Statistical Testing and Multiple Comparisons

Pietro Franceschi
pietro.franceschi@fmach.it

FEM - UBC

Fact sheet

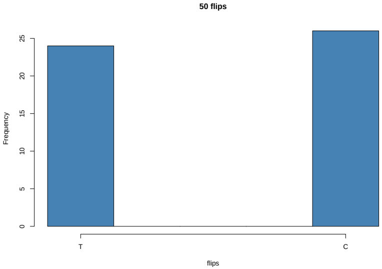
- **Scientific results** should be of *general validity*
- We **infer** general results from limited number of observations
- **Variability** is unavoidable ad can be big
- Due to **Variability** interesting can be the results of chance

The common shape of variability

- What we observe is the result of a “chain” of processes (e.g. gene → protein → metabolite)
- We never observe only one chain (e.g we consider *many* people with similar metabolism)
- The fact that we have noise on the “chain” produces variability in the output
- This variability has a bell shaped profile, which is more often than not **Gaussian**
- We have to measure more than once ;-)

Coin toss

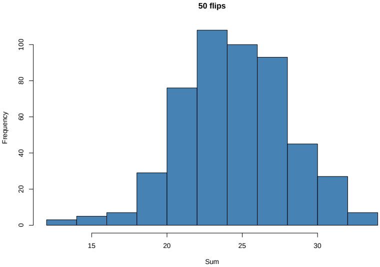
The distribution of the outcomes of 50 tosses of the same coin



This “biological” process is clearly non normally distributed, but has variability

Sum of 50 coin tosses

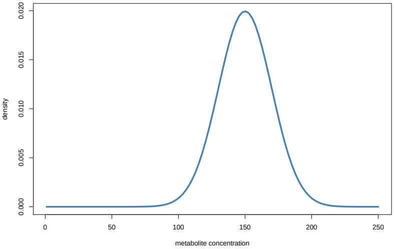
Suppose that now my “biological” process is the result of the sum of 50 coin tosses where T counts as 1 and C as 0. What is the distribution of the results of 500 sums?



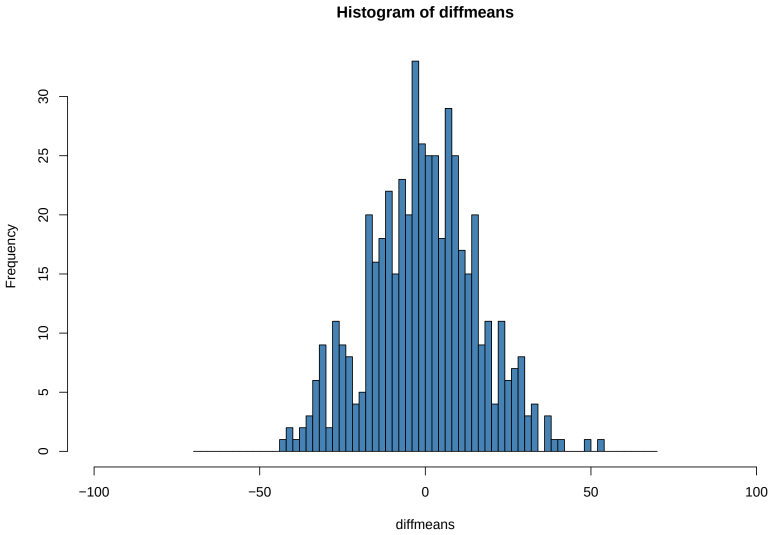
Here the “biological” process yields normally distributed data!

Example

- Let's consider a property (concentration of a metabolite, physical measure, . . .) normally distributed in the population (mean $= 150$, standard deviation $= 20$).
- Let's repeatedly extract two groups of three samples from the population
- Let's consider the difference between the means of the two groups . . .



Distribution of the differences



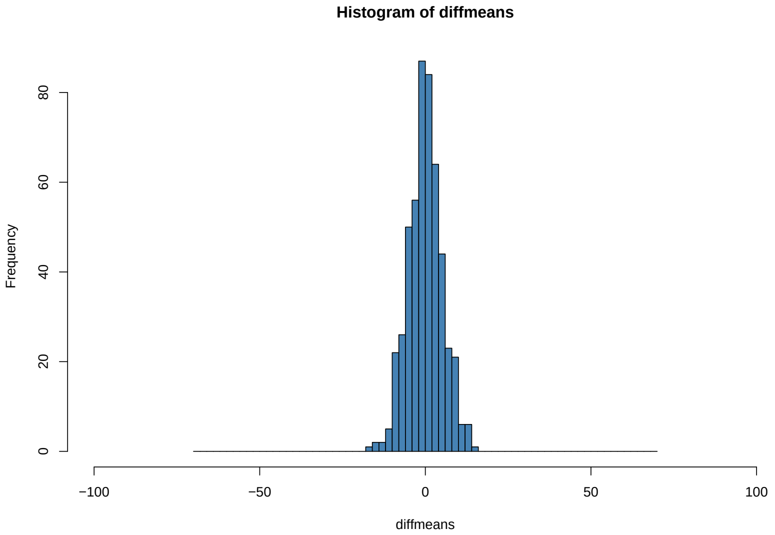
Observations

- The histogram is centered around zero . . .
- We can get differences as large as 50
- 50 is 1/3 of the population mean
- **Bad Luck** is unavoidable

The same holds

- Biomarkers
- Clusters
- Variable correlations
- . . .
- Any type of potentially interesting result

More samples . . .



What can we do . . .

- Forget the problem and live in peace
- Enlarge the sample size . . . yes but
- Develop a framework that allows us to **quantify** our level of **confidence** on the fact that our results are true in general

This process is called **Statistical Testing**

Statistical Testing

- Identify the **property** and the quantity (**statistic**) we can measure on the samples which connected to that property (Eg. mean, range, minimum value, ecc . . .)
- Define the **question** in terms of this property (e.g. the mean of the property in treated and control samples is different)
- Assume that what we observe is **the results of chance alone** (Null Hypothesis or H0)
- Calculate the probability of observing (at least) what we see only by chance (**p-value**)
- Set a reasonable threshold on that probability (0.05, 0.01, . . .)
- Decide if H0 is sufficiently unlikely so it can be rejected

Example: lowering cholesterol

- Suppose that the level of cholesterol in the population is normally distributed with mean 200 and standard deviation 50
- We claim that a new secret drug reduces significantly the cholesterol level in the population
- To prove that we get a sample of 50 people, we treat them with the drug and we measure their cholesterol

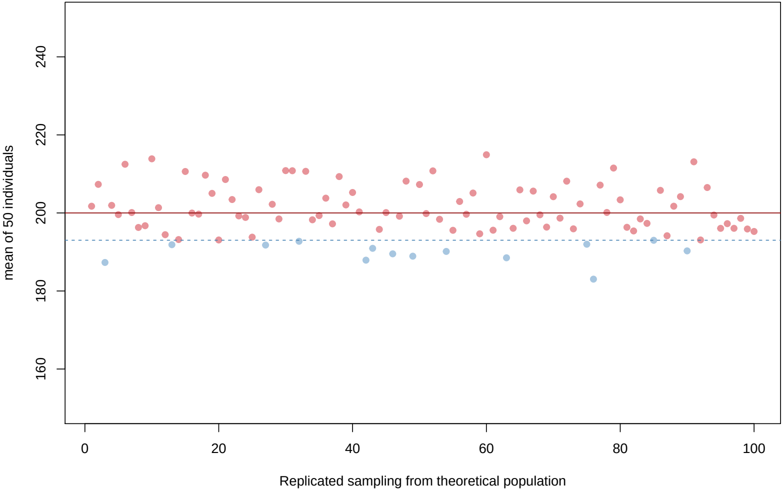
Can we test if this pilot study supporting our claim?

Let's test that!

- We chose the **mean** level of cholesterol as the statistic to be tested. The mean level of cholesterol in our group is 193
- The question. Is the observed mean **sufficiently different** from the population mean?
- We suppose that the **drug has no effect (H0)**, so my 50 people are a random draw from the population and a mean value of 193 is obtained only by chance

- We **calculate the distribution** of the mean level of cholesterol on groups of 50 people (it is not the distribution of cholesterol!) randomly drawn from the population
- We **calculate the probability** of obtaining at least the observed value (p-value) from this distribution
- We reject the null hypothesis if the p-value is lower than the selected threshold

Let's plot it!


mean of 50 individuals
160
180
200
220
240
0
20
40
60
80
100
Replicated sampling from theoretical population

What we see

- The distribution of the means is centered around the population mean!
- The blue line represents the mean of my 50 people
- Blue points represent samples of 50 people showing, by chance, a mean level of cholesterol lower than 193
- Apparently getting at least that value only by chance is not extremely unlikely . . . 14 blue dots out of 100 (0.14 !)
- I cannot reject H0 at the 0.05 level . . .

- We are **never** sure
- . . . even if the threshold is small
- The threshold is **arbitrary**
- Correct phrasing : **“my result is significant at the 0.05 level of confidence”**
- It is fair to change the threshold!

Key point

To calculate the p-value we need to know or estimate the distribution of the statistics we are testing under the null hypothesis

- A priori knowledge
- Estimation from the data ()
- Brute force (e.g. permutation) leading to an **empirical** estimation of the p-value

t-statistics

Let $\widehat{\beta}$ be an estimator of parameter β in some statistical model. Then a *t*-statistic for this parameter is any quantity of the form

$$t_{\widehat{\beta}} = \frac{\widehat{\beta} - \beta_0}{s.e.(\widehat{\beta})}$$

Where β_0 is a known constant, $\widehat{\beta}$ is the estimate of the parameter and $s.e.(\widehat{\beta})$ is the standard error of the estimate.

Student t-test

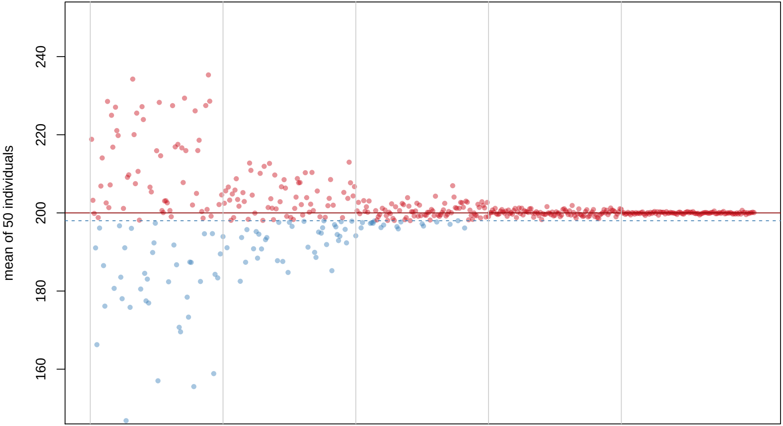
$$t = \frac{\bar{X} - \mu}{\widehat{\sigma}/\sqrt{n}}$$

t statistic follows a *t* distribution

LIVE

Back to our magic drug . . .

Unfortunately it turns out that our drug is not so good . . . apparently it reduces the cholesterol of 0.01%



Do we always need statistics?

- Is a reduction of 0.01% really useful
- Placing an individual within his/her reference population is not a statistical problem
- Big number of samples will make tiny differences statistically significant!
- Statistical significance is not biological/medical significance
- The *p-value* alone cannot be used to judge the relevance of a research . . .

Performing more than one test

With the advent of high throughput *omics*, more often than not the samples are characterized by **multiple measures** (e.g. metabolites, proteins, sensors) so what one want to **test is an hypothesis over a (large) set of variables**

e.g: I'm measuring 1000 metabolites in two groups of samples. Are they different in **at least one metabolite?**

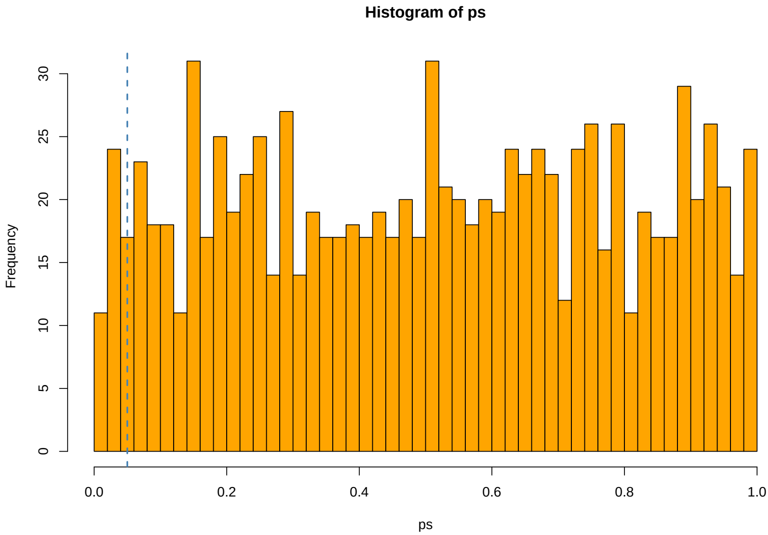
DID YOU KNOW?

What would be better than take the machinery and run it 1000 times on the different metabolites?

Always a dummy dataset . . .

- 20 samples
- 2 classes
- 1000 variables
- random numbers, **no difference**

distribution of p-values



What we see

- p-values are uniformly distributed
- we also have significant differences at the 0.05 level
- Bad luck ;-) !
- Since here I have no differences the distribution of p-values under H0 is uniform

What can we do?

- forget the problem and live in peace ;-)
- reduce the threshold of significance dividing it by the number of tests **Bonferroni** correction
- accept the presence of a fraction of *false positives* tests. This approach is called **False Discovery Rate** control