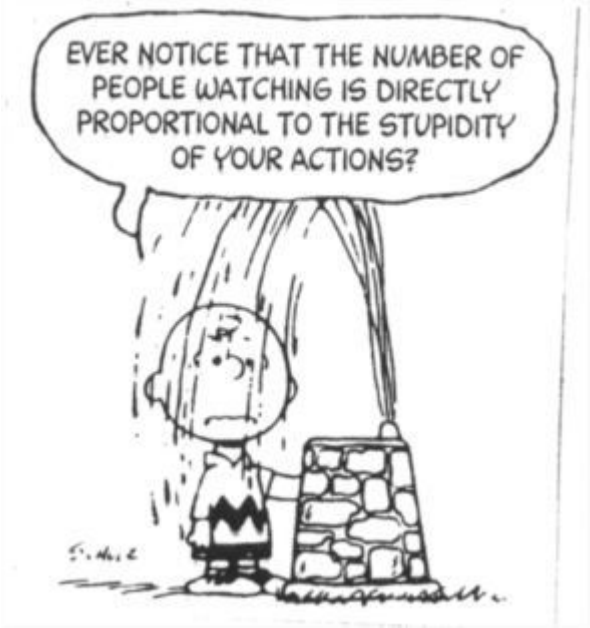
The univariate way

Pietro Franceschi
Computational Biology - Research and Innovation Centre - Fondazione E. Mach

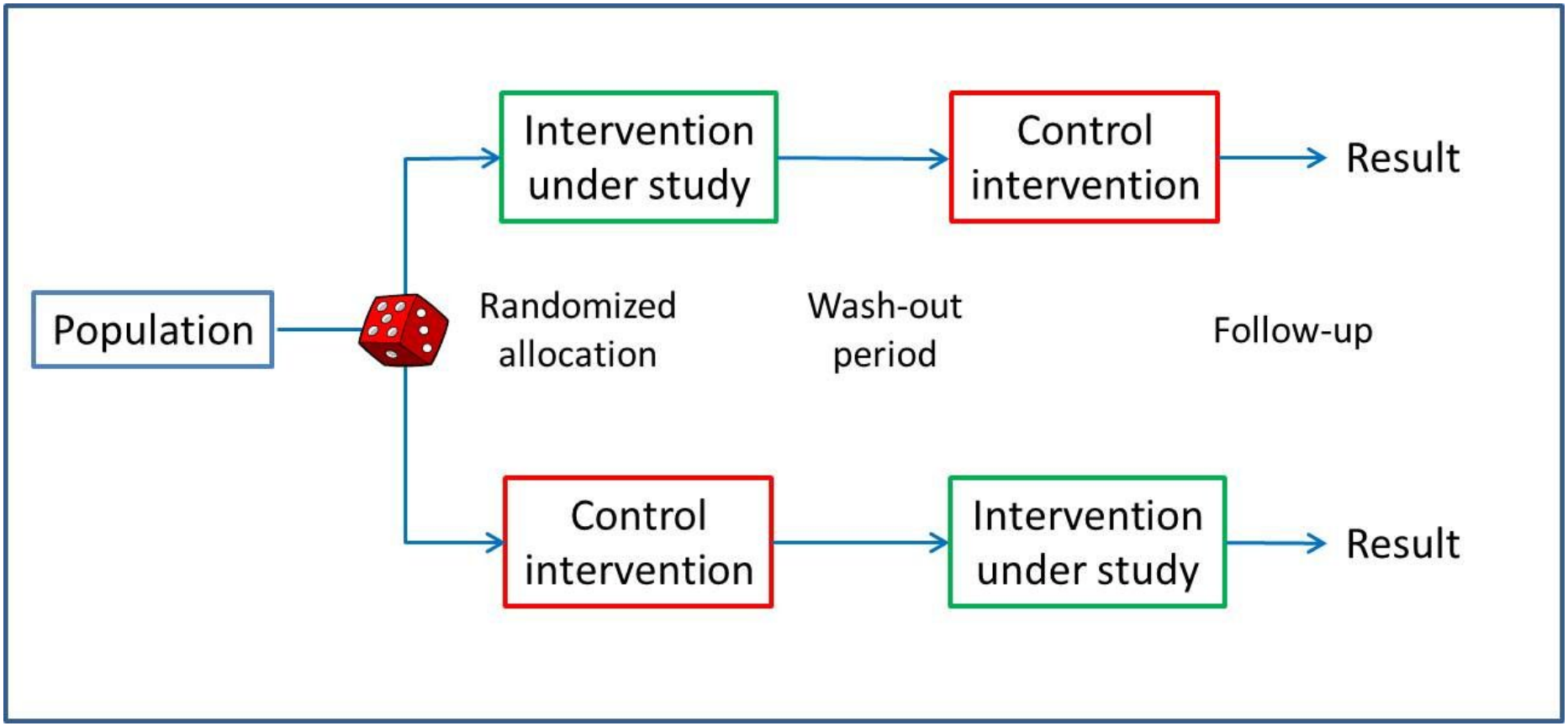
pietro.franceschi@fmach.it

Univariate in -omics

- Easier to interpret
- Easier to adapt to complex experimental designs



An example: Crossover designs

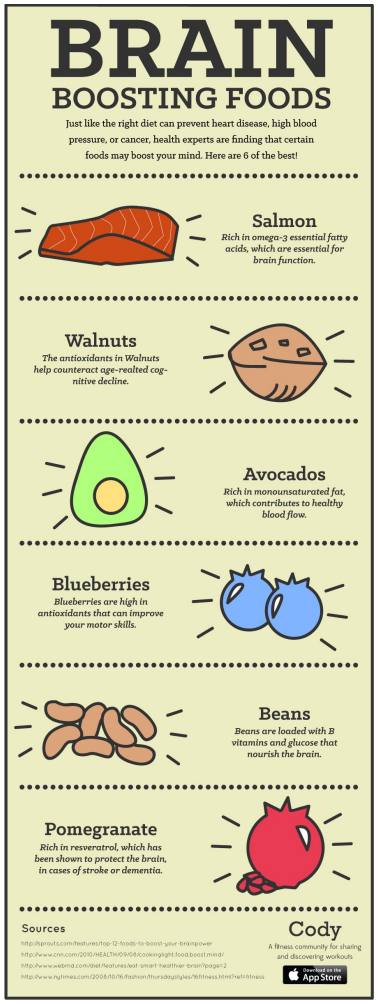


… food for brain

Different individuals with different characteristics (age, sex, BMI, ...)

- Different treatments
- Different baselines

These infos can be “easier” included in classical statistical modeling



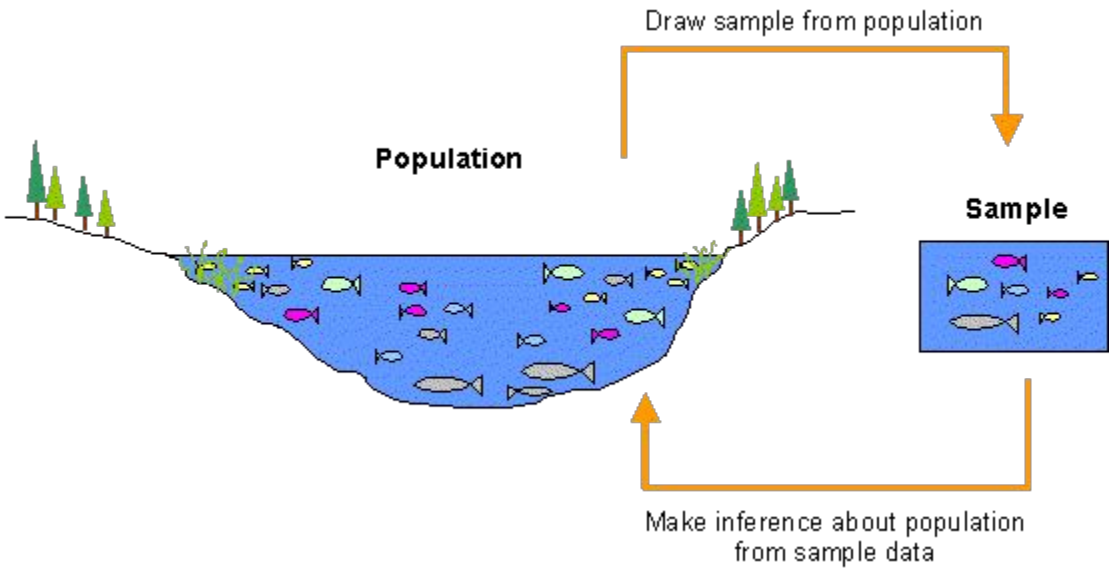
Collection

Statistics for Biologists

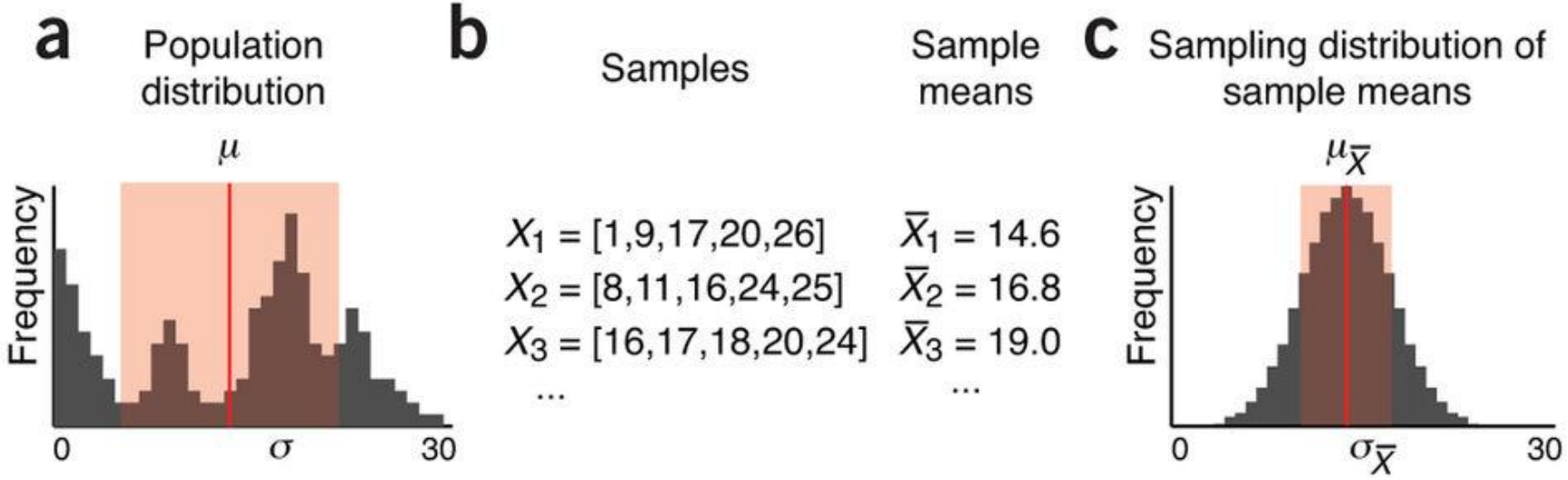
Since September 2013 *Nature Methods* has been publishing a monthly column on statistics called "Points of Significance." This column is intended to provide reseachers in biology with a basic introduction to core statistical concepts and methods, including experimental design. Although targeted at biologists, the articles are useful guides for researchers in other disciplines as well. A continuously updated list of these articles is provided below.

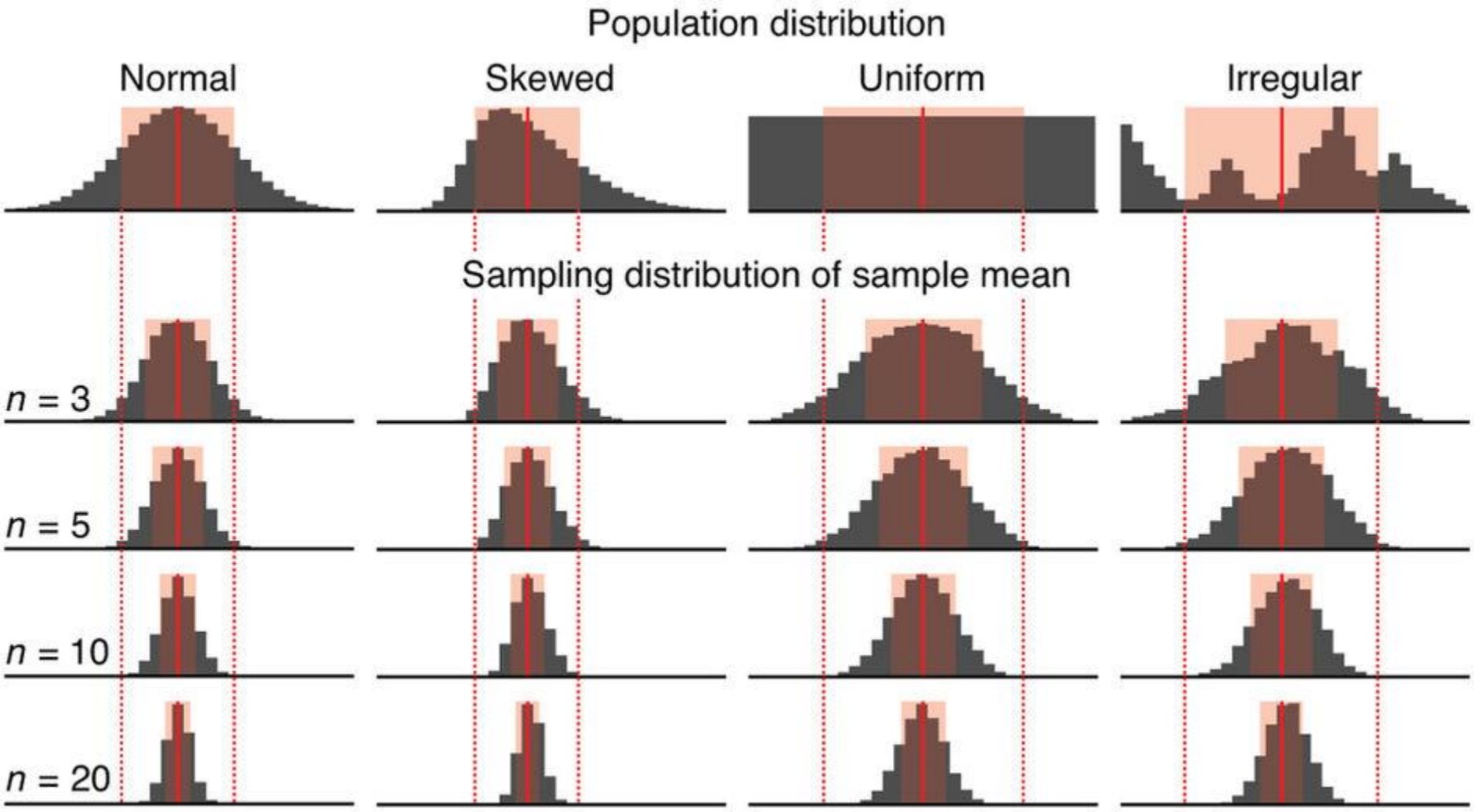
https://www.nature.com/collections/qghhqm/pointsofsignificance

Why is better to measure more samples?

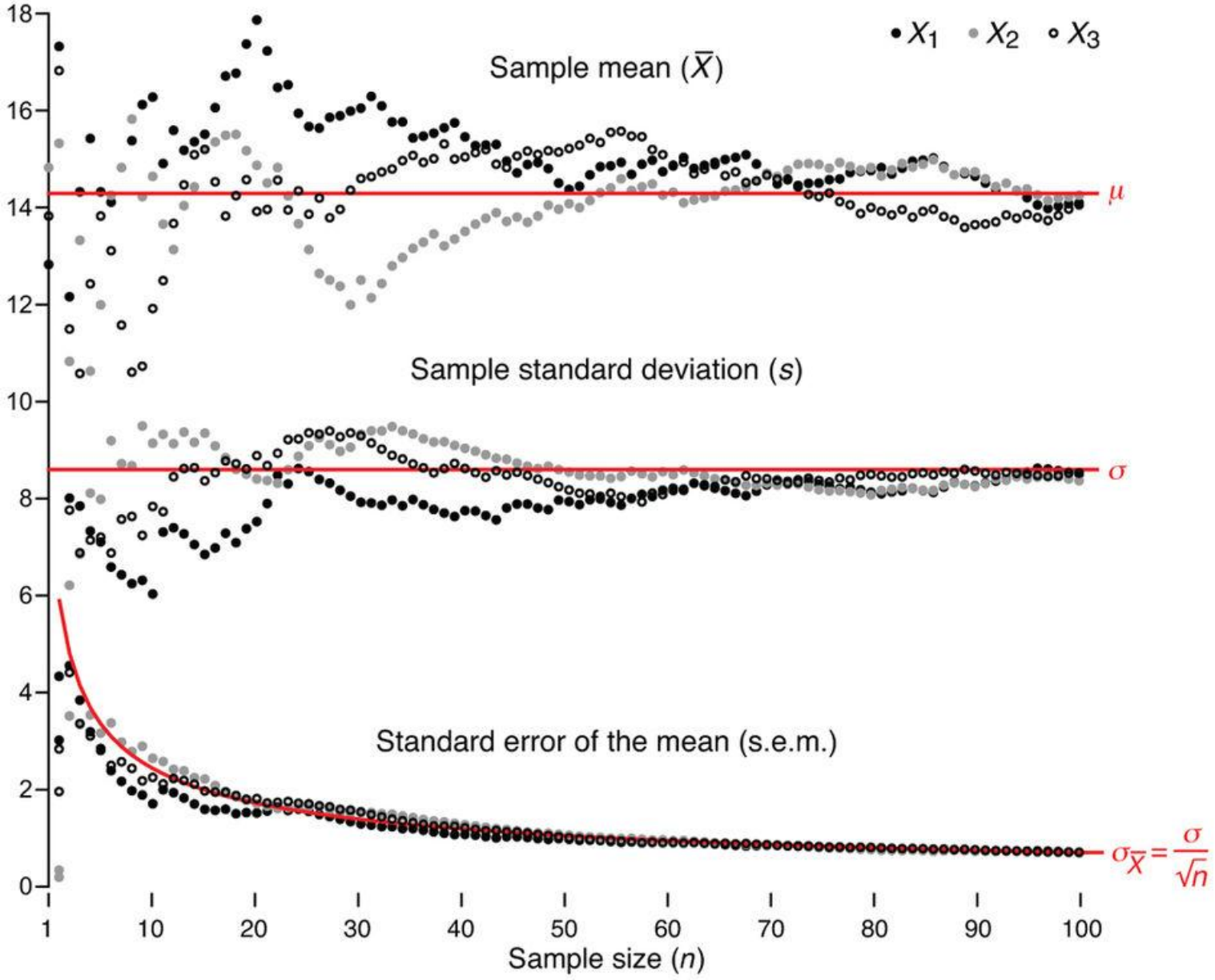


Why is better to measure more samples?



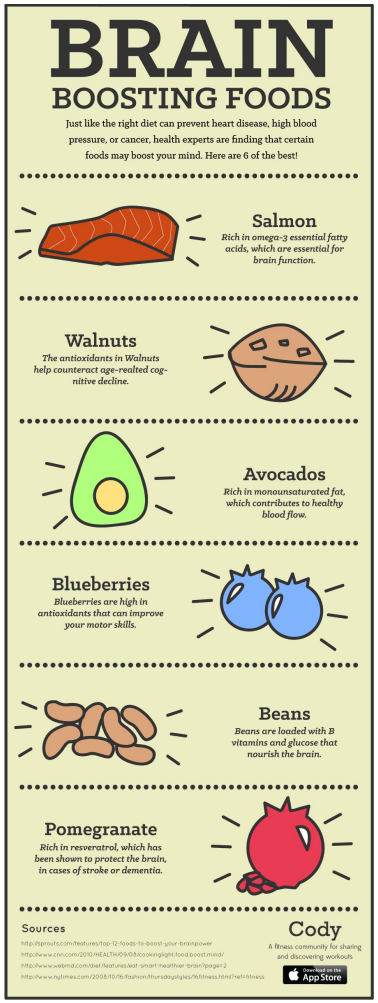


Central Limit Theorem

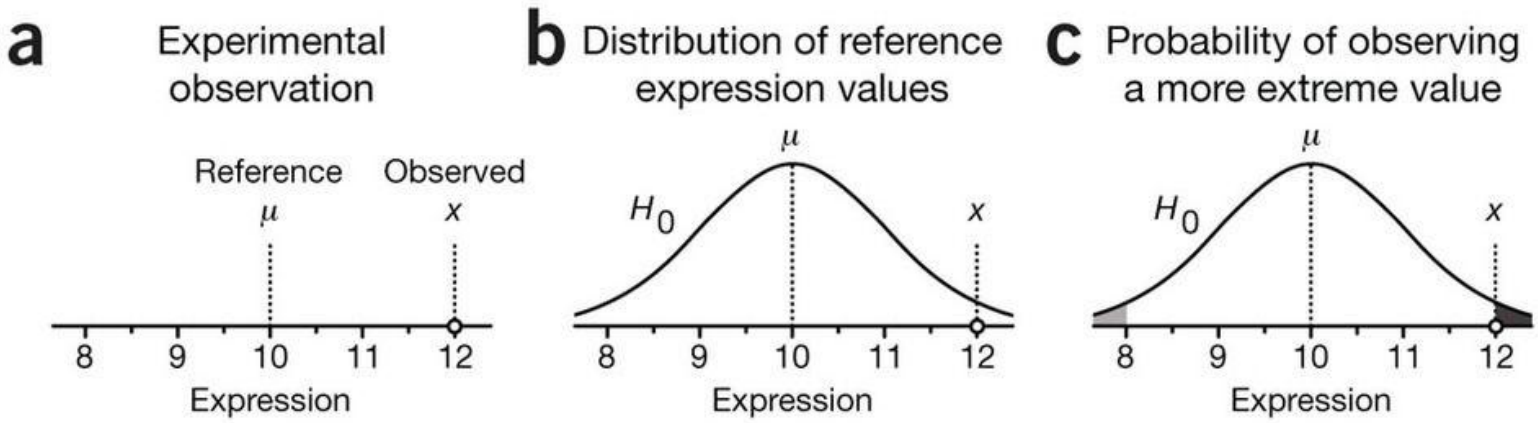

X_1 X_2 X_3
Sample mean ($\overline{X}$)
μ
Sample standard deviation (s)
σ
Standard error of the mean (s.e.m.)
$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$
18
16
14
12
10
8
6
4
2
0
1
10
20
30
40
50
60
70
80
90
100
Sample size (n)

… food for brain

- My estimate of sigma is “not” changing!
- My estimate of the man is “not” changing!
- The standard error of the mean is going to zero ...

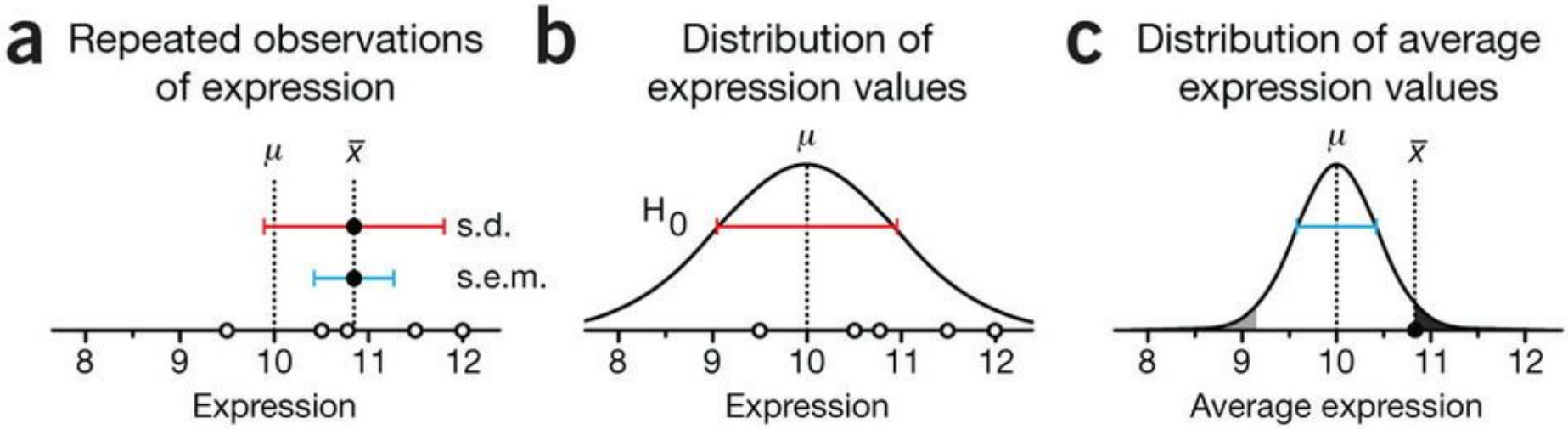


… testing statistical significance for dummies



A. I have a reference and a measure … difficult
B. I have more samples, so I can estimate the H0 distribution!
C. With the assumption that I can estimate the spread of H0 from the data … I can calculate the p-value

… if I measure several times …



A. I estimate sigma from s.d. … and s.e.m. from sd and the number of samples
B. I can construct H0 and derive the distribution of the sample means ;-) !
C. I calculate the p-value … so with more samples the situation gets better ;-)

All this stuff can be done with Student's t-test ...

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}}$$

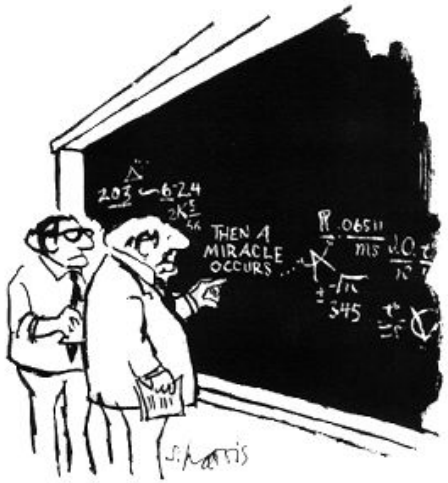
Is distributed wits a t-distribution

- Statistical testing in complex designs
- Effect size
- Power calculation
- Non parametric tests

Multiple tests ...

… univariate tests in -omics

In a typical -*omic* experiments we want to test "univariate" hypothesis over thousands of variables. More often than not the question is "global" … something like "is there something different?"



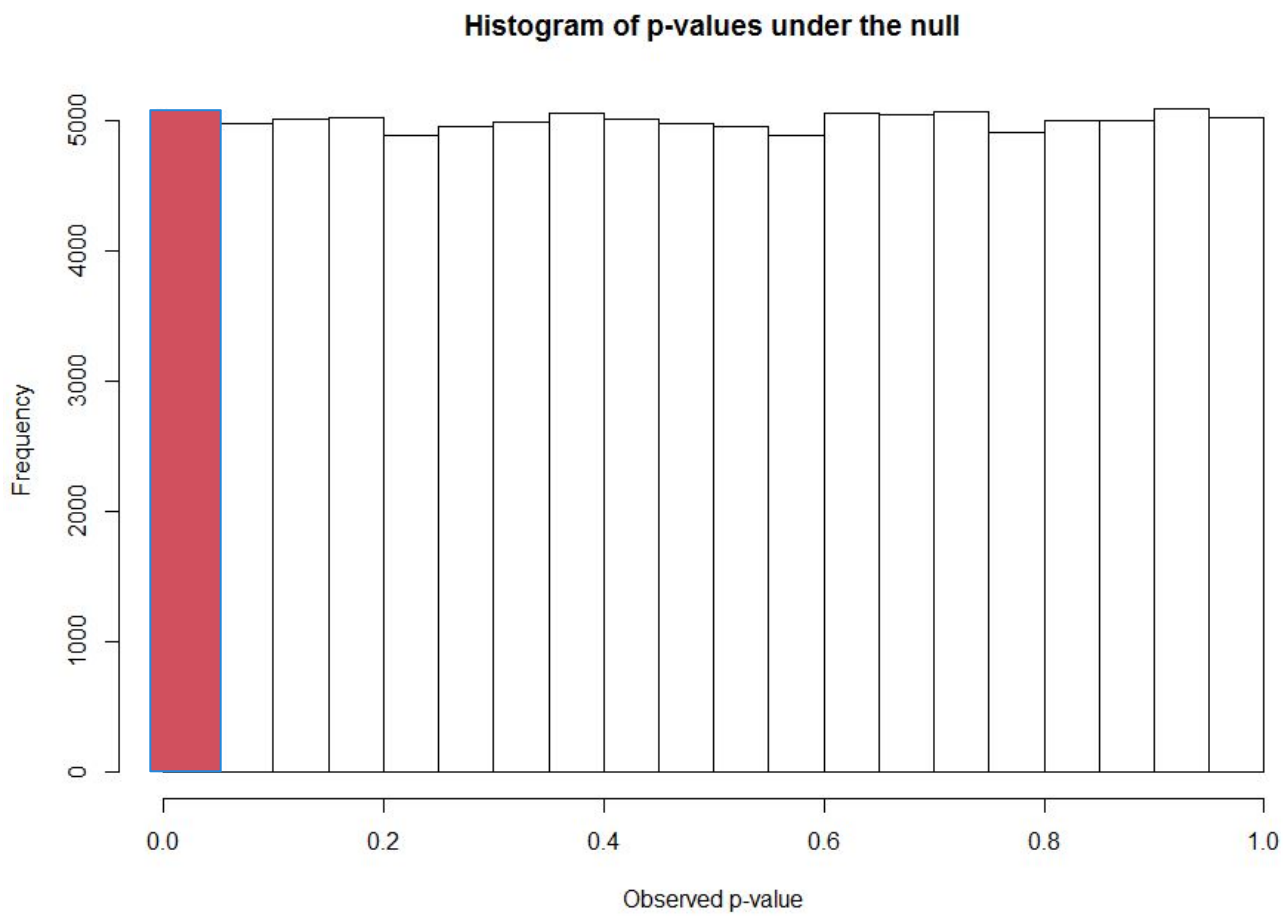
"I think you should be more explicit here in step two."

Something expected?

- We construct a “fake” data matrix of 20 samples and 1000 variables
- We divide them in 2 classes. The first ten versus the second ten
- We perform a t-test on all the variables (1000 p-values)

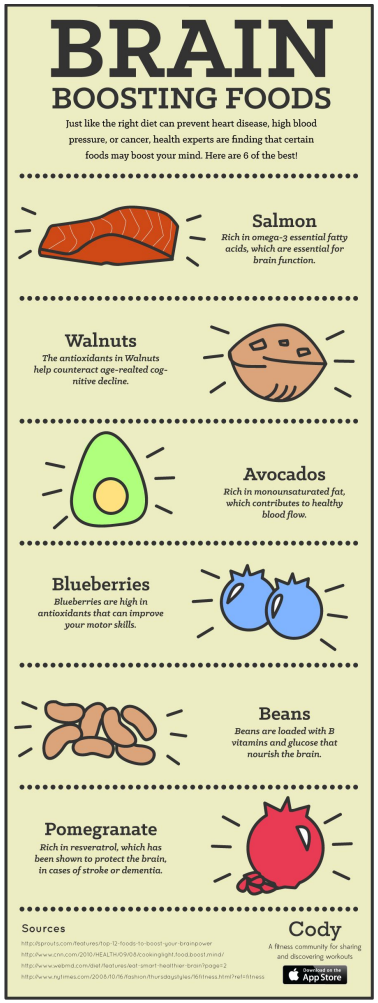
What is, in your opinion, the shape of the distribution of the p-values?

Et voilà ....


Histogram of p-values under the null
Frequency
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
Observed p-value

… food for brain

- I always have low p-values!
- A lower p-value is not telling me that selling this variable as a biomarker is “safer”
- We are back to the problem of “random” organization we discussed yesterday ...



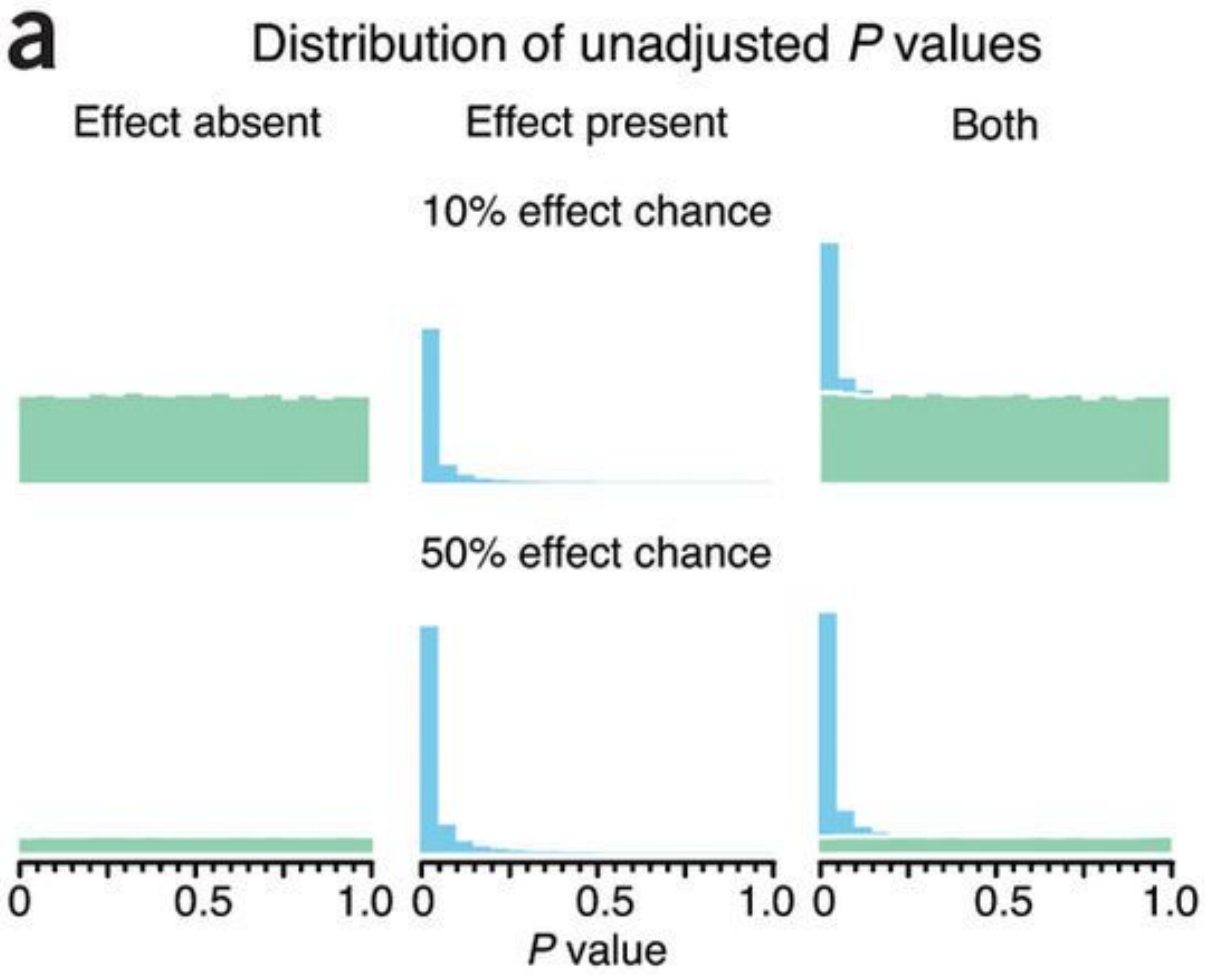
… the multiplicity issue

Each test has its own p-value … but the probability that we wrongly reject at least one null hypothesis grows with the number of tests

What can we do ...

- Forget the problem and live in peace
- Change the threshold of tolerance (the alpha) to control the probability of making one “error” (Control FWER - Bonferroni)
- Accept (and control) that a fraction of my biomarkers will be “false positives” (control False Discovery Rate)

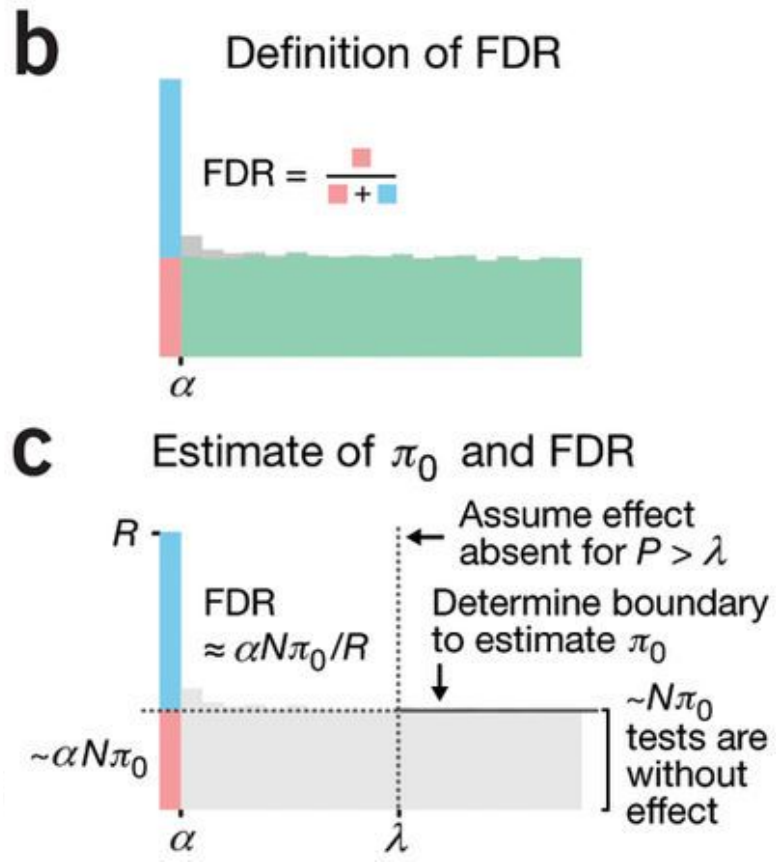
Suppose we know the true biomarkers …



10% Of the variables are markers

50% Of the variables are markers

False Discovery Rate



One uses the “theoretical” distribution of the p-values under H0
To estimate the fraction of false positives

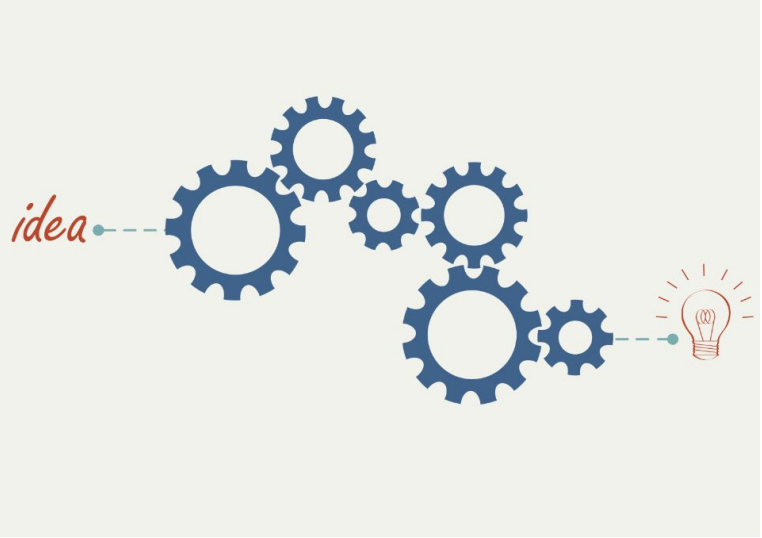
Questions for you ...

- Do I always need statistical significance?
- Do you expect that the correlation among the variables could change this picture?
- Why ?
- ...

Linear Regression

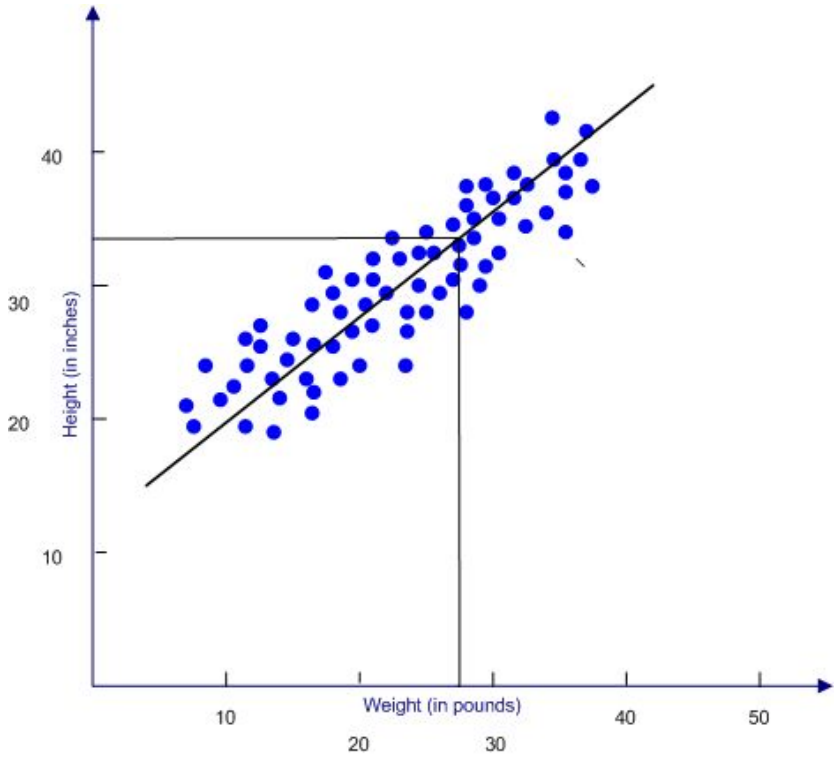
Regression

- Examine the “association” among two or more variables
- In its simplest form, one of the two variables changes as a function of the other with a trend that, on average, is a straight line



Scatter Diagram

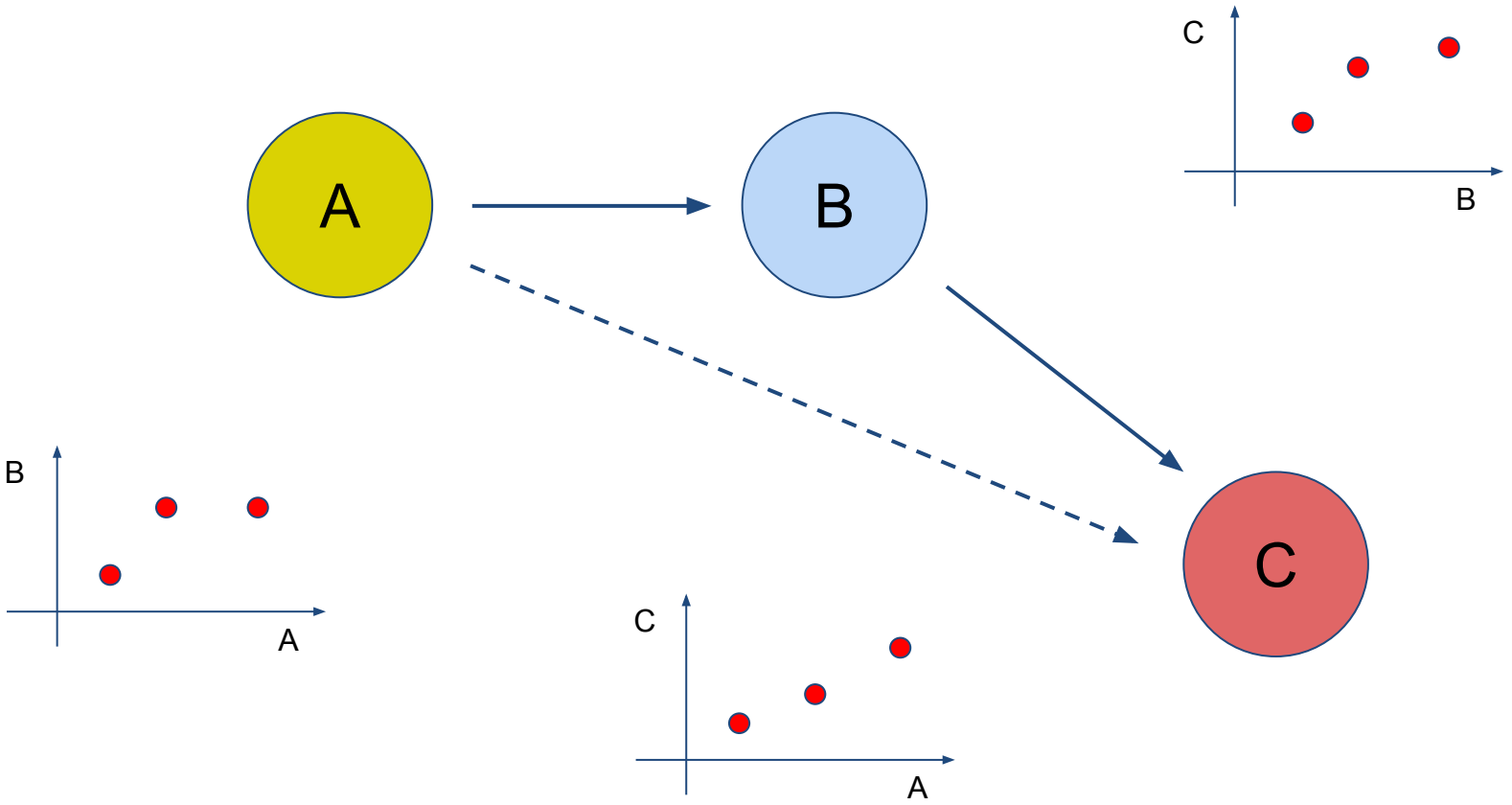
- A calibration curve inside the so called “linear range”
- The relation between height and weight (and size of the shoes ;-))
- ...



Regression analysis: goals

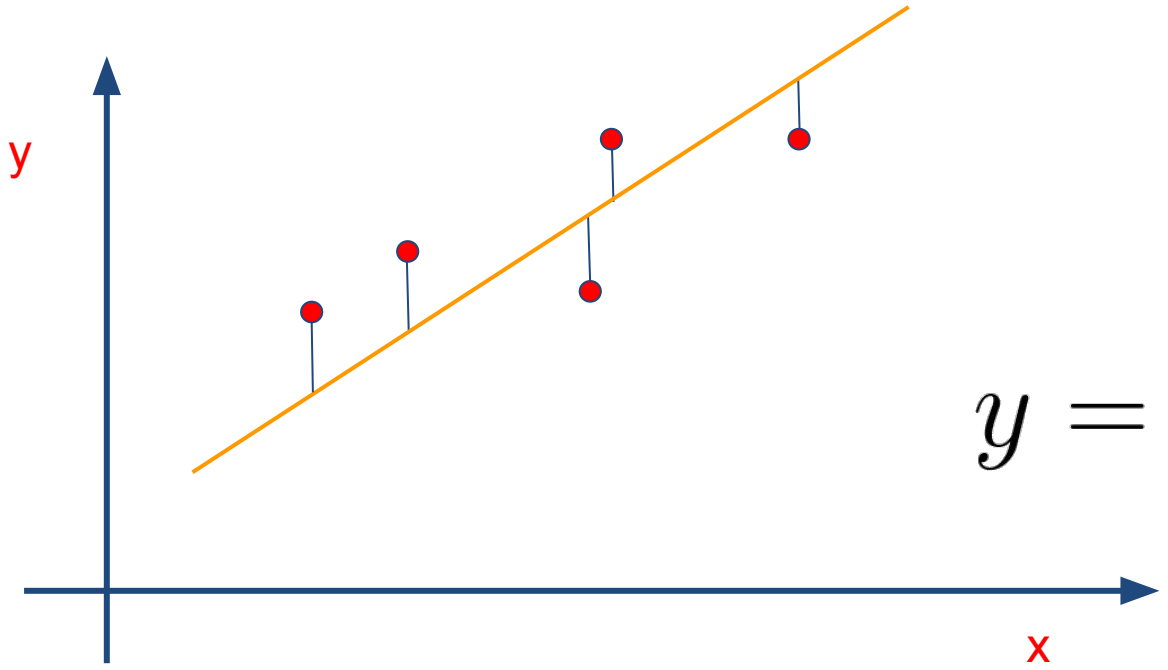
- Determine “objective” parameters to quantify association
- Predict new values knowing the value of the dependent variable
- Assess the significance of a “trend” in the data
- Correct for the “influence” of a variable on the signal of another one ...

Partial Correlation



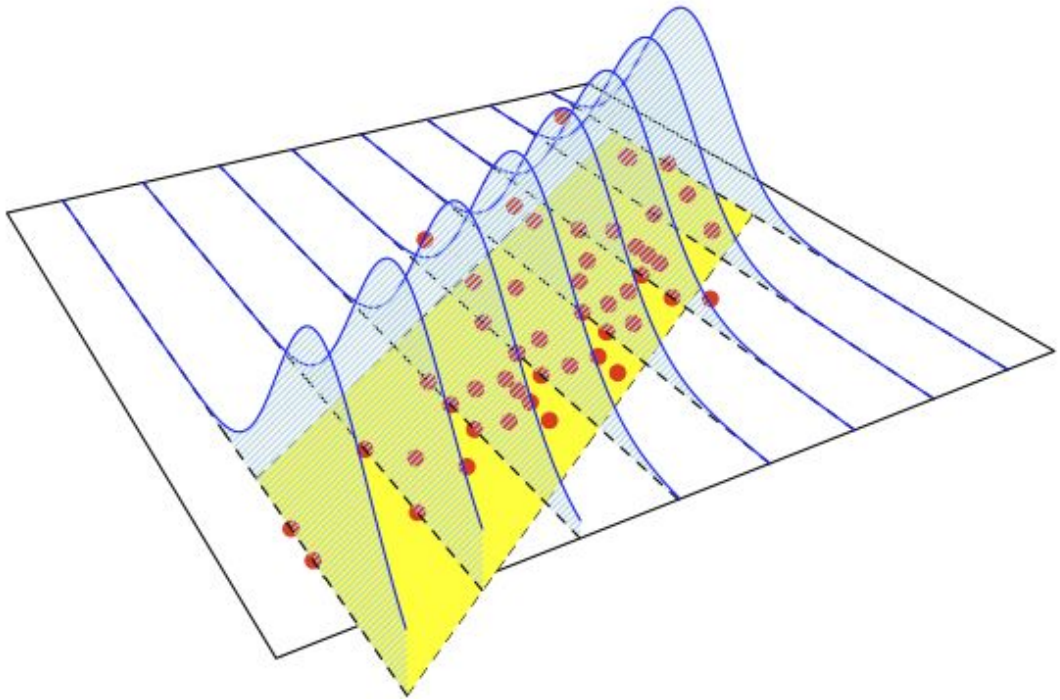
A geometric strategy ...

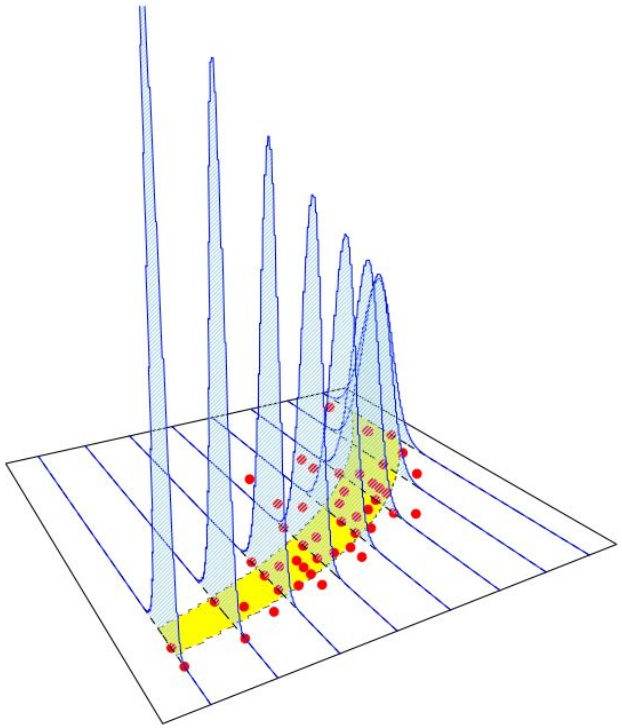
The “best” line is the one that minimizes the (square) vertical distance



$$y = ax + b$$

… a more statistical point of view …





http://freakonometrics.hypotheses.org/category/actuarial-science/ratemaking/glm-ratemaking


#1
LETS GO
LIVE!!!