# BISC-869, Experimental Design

February 1, 2021

How should you allocate replicates to different levels of your experiment? Is it better to have more plots, or more plants within plots? Is it better to have more small families or fewer, larger families?

What is a $p$-value?

What is a "Type I error"?

What is a "Type II error"?

**Statistical Power**: the likelihood that a study will detect an effect when there is an effect there to be detected.

Science is expensive: a low-power study is a waste of resources, and so is a study that is larger than necessary.

Ethics boards and animal care committees require researchers to justify the sample sizes for proposed experiments on animals, humans.

Problems with low power studies

- High chance of a false negative.
- Highly uncertain estimates of effect size (wide confidence intervals).
- If a statistically significant result is obtained in a low power study, the estimate of effect is likely to be exaggerated.
- If a statistically significant result is obtained in a low power study, there is a high chance that the estimated effect is in the wrong direction.

Goals when planning your sample size

- **Plan for precision:** Choose a sample size that yields a confidence interval of specified width. A narrow confidence interval means we have an estimate with high precision.
- **Plan for power:** Involves choosing a sample size that would have a high probability of rejecting $H_0$ ($\geq 80\%$) if the absolute magnitude of the difference between the means, $|\mu_1 - \mu_2|$, is at least as great as a specified value $D$.
- **Compensate for data loss:** Some experimental individuals may die, leave the study, or be lost between the start and the end of the study. The starting sample sizes should be made even larger to compensate.

**Challenges of planning sample size**

- Key quantities to plan sample sizes, such as the within-group standard deviation, $\sigma$, are not known.
- Typically a researcher makes an educated guess for these unknown parameters based on pilot studies or previous investigations.
- If no information is available then consider carrying out a small pilot study first, before attempting a large experiment.
- Note: post-hoc power calculations are useless (i.e., calculating how likely it is that the null hypothesis is true, based on your non-significant outcome is non-sensical; see Colegrave & Ruxton, Behavioral Ecology. 14: 446–450).

In an experimental study the researcher assigns treatments to units or subjects so that differences in response can be compared. There must be at least 2 treatments (or treatment and control).

- Examples: Clinical trials, reciprocal transplant experiments, factorial experiments on competition and predation.

In an observational study, nature does the assigning of treatments to subjects. The researcher has no influence over which subjects receive which treatment (no matter how complex the apparatus needed to measure response)

- Examples: Common garden "experiments", QTL mapping "experiments".

An observational study cannot distinguish between two reasons for an association between an *explanatory variable* and a *response variable*.

Survival of climbers to Mount Everest is higher for individuals taking supplemental oxygen than not.

1. Supplemental oxygen (explanatory variable) increases survival (response variable).
2. Supplemental oxygen has little or no effect. Survival and oxygen are associated because other variables affect both (e.g., greater overall preparedness). Variables (like preparedness) that distort the causal relationship between the measured variables of interest (oxygen use and survival) are called *confounding variables*.

We do experiments **to eliminate confounding variables**.

With an experiment, random assignment of treatments to subjects allows researchers to tease apart the effects of the explanatory variable from those of confounding variables.

With random assignment, no confounding variables will be associated with treatment except by chance.

If a researcher could assign supplemental oxygen/no-oxygen randomly to Everest climbers, this will break the association between oxygen and degree of preparedness. Random assignment will roughly equalize the preparedness levels of the two oxygen treatment groups.

In this case, any resulting difference between oxygen treatment groups in survival (beyond chance) must be caused by treatment.

An experimental study in which two or more treatments are assigned to human subjects.

The design of clinical trials has been refined because the cost of making a mistake with human subjects is so high.

Experiments on nonhuman subjects are simply called "laboratory experiments" or "field experiments", depending on where they take place.

ARTICLES

# Effectiveness of COL-1492, a nonoxynol-9 vaginal gel, on HIV-1 transmission in female sex workers: a randomised controlled trial

*Lut Van Damme, Gita Ramjee, Michel Alary, Bea Vuylsteke, Verapol Chandeying, Helen Rees, Pachara Sirivongrangson, Léonard Mukenge-Tshibaka, Virginie Ettiègne-Traoré, Charn Uaheowitchai, Salim S Abdool Karim, Benoît Mâsse, Jos Perriëns, Marie Laga, on behalf of the COL-1492 study group**

Transmission of the HIV-1 virus via sex workers contributes to the rapid spread of AIDS in Africa.

The spermicide nonoxynol-9 had shown in vitro activity against HIV-1, which motivated a clinical trial by van Damme et al. (2002). They tested whether a vaginal gel containing the chemical would reduce the risk of acquiring the disease by female sex workers.

Data were gathered on a volunteer sample of 765 HIV-free sex-workers in six clinics in Asia and Africa.

Two gel treatments were assigned randomly to women at each clinic. One gel contained nonoxynol-9 and the other contained a placebo (an inactive compound that subjects could not distinguish from the treatment of interest).

Neither the subjects nor the researchers making observations at the clinics knew who had received the treatment and who had received the placebo (A system of numbered codes kept track of who got which treatment.)

Results of the clinical trial:

| | Nonoxynol-9 | | Placebo | |
|---|---|---|---|---|
| **Clinic** | ***n*** | **Number infected** | ***n*** | **Number infected** |
| Abidjan | 78 | 0 | 84 | 5 |
| Bangkok | 26 | 0 | 25 | 0 |
| Cotonou | 100 | 12 | 103 | 10 |
| Durban | 94 | 42 | 93 | 30 |
| Hat Yai 2 | 22 | 0 | 25 | 0 |
| Hat Yai 3 | 56 | 5 | 59 | 0 |
| Total | 376 | 59 | 389 | 45 |

*"This study did not show a protective effect of COL-1492 on HIV-1 transmission in high risk women. Multiple use of nonoxynol-9 could cause toxic effects enhancing HIV-1 infection. This drug can no longer be deemed a potential HIV-1-prevention method."*

To reduce **bias**, the experiment included:

- *Simultaneous control group:* the women receiving the placebo.
- *Randomization:* treatments were randomly assigned to women at each clinic.
- *Blinding:* neither the subjects nor the clinicians knew which women were assigned which treatment.

To reduce the effects of **sampling error**, the experiment included:

- *Replication:* the study was carried out on multiple independent subjects.
- *Balance:* the number of women was nearly equal in the two groups at every clinic.
- *Blocking:* subjects were grouped according to the clinic they attended, yielding multiple repetitions of the same experiment in different settings ("blocks").

**Simultaneous control group**

- A study lacking a control group for comparison cannot determine whether the treatment of interest is the cause of any of the observed changes.
- The health of human subjects often improves after treatment merely because of their expectation that the treatment will have an effect, a phenomenon known as the "placebo effect".
- Control subjects should be perturbed in the same way as the other subjects, except for the treatment itself (as far as ethical considerations permit). The "sham operation", in which surgery is carried out without the experimental treatment itself, is an example.
- In field experiments, applying a treatment of interest may physically disturb the plots receiving it and the surrounding areas, perhaps by trampling the ground by the researchers. Ideally, the same disturbance should be applied to the control plots.

**Randomization**

- The researcher should randomize assignment to units or subjects.
- Randomization means that treatments are assigned to units at random, such as by flipping a coin or using random numbers. Other ways of assigning treatments to subjects are inferior. "Haphazard" assignment has repeatedly been shown to be non-random and prone to bias.
- Randomization breaks the association between possible confounding variables and the explanatory variable, allowing the causal relationship between the explanatory and response variables to be assessed.
- Randomization doesn't eliminate the variation contributed by confounding variables, only their correlation with treatment.
- A completely randomized design is an experimental design in which treatments are assigned to all units by randomization.

## Blinding

- Blinding is the process of concealing information from participants (sometimes including researchers) about which subjects receive which treatment.
- In a **single-blind** experiment, the subjects are unaware of the treatment that they have been assigned. Not much of a concern in non-human studies.
- In a **double-blind** experiment, those administering the treatments and measuring the response are also unaware of which subjects are receiving which treatments.
- Blinding prevents subjects and researchers from changing their behavior, consciously or unconsciously, as a result of knowing which treatment they were receiving or administering.
- Medical studies without double-blinding exaggerated treatment effects by 16% on average, compared to studies without double-blinding (Jüni et al. 2001).
- Experiments on non–human subjects are also prone to bias from lack of blinding.
- Bebarta et al. (2003) reviewed 290 two-treatment experiments carried out on animals or on cell lines. The odds of detecting a positive effect of treatment were more than threefold higher in studies without blinding than in studies with blinding. (Experiments without blinding also tend to have other problems such as a lack of randomization.)
- Blinding can be incorporated into experiments on nonhuman subjects using coded tags that identify the subject to a "blind" observer without revealing the treatment (and who measures units from different treatments in random order).

The goal of experiments is to estimate and test treatment effects against the background of variation between individuals ("noise") caused by other variables.

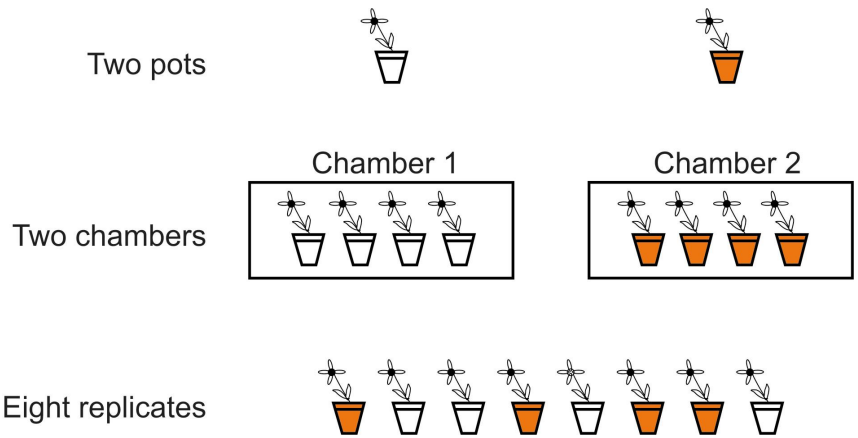
One way to reduce noise is to make the experimental conditions constant. Fix the temperature, humidity, and other environmental conditions, for example, and use only subjects that are the same age, sex, genotype, and so on. In field experiments, constant experimental conditions might not be feasible.

Constant conditions might not be desirable, either. By limiting the conditions of an experiment, we also limit the generality of the results - that is, the conclusions might apply only under the conditions tested and not more broadly.

Another way to make treatment effects stand out is to include extreme treatments.

## Replication

- Replication is the assignment of each treatment to multiple, independent experimental units.
- Studies that use more units (i.e., larger sample sizes) will have smaller standard errors and a higher probability of getting the correct answer from a hypothesis test.
- Larger samples mean more information, and more information means better estimates and more powerful tests.
- Replication is not about the number of plants or animals used, but the number of independent units in the experiment. An "experimental unit" is the independent unit to which treatments are assigned.
- The figure shows three experimental designs used to compare plant growth under two temperature treatments (indicated by the shading of the pots). The first two designs are unreplicated.

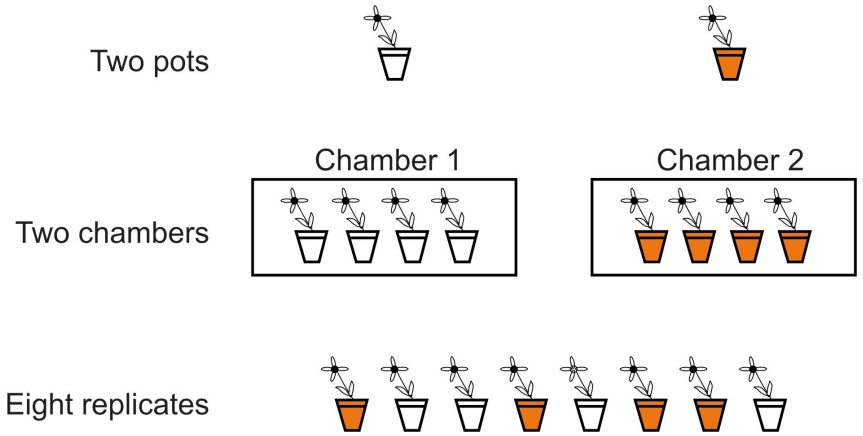


**Replication**

- An experimental unit might be a single plant/animal, if individuals are randomly sampled and assigned treatments independently.
- An experimental unit might be a batch of individual organisms treated as a group, such as a field plot containing multiple individuals, a cage of animals, a household, a Petri dish, or a family.
- Multiple individual organisms belonging to the same unit (e.g., plants in the same plot, bacteria in the same dish, members of the same family, and so on) should be considered together as a single replicate. This is because they are likely to be more similar to each other, on average, than to individuals in separate units (apart from the effects of treatment).
- Erroneously treating the single organism as the independent replicate when the chamber or field plot is the experimental unit is *pseudoreplication*.

**Interspersion**

– Treatments must always be *interspersed* with each other in space and time.




– Randomization is one way this is usually implemented.

**Balance**

- A design is balanced if all treatments have the same sample size.
- Balance helps reduce the influence of sampling error on estimation. To appreciate this, look at the equation for *the standard error of the difference between two treatment means*.

$$\sigma_{m_1-m_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Let's assume equal variances ($\sigma_1^2 = \sigma_1^2$). Then, this equation reduces to

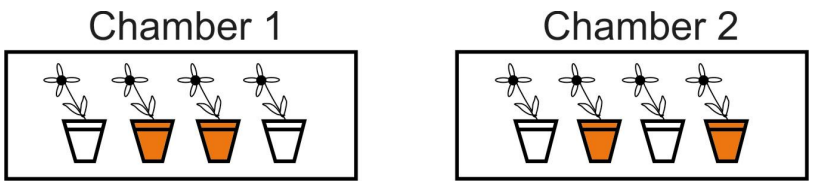
$$\sigma_{m_1-m_2} = \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

For a fixed total number of experimental units, $n_1 + n_2$, the standard error is smallest when the quantity $\frac{1}{n_1} + \frac{1}{n_2}$ is smallest, which occurs when $n_1$ and $n_2$ are equal.

- Balance is not as important as replication (i.e., $n_1 + n_2$).

**Blocking**

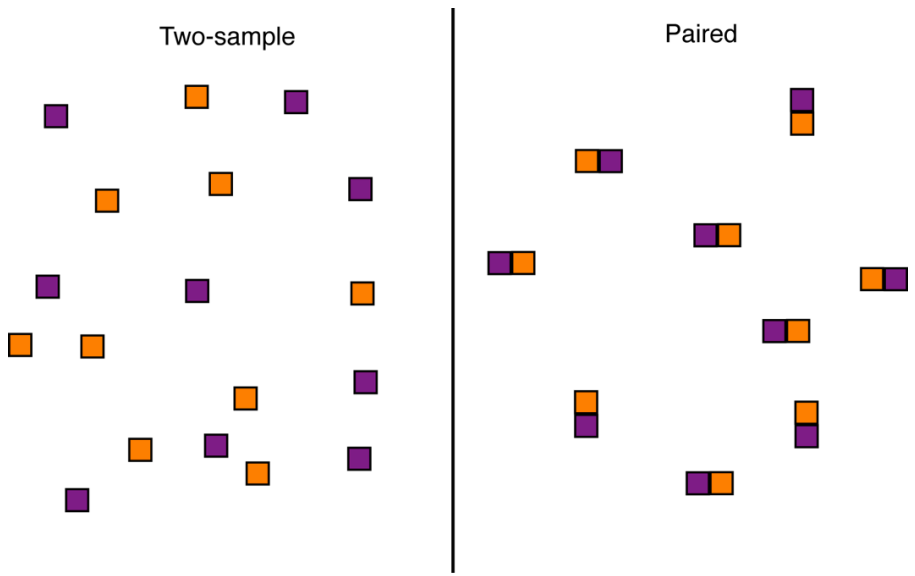
- Blocking is the grouping of experimental units that have similar properties. Within each block, treatments are randomly assigned to experimental units.
- Blocking essentially repeats the same, completely randomized experiment multiple times, once for each block.
- Differences between treatments are only evaluated within blocks, and in this way the component of variation arising from differences between blocks is discarded.




- Block (here, chamber) must be included as a (random) factor in the statistical analysis. Analysis follows design. We'll talk about this more when we apply mixed effects models.

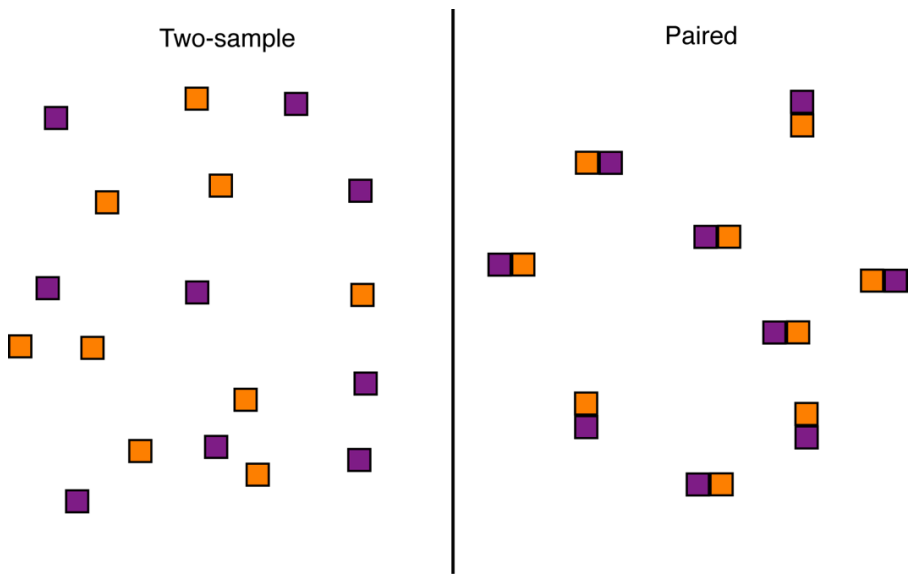
**Blocking: Paired design**

- For example, consider the design choices for a two-treatment experiment to investigate the effect of clear cutting on salamander density.
- In the completely randomized ("two sample") design, we take a random sample of forest plots from the population and then randomly assign either the clear-cut treatment or the no clear-cut treatment to each plot.
- In the paired design we take a random sample of forest plots and clear-cut a randomly chosen half of each plot, leaving the other half untouched.



**Blocking: Paired design**

- In the paired design, measurements on adjacent plot-halves are not independent. This is because they are likely to be similar in soil, water, sunlight, and other conditions that affect the number of salamanders.
- As a result, we must analyze paired data differently than when every plot is independent of all the others, as in the case of the two-sample design.
- The paired design is usually more powerful than completely randomized design, because it controls for a lot of the extraneous variation between plots or sampling units that might obscure the effects we are estimating.



**Blocking: Randomized complete block design**

- Paired designs are a special case of RCB design (which allows more than two treatments). Each treatment is applied once to every block.
- By accounting for some sources of sampling variation, such as the variation among trees, blocking can make differences between treatments stand out.
- Blocking is worthwhile if units within blocks are relatively homogeneous, apart from treatment effects, and units belonging to different blocks vary because of environmental or other differences.
- In the example of a clinical trial, "Clinic" was a blocking variable.

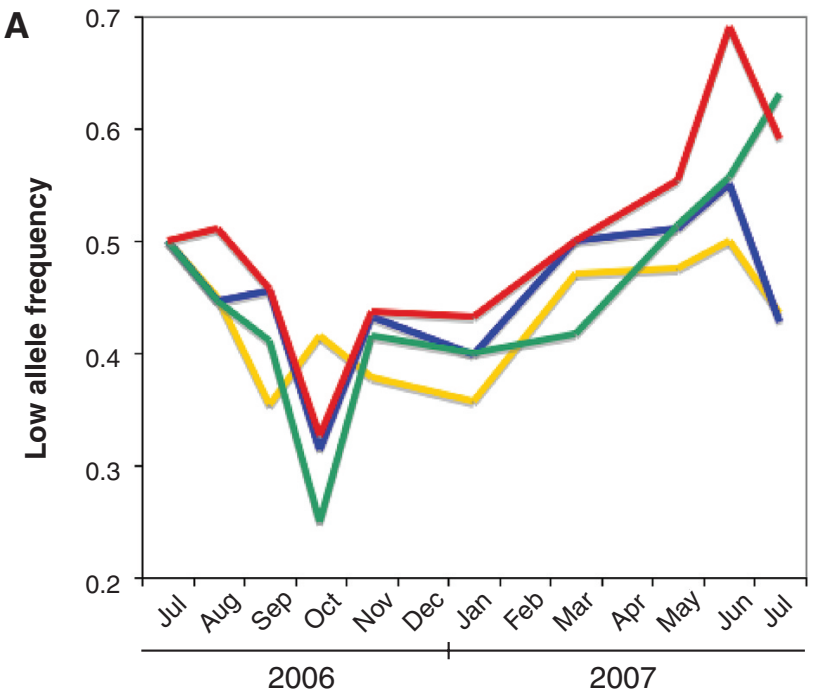
**Experiments with more than one factor**

- A factor is a single treatment variable whose effects are of interest to the researcher.
- The *factorial* design is the most common experimental design for more than one treatment variable, or factor. In a factorial design *every combination of treatments* from two (or more) treatment variables is investigated.
- The main purpose of a factorial design is to evaluate possible interactions between variables. An interaction between two explanatory variables means that the effect of one variable on the response depends on the state of a second variable.
- Even if there are no interactions, a factorial design can be an efficient way to collect information on the effects of more than one treatment variable.

Must account for repeated measures of the same subjects (plots)

## Natural Selection on a Major Armor Gene in Threespine Stickleback

Rowan D. H. Barrett,* Sean M. Rogers, Dolph Schluter



## Analysis follows design

The structure of your analysis should reflect the structure of study design.

Remember, pseudoreplication is a problem of analysis, not design. It can happen when the analysis doesn't follow the experimental design.

For example, if subjects are grouped (fish in aquaria; colonies in a Petri dish; repeated measurements of the same individuals), then your analysis needs to include a (random) group level variable in the statistical model.

Grouping variables are incorporated using "mixed effects models", which we will learn about.

Recognizing how you will analyze the data when you design your study is a prerequisite for planning the sample sizes you will need.

To plan an experimental design and the sample sizes required to achieve your experimental goals, use R to make up (simulate) data. Then use R to analyze the data.

Repeat this many times and you will acquire estimates of power and precision for alternative plans.

Experimental studies are not always feasible, in which case we must fall back upon observational studies.

- The best observational studies incorporate as many of the features of good experimental design as possible to minimize bias (e.g., simultaneous controls, blinding) and the impact of sampling error (e.g., replication, balance, blocking, and even extreme treatments) except for one: randomization. Randomization is out of the question, because in an observational study the researcher does not assign treatments to subjects.
- Two strategies are used to limit the effects of confounding variables on a difference between treatments in a controlled observational study: matching; and statistically adjusting for known confounding variables.

## Two extra suggestions when planning data collection

- Always record raw untransformed data (transformations can always be done later, but some transformations cannot be undone).
- Always try to think of additional "easy to collect" data.