to model effect of genotypes on phenotypes

$$phenotype \sim \beta \times genotype + \epsilon$$

$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix}$

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

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

$\begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix}$

$= \{0,1,2\}$ (genotype, directly typed $\begin{bmatrix} 1 \\ \vdots \\ 2 \end{bmatrix}$

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

Modelling and predicting

- If we estimate the effect of one variable on another variable, we do modelling
- When we apply this effect to new observations of the variable, we do prediction
- Process is called machine learning, predictive modelling or predictive analysis
- In human genetics, main task is to model effect of genotypes on phenotypes

Usually, we do not predict (except PRS)

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

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix}$$

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

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

$$\begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ \vdots \\ 2 \end{bmatrix}$$

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

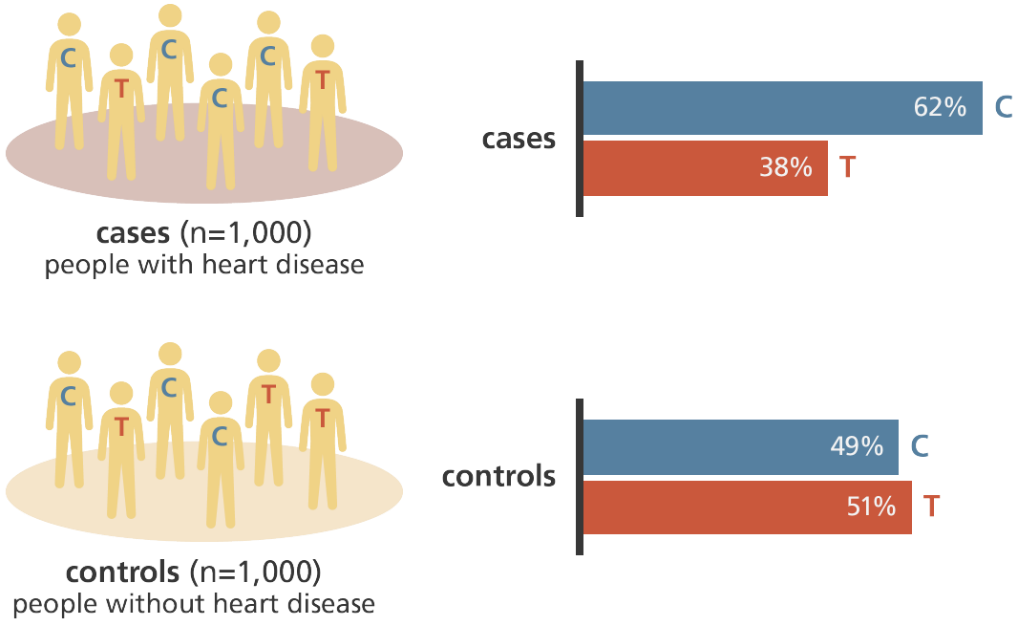
$\in [0,2]$ (dosage, imputed) $\begin{bmatrix} 0.965 \\ \vdots \\ 1.816 \end{bmatrix}$

Case/control

- Estimated effect: odds ratio (OR)

"how much more likely are you to be a case if you carry the risk allele?"

per genotype, calculate the odds $O = \frac{p}{1-p}$



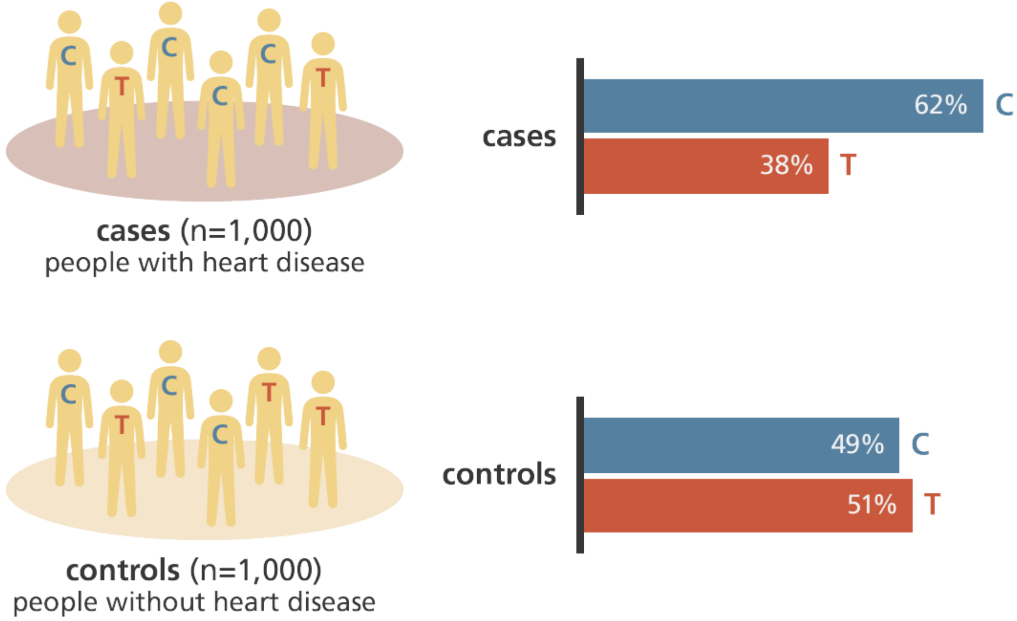
Data input

Case/control

- Estimated effect: odds ratio (OR)

  "how much more likely are you to be a case if you carry the risk allele?"

  per genotype g and for a disease Y, calculate the odds $O = \frac{p_{Y=1|g}}{1-p_{Y=1|g}}$



	cases	controls
T	380	490
C	620	510

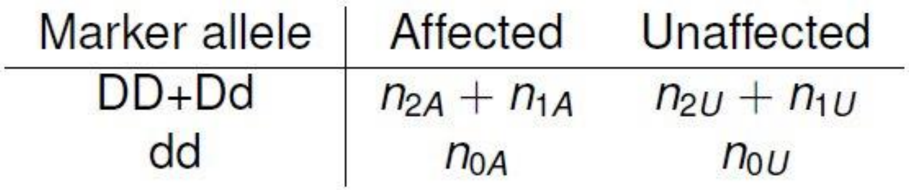
$$O_T = \frac{380/n_T}{490/n_T} \qquad O_C = \frac{620/n_C}{510/n_C}$$

$$OR_{C/T} = \frac{620 \times 490}{510 \times 380} = 1.56$$

Data input

Case/control

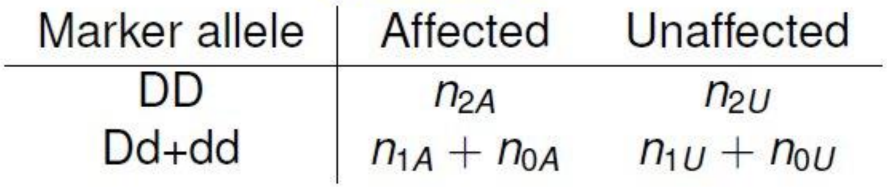
Dominant



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

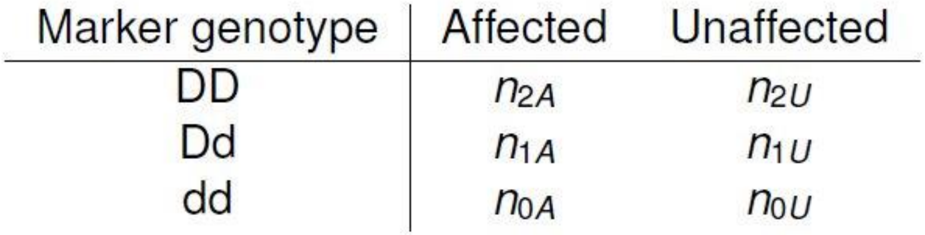
$$OR = \frac{n_{affected\ carriers} \times n_{healthy\ non-carriers}}{n_{healthy\ carriers} \times n_{affected\ non-carriers}}$$

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

Data input

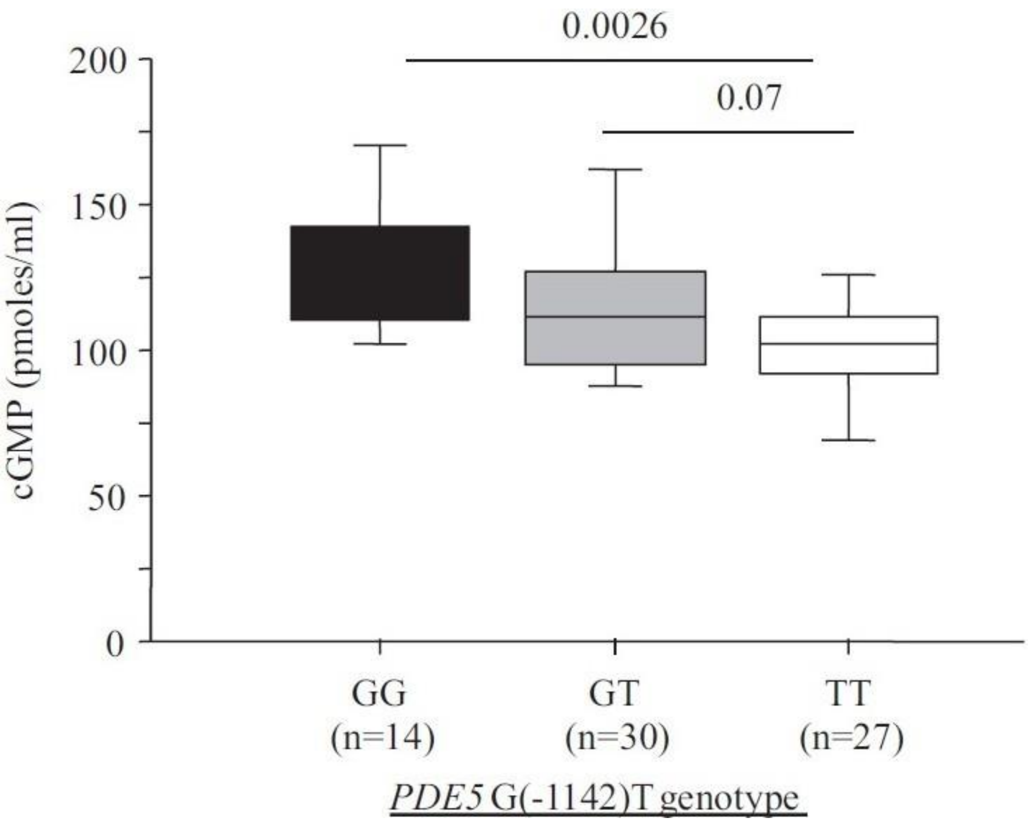
Case/control

- Output: OR and 95% confidence interval of the OR
- Test: is it significantly different from 1?
- Tests: Fisher's exact test or Chi-squared
- In case of dosages or covariates: logistic regression

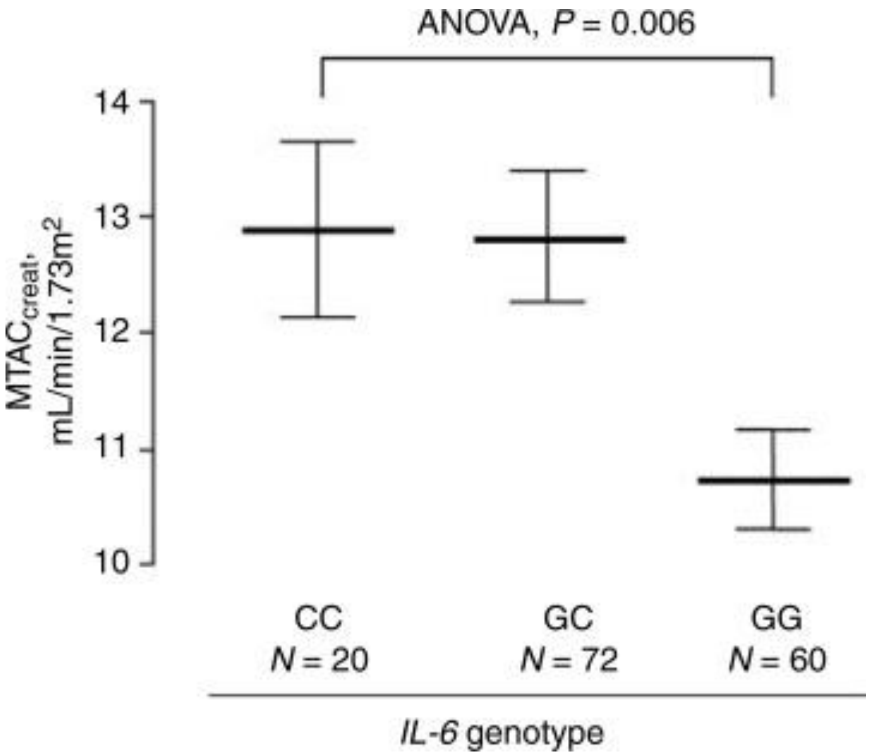
Data input

Continuous trait

- For directly typed (0,1,2): ANOVA



additive

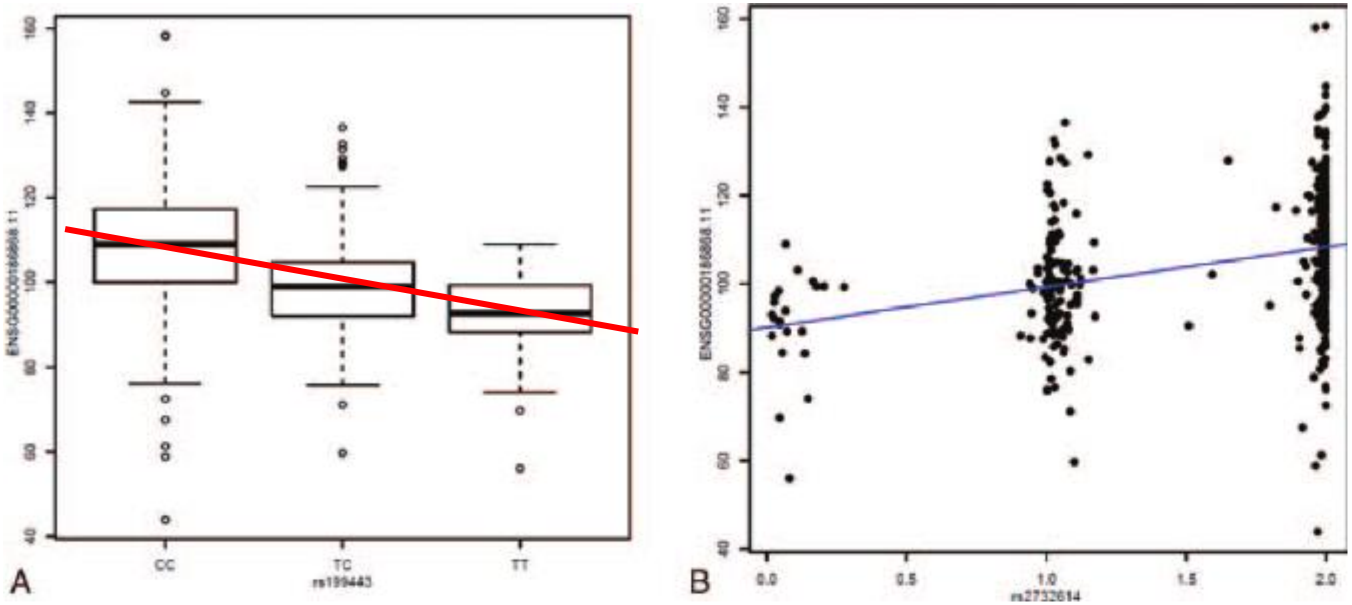


recessive

Data input

Continuous trait

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

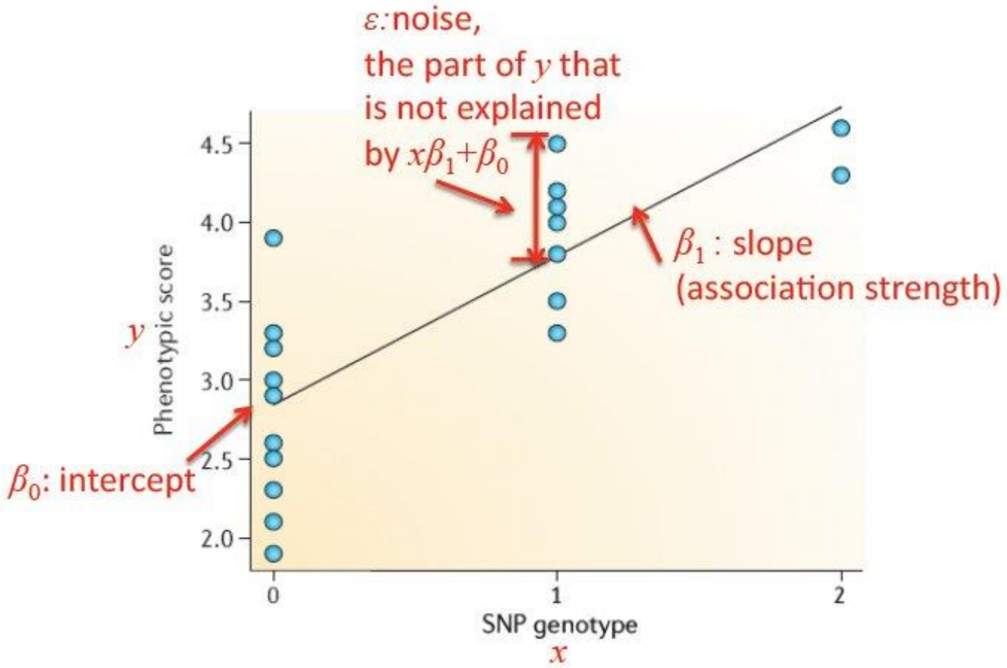


Data input

Continuous trait

A linear regression model is defined as

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



Data:

- y: a continuous trait
- x: SNP genotype at a given locus

Parameters:

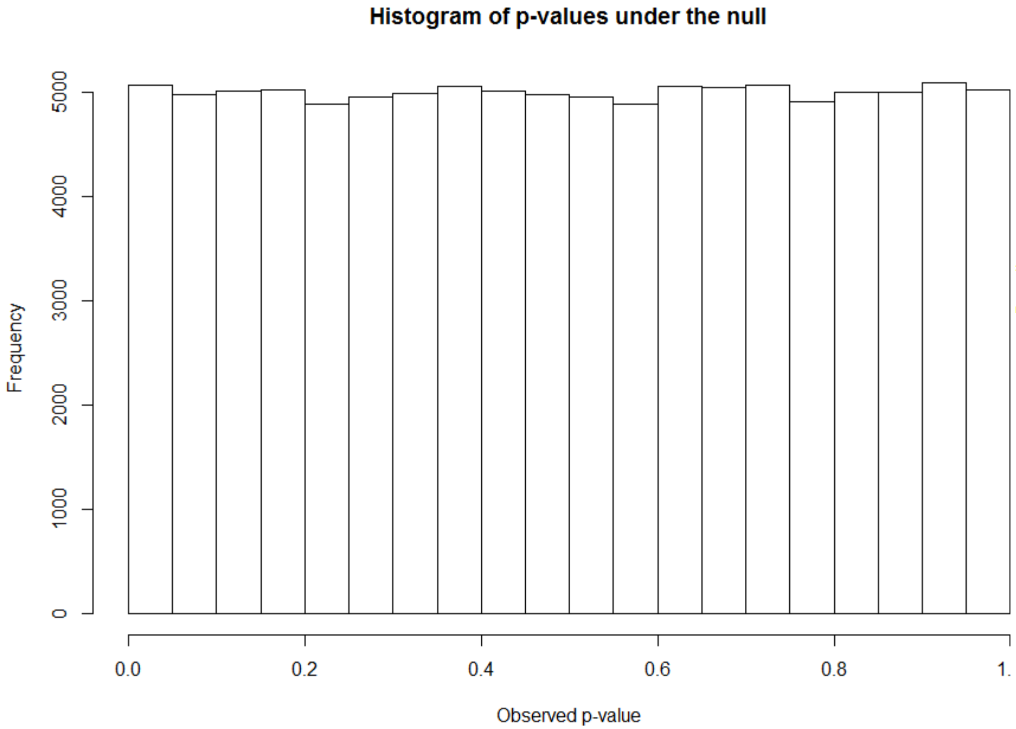
- β_1: regression coefficient, represents the strength of association between x and y
- β_0: intercept term (is 0 or ignored)
- ε: noise or the part of y that is not explained by x (e.g., environmental effect)

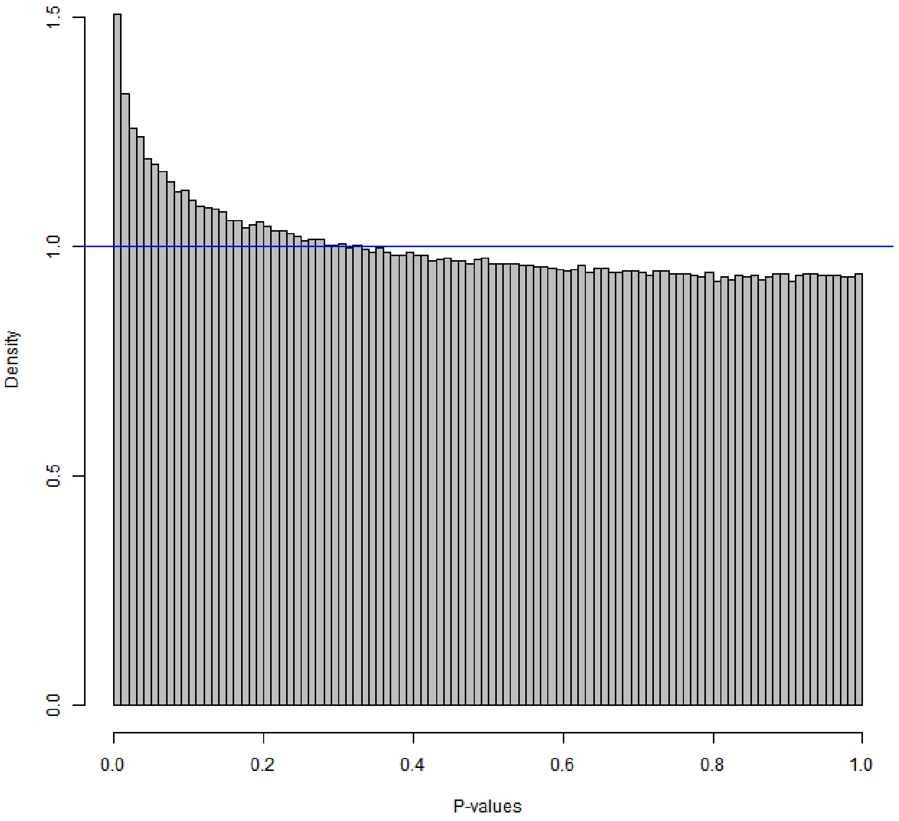
Assumptions:

- The individuals in the study are not related
- The phenotype y has a normal distribution

Checking results

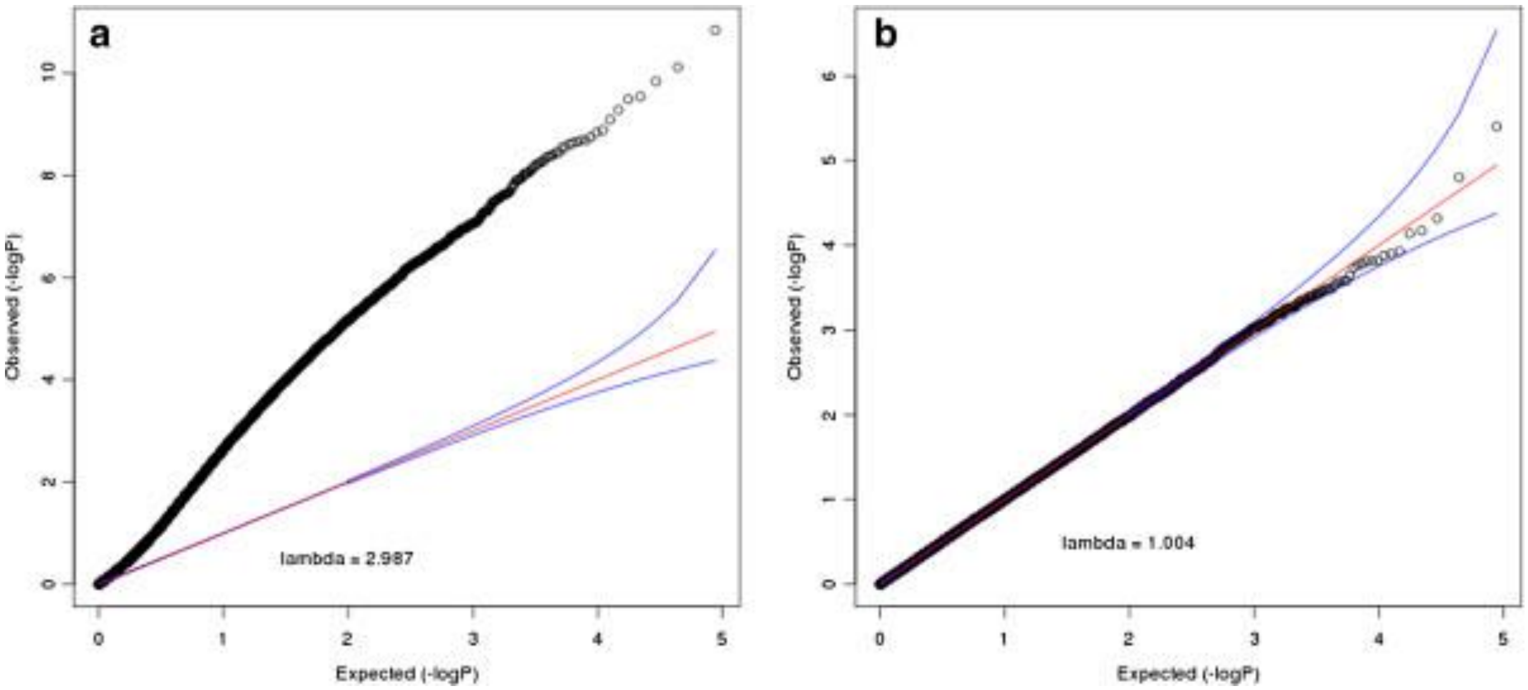
- QQ-plot
  - Distribution of p-values is uniform [0,1] under the null
  - If we have much signal, more around 0
  - Compare quantiles with expected ones : QQ-plot
  - In R: `qqunif`





Checking results

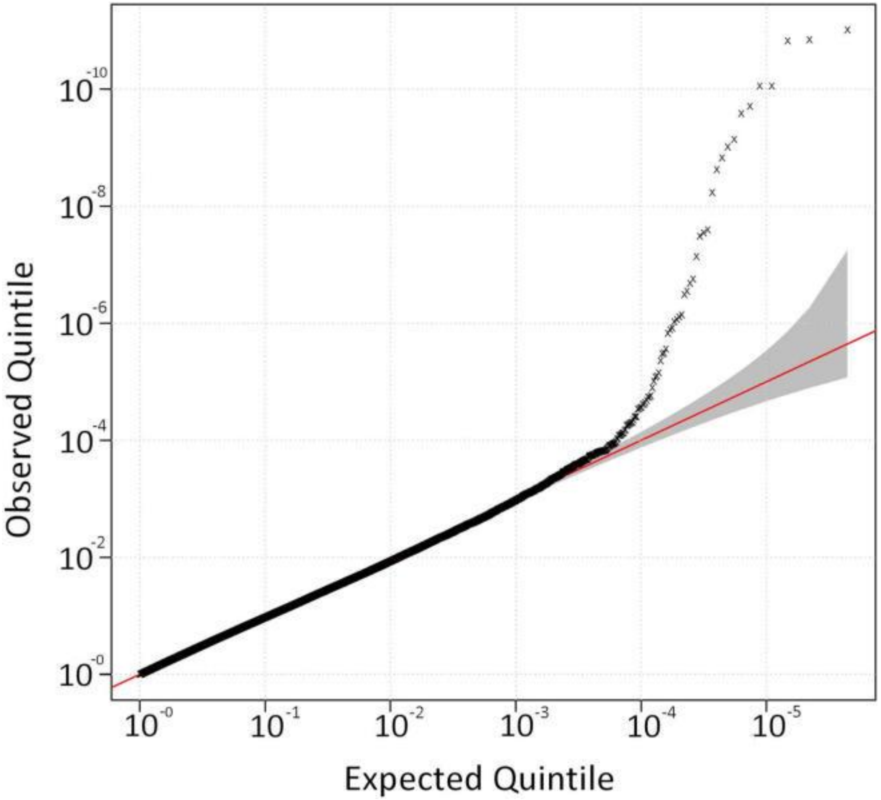
- QQ-plot
  - Distribution of p-values is uniform [0,1] under the null
  - If we have much signal, more around 0
  - Compare quantiles with expected ones : QQ-plot
  - In R: `qqunif`



Data input

Checking results

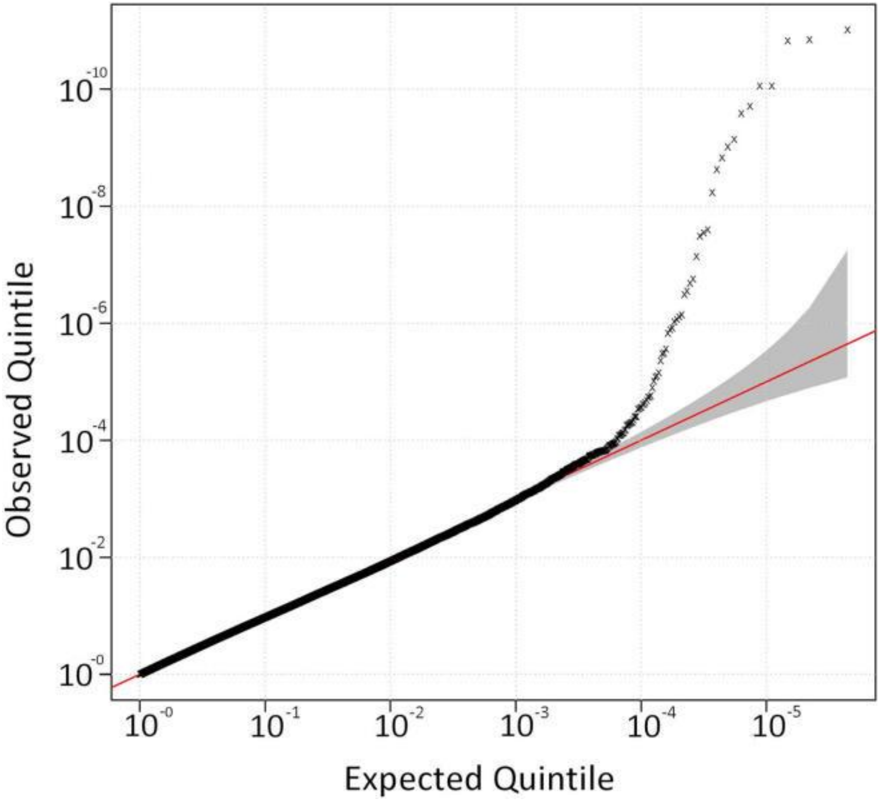
- QQ-plot
  - Distribution of p-values is uniform [0,1] under the null
  - If we have much signal, more around 0
  - Compare quantiles with expected ones : QQ-plot
  - In R: `qqunif`



- Inflation: too much signal
- Measured visually, but also lambda

Checking results

- QQ-plot
  - Distribution of p-values is uniform [0,1] under the null
  - If we have much signal, more around 0
  - Compare quantiles with expected ones : QQ-plot
  - In R: `qqunif`



Appearances can be deceiving:

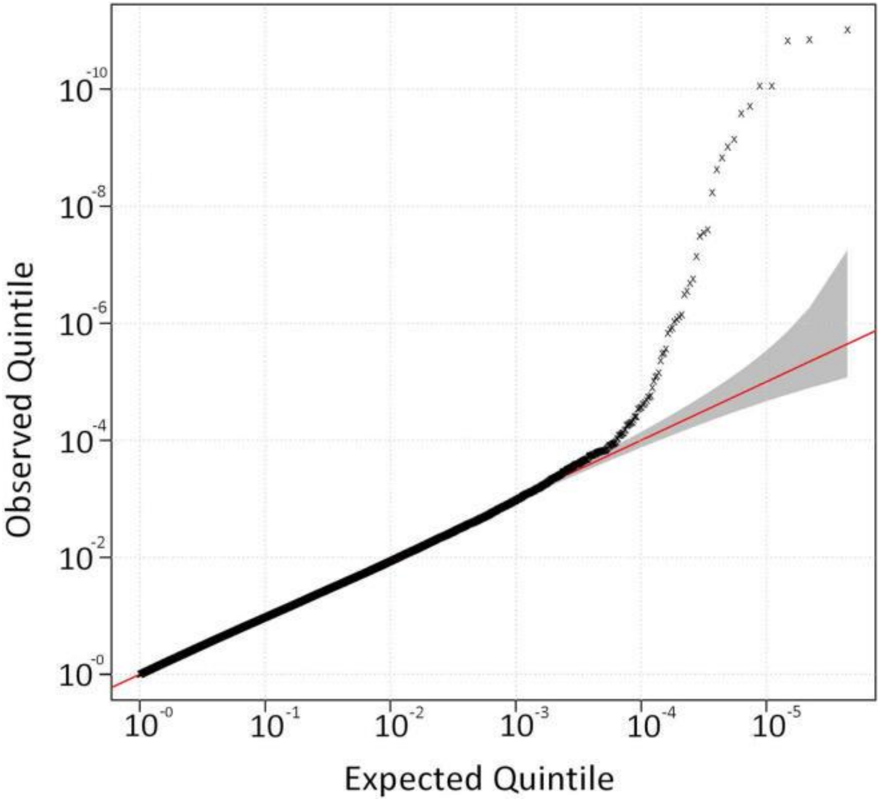
- A QQ-plot can look inflated when it isn't (just a lot of signal)
- And conversely
- We calculate the genomic inflation factor

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

(median of χ^2 test statistics divided by median of χ_1^2)

Checking results

- QQ-plot
  - Distribution of p-values is uniform [0,1] under the null
  - If we have much signal, more around 0
  - Compare quantiles with expected ones : QQ-plot
  - In R: `qqunif`



Appearances can be deceiving:

- A QQ-plot can look inflated when it isn't (just a lot of signal)
- And conversely
- We calculate the genomic inflation factor

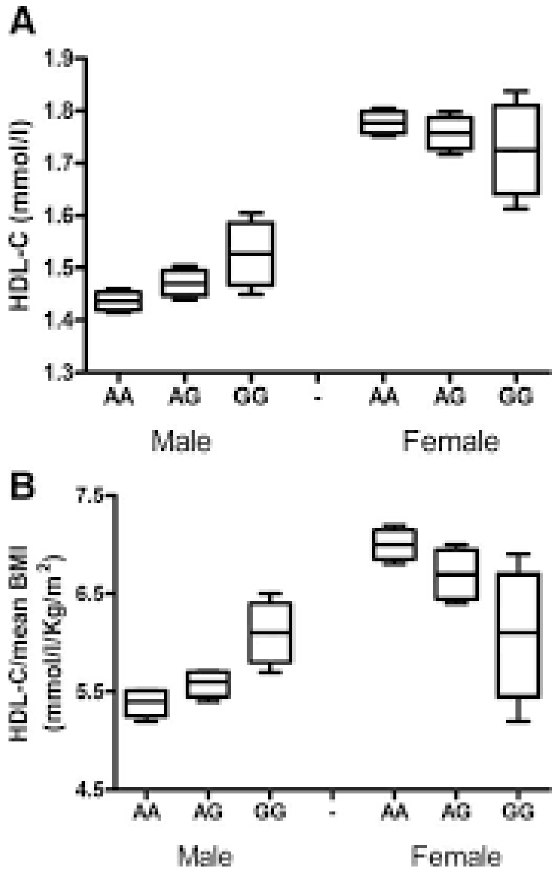
$$\lambda = \frac{median(Q_{\chi^2}(p))}{0.45}$$

(median of χ^2 test statistics divided by median of χ_1^2)

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

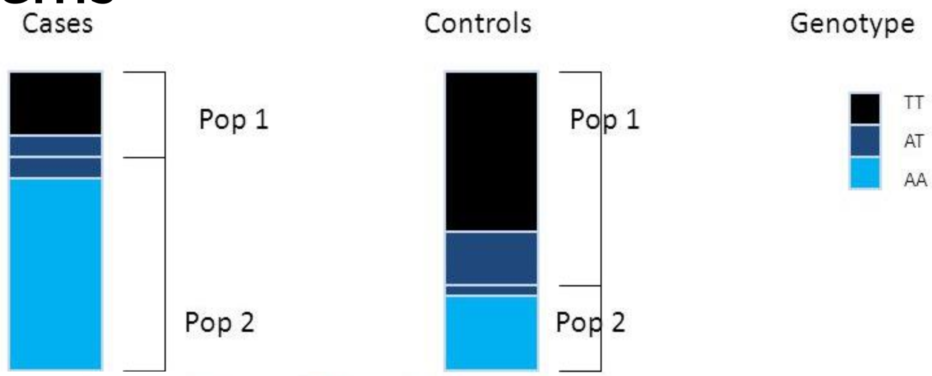
Problems

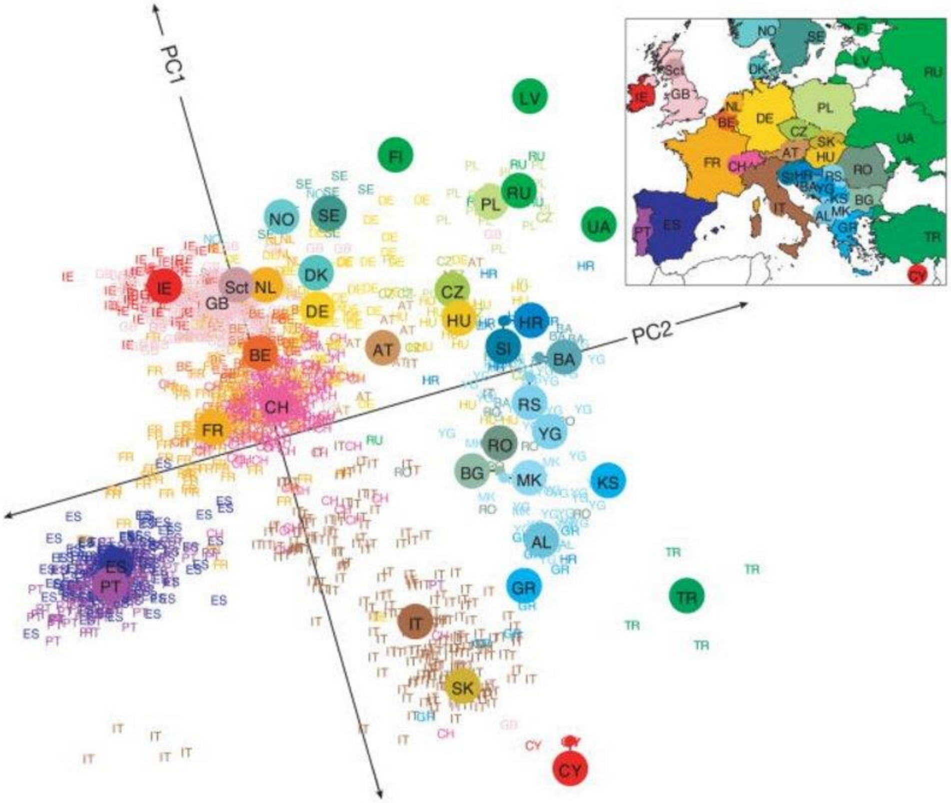
- Covariates: sex, batch effects, chip effects



Problems

- Covariates: sex, batch effects, chip effects
- Structure: villages or subpopulations: linear mixed models can model the intra-group effect

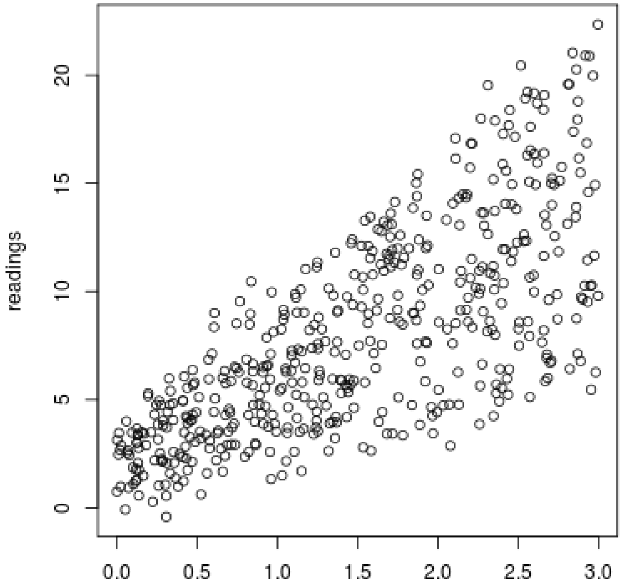


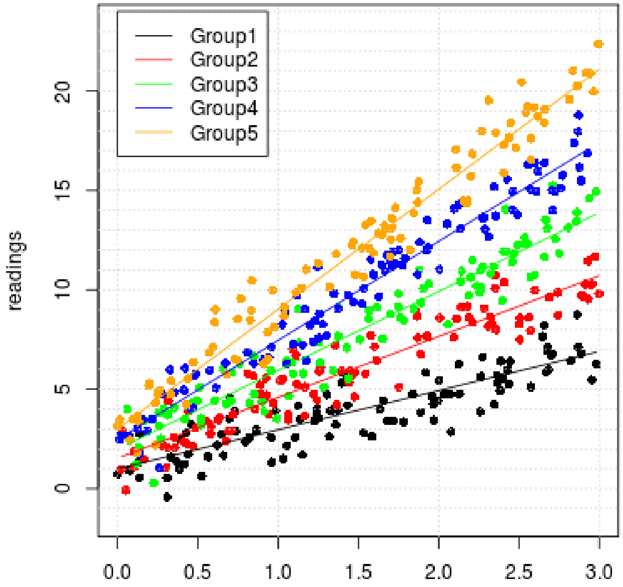


Problems

- Covariates: sex, batch effects, chip effects
- Structure: villages or subpopulations: linear mixed models can model the intra-group effect

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


readings


Group1
Group2
Group3
Group4
Group5
readings