

Designing experiments

Outline for today

- What is an experimental study
- Why do experiments
- Clinical trials
- How to minimize bias in experiments
- How to minimize effects of sampling error in experiments
- Experiments with more than one factor
- What if you can't do experiments
- Planning your sample size to maximize precision and power

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.
 - 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).

What is an experimental study

- In an experimental study, there must be at least two treatments
- The experimenter (rather than nature) must assign treatments to units or subjects.
- The crucial advantage of experiments derives from the random assignment of treatments to units.
- Random assignment, or randomization, minimizes the influence of confounding variables, allowing the experimenter to isolate the effects of the treatment variable.

Why do experiments

- By itself an observational study cannot distinguish between two reasons behind an association between an *explanatory variable* and a *response variable*.
- For example, survival of climbers to Mount Everest is higher for individuals taking supplemental oxygen than those not taking supplemental oxygen.
- One possibility is that supplemental oxygen (explanatory variable) really does cause higher survival (response variable).
- The other possibility is that supplemental oxygen has little or no effect on survival. The two variables are associated because other variables affect both supplemental oxygen and survival at the same time. For example, use of supplemental oxygen might be a benign indicator of a greater overall preparedness of the climbers that use it, and greater preparedness rather than oxygen use is the main cause of the enhanced survival.
- Variables (like preparedness) that distort the causal relationship between the measured variables of interest (oxygen use and survival) are called *confounding variables*.

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.
- For example, assigning supplemental oxygen/no-oxygen randomly to Everest climbers 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

- The gold standard of experimental designs is the clinical trial. Experimental design in all areas of biology have been informed by procedures used in 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.

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

Clinic	Nonoxynol-9		Placebo	
	<i>n</i>	Number infected	<i>n</i>	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

Design components of clinical trial

- The goal of experimental design is to eliminate bias and to reduce sampling error when estimating and testing effects of one variable on another.
- To reduce bias, the experiment included:
 - Simultaneous control group: the study included both the treatment of interest and a 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 control group is a group of subjects who are treated like all of the experimental subjects except do not receive the treatment of interest.
- A study lacking a control group for comparison cannot determine whether the treatment of interest is the cause of any of the observed changes.
- There are several possible reasons for this, including the following:
 - Sick human subjects selected for a medical treatment may tend to “bounce back” toward their average condition regardless of any effect of the treatment.
 - Stress and other impacts associated with administering the treatment (such as surgery or confinement) might produce a response separate from the effect of the treatment of interest.
- 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.

Simultaneous control group

- In clinical trials either a placebo or the currently accepted treatment should be provided. A placebo is an inactive treatment that subjects cannot distinguish from the main treatment of interest.
- In experiments requiring intrusive methods to administer treatment, such as injections, surgery, restraint, or confinement, the 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









- Once treatments are chosen, 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. Chance rather than conscious or unconscious decision determines which units end up receiving the treatment of interest and which receive the control.
- A *completely randomized design* is an experimental design in which treatments are assigned to all units by randomization.

Randomization

- 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.
- It ensures that variation from confounding variables is similar between the different treatment groups.

Randomization

- Randomization should be carried out using a random process, for example:
 - List all n subjects, one per row, in a computer spreadsheet.
 - Use the computer to give each individual a random number.
 - Assign treatment A to those subjects receiving the lowest numbers and treatment B to those with the highest numbers.

Experimental unit								
Random number	11	18	87	55	76	70	90	4
Treatment	A	A	B	A	B	B	B	A

- Other ways of assigning treatments to subjects are almost always inferior because they do not eliminate the effects of confounding variables.
- “Haphazard” assignment, in which the researcher chooses a treatment while trying to make it random, has repeatedly been shown to be non-random and prone to bias.

Blinding

- Blinding is the process of concealing information from participants (sometimes including researchers) about which subjects receive which treatment.
- 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

- In a *single-blind* experiment, the subjects are unaware of the treatment that they have been assigned. Treatments must be indistinguishable to the subjects, which prevents subjects from responding differently according to their knowledge of their treatment.
- 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.
 - Researchers sometimes have pet hypotheses, and they might treat experimental subjects in different ways depending on their hopes for the outcome.
 - Many response variables are difficult to measure and require some subjective interpretation, which makes the results prone to a bias.
 - Researchers are naturally more interested in the treated subjects than the control subjects, and this increased attention can itself result in improved response.

Blinding

- Reviews of medical studies have revealed that 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. (This probably overestimates the effects of a lack of blinding, because the 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).

[break]

Minimizing the effects of sampling error

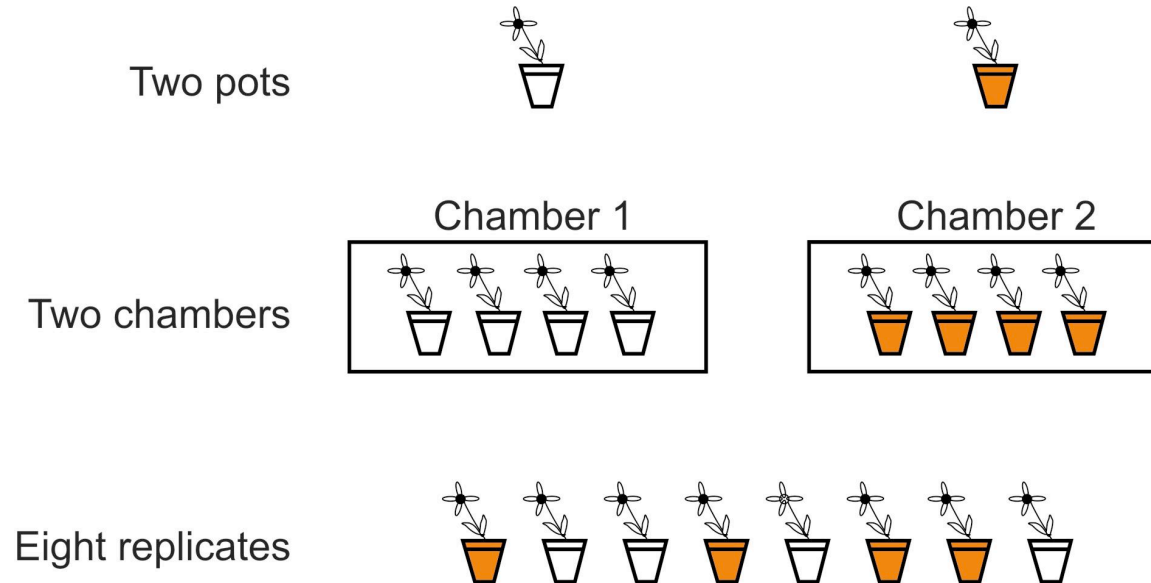
- The goal of experiments is to estimate and test treatment effects against the background of variation between individuals (“noise”) caused by other variables.
- One way to reduce noise is to make the experimental conditions constant. Fix the temperature, humidity, and other environmental conditions, for example, and use only subjects that are the same age, sex, genotype, and so on. In field experiments, however, 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.
- Without replication, we would not know whether response differences were due to the treatments or just chance differences between the treatments caused by other factors.
- Studies that use more units (i.e., that have 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.
- 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 on average to each other 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

Replication

- From the standpoint of reducing sampling error, more replication is always better.
- As proof, examine the formula for the standard error of the difference between two sample mean responses to two treatments, $\bar{Y}_1 - \bar{Y}_2$

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- Increasing n_1 and n_2 directly reduces the standard error, increasing precision.
- Increased precision yields narrower confidence intervals and more powerful tests of the difference between means.
- On the other hand, increasing sample size also has costs in terms of time, money, and even lives.

Balance

- A study design is balanced if all treatments have the same sample size. Conversely, a design is unbalanced if there are unequal sample sizes between treatments.
- Balance is a second way to reduce the influence of sampling error on estimation and hypothesis testing. To appreciate this, look again 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

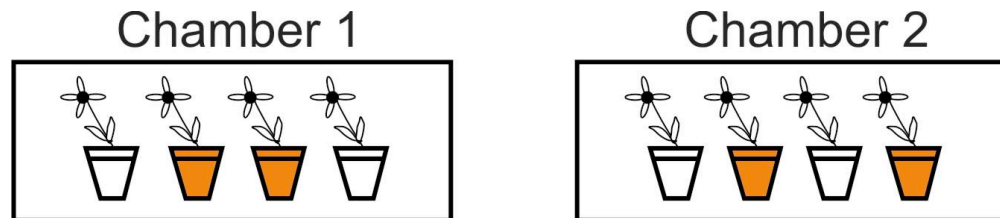
$$\left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

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.

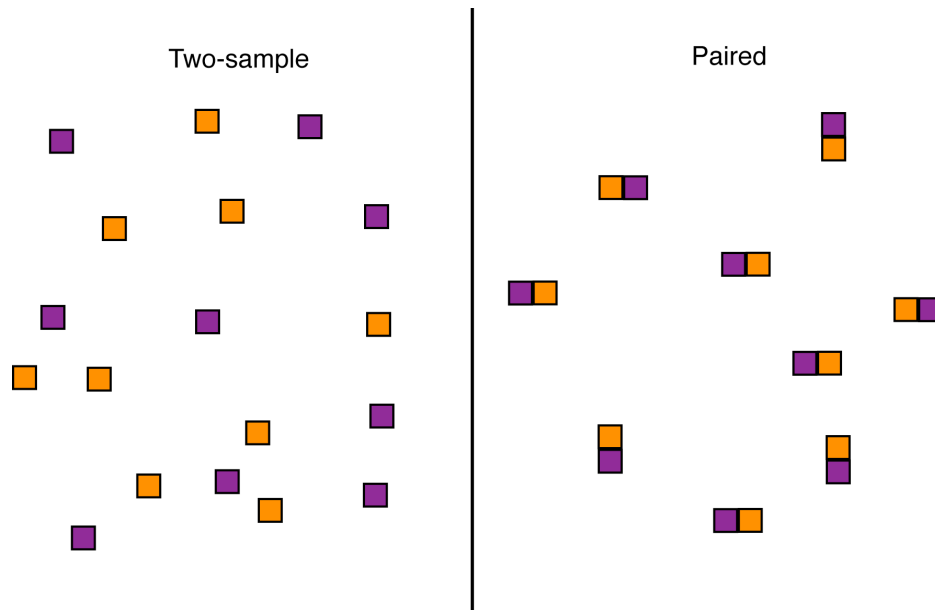
Blocking

- 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.



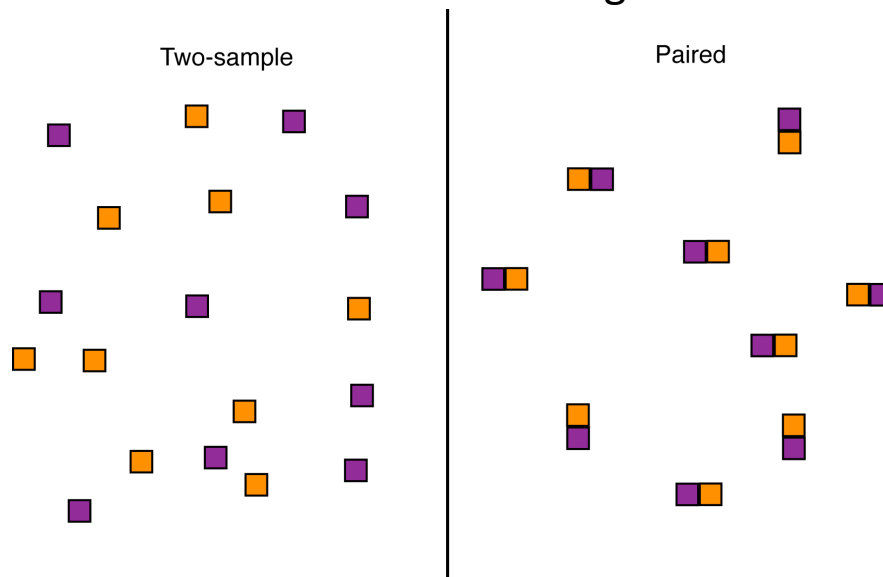
Blocking: Paired design

- 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 each plot to either the clear-cut treatment or the no clear-cut treatment.
- 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.
- 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 sometimes obscures the effects we are looking for.



Blocking: Randomized Complete Block design

- RCB design is analogous to the paired design, but may have more than two treatments. Each treatment is applied once to every block.
- As in the paired design, treatment effects in a randomized block design are measured by differences between treatments exclusively within blocks, a strategy that minimizes the influence of variation among blocks.
- 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.

Blocking: Randomized block design

- For example, Srivastava and Lawton (1998) made artificial tree holes from plastic that mimicked the buttress tree holes of European beech trees to examine how the amount of decaying leaf litter affected the number of insect eggs deposited (mainly by mosquitoes and hover flies) and the survival of the larvae.
- In one treatment (LL), a low amount of leaf litter was provided. In a second treatment (HH), a high level of debris was provided. In the third treatment (LH), leaf litter amounts were initially low but were then made high after eggs had been deposited.
- A randomized block design was used in which artificial tree holes were laid out in triplets (blocks). Each block consisted of one LL tree hole, one HH tree hole, and one LH tree hole.
- The location of each treatment within a block was randomized.



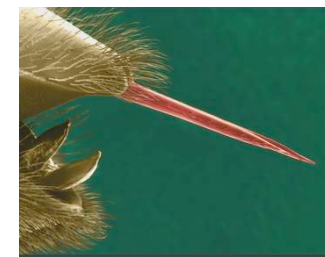
Blocking: Randomized block design

- For example, Blaustein et al. (1997) used a field experiment to investigate whether UV-B radiation was a cause of amphibian deformities. They measured long-toed salamanders either exposed to natural UV-B radiation or under UV-B shields. It was not possible to carry out all replicates simultaneously, so the researchers carried them out over several days.
- They made sure that both treatments were included on each day. In their analysis they grouped replicates together that were carried out on the same day into blocks.



Example of 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, and that continue to pump venom into the wound: scraping off with a credit card or pinching off with thumb and index finger.
- A total of 40 stings was induced on volunteers. Twenty were removed with the credit card method, and 20 were removed with the pinching method. The size of the subsequent welt by each sting was measured after 10 minutes. All 40 measurements were combined to estimate means, standard errors, and the P-value for a two-sample t-test of the difference between treatment means. Pinching led to a slightly smaller average welt, but the difference between methods was not significant.
- However, all 40 measurements came from two volunteers (both authors of the study), each of whom received one treatment ten times on one arm and the other treatment ten times on the other arm.
- 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.
- One reason to consider experiments with multiple factors is that the factors might interact.
- The factorial design is the most common experimental design used to investigate more than one treatment variable, or factor, at the same time. 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.

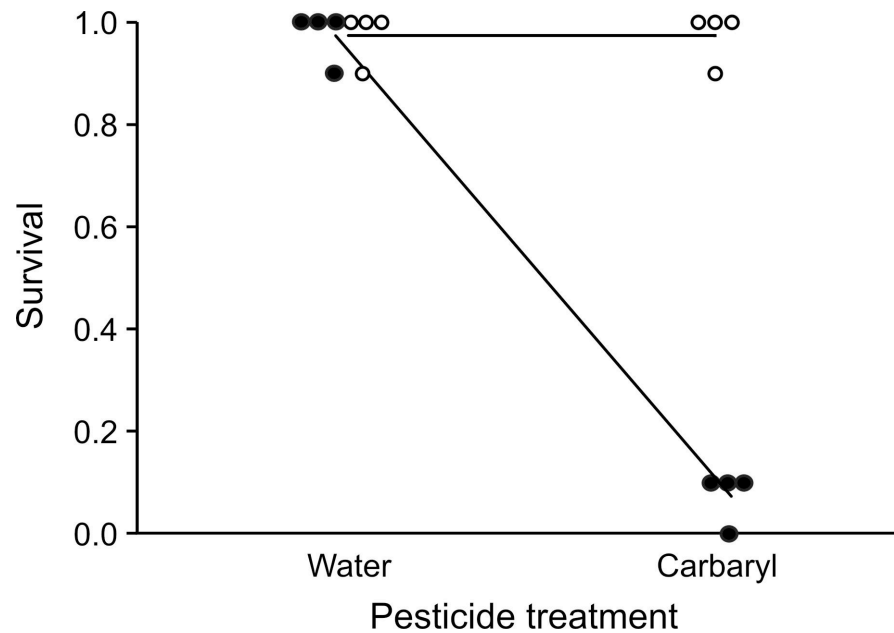
Factorial experiments

- For example, Relyae (2003) looked at how a moderate dose (1.6 mg/L) of a commonly used pesticide, carbaryl (Sevin), affected bullfrog tadpole survival. In particular, the experiment asked how the effect of carbaryl depended on whether a native predator, the red-spotted newt, was also present. The newt was caged and could cause no direct harm, but it emitted visual and chemical cues that are known to affect tadpoles.
- The experiment was carried out in 10-L tubs (experimental units), each containing 10 tadpoles. The four combinations of pesticide treatment (carbaryl vs. water only) and predator treatment (present or absent) were randomly assigned to tubs. For each combination of treatments, there were four replicate tubs.



Factorial experiments

- The results showed that survival was high except when pesticide was applied together with the predator—neither treatment alone had much effect. Thus, the two treatments, predation and pesticide, seem to have interacted



○ predator absent; ● predator present

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).
- Unlike randomization, matching in an observational study does not account for all confounding variables, only those explicitly measured. Thus, while matching reduces bias, it does not eliminate bias.
- Matching also reduces sampling error by grouping experimental units into similar pairs, analogous to blocking in experimental studies.
- 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, a statistical method such as analysis of covariance (a type of linear model) is used to correct for differences between treatment and control groups in suspected confounding variables.

Planning your sample size

- Ethics boards and animal care committees require researchers to justify the sample sizes for proposed experiments on animals, humans.
- Sample size planning involves two objectives: to achieve a predetermined level of precision of an estimate of treatment effect; or to achieve a predetermined power in a test of the null hypothesis of no treatment effect.
- Planning for precision involves choosing a sample size that yields a confidence interval of expected width. Typically, we hope to set the bounds as narrowly as we can afford
- Planning for power involves choosing a sample size that would have a high probability of rejecting H_0 if the absolute magnitude of the difference between the means, $|\mu_1 - \mu_2|$, is at least as great as a specified value D .

Plan for precision

- For example, consider a comparison of means of two-treatments.
 μ_1 = unknown mean of the treatment group,
 μ_2 = unknown mean of the control group.
- When the results are in we will compute the sample means \bar{Y}_1 and \bar{Y}_2 and use them to calculate a 95% confidence interval for $\mu_1 - \mu_2$, the difference between the population means of the treatment and control groups.
- To simplify, assume that the sample sizes, n are equal, and that the measurement in the two populations are normally distributed with equal standard deviation, σ .
- In this case, a 95% confidence interval for $\mu_1 - \mu_2$ will take the form $\bar{Y}_1 - \bar{Y}_2 \pm \text{uncertainty}$, where “uncertainty” is half the width of the confidence interval.

Plan for precision

- Planning for precision involves deciding the uncertainty we can tolerate in advance. Once we've decided that, then the sample size needed in each group is approximately

$$n = 8 \left(\frac{\sigma}{\text{uncertainty}} \right)^2$$

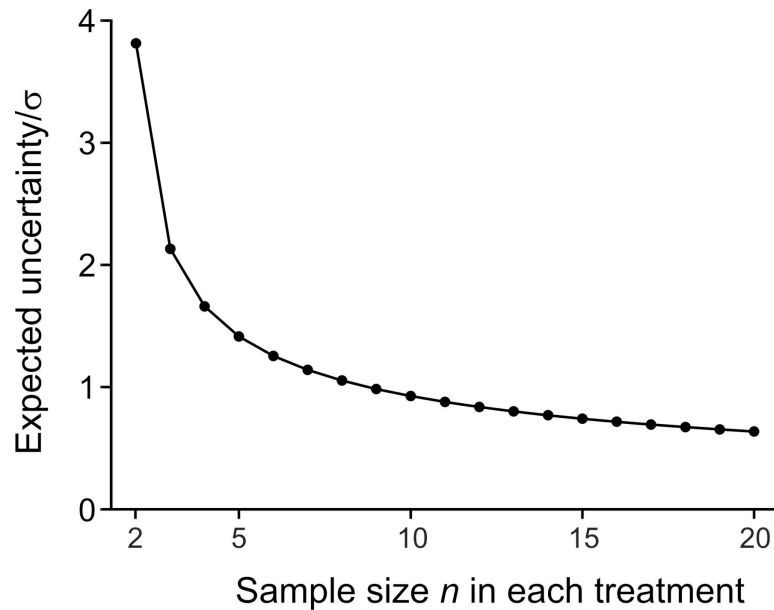
- A larger sample size is needed if σ , the standard deviation within groups, is large than if it is small.
- A larger sample size is needed to achieve a high precision (a narrow confidence interval) than to achieve a lower precision
- A major challenge is that key quantities like σ are not known. Typically 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.

Plan for precision

- After planning, imagine that the experiment is run and you now have your data. Will the confidence interval calculate have the precision you planned for? Probably not for two reasons.
- You only had an educated guess for σ .
- Second, the within-treatment sample standard deviation s from the experiment will not equal σ because of sampling error. The resulting confidence interval will be narrower or wider accordingly.
- The probability that the resulting confidence interval is less than or equal to the desired precision is only about 0.5. To increase this probability you would need an even larger sample size.

Plan for precision

- The graph below shows expected precision of the 95% confidence interval for the difference between two treatment means. The vertical axis is given in standardized units, uncertainty/ σ .
- Very small sample sizes lead to very wide interval estimates of the difference between treatment means. More data gives better precision.
- Precision initially declines rapidly with increasing sample size, but it then declines more slowly. Thus, we get diminishing returns by increasing the sample size past a certain point



Plan for power

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0.$$

- The power of a test is the probability of rejecting H_0 if it is false.
- Planning for power involves choosing a sample size that would have a high probability of rejecting H_0 if the absolute magnitude of the difference between the means, $|\mu_1 - \mu_2|$, is at least as great as a specified value D .
- D is just the minimum we care about. By specifying a value for D in a sample size calculation we are deciding that we aren't much interested in rejecting the null hypothesis of no difference if $|\mu_1 - \mu_2|$ is smaller than D .
- A conventional power to aim for is 0.80. That is, if H_0 is false, we aim to demonstrate that it is false in 80% of experiments.
- Power calculations made once the experiment is over, to determine in retrospect how powerful the experiment was, are useless and should be avoided (by definition, a test that failed to reject H_0 will always be found to have had low power).

Plan for power

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0.$$

- If we aim for a power of 0.8 and a conventional significance level of $\alpha = 0.05$, then a quick approximation to the planned sample size n in each of two groups is

$$n \cong 16 \left(\frac{\sigma}{D} \right)^2$$

(Lehr 1992). This formula assumes that the two populations are normally distributed and have the same standard deviation (σ), which we are forced to assume is known.

- For a given power and significance level, a larger sample size is needed when the standard deviation σ within groups is large, or if the minimum difference that we wish to detect is small.

- Sample size formulas for desired precision and power are available for one- and two-sample means, proportions, and odds ratios. We can also use R to calculate power using simulation (workshop next week).

Plan for data loss

- The methods described here for planned sample sizes refer to sample sizes at the end of the experiment.
- But 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.

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:

Moderators: