Multiple Testing
GWAS

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

HerrFalloppio

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)

Oscar González-Recio

OscarGenomics

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)

simple testing

- **inference** → is there a **difference** between groups?
  - e.g. AA vs AB vs BB
- **significance** is related to the **size** and **variance** of this difference
- ***p*-value**: prob of obtaining such an extreme t statistics under H_0 given we repeat the experiment an infinite number of times
  - P-value < α→ small likelihood of the data under H_0 → significant difference
  - P-value >α→ there is a high chance of observing these data if there is no difference between groups
- α = 0.05 → threshold: 5% of rejecting H_0 when it is true (Type I error).
  - **false positive**: significant result when there is no difference (H_0 is true)

multiple testing

- **many tests → many false positives**
  - e.g. 2000 (independent) tests, α=0.05 → How many expected false positives?

    100 false positives by chance alone
- multiple testing problem

- A typical GWAS conducts hundreds of thousands to millions of tests independently, each for a single marker and with its own false-positive probability.
  - many SNPs, many statistical tests, many p-values (large p, small n problem)

How to cope with the problem

1. Increase the sample size
   (e.g. Bio Banks)
2. Reduce the number of tests
   - Based on LD
   - Choose relevant regions (functional analysis)
3. Decrease the significance threshold
   - **Arbitrary significance level** (e.g. $5x10^{-8}$)
   - **Bonferroni correction**
   - **False discovery rate**
   - ***q* values** (*important pitfalls*)
   - **Permutation analysis**
   - **Go Bayesian...**

Bonferroni correction

- Bonferroni, mathematician (1892 - 1960)
- **adjust** the significance threshold:
    - **New significance threshold $\leqslant \alpha/m$**

      [m: number of tests (markers)]

- Bonferroni correction tends to be too conservative
    - few false positives
    - many false negatives

False discovery rate (FDR)

- Decrease the significance threshold

0.010
0.025
0.026
0.031
0.042
0.049
0.050
0.065
0.078
0.101
0.125
0.128
...

List of ordered p-values

1) If I apply a threshold alpha to decide on significance, how much can I trust the results?
2) Where should I draw a line (threshold) of significance so that at most e.g. 10% of results are false positives?

False discovery rate (FDR)

- **FDR:** how many of the positive results are false positives?

- Benjamini & Hochberg (1995), Storey (2002), Storey & Tibshirani (2003)

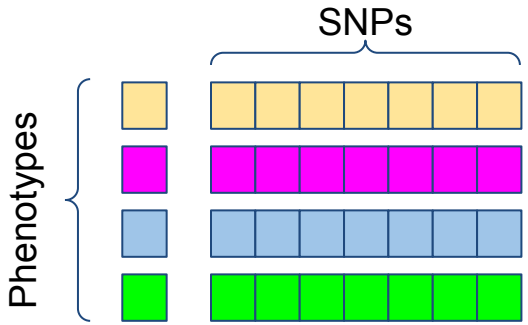
- **Significance level = 0.05** → 5% of **all** tests on average will be false positives (assuming independency)
- **FDR = 0.05** → 5% of **significant** tests will on average be false positives

fewer false positives!

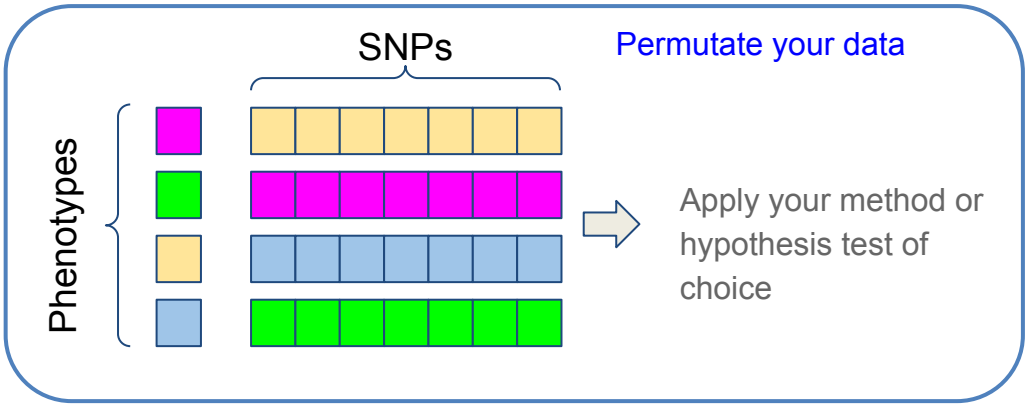
Permutation tests

- Determine the significance of a result by randomly reshuffling the data and recalculating the test statistic.
  - This allows to test the null hypothesis that the observed difference between two groups is due to chance, rather than a real difference between the groups.
- Permutation tests are often used when the assumptions of traditional parametric tests are not met, or when the sample size is small.
- They are also useful when the data is not normally distributed, or when the groups being compared are not independent.

Permutation tests



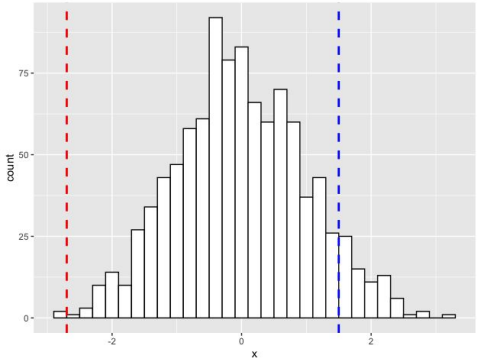
Apply your method or hypothesis test of choice



Permutate your data

Apply your method or hypothesis test of choice

x large number of times (e.g. x1000)



Non-significant, within 95%HPD

Significant, out of 95%HPD

Bayesian inference

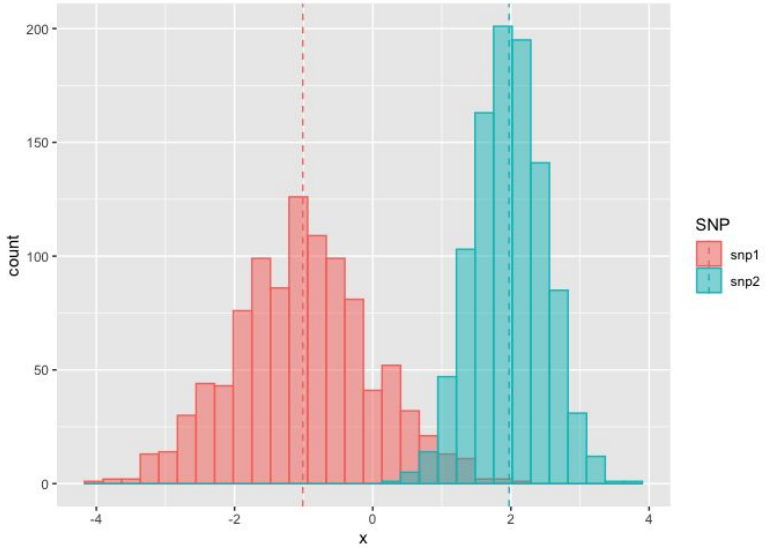
- In Bayesian inference, the probability of a hypothesis is updated using Bayes' theorem as follows

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

  - where P($\boldsymbol{\theta}$ |y) is the updated probability of the effect given the new evidence, P(y|$\boldsymbol{\theta}$) is the likelihood of the evidence given the effect, P($\boldsymbol{\theta}$) is the prior probability of the effect, and P(y) is the probability of the evidence.
  - Make inferences of the posterior distribution using McMC algorithms (Gibbs sampling, acceptance rejection, Metropolis-Hasting)

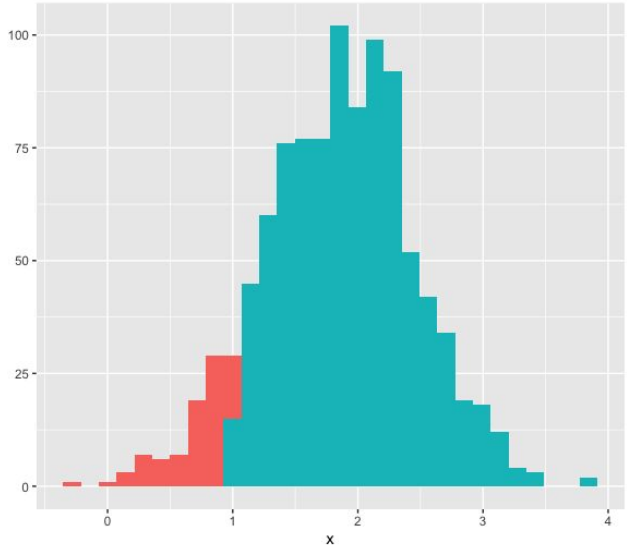
Bayesian inference

- What is the mean of the posterior distribution and its standard deviation?
- Does it contain zero?



Bayesian inference

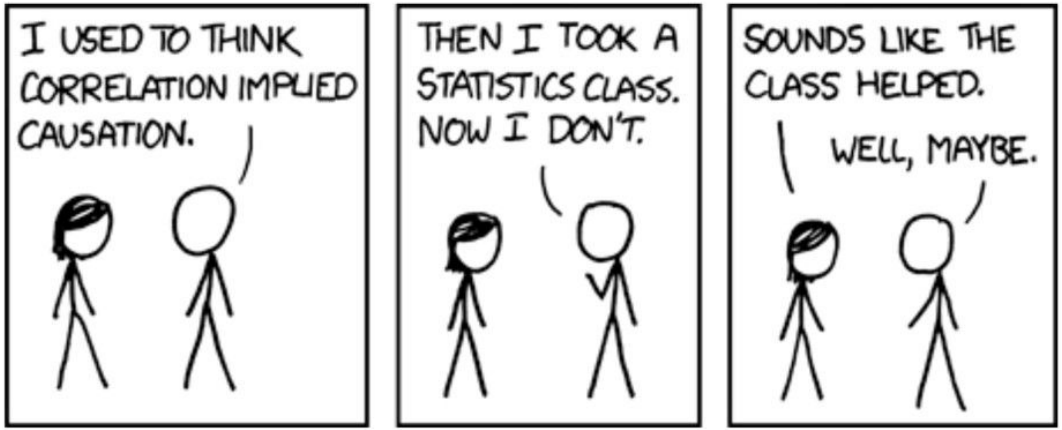
- What is the probability of the effect being larger than a relevant magnitude (e.g. +1).
- Is it a sufficient probability (e.g. 80%)



Possible to combine Bayesian inference and permutation test

REMEMBER

- Correlation does not imply causation



https://xkcd.com/552/

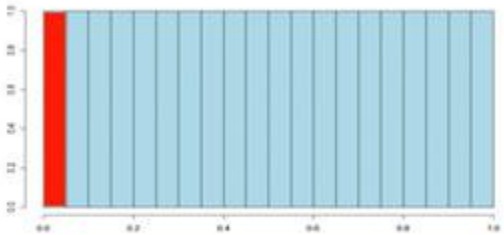
Make your rationale choice

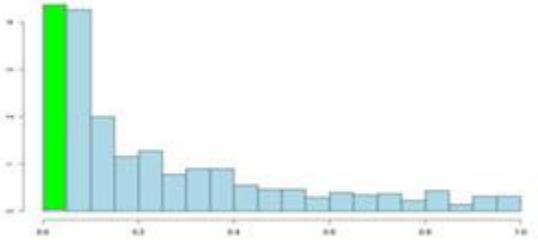
NEXT LECTURE

Power of GWAS experiments

q-values

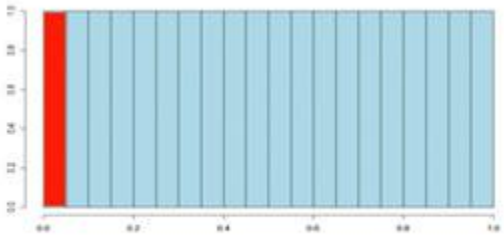
- q-values: proxies for FDR based on the **distribution of p-values**

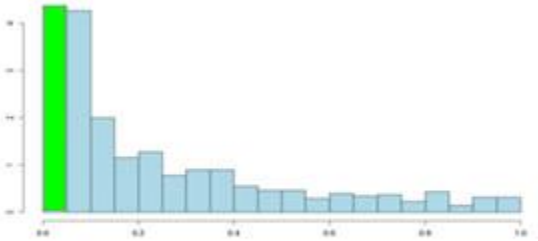




q-values

- q-values: proxies for FDR based on the **distribution of p-values**



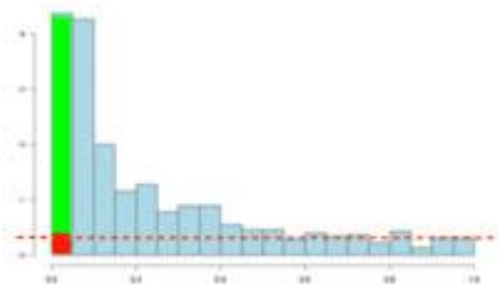


no significant differences

significant differences

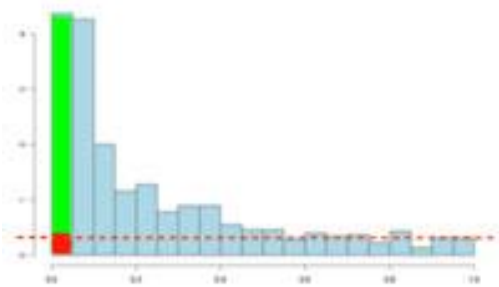
q-values

- the q-value approach tries to find the proportion of significant results which are likely to be false positives
- intuitively, it finds the height (density) at which the distribution of p-values flattens out

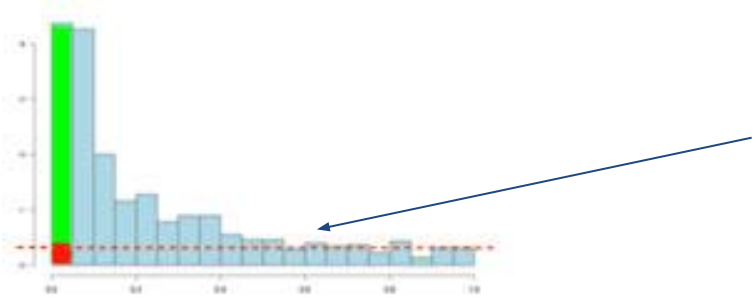


q-values

- the q-value approach tries to find the proportion of significant results which are likely to be false positives
- intuitively, it finds the height (density) at which the distribution of p-values flattens out



q-values



here the distribution is similar to the case where there is no actual difference

- this proportion of false positives is then incorporated in the calculation of adjusted p-values (**q-values**)

interpretation of q-values

- *Significance level* = 0.01 → probability of the p-value under H_0
- q-value = 0.02 → probability of the SNP being a false positive

interpretation of the single SNP

- *Significance level* = 0.01 → 1% chance of false positives (e.g. 7900 SNPs → 79 false positives expected)
- q-value = 0.02 → 2% of positive results may be false positives (e.g. 800 SNPs with q-value ≤ 0.02 → 16 false positives expected)

interpretation of the distribution of SNPs

q-values

- What's **wrong** with **q-values**?
  - They assume p-value is the probability of rejecting the null hypothesis when it is true
  - They do not consider that p-values are drawn from a probability distribution, and assume an infinite repetition of the experiment (obtaining different p-values for each experiment).