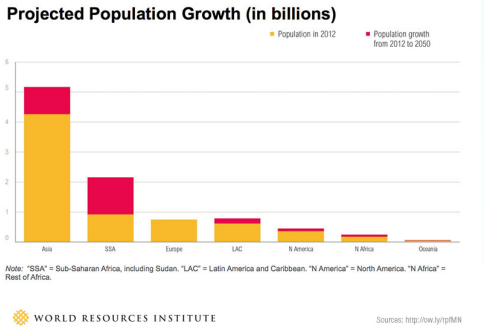
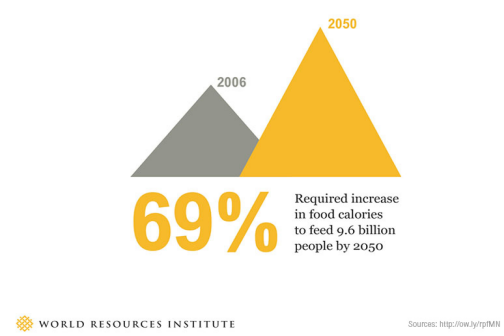
Concepts

- The need for increased food with equal or less resources.
- Humans have addressed this so far with domestication, artificial selection and improved culture techniques.
- How bioinformatics can help.

Global Food Challenge

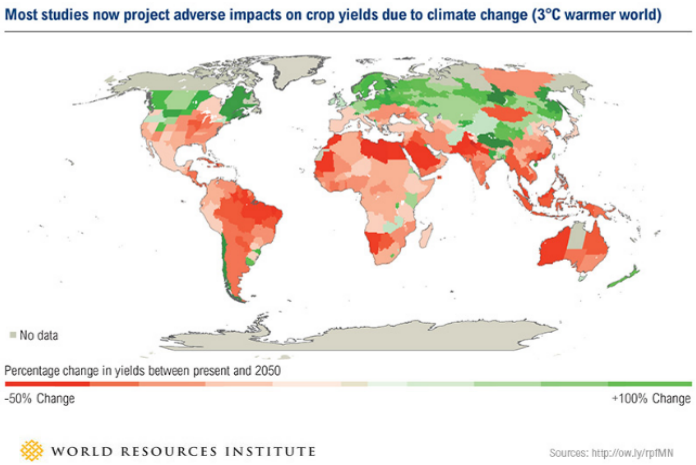




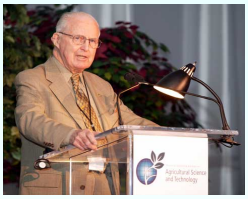
http://www.wri.org/

Climate Change and Water Stress Exacerbate the Challenge

Climate change is expected to negatively impact crop yields, particularly in the hungriest parts of the world, such as sub-Saharan Africa.



The Green Revolution



Norman Borlaug (Peace Nobel prize, 1970): He developed semi-dwarf, high-yield, disease-resistant wheat varieties in Mexico. During the mid-20th century, Borlaug led the introduction of these high-yielding varieties combined with modern agricultural production techniques to Mexico, Pakistan, and India. As a result, Mexico became a net exporter of wheat by 1963. Between 1965 and 1970, wheat yields nearly doubled in Pakistan and India, greatly improving the food security in those nations.

Borlaug was often called "the father of the Green Revolution" and is credited with saving over a billion people worldwide from starvation

https://commons.wikimedia.org/w/index.php?curid=127051

Ultimate goals

1. Determine the genetic basis of 'complex' traits.
2. Improve genetic and phenotypic prediction.

using high throughput genomic technologies

BROAD ADVICES

1. DONT BE SCARED, BUT BE RESPECTFUL: **QUALITY CONTROL IS A MUST.**
2. LEARN LINUX, AWK & R as much as you can
3. LEARN A SCRIPTING LANGUAGE: PYTHON SUGGESTED.

Scenario

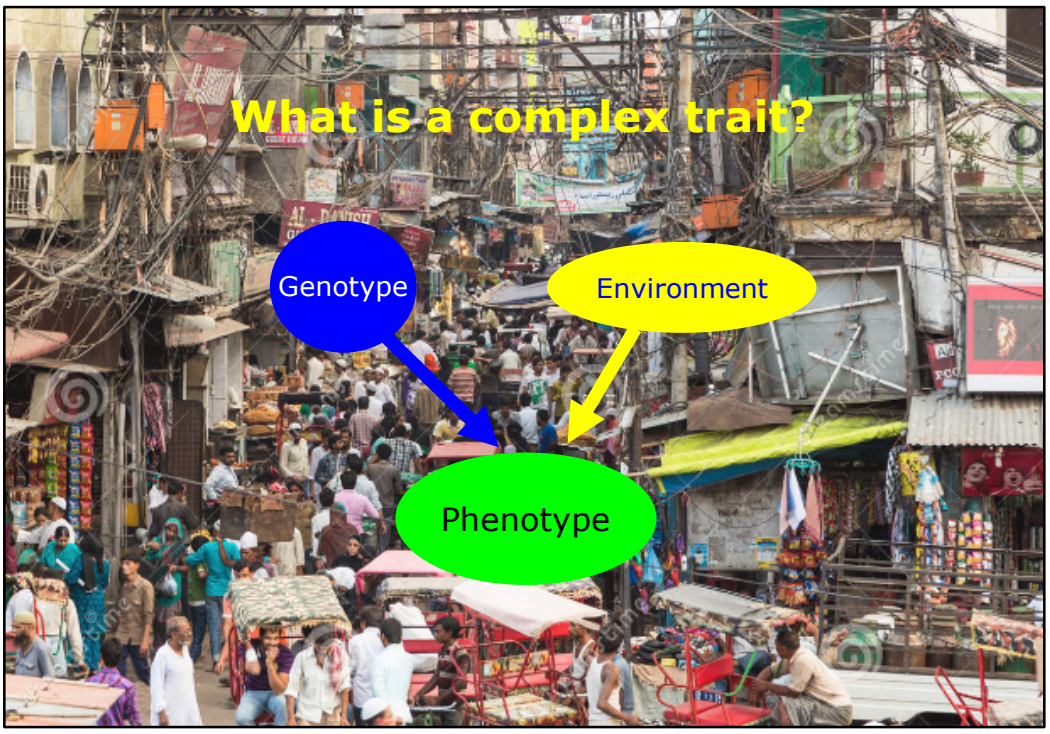
- ✓ Biology has become a data rich science, where the limiting step is the data analysis, rather than obtaining the data.
- ✓ Nevertheless, large data sets with reliable phenotypic measurements and genomic data are still rare.
- ✓ As a result, a combination of computer and experimental approaches is key to success.

What is a simple trait?

A ⟷ ▲

T ⟷ ◆

1 Genotype ⟷ 1 Phenotype




AL - DANISH
GUEST HOUSE
In general, we simplify complexity by taking the linear (additive) part only


AL - DANISH
GUEST HOUSE
Our approach to complexity is indeed VERY simple!

Why traits are 'complex'?

- Continuous dialogue between the individual's genotype and the environment.
- The genome itself is complex, it is full of motifs (structural, regulatory,...) and only a tiny part is identified and characterized.
- Genome is highly unstable but also resilient, it bears a large number of different kinds of polymorphisms.

Main consequences of complexity

1. There is no unique relationship between phenotype and genotype. The relationship is measured instead in probabilistic terms.
2. As a result, we cannot escape from Statistics and Mathematics.
3. It is going to be (very) difficult to prove causality.

Main genomic data types

- ✓ DNA Sequence (e.g., GenBank)
- ✓ DNA polymorphism (marker, e.g., dbSNP)
- ✓ RNAseq (functional genomics, e.g., GEO), epigenomics…
- ✓ Annotation data, including interactions, pathways,…

http://www.ncbi.nlm.nih.gov/sites/entrez

Outline of this course

1. Background
2. Fundamentals of Breeding
3. Genomic Selection
4. Big Data and Machine Learning

BACKGROUND:
Population Genetics Concepts

Genetic Variability

Measures of variability

- Number of pairwise differences per length DNA sequenced (Tajima's)
- Number of SNPs per length DNA sequenced (Watterson's)

Genetic Variability

Factors that affect variability

- Mutation: the ultimate source of all variability
- Recombination
- Genetic drift
- Selection
- Migration

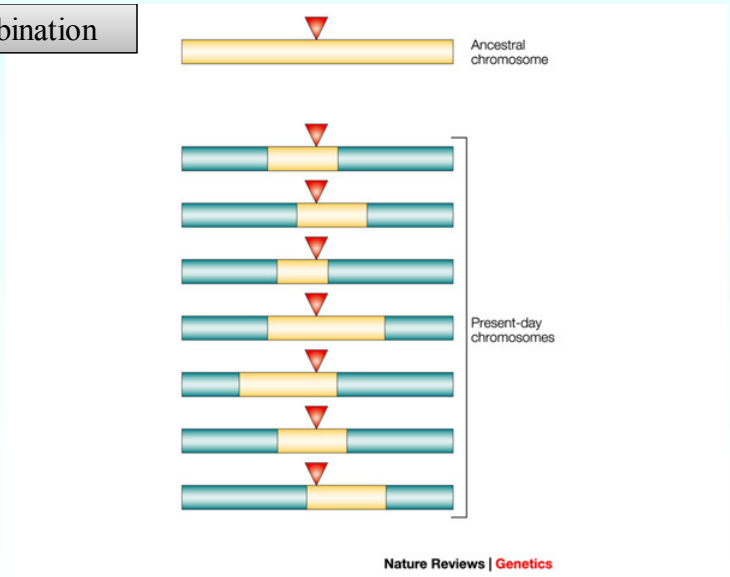
Sources of variability: Mutation

- It is caused by errors in DNA replication.
- DNA variability would soon become exhausted if mutation did not exist.
- Mutations cause different kinds of polymorphisms: SNPs, microsatellites, copy number variations ...
- Mutation rate varies along the genome, between sexes and across evolutive lineages.

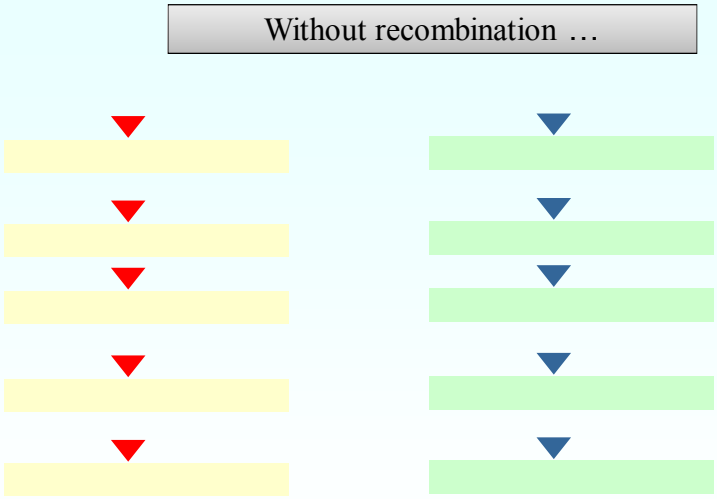
Sources of variability: Recombination

- Its frequency is not constant along the genome and differs between sexes.
- It is also affected by the environment and is partially under genetic control.
- It breaks down the genealogical history shared between two linked loci, making it possible association mapping, i.e., it establishes a link between physical distance and linkage disequilibrium.

recombination



Ardlie et al. 2002



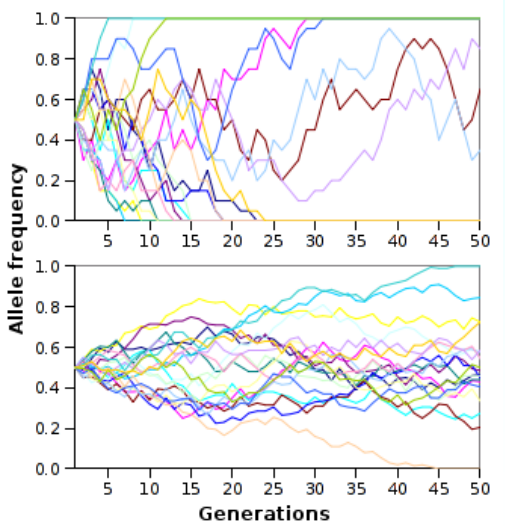
NO GWAS!!

Sources of variability: Drift

- Genetic drift it refers to the changes in allele frequencies caused by sampling. It is an unavoidable consequence of the finite size of populations.
- Given a population of size N and an allelic frequency p, the expected p in the next generation is given by the binomial distribution:

$$P(p_1 = p_0 \mid N, p_0) = \binom{N}{Np_0} p_0^{Np_0} (1 - p_0)^{N(1-p_0)}$$

Genetic drift



N=10

N = 100

wikipedia

Exercise: drift.R script

http://scit.us/redlynx/

```
#----- simulates drift
drift <- function(N,f) {
   t=0
   while(f>0 & f<1) {
      genotypes <- rbinom(N,1,f)
      f<-mean(genotypes)
      t=t+1
   }
   c(t,f)
}
```

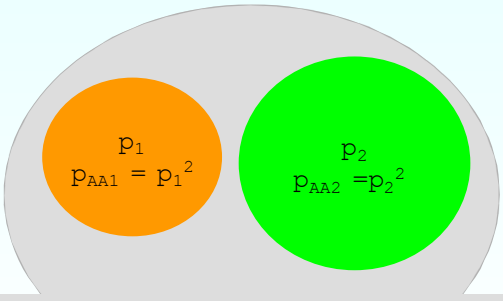
Sources of variability: Selection

- It is usually the main force we are interested in, because selection targets the loci for phenotypes of economic interest (artificial selection).
- It affects only a small subset of loci, whereas drift influences the whole genome.

Sources of variability: Admixture

- Neutral alleles will evolve differently in each population (because it is a Markovian process), causing a differentiation between populations.
- Admixture or migration between individuals of different populations cause a departure from expectation of the null model.
- A common phenomenon is analyzing a pool of populations ignoring population structure (Wahlund's effect).

Sources of variability: Admixture



$$p_T = (p_1N_1+p_2N_2) / (N_1+N_2)$$

$$p_{AAT} = (p_{AA1}N_1+p_{AA2}N_2) / (N_1+N_2) > p_T^2$$

Wahlund's principle: apparent excess of homozygotes (deficit of heterozygotes)

Measures of admixture: F statistics

Population structure (subdivision) results in an increased inbreeding – or decreased heterozygosity – relative to a single population of equal sum of sizes.

Three levels can be distinguished:

- H_I = Individual heterozygosity in a population, observed heterozygosity averaged across subpopulations.
- H_S = Expected individual heterozygosity in an equivalent random mating subpopulation.
- H_T = Expected individual heterozygosity in an equivalent random mating total population.

Perhaps the most surprising result of modern genomics is how dramatic phenotypic changes are accompanied by barely undetectable changes in DNA sequence

This is called 'The infinitesimal model'

It presupposes that quantitative traits are explained by a large number of genes, each acting individually and of small effect. In addition, quantitative traits are also modified by the environment.

XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. ARTHUR THOMSON. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

Practicals: chapter 1

BACKGROUND:
Statistics Concepts

In God we trust, all others bring data*

* Found in Tibshirani et al., attributed to both Deming and Heyden

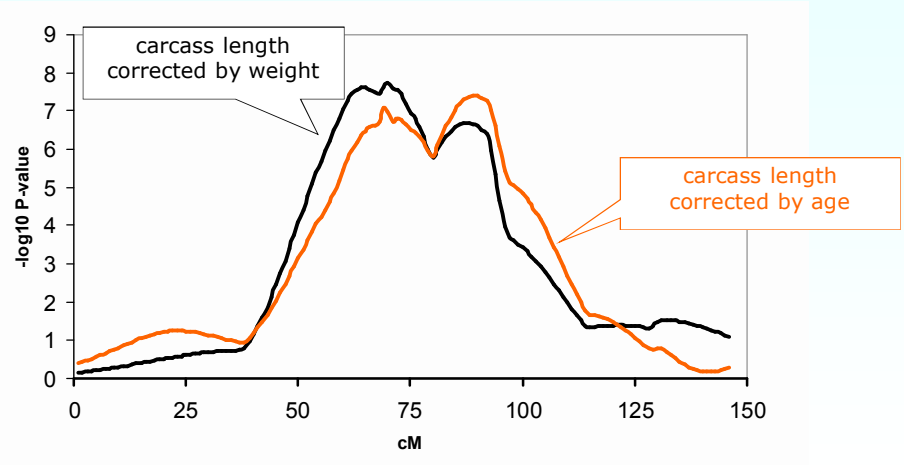
Statistics

- The term 'Statistics' and 'State' share the same etimology, as Statistics was initially (end 18th century) a means for the State to have censuses. This is now called **descriptive** statistics.
- But the real importance of Statistics in science lies in its role as **quantifying uncertainty**.
- The concepts of **inference** and **model** are central to this goal.
- Statistics is not a coherent and unified framework, there coexist many different schools (frequentist, Bayesian, non parametric) where many angry disputes have been witnessed.

Concept of Model

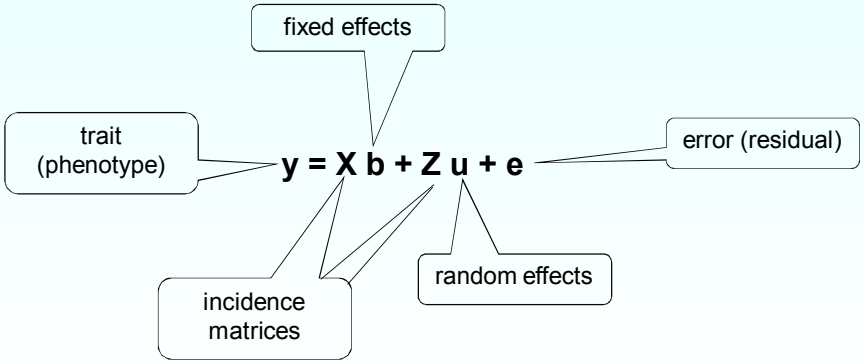
- A model is an abstraction of reality, it never exists a *true* model.
- 'The value of a model is that it often suggests a simple summary of data in terms of the major systematic effects together with a summary of the nature and amount of unexplained variation' (McCullagh and Nelder, 1983).
- Two desirable characteristics: **parsimony** and **goodness of fit**.
- In Statistical Genetics, a model is also a definition of the trait.

A model is a definition of a trait



Iberian x Landrace pig cross, SSC4

Usual models: Mixed Linear Models



Usual models: Mixed Linear Models

$$\mathbf{y = X\,b + Z\,u + e}$$

$$\begin{pmatrix} y \\ u \\ e \end{pmatrix} \sim N\left[\begin{pmatrix} Xb \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} ZGZ'\sigma_u^2 + I\sigma_e^2 & ZG\sigma_u^2 & I\sigma_e^2 \\ GZ'\sigma_u^2 & G\sigma_u^2 & 0 \\ I\sigma_e^2 & 0 & I\sigma_e^2 \end{pmatrix} \right]$$

Normality and additivity are two tightly linked phenomena

Inference methods: General

- Statistical inference means quantifying the values of the parameters in the model, e.g., b in previous formulas, or θ in general terms.
- It is about prediction rather than description.
- There are several inference frameworks (frequentist, Bayesian...) and several criteria within each framework (maximum likelihood, minimum least squares, maximum a posteriori, ...).

Inference methods: Maximum Likelihood

$$\mathbf{y = X\,b + Z\,u + e}$$

$$V = ZGZ'\sigma_u^2 + I\sigma_e^2$$

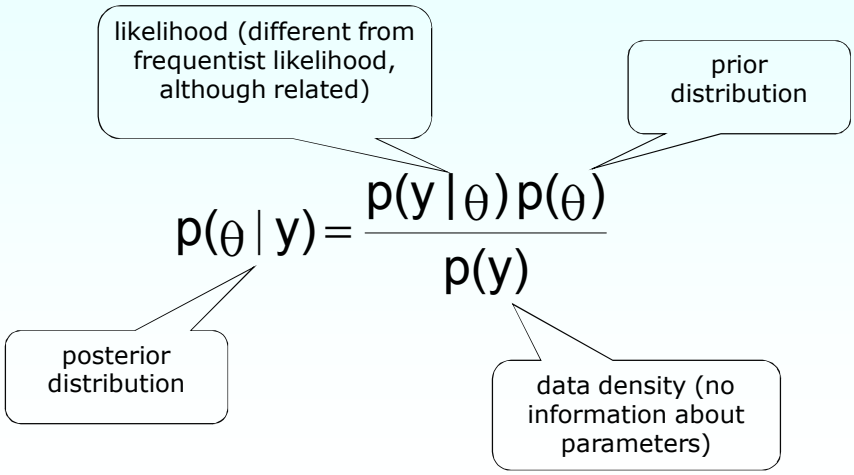
$$L(y) = k\,|V|^{-1/2}\,\exp\{-0.5(y-Xb)'\,V^{-1}\,(y-Xb)\}$$

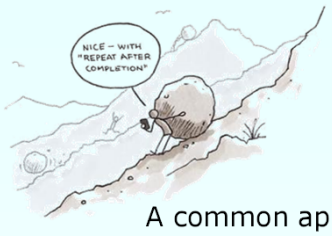
- A maximum likelihood estimate is the the set of values b and variances (σ^2) that maximizes the (log) likelihood.
- Nice properties but difficult to compute in most cases.

Inference methods: Bayesian methods

- In a frequentist framework (like the maximum likelihood estimate) parameters are evaluated in terms of how good is its performance under conceptual repeated sampling.
- In the Bayesian approach, inference is conditioned on the actual data, and parameters are evaluated in terms of decision theory.
- While the parameters are fixed unknown quantities that we want to know with as least error as possible in the frequentist paradigm, for Bayesian theory the parameters follow a distribution that reflect the uncertainty that we have about them.
- Bayesian methods are generally intensive computationally, and depends on fine tuning that requires a certain expertise.

Inference methods: Bayesian methods





Inference methods: Bayesian methods

A common approach for inference (obtaining the marginal posterior) for each parameter is Gibbs sampling.

Sample $p(\theta_1|\theta_2,\theta_3,..\theta_n,y)$
Sample $p(\theta_2|\theta_1,\theta_3,..\theta_n,y)$
…
Sample $p(\theta_n|\theta_2,\theta_3,..\theta_{n-1},y)$

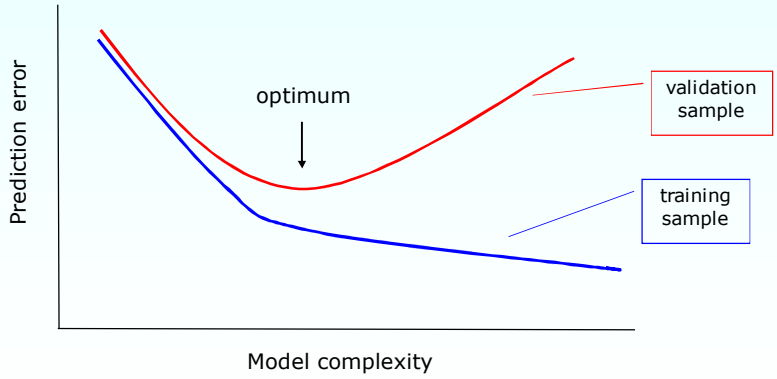
→

$p(\theta_1|y)$
$p(\theta_2|y)$
…
$p(\theta_n|y)$

conditional posterior

marginal posterior

Model comparison: A modeling tradeoff



Hastie et al. Elements of Statistical Learning

Model comparison

- Goodness of fit (distance between prediction and observation) is a good thing.
- But an increase in number of parameters should be penalized. This means that parsimony is also a desirable property.
- Model comparison is basically about choosing a compromise between these two concepts. There are many criteria to choose the best model. The ranking of models will depend on the criterion chosen.

Multiple comparisons: Bonferroni correction

- Recall the definition of type I error, α, it is the number of times the null hypothesis is rejected and is true. At the usual α=0.05 the number of times H_0 is rejected can be very high if the number of tests say SNPs is very high.
- A traditional approach has been to control the total number of false positives by raising the significance threshold.
- Bonferroni correction is to set the threshold to α/n where n is the total number of tests. Bonferroni is highly conservative as it assumes independence between tests.

Multiple comparisons: Permutation tests

- Under H_0 the relationship between the data and the parameter is circumstancial.
- Thus, one can generate the distribution of a given statistic by randomly shuffling the data and the explanatory variable(s).
- Advantages: robust, it does not assume any null distribution, easy to compute.
- Inconveniences: difficult to apply with hierarchical tests (e.g., 2 QTL vs. 1), when there is an additional relation between individuals (e.g., relationship matrix).

Multiple comparisons: False discovery rate (FDR)

- FDR is the number of false significant tests divided by the total (false + true) number of significant tests.
- Suppose a number n of P-values ranked from lowest to highest $P_1 < P_2 < \ldots. P_n$, then a β FDR is attained by selecting P-values below the threshold (find the largest k s.t.):

$$P_k \leq \frac{k}{n}\beta$$

Multiple comparisons: Example

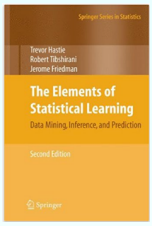
α=0.05/10

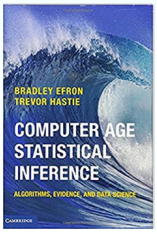
i	P-value	i/n	i/n 0.05	FDR=0.05	i/n 0.5	FDR=0.5	Bonferroni
1	0,002	0,1	0,005	-0,003	0,05	-0,048	0,002
2	0,001	0,2	0,01	-0,009	0,1	-0,099	0,001
3	0,003	0,3	0,015	-0,012	0,15	-0,147	0,003
4	0,006	0,4	0,02	-0,014	0,2	-0,194	0,006
5	0,091	0,5	0,025	0,066	0,25	-0,159	0,091
6	0,3	0,6	0,03	0,27	0,3	0	0,3
7	0,34	0,7	0,035	0,305	0,35	-0,01	0,34
8	0,42	0,8	0,04	0,38	0,4	0,02	0,42
9	0,54	0,9	0,045	0,495	0,45	0,09	0,54
10	0,73	1	0,05	0,68	0,5	0,23	0,73

Multiple comparisons: Important remark

Unfortunately, one of the consequences of complexity is that raising the significance threshold over a given (unknown) optimum does not reduce the rate of false positives but does decrease power.

BACKGROUND:
Statistics vs. Machine Learning





- Machine Learning is a wide field related to developing algorithms that can automatically identify patterns in data (and use them for prediction of future records).
- Statistics and ML can be highly interrelated with many shared concepts and procedures (albeit often with distinct vocabulary).
- Historically, ML has been developed by Computer Scientists whereas Statistics has been linked to Mathematical Faculties. This is no longer true, especially with intensive use of computers by modern Statistics.

Statistics	Machine Learning
➢ Focused on inference	➢ The target is prediction
➢ Based on Models	➢ Model free
➢ Theoretically founded	➢ Pragmatic
➢ Problem constrained	➢ Data heterogeneity is no problem
➢ Clear interpretation	➢ Often cannot be interpreted
➢ General solutions	➢ Specific solutions