Outline for today

- Plan your sample size
- Experiments vs observational studies
- Why do experiments
- Clinical trials: experiments on people
- Design experiments to minimize bias and effects of sampling error
- What if you can’t do experiments: think like an experimentalist

Plan your sample size

- Ethics boards and animal care committees require researchers to justify the sample sizes for proposed experiments on animals, humans.
- Science is expensive: a low-power study is a waste of resources, and so is a study that is larger than necessary
- How to allocate replicates to different levels of the 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?
- Tools in R. This week's workshop will explore some using simulation and canned packages.

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 of high precision.
- ***Plan for power.*** Involves choosing a sample size that would have a high probability of rejecting H_0 (≥ 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, σ, 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.
- B.t.w., post-hoc power calculations are useless (see this week's assigned reading).

Experiment vs observational study

What is an experimental study?

- 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).
    - Clinical trials, reciprocal transplant experiments, factorial experiments on competition and predation, etc. are examples of experimental studies.
- In an *observational study*, nature does the assigning of treatments to subjects. The researcher has no influence over which subjects receive which treatment.
    - Common garden “experiments”, QTL “experiments”, etc, are examples of observational studies (no matter how complex the apparatus needed to measure response).

Why do experiments

- An observational study cannot distinguish between two reasons behind 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. Perhaps supplemental oxygen (explanatory variable) increases survival (response variable).
- Or, 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*.

http://www.everest-2002.de/home_e.html

Why do experiments

- 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.

Clinical trials

- 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.

Example of an experiment (clinical trial)

THE LANCET • Vol 360 • September 28, 2002 • www.thelancet.com

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**

Example of an experiment (clinical trial)

- 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.)

Example of an experiment (clinical trial)

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."

Design components of clinical trial

- 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”).

OPEN ACCESS Freely available online

PLOS | BIOLOGY

Perspective

Whole ~~Animal~~ Organism Experiments Should Be More Like Human Randomized Controlled Trials

Beverly S. Muhlhausler[1]*, Frank H. Bloomfield[2,3,4], Matthew W. Gillman[5,6]

1 FOODplus Research Centre, School of Agriculture Food and Wine, The University of Adelaide, Australia, **2** Liggins Institute, University of Auckland, Auckland, New Zealand, **3** Department of Paediatrics: Child and Youth Health, University of Auckland, Auckland, New Zealand, **4** Gravida, National Centre for Growth and Development, New Zealand, **5** Obesity Prevention Program, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, United States of America, **6** Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, United States of America

… the reporting of animal studies received comparatively little attention until the publication of the ARRIVE [Animal Research: Reporting In Vivo Experiments] guidelines in 2010 [4]. These guidelines were spurred by a survey of 271 studies reporting original research on rats, mice, and non-human primates carried out in the United Kingdom and the United States of America [5]. The results painted a poor picture of the quality of reporting in animal research. Only 59% of the 271 articles stated the hypothesis or objective of the study, the number of animals used, and characteristics of the animals. **Few of the papers surveyed reported using random allocation** to treatment group (13%) **or blinding** of outcome assessment (14%), and statistical methods were not described adequately in 30% of the publications [5]. In a similar review of animal studies published in Cancer Research, only 28% reported random allocation of animals to treatment groups, only 2% reported blinding of observers to this allocation, and none reported methods to determine sample size [6].

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 the researchers 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.
- For example, studies showing that acupuncture has a significant effect on back pain are limited to those without blinding (Ernst and White 1998).

Blinding

- Medical studies carried out without double-blinding exaggerated treatment effects by 16% on average, compared with studies carried out with 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).

Minimizing the effects of sampling error

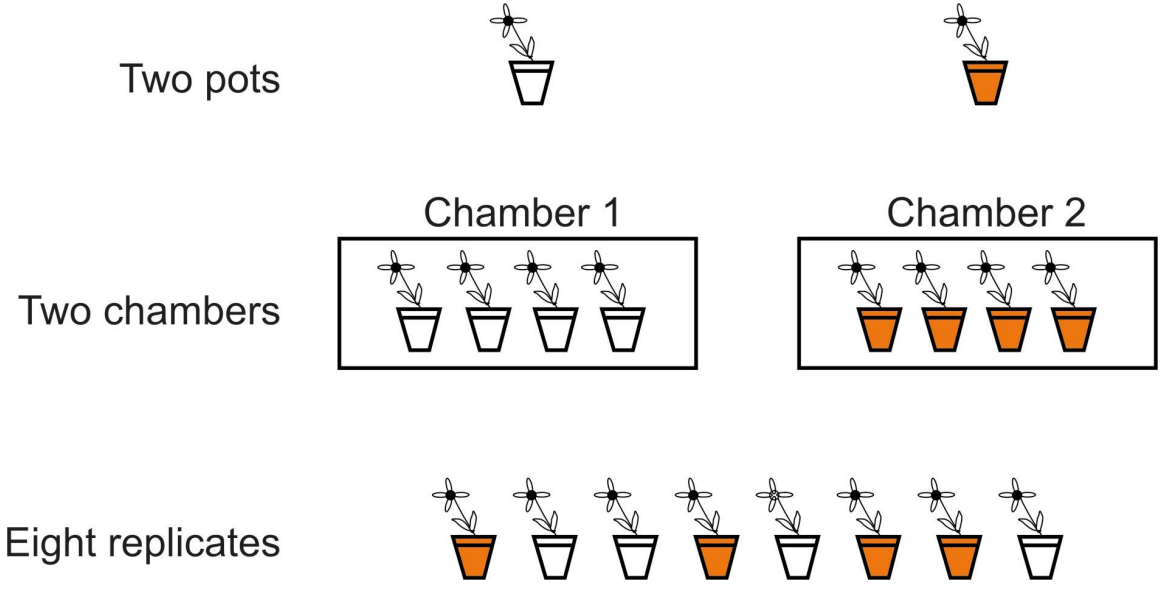
- The goal of experiments is to estimate and test treatment effects against the background of variation between individuals ("noise") caused by other variables.
- One way to reduce noise is to make the experimental conditions constant. Fix the temperature, humidity, and other environmental conditions, for example, and use only subjects that are the same age, sex, genotype, and so on. In field experiments, highly 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

- 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 (typically, the *unit that is interspersed*).
- 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 animal or plant if individuals are randomly sampled and assigned treatments independently.
- Or, an experimental unit might be made up of 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*

Balance

- A study design is balanced if all treatments have the same sample size.
- Balance helps to reduce the influence of sampling error on estimation and hypothesis testing. To appreciate this, look at the equation for the standard error of the difference between two treatment means. For a fixed total number of experimental units, $n_1 + n_2$, the standard error is smallest when the quantity

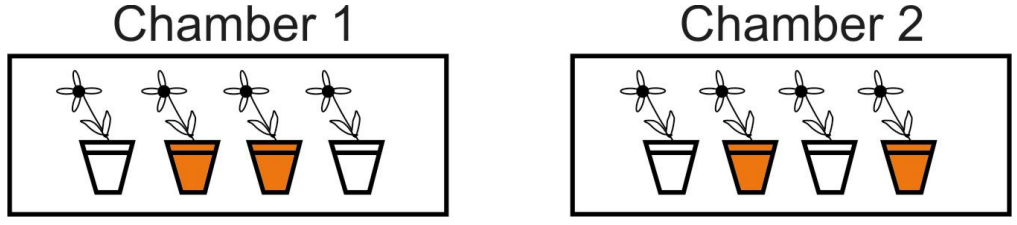
$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

  is smallest, which occurs when n_1 and n_2 are equal.
- Balance has other benefits. For example, ANOVA is more robust to departures from the assumption of equal variances when designs are balanced or nearly so.
- However, greater balance is not as important as greater 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.

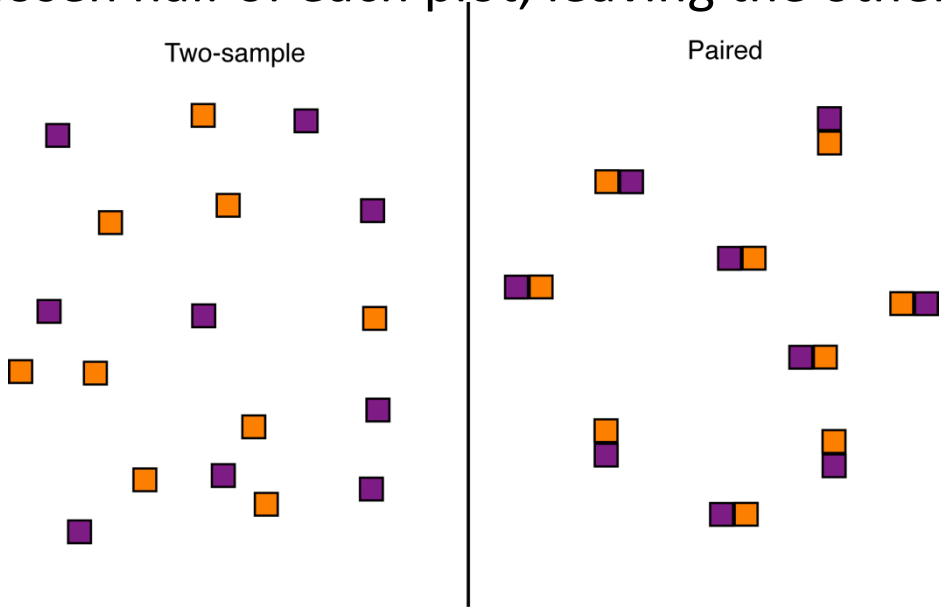
`include block as a random effect in a mixed model`



- Block (here, chamber) must be included as a (random) factor in the statistical analysis.

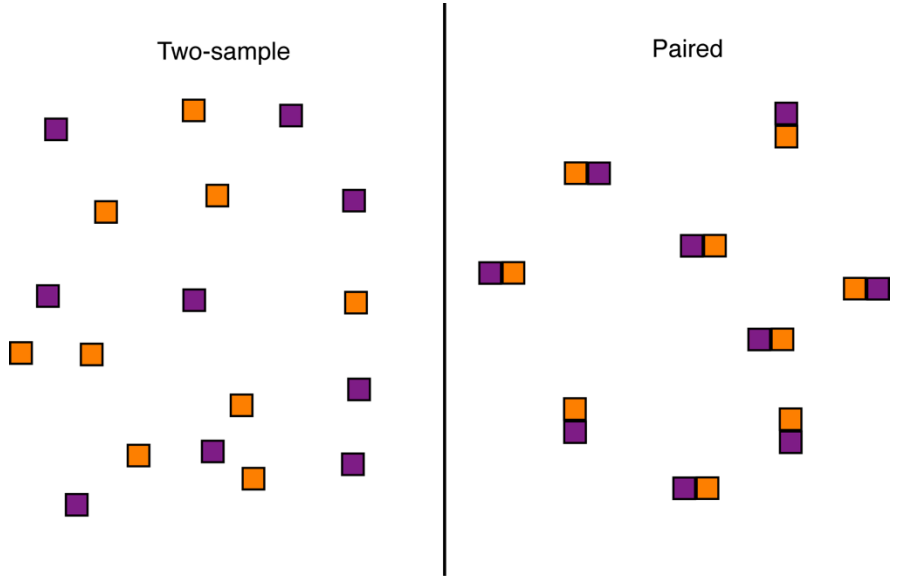
Blocking: Paired design

- 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.

Pseudoreplication

- For example, Visscher et al. (1996) compared the effects of two methods of removing the barbed stinger, poison sac, and muscles left behind after a honeybee stings its victim (that continue to pump venom): scraping off with a credit card or pinching off with thumb and index finger.
- *n* = 40 stings, 20 removed with the credit card method, and 20 with the pinching method. Data were the size of the welt after 10 minutes. All 40 measurements were combined to estimate means, standard errors, and the *P*-value for a test of treatment effect. Pinching led to a slightly smaller average welt, but the difference was not statistically significant.
- However, all 40 measurements came from two volunteers.
- Pseudoreplication will lead to calculations of standard errors and *P*-values that are too small.

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.

Analysis follows design

- The structure of your analysis will reflect the structure of study design.
- Pseudoreplication is a problem of analysis, not design.
- For example, if subjects are grouped, then your analysis needs to include a group level variable in the statistical model.
- Grouping variables are incorporated using “mixed models”, which we will learn about in a few weeks.
- Recognizing how you will analyze the data when you design your study is a prerequisite for planning the sample sizes you will need.

What if you can't do experiments?

- 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. Instead, the subjects come as they are.
- Two strategies are used to limit the effects of confounding variables on a difference between treatments in a controlled observational study: matching; and adjusting for known confounding variables.

Matching

- A strategy commonly used in epidemiological studies.
- With matching, every individual in the target group with a disease or other health condition is paired with a corresponding healthy individual that has the same measurements for known confounding variables such as age, weight, sex, and ethnic background (Bland and Altman 1994).
- In a weaker version of this approach, a comparison group is chosen that has a similar frequency distribution of measurements for each confounding variable as the treatment group, but no pairing takes place.

Adjusting for known confounding variables

- With adjustment, analysis of covariance (a type of linear model) is used to correct for differences between treatment and control groups in suspected confounding variables.

Discussion paper:

Colegrave and Ruxton (2003). Confidence intervals are a more useful complement to nonsignificant tests than are power calculations

(might also want to look at Hoenig and Heisey (2001), which they cite)

Download from "**assignments**" tab on course web site.

Presenters: Need two volunteers!

Moderators: Need two volunteers!