Tests, effects amd power

Pietro Franceschi
pietro.franceschi@fmach.it

FEM - UBC

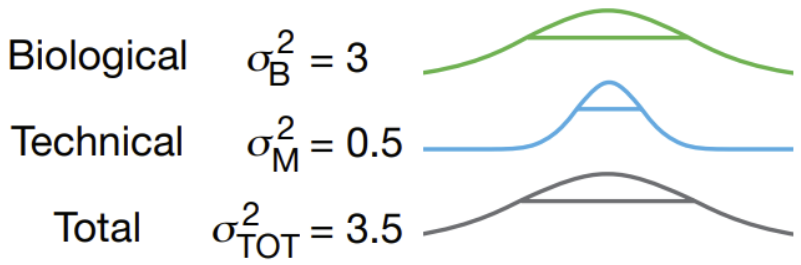
The BIG question

Is what I'm observing **true beyond my sample**.
Can I draw general conclusions from a limited set of samples?

In presence of variability, there will be always the possibility that what I observe in my data cannot be generalized at the population level.

- measure more sample
- validate
- give a measure of my **confidence** on the results

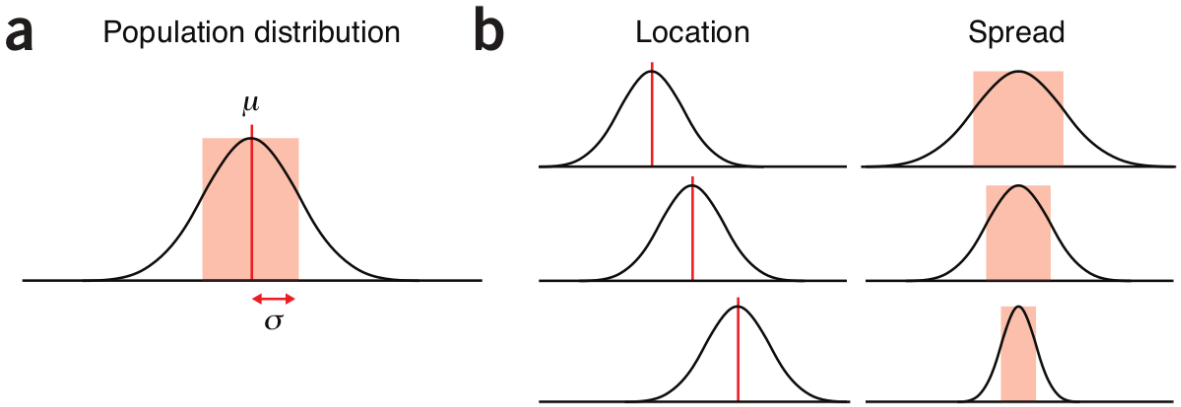
Variability



On Variability

The overall variability comes from the *sum* of the different sources of variability

Distributions



On Variability

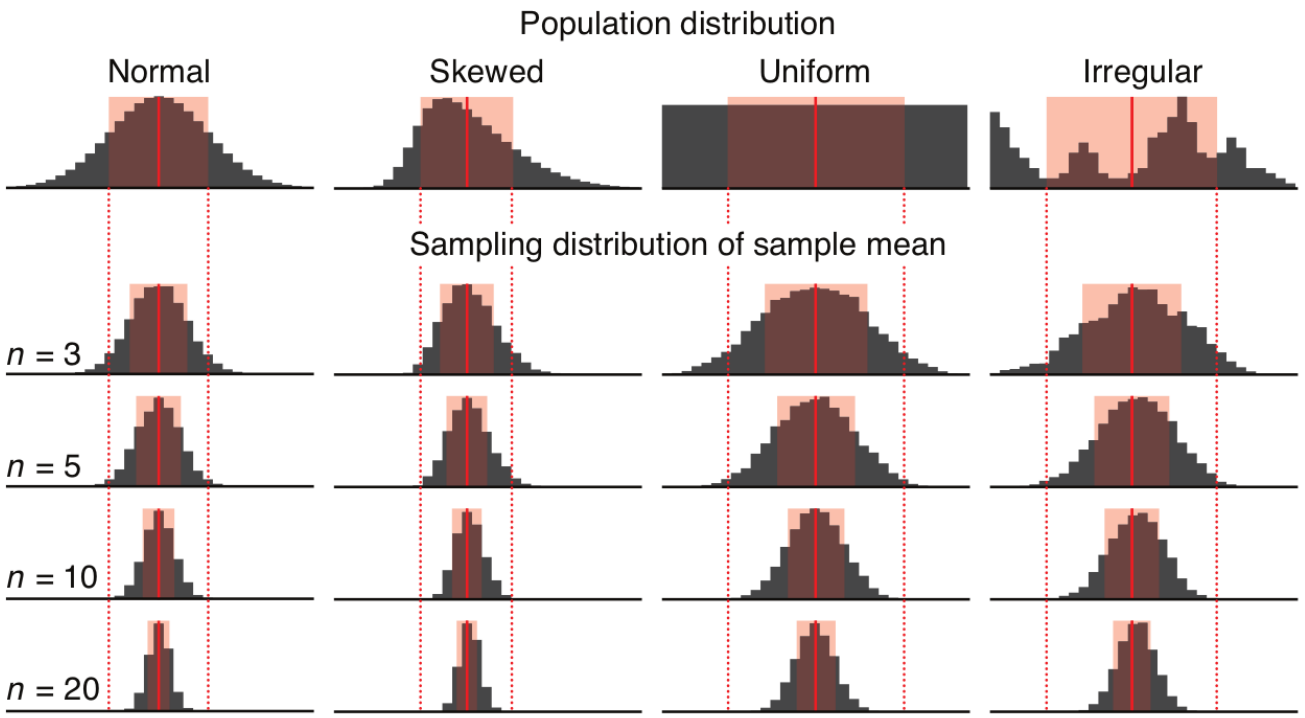
Due to variability each property can have different values with different probabilities, this result in a *distribution*. Each distribution can be characterized by:

- location (e.g. mean)
- spread (e.g variance)

Sampling Distribution

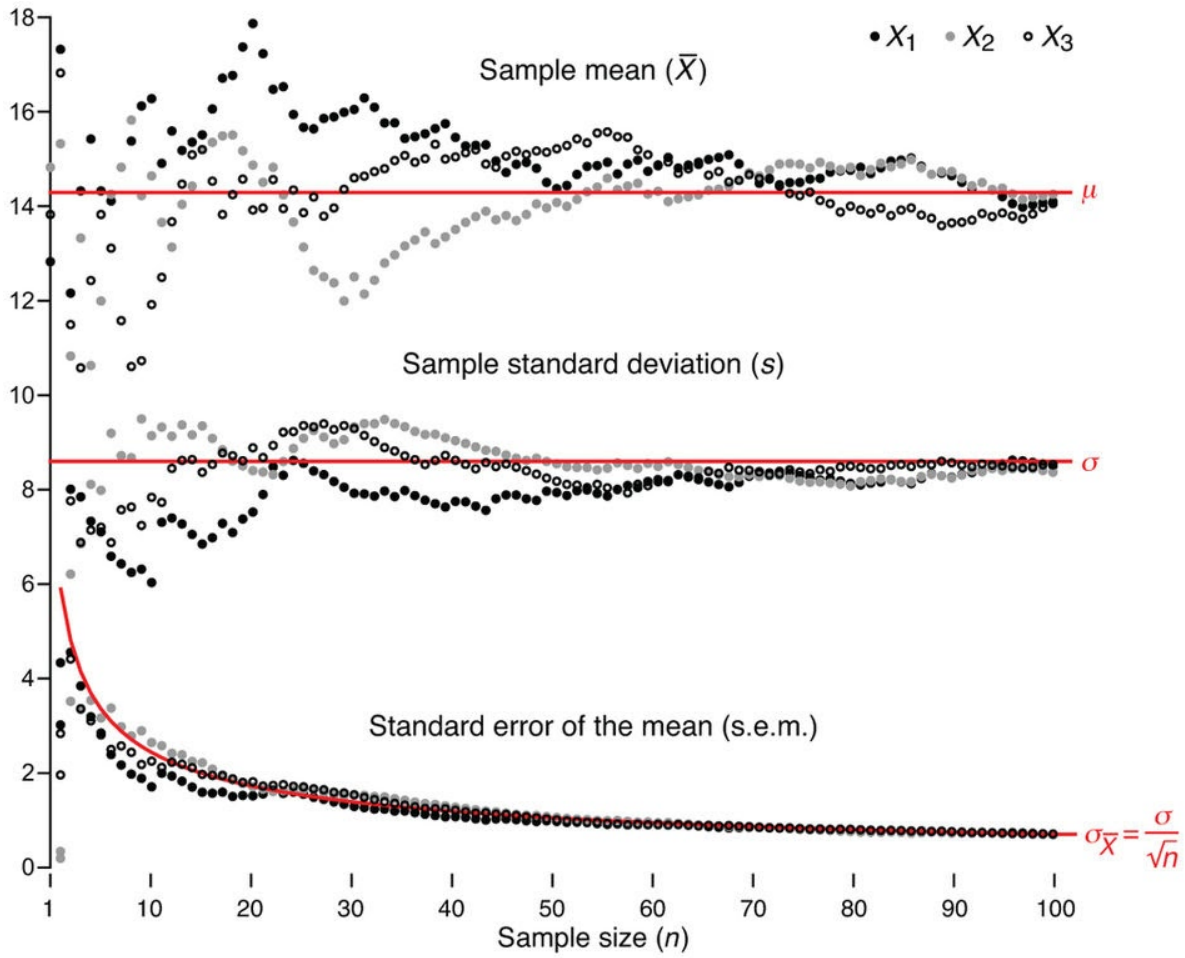
Definition

The probability distribution of a given random-sample-based statistic. E.g. The distribution of the mean of a group of samples upon repeated sampling from a population



The sampling distribution of the mean is always normally distributed

More samples . . .



Statistical testing

Due to variability, it is impossible to get **certain** answers from an experiment.
The best one can do is to try to **quantify the level of confidence**.
Statistical testing is a procedure which allows us to quantify this level of confidence

Statistical testing

E.g. I set up an experiment to test a new pruning strategy, which should improve apple productivity. Is the productivity I measure significantly different from what I observe for the "standard practice"?

Measuring confidence

- Suppose that what we observe is the result of chance alone (Null Hypothesis - H0)
- Use statistics to calculate the probability of getting at least what we observe under H0 (by chance!) (*p-value*)
- Set a threshold of reasonable confidence (0.05,0,01, . . .)

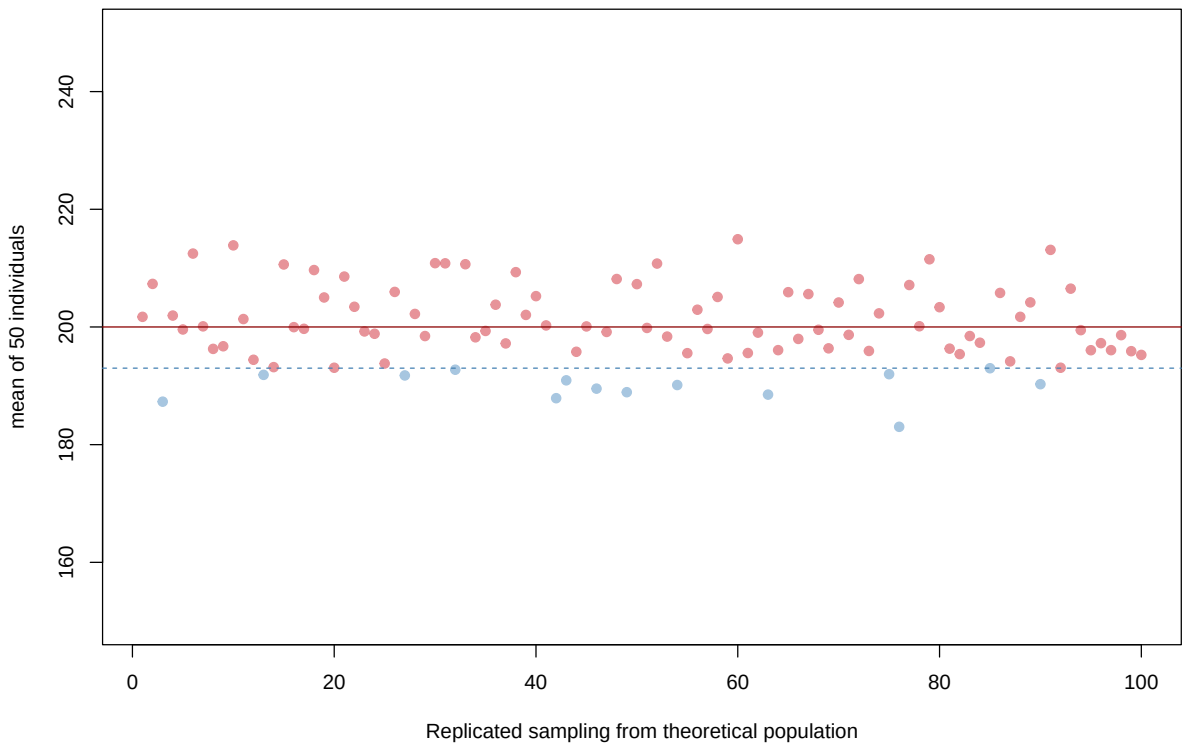
Example: lowering cholesterol

- 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 average cholesterol level. The mean level turns out to be 193.
- Is this pilot study supporting my claim?

Let's test that!

1. Suppose that the drug has no effect (H0), so my 50 people are a random draw from the population of people treated with the standard drug.
2. Calculate the distribution of the mean level of cholesterol on groups of 50 people (it is not the distribution of cholesterol in the population!)
3. Calculate the probability of obtaining at least the observed value from this distribution (*p-value*)
4. Reject H0 if the p-value is lower than the selected threshold

Let's plot it!



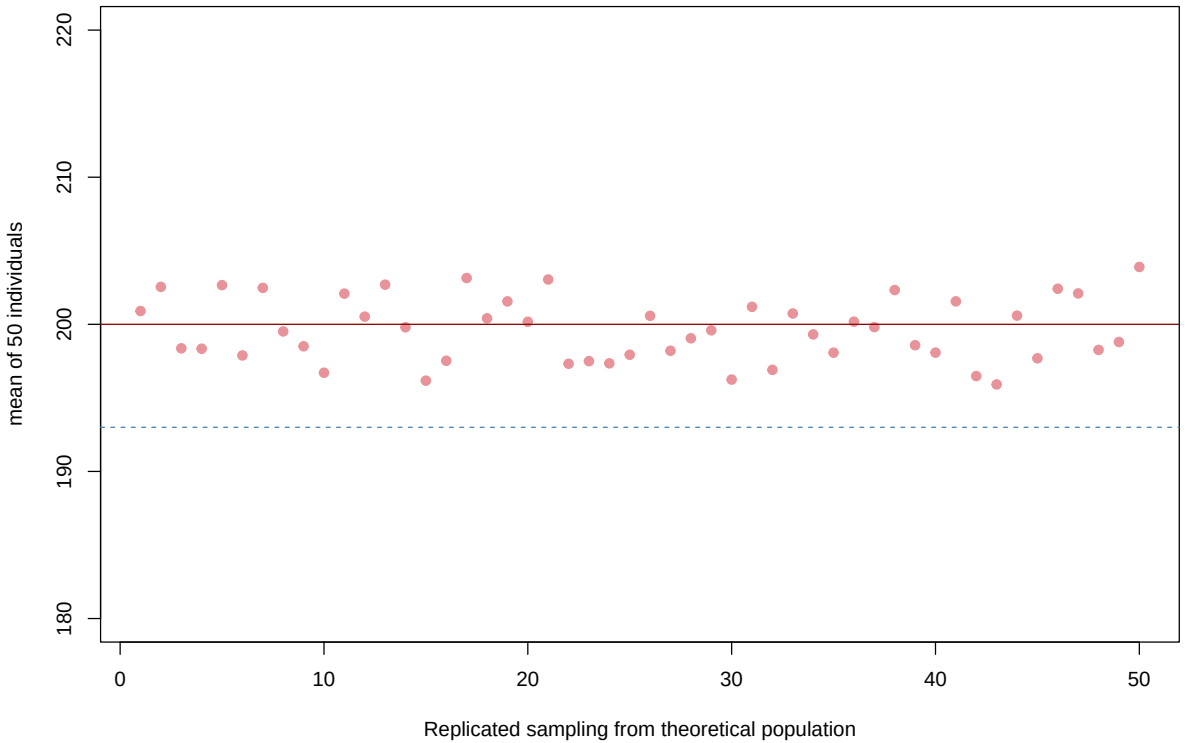
- Each dot shows the value of the mean of an independent sample of 50 people extracted from the population
- The blue line represents the mean of my sample 50 people (193)

What we see

- The distribution of the means is nicely centered around the population mean!
- Apparently getting at least 193 only by chance is not extremely unlikely . . . 14 blue dots out 100 (p *value* = *0.14 !)
- I cannot reject H0 at the 0.05 level . . . but I could at the 0.15 level of confidence!

More samples!

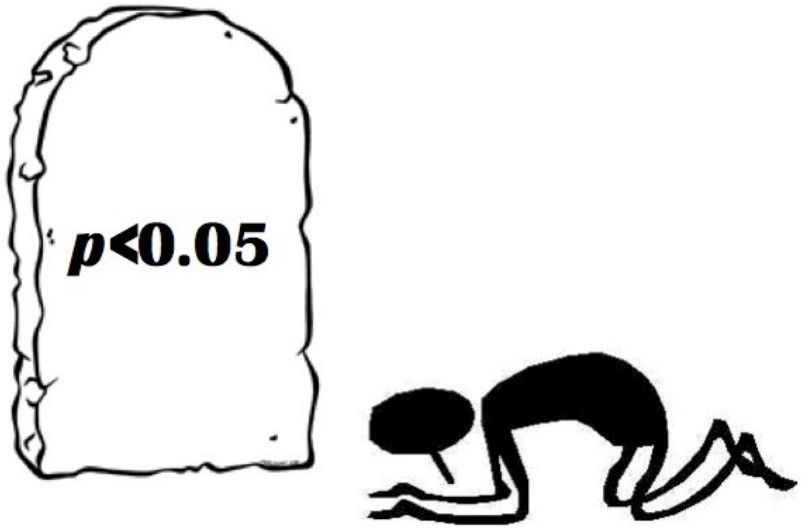
I'm stubborn . . . We redo the same study with a test group of 500 people . . .



Magic! Now it is significant! No more blue dots, so it is extremely unlikely to get an average cholesterol level of 193 in a group of 500 people if the drug has no lowering effect.

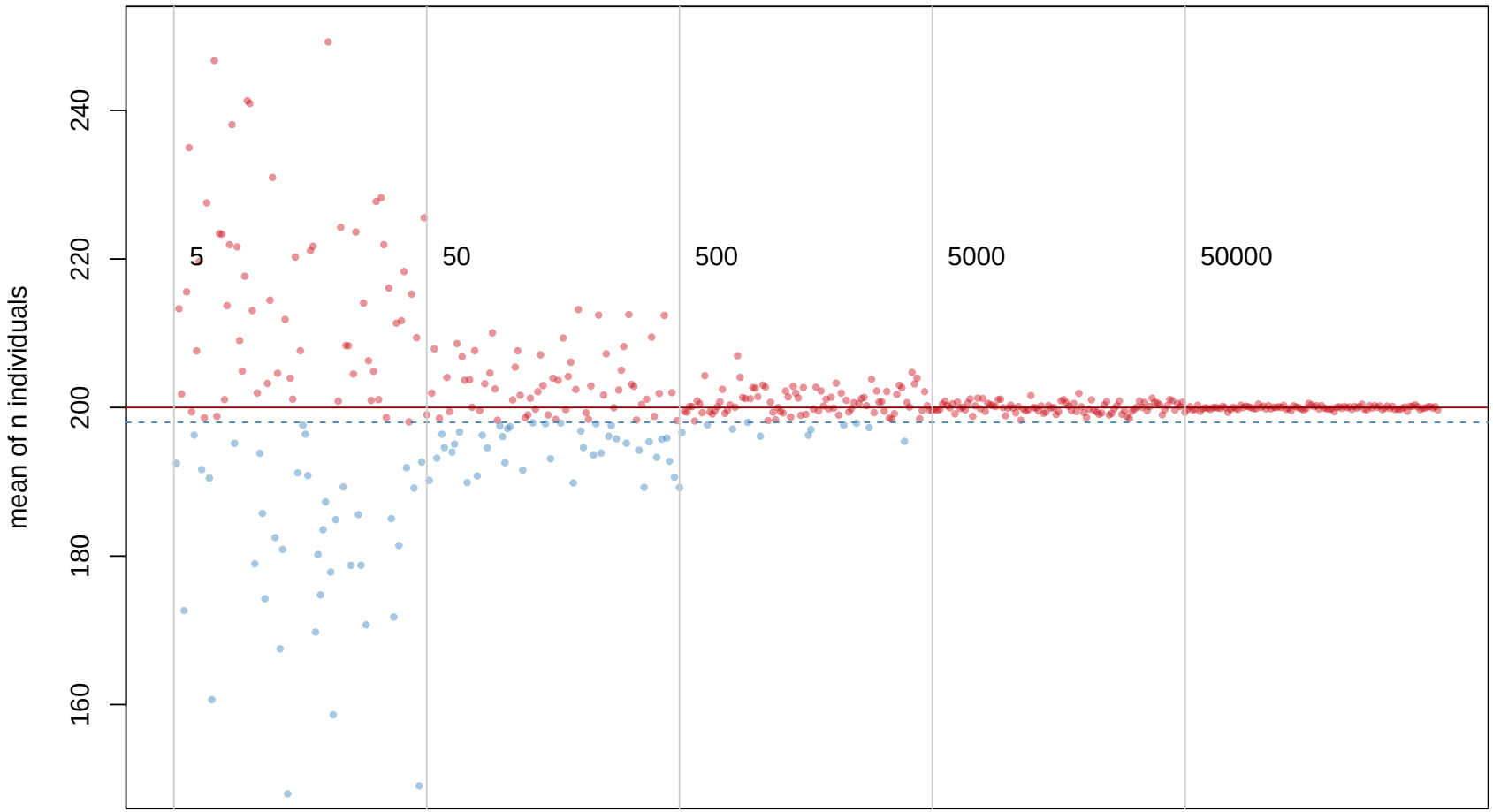
Take home messages

- The choice of a threshold (0.05) for statistical significance is arbitrary
- With more samples we can "see" smaller differences
- We are never sure!
- Is statistical significance the only thing we are looking for?


p<0.05

Back to our magic drug . . .

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



Replicated sampling from theoretical population

Is a low *p-value* the only thing we need?

- Is a reduction of 0.01% really useful/relevant?
- Big number of samples will make tiny differences statistically significant!
- Statistical significance does not mean biological/agronomic relevance
- The *p-value* alone cannot be used to judge the relevance of a research ...

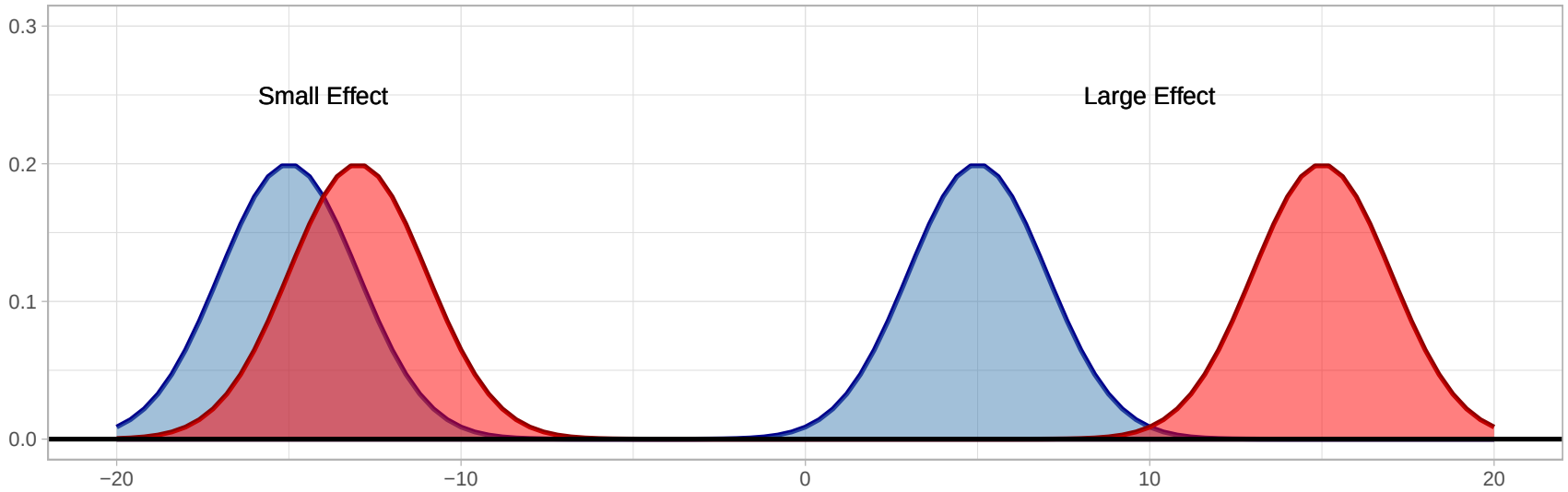
Is a low *p-value* the only thing we need?

- Is a reduction of 0.01% really useful/relevant?
- Big number of samples will make tiny differences statistically significant!
- Statistical significance does not mean biological/agronomic relevance
- The *p-value* alone cannot be used to judge the relevance of a research ...

Erroneous . . .

- *p-values* deals with probability of obtaining by chance, not with the strength of an effect
- strong effects with low variability **will** result in low *ps*
- the reverse is not necessarily true!
- "look, I have a low p value!" is not the only thing that matters

Measuring the *effect size*



Notes

1. The difference in means is not sufficient
2. The measure should take into account of the variability
3. The variability of the population not the one of he sampling distribution ;-)

Fold Change and Effect Size

The *fold change* is **not** a good measure of the effect size . . .

Cohen's d

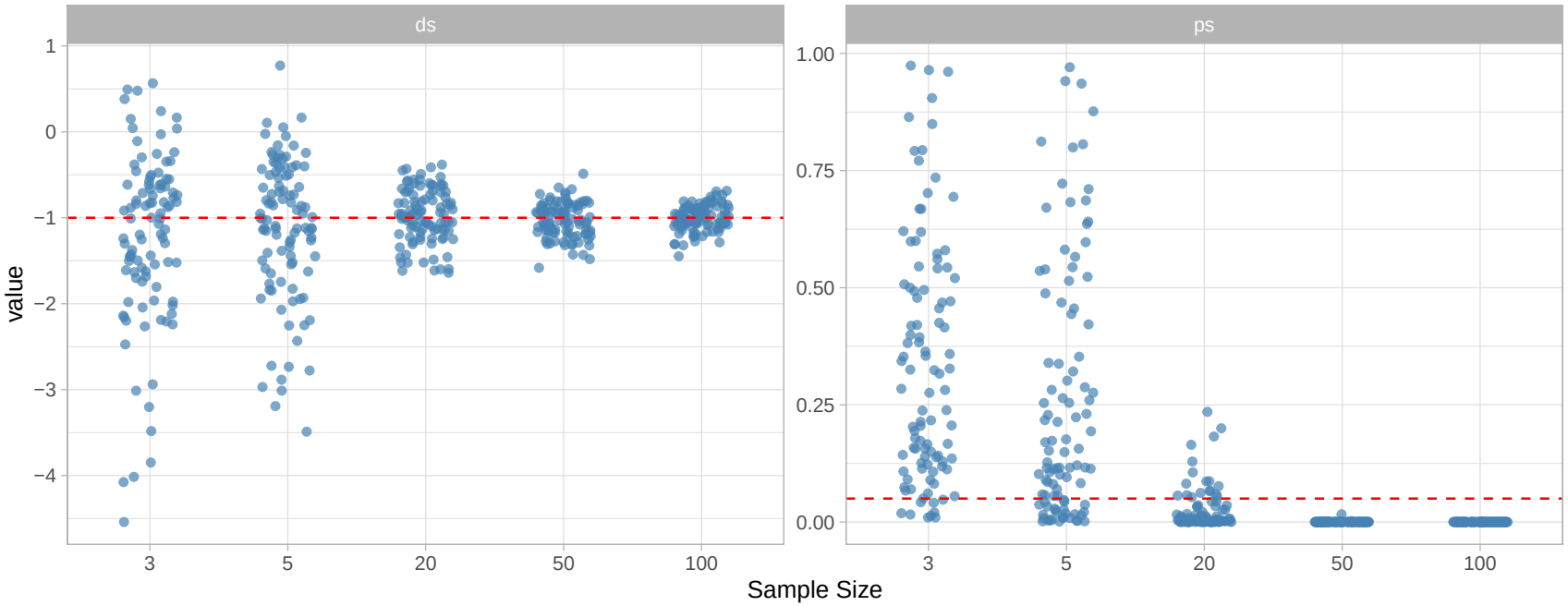
$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Where

- $\bar{x}$ are the estimates of the population means
- s is the estimate of the population standard deviation (pooled)

Let's *see*

- 2 populations: $\mu_1 = 5$, $\mu_2 = 10$, $\sigma = 5$
- *t-test* to test the difference
- different sample sizes



Notes

- With three samples variability in *p values* and estimated *d* is large
- Also the probability of calling non significant a real difference is large . . .
- The average estimate of *d*, however, is not changing with the sample size
- The reason for that is that *d* is calculated by using the standard deviation of the population and not the standard deviation of the sampling ditribution of the mean.
- The effect size does not tell to me if something is biologically/ecologically relevant

H0, Ha and Power Calculation

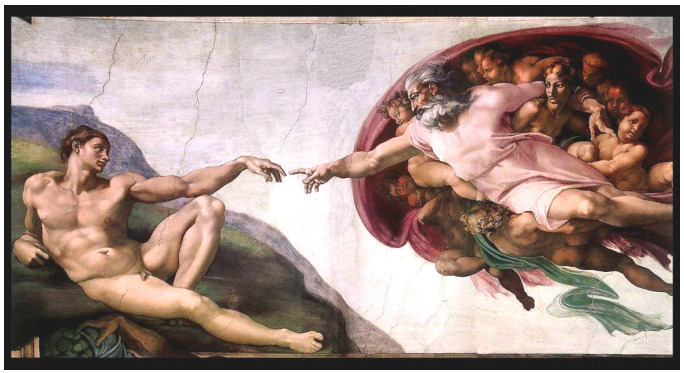
Alternative hypothesis

In statistical hypothesis testing, the alternative hypothesis is a position that states **something is happening**, a new theory is preferred instead of an old one (null hypothesis).
It is usually consistent with the research hypothesis because it is constructed from literature review or previous studies . . . or technical/economical considerations

In presence of Ha a sensible question is: *how many sample should I measure to assess if my data support Ha?*

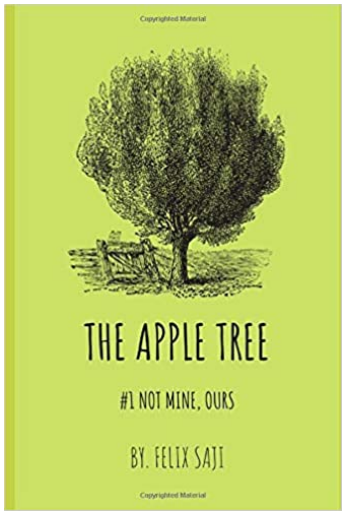
Eg. My new pruning strategy is worth being introduced only if it is increasing the productivity of 10 %

The origin Ha



- Biological/agronomic knowledge
- Technical objective (it is worth introducing a practice only if it is x% better then the old one)
- Literature search
- Preliminary experiment

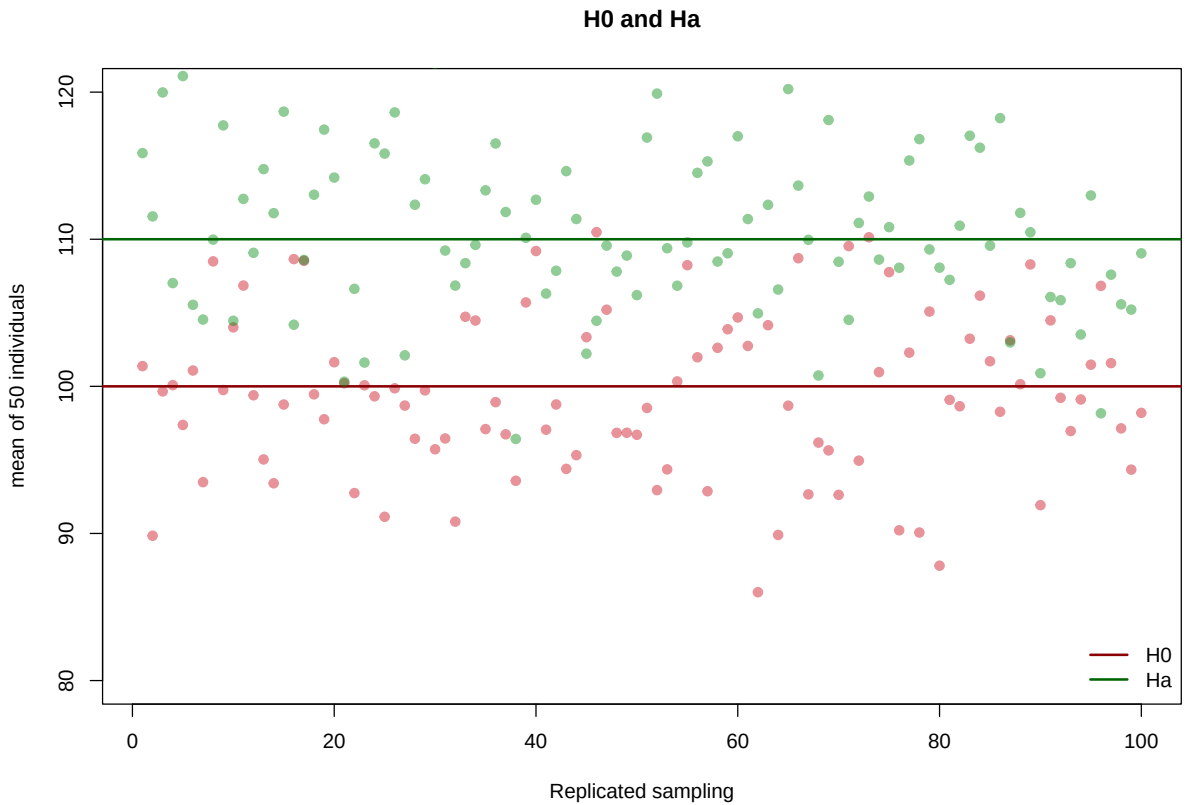
Power calculation by hand: the tree problem



- Standard productivity 100 kg/tree
- Expected improvement with new pruning: 10% (Ha = 110 kg/tree)
- Variability 40 kg/tree
- I propose to test everything in a CRD with 50 trees

Is this number of trees appropriate to be able to reliably see the expected difference?

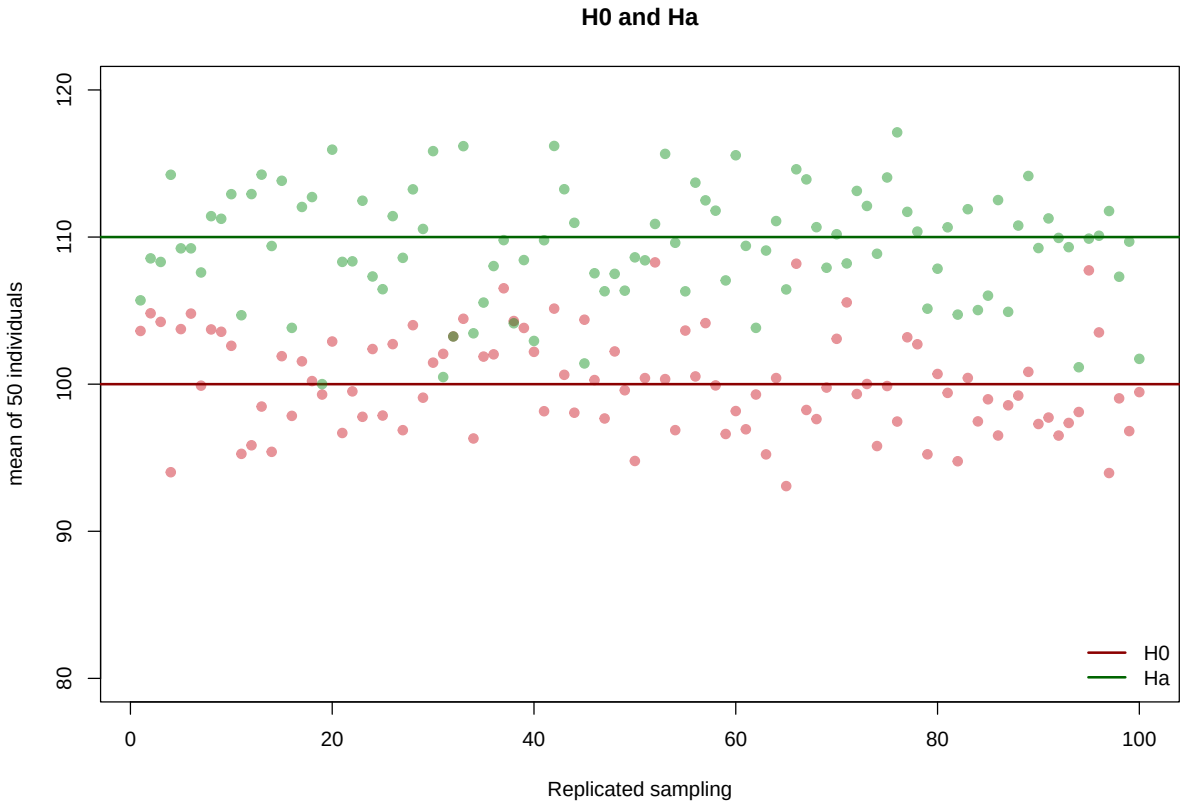
What About Ha?



- Each dot represents the possible outcome of a real experiment: red under H0, green under Ha
- Even if Ha is true, in many cases I could obtain a productivity similar to the one under H0 (green dots mixed with red dots)
- My setting is not suitable to disentangle H0 and Ha

100 trees

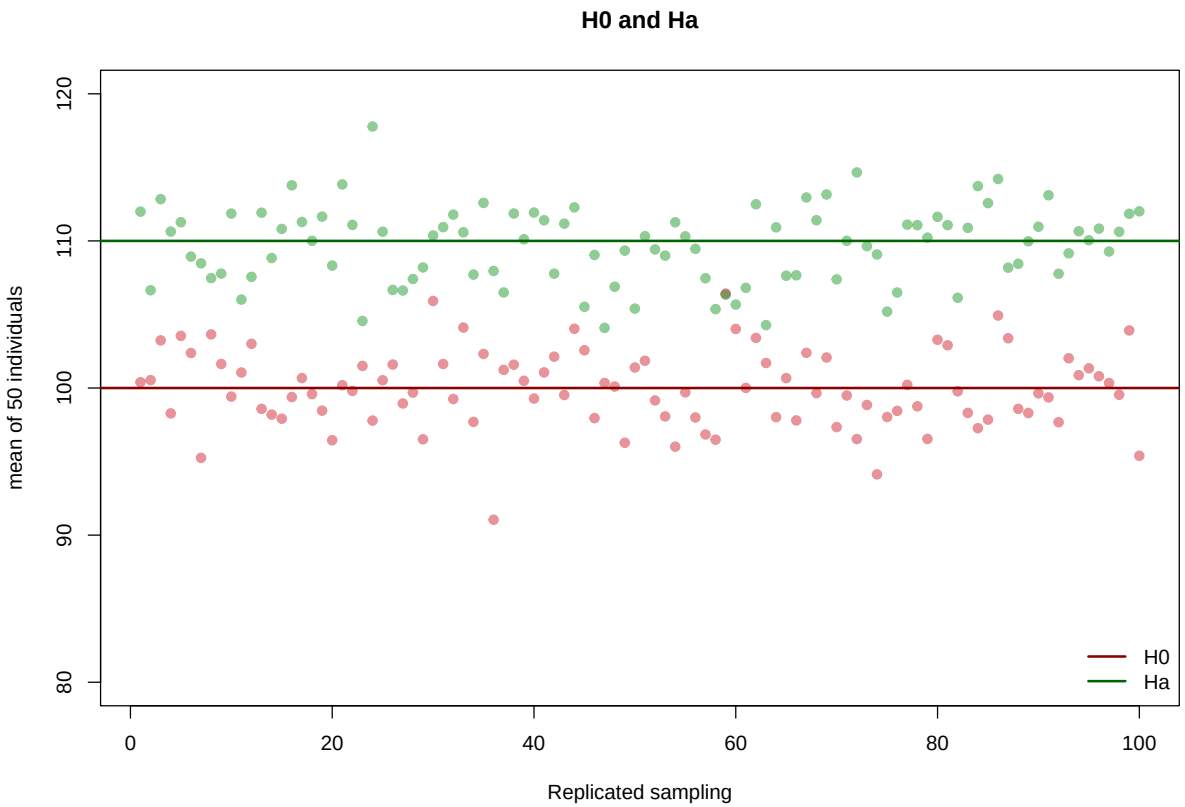
More trees/samples!



- Better, but not sufficient yet . . .

175 trees

More trees/samples!



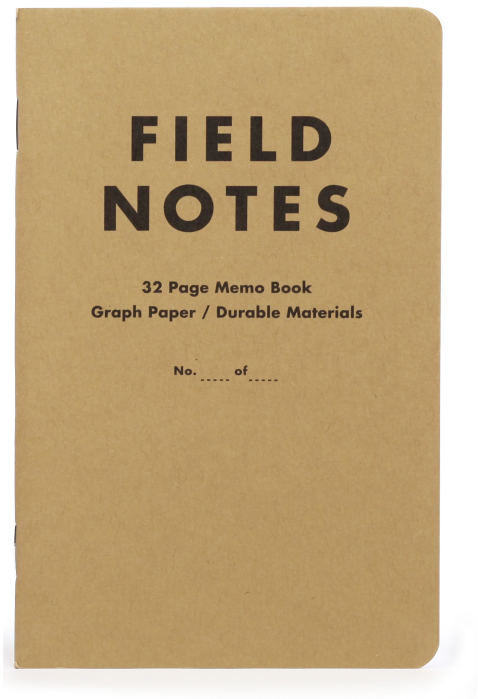
- Yes! Now the "contrast" is sufficient
- If I want to be able to call the difference of 10% significant I need at least 175 plants
- In real life you strike a balance between level of significance and number of samples

Important

Power Calculation

Power calculation does not ensure that the new pruning works (successful experiment) . . . but ensures that our design would be able to see (at least) the target effect

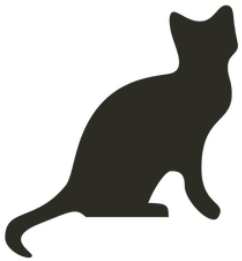
Summary



- In real life situations the standard deviation of the population is unknown!
- We developed ideas without any mathematical formalism
- The same line of reasoning can be extended to various types of scenarios
- Low *p-value* is not the only thing we need!

Web Resources

- http://www.gpower.hhu.de/en.html
- https://www.stat.ubc.ca/~rollin/stats/ssize/
- https://glimmpse.samplesizeshop.org/



THANKS FOR LISTENING!

IT'S

TIME FOR QUESTIONS!