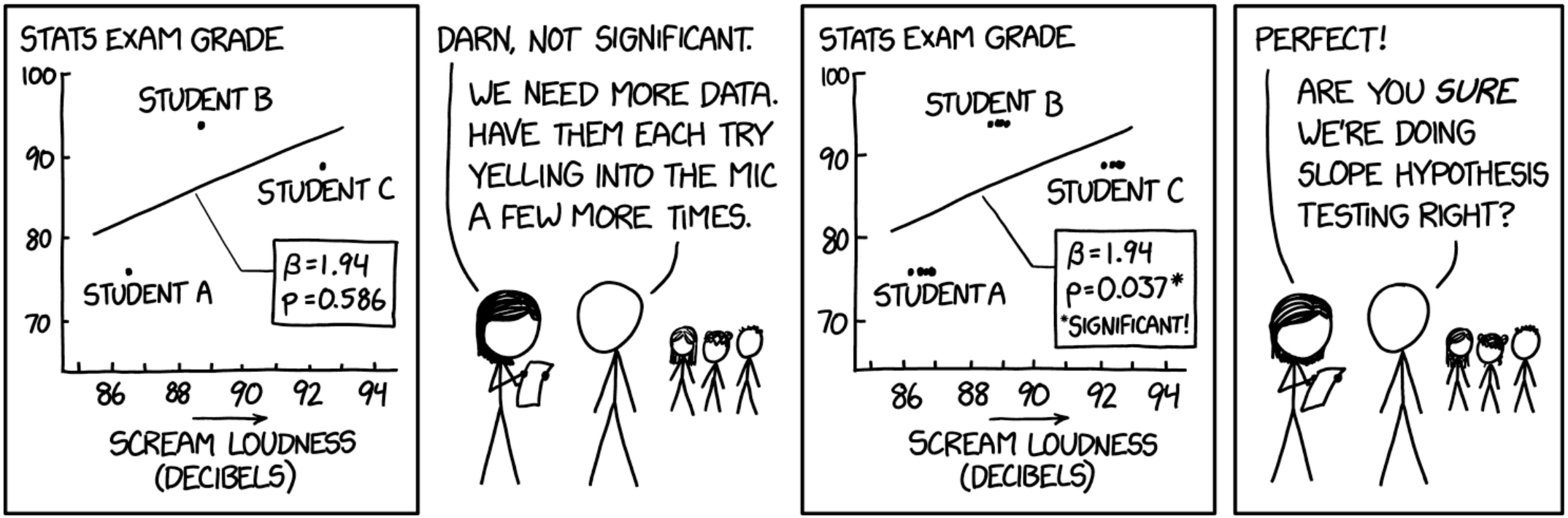


https://xkcd.com/2533/

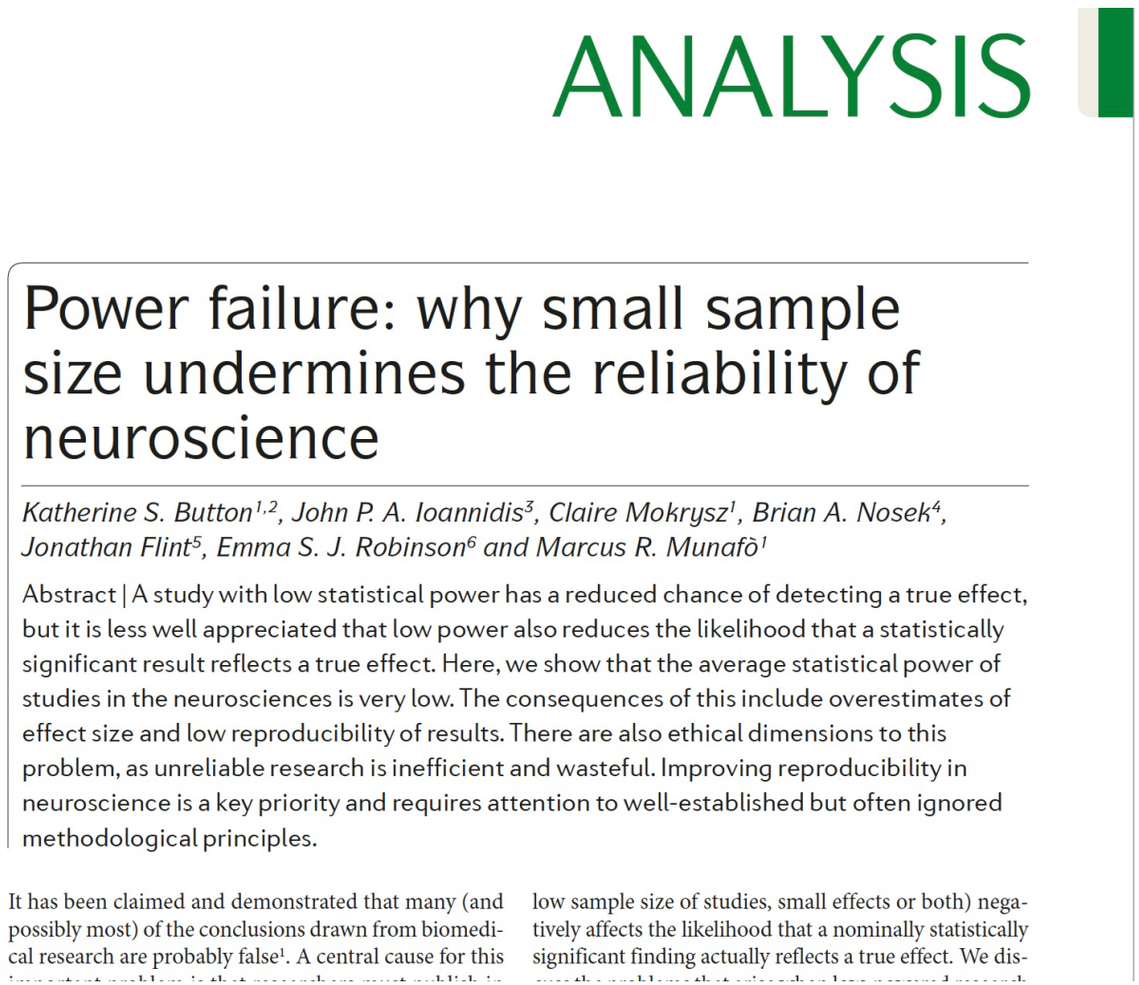
Outline for today

- 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
- Analysis follows design
- What if you can’t do experiments?

Plan your sample size

- Ethics boards and animal care committees require researchers to justify the sample sizes for proposed experiments on humans and other animals.
- Science is expensive: a low-power study is a waste of resources, and so is a study that is larger than necessary.
- Power (probability of detecting a true effect) of studies is often low in biology.
- *"We optimistically estimate the median statistical power of studies in the neuroscience field to be between ~8% and ~31%."*

Button et al (2013) *Nature Neuroscience*



ANALYSIS

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

It has been claimed and demonstrated that many (and possibly most) of the conclusions drawn from biomedical research are probably false[1]. A central cause for this

low sample size of studies, small effects or both) negatively affects the likelihood that a nominally statistically significant finding actually reflects a true effect. We dis-

Problems with low power studies

- High chance of a false negative: failing to detect a true effect.
- Low replicability. (Low power from small sample sizes was regarded as one reason behind the low replicability discovered by the *Science* study mentioned in Lecture 01 that repeated 100 psychology studies.)
- When 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.
- High uncertainty of estimates of effect size (wide confidence intervals) in studies with small sample size.
- “Winner’s curse” and publication bias.

Problems with low power studies

- If a statistically significant result is obtained in a low power study ("evidence for an effect"), the estimate of effect is likely to be exaggerated ("winner's curse").
- This "curse" can affect the replicability of published study results when statistically significant results are more likely to be published.
- We saw this effect in the *Science* study that repeated 100 psychology studies.

Button et al (2013) *Nature Neuroscience*

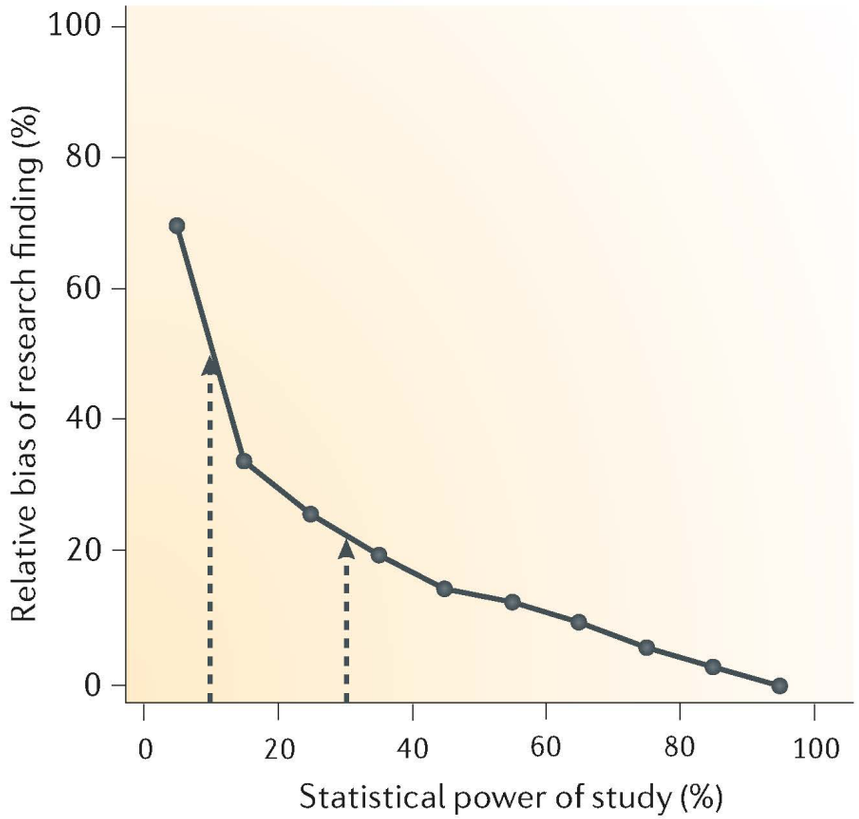


Figure 5 | **The winner's curse: effect size inflation as a function of statistical power.** The winner's curse refers to the phenomenon that studies that find evidence of an effect often provide inflated estimates of the size of that effect. Such inflation is expected when an effect has to pass a certain threshold — such as reaching statistical significance — in order for it to have been 'discovered'. Effect inflation is worst for small, low-powered studies, which can only detect effects that happen to be large.

Goals when planning your sample size

- ***Plan for sufficient precision.*** Choose a sample size to yield a confidence interval of specified width when estimating an effect. A narrow confidence interval indicates that an effect is estimated with high precision.
- ***Plan for sufficient power (e.g, 80%).*** Let's say that you are testing the difference between two treatment means $\mu_1 - \mu_2$ and only care if the difference is a specified value *D* or greater. Choose a sample size to yield a high probability of rejecting H_0 (≥ 80%) if the absolute value of the difference between the means, $|\mu_1 - \mu_2|$, is at least *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.

Hard decisions

- To achieve these goals, how should you 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?
- R is an amazing tool for simulating data to help plan sample size for any experimental design (workshop this week!).
- Power calculations are available in R for standard experimental designs (e.g., `pwr` package).

Challenges of planning sample size

- Key quantities needed to plan sample sizes, such as how much variation is present within groups (σ, the within-group standard deviation) are not known.
- Typically, a researcher makes an educated guess for these unknown parameters based on pilot studies or previous investigations.
- Parameter estimates based on the published literature may be biased, if they are derived from low power studies. Expect the real effects to be smaller.
- If no information is available then consider carrying out a pilot study first, before attempting a large experiment.
- Note: post-hoc power calculations are useless (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 (e.g., treatment and control).
    - Clinical trials, reciprocal transplant experiments, factorial experiments, 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” without two treatments, QTL “experiments”, etc, are examples of observational studies (no matter how complex the apparatus needed to measure response).

Why do experiments

- Because an observational study cannot distinguish between two possible reasons for an association between an *explanatory variable* and a *response variable*.
- For example, survival of climbers of Mount Everest was higher for individuals taking supplemental oxygen vs not taking supplemental oxygen.

  Reason 1) Supplemental oxygen (explanatory variable) increased survival (response variable).

  Reason 2) Supplemental oxygen had little or no effect on survival. Instead, survival and oxygen were associated because other variables affected 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.
- That's because with random assignment, no confounding variables will differ between treatments (except by chance).
- Assign supplemental oxygen/no-oxygen randomly to Everest climbers would break the correlation 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

- A clinical trial is 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.

Our experiments should be more like clinical trials

OPEN ACCESS Freely available online

PLOS | BIOLOGY

Perspective

Whole Animal 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].

Example of an experiment (clinical trial)

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 AUGUST 6, 2020 VOL. 383 NO. 6

A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19

D.R. Boulware, M.F. Pullen, A.S. Bangdiwala, K.A. Pastick, S.M. Lofgren, E.C. Okafor, C.P. Skipper, A.A. Nascene, M.R. Nicol, M. Abassi, N.W. Engen, M.P. Cheng, D. LaBar, S.A. Lother, L.J. MacKenzie, G. Drobot, N. Marten, R. Zarychanski, L.E. Kelly, I.S. Schwartz, E.G. McDonald, R. Rajasingham, T.C. Lee, and K.H. Hullsiek

https://www.cnbc.com/2020/07/28/trump-says-he-still-thinks-hydroxychloroquine-works-in-treating-early-stage-coronavirus.html

Example of an experiment (clinical trial)

- Covid-19 disease occurs after infection by SARS-CoV-2. Can hydroxychloroquine prevent symptomatic infection?
- Hydroxychloroquine is used to treat autoimmune diseases. Closely-related chloroquine is used to treat malaria. A lab study with cultured cells found that chloroquine can block the coronavirus from invading cells.
- Data were gathered on a volunteer sample of 821 participants in Canada and the USA (within 4 days of exposure to virus) but asymptomatic.
- Two treatments were assigned randomly to subjects at each clinic. One contained hydroxychloroquine 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:

The incidence of new Covid-19 illness did not differ significantly between those receiving hydroxychloroquine (49 of 414 [11.8%]) and those receiving placebo (58 of 407 [14.3%])

95% confidence interval, −7.0 to 2.2%.

P = 0.35.

"This randomized trial did not demonstrate a significant benefit of hydroxychloroquine as postexposure prophylaxis for Covid-19."

Design components of clinical trial

- To reduce *bias*, the experiment included:
  - Simultaneous control group (the participants receiving the placebo).
  - Randomization: treatments were randomly assigned to participants in each geographic region.
  - Blinding: neither the subjects nor the clinicians knew which participants were assigned which treatment.
- To reduce the *effects of sampling error*, the experiment included:
  - Replication: the study was carried out on multiple independent subjects.
  - Blocking: subjects were grouped according to country (Canada vs USA), yielding a repetition of the same experiment in different settings ("blocks").
  - Balance: the number of participants was nearly equal in the two groups within every clinic.

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* treatment 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. It eliminates 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. Can be assumed in most 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.

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 (who should also measure units from different treatments in random order).

Minimizing the effects of sampling error

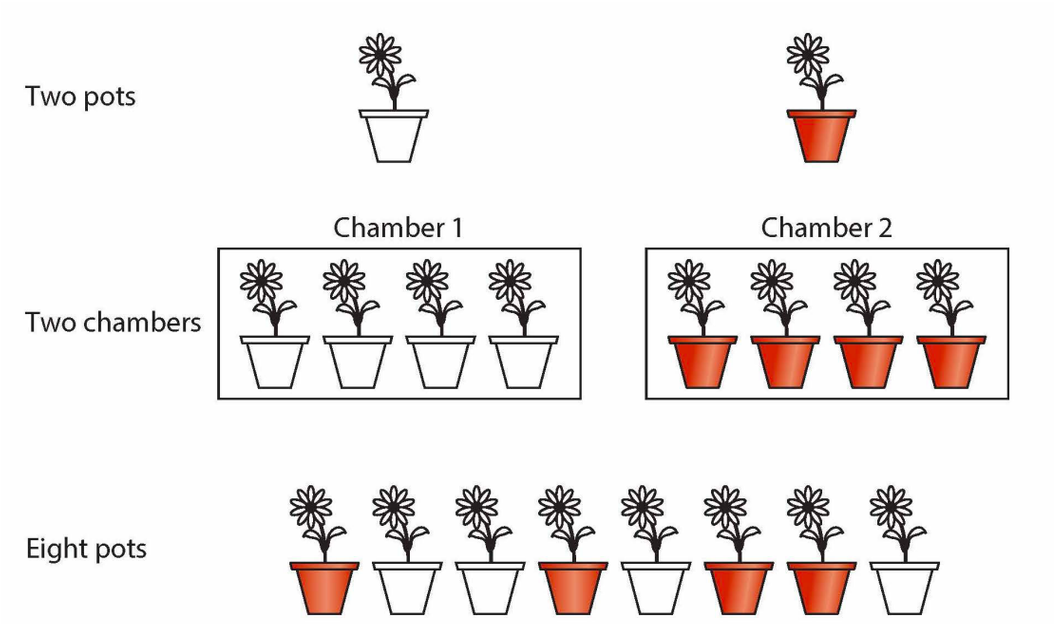
- The goal of experiments is to estimate and test treatment effects against the background of variation between individuals (“noise”) caused by other variables.
- One way to reduce noise is to make the experimental conditions constant. Fix the temperature, humidity, and other environmental conditions, 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 treatments 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 more precise estimates (less magnitude uncertainty) 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, it is 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 group of individual organisms, such as a field plot or pond containing multiple individuals, a cage of animals, a household, a family, a Petri dish, or an aquarium.
- Multiple individual organisms belonging to the same unit (e.g., plants in the same plot, bacteria in the same dish, family members) are likely to be more similar to each other, on average, than are 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*.
- Future lecture: *mixed effects models* can be used to analyze such data while avoiding pseudoreplication.

Balance

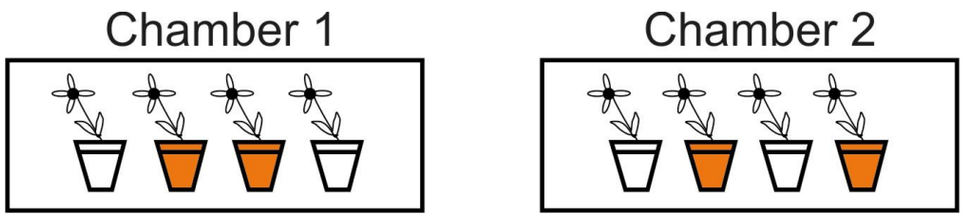
- A study design is balanced if all treatments have the same sample size.
- Balance helps to reduce the influence of sampling error on estimation and hypothesis testing. To appreciate this, look at the equation for the standard error of the difference between two treatment means. For a fixed total number 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 (assuming equal variances).
- 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

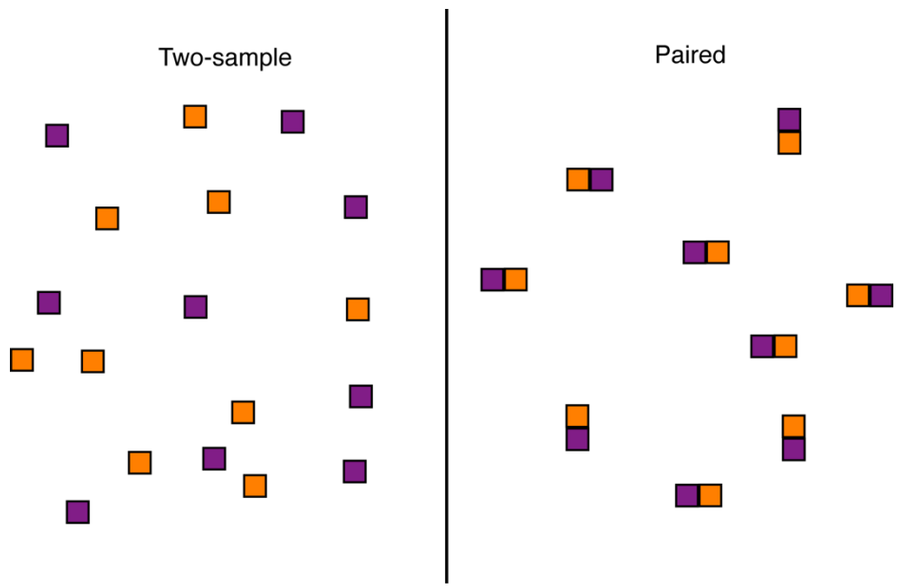
- Blocking is the grouping of experimental units that have similar properties (e.g., standards of medical care). Within each block, treatments are randomly assigned to experimental units.
- Blocking essentially repeats the same, completely randomized experiment multiple times, once for each block.
-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



- *Analysis follows design*. The structure of the analysis should reflect the structure of the experiment. Block (here, chamber) must be included as a (random) factor in the statistical analysis of the data results.

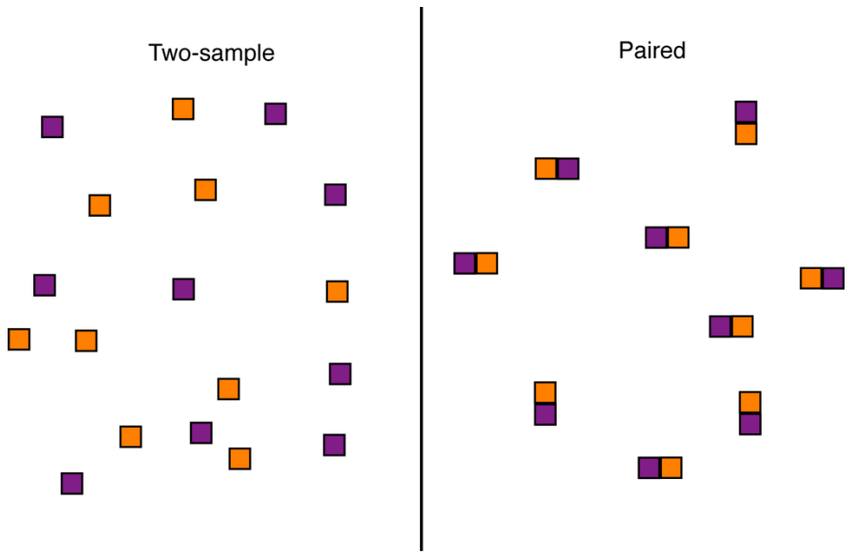
Blocking: Paired design

- For example, consider the design choices for a hypothetical 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 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 (RCB) design

- The paired design is a special case of the RCB design, which includes two or more than two treatments within blocks. Every block receives each treatment.
- By accounting for some sources of sampling variation, such as the variation among sites, blocking can make differences between treatments stand out.
- Blocking is worthwhile if units within blocks are relatively homogeneous, apart from treatment effects.
- In the example of a clinical trial, “country” 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.

Warning about experiments with time as a factor

Example experiment in desert ants

- Involves repeated measures of the same subjects (plots)
- Modeling time as a factor likely violates sphericity assumption of ANOVA and introduced autocorrelation of residuals.

Ecology, 65(6), 1984, pp. 1780–1786

GRANIVORY IN A DESERT ECOSYSTEM: EXPERIMENTAL EVIDENCE FOR INDIRECT FACILITATION OF ANTS BY RODENTS[1]

D. W. DAVIDSON
Department of Biology, University of Utah, Salt Lake City, Utah 84112 USA

AND

R. S. INOUYE[2] AND J. H. BROWN
Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721 USA

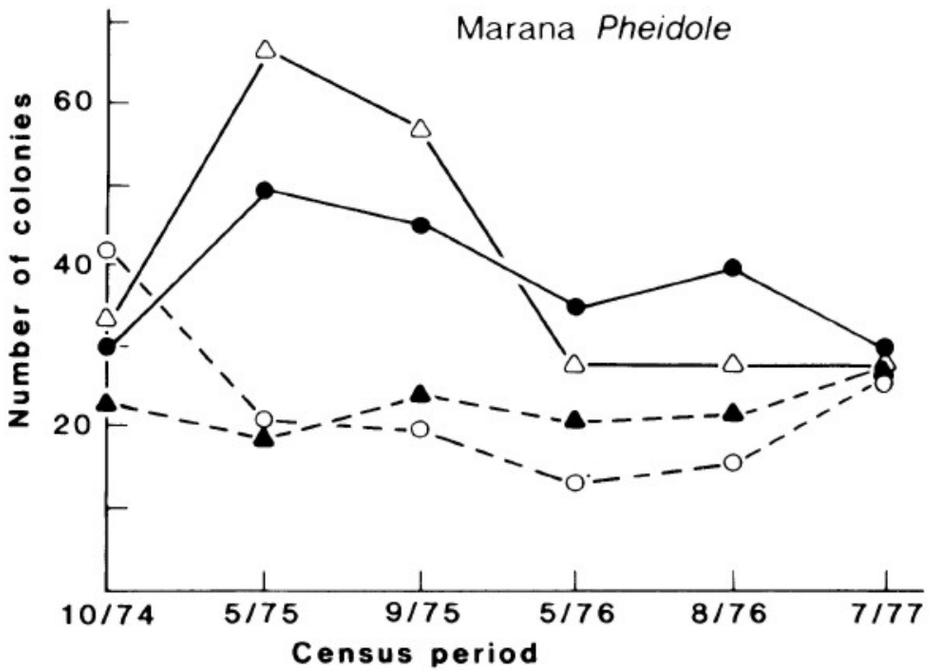


FIG. 3. Changes in density of *Pheidole* spp. (including *P. xerophila tucsonica, P. sitarches,* and *P. gilvescens*) on two rodent removal plots (——) and two control plots (— —) at Marana, Arizona over a 2¾-yr period.

Many experimental designs have been developed

- How to choose the best design for your purposes?
- R package `Agricolae` will help model most of the designs typically used in agricultural research (e.g., completely randomized design, randomized complete block design, balanced incomplete block designs, split-plot design, factorial design, latin square design).
- You can specify the randomization algorithm (and the random seed, to make it reproducible).
- The package claims to generate information about parameters, the resulting field book, summary statistics on the design like efficiency index, and even a sketch showing the distribution of plots in the field.
- Let me know if you try it!
- CRAN task view:
    - cran.r-project.org/web/views/ExperimentalDesign.html

Analysis follows design

- The structure of your analysis should reflect the structure of study 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.
- Mixed models (coming up!) allow you to do this, and avoid pseudoreplication.
- Remember that pseudoreplication is a problem of analysis, not design. It happens when the structure of the analysis doesn't match that of the experimental design.

Analysis follows design

- Recognize how you will analyze the data when you design your study – this is a prerequisite to plan the sample sizes you will need.
- To plan an experimental design and the sample sizes required to achieve goals, use R to simulate data according to parameters you provide. Then use R to analyze the simulated data.
- Iterative loops in your simulations allow you to undertake the sampling process and analysis many times. This will help you to generate estimates of power and precision no matter what your study design.
- You will attempt this in the workshop on Thursday!

What if you can't do experiments?

- Experimental studies are not always feasible, in which case we must fall back upon observational studies.
- To minimize bias and the impact of sampling error, the best observational studies incorporate as many of the features of good experimental design as possible (e.g., simultaneous controls, blinding; 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 (best)
  - statistically adjusting for known confounding variables (requires matching)

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 "**handouts**" tab on course web site.

Presenters: _____& _____

Moderators: _____& _____