A few remarks on Experimental Design

Simon Anders
• How many replicates do I need?

• What is a suitable replicate?

• What effects can I hope to see?
A preliminary: the word “sample”

(singular!)

Statistician: “Our sample comprises 20 male adult subjects chosen at random from the patient population.”

Biologist: “For our experiment, we used blood samples from 20 patients chosen at random.”

(plural!)
An observed effect is called **statistically significant**.

What does this mean?

\[ p < 0.05 \]

If the experiment were repeated with new, independently obtained samples, the effect would likely be observed again.
If the experiment were repeated with new, independently obtained samples, the effect would likely be observed again.

When can we claim such a thing?

Only if we tried more than once!
Replication
How many replicates?
What is the average height of a man in Germany?

Each dot represents the height of one subject.
What is the average height of a man in Germany?

![Graph showing the relationship between sample mean and sample size]

- Each dot represents the average of $n$ subjects.
- The standard error of mean is given by:
  \[
  \text{standard error of mean} = \frac{\text{standard deviation of observations}}{\sqrt{\text{sample size}}}
  \]
Purposes of replication:

1. More replicates allow for more precise estimates of effect sizes.
What is the average height of a man in Germany?

Each dot: standard deviation, estimated from $n$ subjects

Standard error of mean $= \frac{\text{standard deviation of observations}}{\sqrt{\text{sample size}}}$
Purposes of replication:

1. More replicates allow for more precise estimates of effect sizes.

2. Replicates allow to estimate how precise our effect-size estimates are.
Two Martian scientists, Dr Nullor and Dr Alton, ask:

Is body height correlated with hair colour in human males within a city’s population?

Dr Nullor’s UFO observes Helsinki:
\[ \mu(\text{dark}) = \mu(\text{blond}) = 180 \text{ cm}. \]

Dr Alton’s UFO observes Beijing:
\[ \mu(\text{dark}) = 170 \text{ cm} \]
\[ \mu(\text{blond}) = 175 \text{ cm} \quad [\text{European immigrants}] \]
How many replicates?

Your **power** to detect an effect depends on

- **effect size** (difference between group means)
- **within-group variance**
- **sample size**
Two extremes

• effect size $\ll$ within-group SD
  many replicates to get precise mean estimates

• effect size $\gg$ within-group SD
  few replicates sufficient, just to verify that SD is small
Correlation and causation

Dr Alton:

“In Beijing, blond men tend to be taller than men with dark hair.”

“In Beijing, having blond hair causes men to be taller.”

Beware of third variables.
Randomization

• Dr Nullor’s student again observes men in Helsinki.

• He runs into a group of Chinese tourists.

• He should have made sure to pick men at random from all over the city.
Blocking

male

female

mixed
So far, our data was normally distributed ...
Purposes of replication:

1. More replicates allow for more precise estimates of effect sizes.
2. Replicates allow for estimating how precise our effect-size estimates are.
3. Enough replicates allow for proper randomization.
4. Enough replicates reduce the need for assumptions on distribution.
Outliers?
Purposes of replication:

1. More replicates allow for more precise estimates of effect sizes.

2. Replicates allow for estimating *how precise* our effect-size estimates are.

3. Enough replicates allow for proper randomization.

4. Enough replicates reduce the need for assumptions on distribution.

5. Replicates allow to spot outliers.
Controls and replicates
Example: Study of a species of sea grass.

• One RNA-Seq library prepared from 20 plants from the North Sea

• and one with 20 plants from the Mediterranean Sea
Control samples should be equal to treated samples in all aspects except for the treatment.

Replicates must differ from each other in all aspects in which controls differ from treated samples.
Example: Katz et al. compare splicing between colon tumour cells:

- one sample of from a patient with a 5FU-sensitive tumour
- and one from a patient with a 5FU-resistant tumour.

[Katz et al., Nature Methods, 2010, MISO method]
Example: Mice are treated with a drug and compared to mice experiencing sham treatment.

All mice are from the same litter.

• Can we generalize to other litters?
• To other strains?
• To other species?
• Experimental units are considered drawn at random from a population.

• Results may be generalized to this population.

• Generalizing further requires a leap of faith, namely:
  Declaring the used system a “model”.

Pseudoreplication

Replicates must differ from each other in all aspects in which controls differ from treated samples.

Otherwise, you have pseudoreplicates.
Levels of replication

• Sequence same prepared library multiple times.
• Prepare multiple libraries from same sample (or: from the same mouse).
• Prepare multiple samples from the same cell culture (or: from the same litter/strain).
• Prepare samples from independently generated cell cultures (or: from outbred mice).
Why do people use less samples in HTS than in classical experiments?
Shrinkage estimation in HTS

Estimating a SD from just two subjects is pointless

unless we measure many similar but independent things.
How many replicates

The more the better?
How many replicates?

Controlled experiment: variation $\ll$ effect size
• single measure: maybe 5 – 15 per group
• RNA-Seq etc.: 3 per group

Study with strong inter-subject variation:
• Dozens to hundreds of subjects!
Common objections
“I cannot afford replicates.”

Use multiplexing:

5 samples, sequenced to 20M reads each offer more power than 2 samples, sequenced to 50M reads each.
“I cannot afford replicates.”

Use multiplexing:

Power depends on:
- number of libraries per group
- total number of reads per group
  (not: per library)
“I know I need at least 50 samples, but I cannot get hold of more than 10. So I use what I have.”

Performing underpowered experiments is a waste of time.
“I know that my within-group variability is much smaller than the effect size.”

Then prove it.
“It’s only a pilot study. I’ll do replication in the main study.”

A pilot study is the perfect opportunity to assess reproducibility.
“I’ll validate with qPCR.”

Very good. But use new samples.
Power prediction

How to decide on the sample size
Read up on the topic: