Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

February 24, 2016

The Voinovich School of Leadership and Public Affairs

Table of Contents

1. Elements of Good Research Designs
2. Three Case Studies
3. Matching
4. Choosing Needed Sample Size
5. Planning for Power

Elements of Good Research Designs

Experiments

Experiments (the gold standard) – very powerful at isolating cause-and-effect because they can leverage the benefits of random assignment and hence minimize the influence of confounding variables ... variables are confounded if their influence on the outcome cannot be separated from one another

For example: Ice-cream consumption in a city appears to be correlated with Crime in the city. Reality: Both go up in warm weather so if one controls for temperature, correlation between ice-cream consumption and crime disappears

For some more interesting running tabulations of spurious correlations

Researchers do have to minimize the probability of experimental artifacts ... something about the experiment itself that taints the outcome

Example

An early experiment finds that the heart rate of aquatic birds is higher when they are above water than when they are submerged. Researchers attribute this as a physiological response to conserve oxygen. In the experiment, birds are forcefully submerged to have their heart rate measured. A later experiment uses technology that measures heart rate when birds voluntarily submerge, and finds no difference in heart rates between submerged and above water groups. This suggests that the stress induced by forceful submersion rather than submersion itself caused the lowering of heart rate in the birds.

Quasi-Experiments

Quasi-Experimental (aka Observational) designs lack this leverage and hence must (a) at best establish an association between X and Y, and (b) struggle with the influence of confounding variables

For example, in assessing risk of accidents or adverse health outcomes one has to control for age, sex, income, race/ethnicity, etc. because one cannot, unlike an experiment, randomly assign individuals to a particular age-group, race/ethnic group, and so on

Key goal becomes to have the treatment and control groups be as similar as possible on all pre-outcome dimensions

Very difficult goal to achieve unless (a) you have enough substantive knowledge and (b) you have good measurements to work with

Control group – A group that do not receive the treatment but otherwise experience similar conditions as other units in the experiment or the quasi-experimental study

Three Case Studies

Starling Song

- Male starlings sing in the spring when they try to attract female mates and to keep other males at bay. In the fall they sing when in flocks of other males. But how do you tell the two songs apart?
- A researcher randomly assigned 24 starlings into two groups of 12 each
- The spring group was kept in a spring-like environment with more light, a nest box, and a nearby female starling
- The fall group was kept in a fall-like environment with less light, no nest boxes, and in the proximity of other male birds
- Each bird was observed and the length of each song was recorded for ten hours
- Each bird sang from between 5 and 60 songs

Cattle Diet

- Researchers studying dairy cow nutrition have access to 20 dairy cows in a research herd. Response variables include milk yield
- Want to compare a standard diet (A) with three other diets (B, C, D), each with varying amounts of alfalfa and corn.
- Cows are randomly assigned to four groups of 5 cows each
- Each group receives each of the four diet treatments for a period of three weeks; first week involves no measurements so that the cow can adjust to the new diet
- Diets are rotated according to a Latin Square design so that each group has a different diet at the same time.

Cow Group	Time 1	Time 2	Time 3	Time 4
1	A	B	C	D
2	C	A	D	B
3	B	D	A	C
4	D	C	B	A

The HIV Transmission Study

Volunteer samples of sex-workers were recruited from 3 clinics in Asia (Thailand) and 3 in Africa (Benin, Côte d'Ivoire and South Africa). Two gel treatments were assigned randomly to the women, one containing Nonoxynol-9, believed to reduce the likelihood of HIV-1, and the other a placebo.
Neither the subjects nor the researchers knew who was getting which of the two gels ... double blinding
Each clinic had a control group
Each clinic had balanced (i.e., roughly equal sized) treatment and control groups
Subjects were blocked (i.e., grouped) within each clinic

	Nonoxynol-9		Placebo	
Clinic	n	No. Infected	n	No. 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

Randomization

Randomization works because without prejudice you end up assigning units to the treatment versus the control groups

There is then no systematic way that you could bias the make-up of each group because even if there are confounding variables, these should end up being evenly distributed the treatment and control groups

May have to use Stratified Randomization

Example

In the dairy cow example it was known that there were 8 cows in their first milking and 12 cows not in the first milking, the 8 primiparous cows could be randomly assigned to two to each group and the 12 multiparous cows could be randomly assigned three to each group.

Blocking and Balance

Blocking puts sampling units into groups that are similar with respect to one or more covariates (for e.g., neighborhoods, plots of land, some portion of a stream, etc). Treatments are assigned at random within the blocks

- The paired design is an extreme form of blocking where each pair of measurements form a block of size two
- Blocking is an attempt to directly control for the effects of a factor
- Blocking on the basis of one factor assures that the one factor is close to balanced in each treatment group
- If you attempt to block on multiple factors, the number of blocks grows large and there may be insufficient units that can be placed into each block
- Blocking and randomization are two methods to reduce bias from confounding factors, but there is a tension between them: the more you need to block the less of the sample left over to be randomized across the blocks

Balance requires that the number of units be equal in each treatment group

- When σ are equal across groups, the standard error for the difference is smallest when $n_1 = n_2$. With unequal population standard deviations it may help to sample more individuals from groups with higher σ^2
- In the cow diet example, balance is ensured because each cow receives each treatment and is measured during each time period

Randomized Block Designs

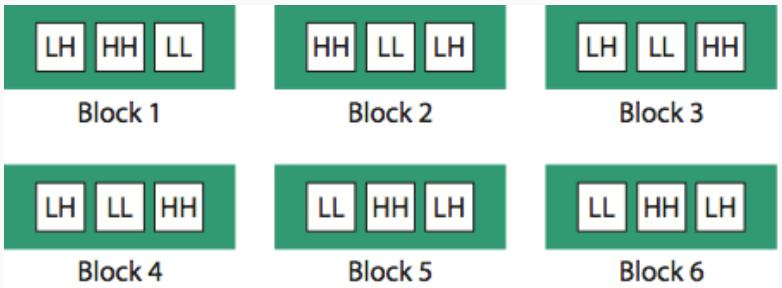
Blocks = groups that share common features. Ideally you want to have every treatment condition randomly assigned within each block

Example 1: A fast food franchise is test marketing 3 new menu items.

- To find out if they have the same popularity, 6 franchisee restaurants are randomly chosen for participation in the study.
- In accordance with the randomized block design, each restaurant will be test marketing all 3 new menu items.
- Furthermore, a restaurant will test market only one menu item per week, and it takes 3 weeks to test market all menu items.
- The testing order of the menu items for each restaurant is randomly assigned as well.

Example 2: Tree-hole study to see if amount of decaying leaf litter typically present in water-filled tree holes influences the number of insect eggs deposited and survival of larvae emerging from these eggs

- Researchers made artificial tree holes from plastic that mimicked the buttress tree holes of European beech trees.
- These plastic holes were placed next to trees in a forest in southern England.
- Three treatment conditions
  1. Low level of leaf litter (LL)
  2. High level of leaf litter (HH)
  3. Low levels initially but increased once eggs were deposited (LH)
- Six blocks, each with three plastic holes, one per treatment, placement randomized within each block



Latin Square Designs

- These designs use one Treatment and two blocking factors
- For e.g., testing 4 diets on four cow groups
- Think of blocking factors as sources of variability – here the cows (each could be slightly different) and the diet sequence (might make a difference)

Cow Group	Time 1	Time 2	Time 3	Time 4
1	A	B	C	D
2	C	A	D	B
3	B	D	A	C
4	D	C	B	A

- Note: If the four groups are made up of roughly similar cows then even if the order of the diets presented influences outcomes, this influence is being nullified since the order of the diets is randomized across the four groups
- Latin Squares can be of any size so long as each treatment occurs only once in each row and in each column

Replication and Pseudo-Replication

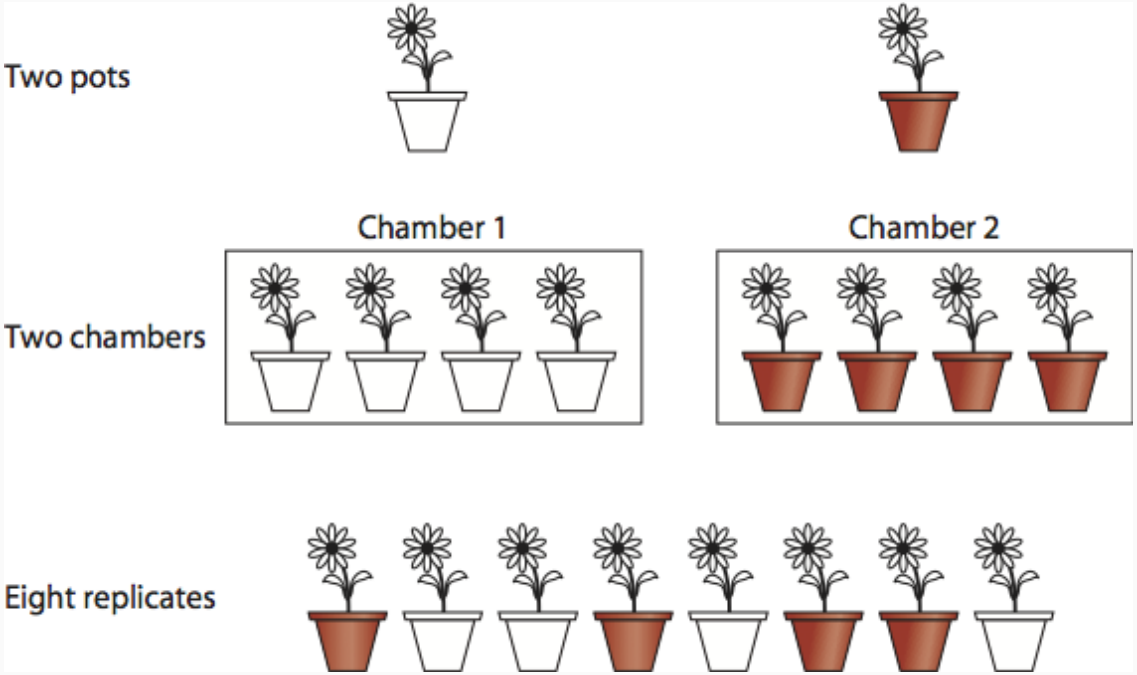
Replication involves exposing multiple independent units to each treatment

- If each treatment is run on only one or a few units then you don't have enough variation within and across treatments to decipher if the treatments are really having an impact
- Think of this as needing, for each treatment, both the mean and the standard deviation; if you have only one unit per treatment then you cannot calculate the standard deviation

Pseudo-Replication occurs when multiple units are not really independent but are treated as such

- The cormorants example – the same birds were made to dive multiple times and each dive measured (falsely) as an independent measurement
- The songs of each Starling are not independent
- We may have four separate measurements for each cow but these are not four independent measurements – they have one common factor, the cow so the sample size is really 20 and not 80

Replication: Two Fertilizers, Two Temperatures, and Plant Growth



- Panel 1 has no replication; just one plant per treatment
- Panel 2 seems to have replicates but plants within a chamber are not independent so they are not true replicates
- Panel 3 randomly assigns the two fertilizers to the plants and randomizes the plants across the chambers

Blinding and Double-Blinding

Subtle biases creep into a study if the investigators, and/or the participants, and/or the data analysts know which unit was received which treatment; one tends to look for what one hopes to find

Blinding refers to instances where the investigators have no idea about how the units were allocated to the various groups

Double-Blinding refers to instances where both the investigators and the participants have no clue who received which treatment. This is especially important in medical trials because the placebo effect has been well established

Triple-Blinding refers to instances where the investigators, the participants, and the data analysts are clueless as to who received which treatment. This of course assumes that the analysts are separate from the investigators, but this is not always the case so triple-blinding is relatively rare

Example

"The study was double-blinded – that is, neither the women nor the study staff (including the biostatisticians) ... knew which group was using the nonoxynol 9 film. ... The films were identical in appearance, packaging, and labeling."

"We asked 126 staff members their opinions of which film was the placebo. Some 18% thought film A (the placebo) was the placebo, 13% thought film B (nonoxynol 9) was the placebo, and 69% had no opinion ... Of the 68 peer educators (the staff members most likely to reflect the opinion of the participants), 16% thought film A was the placebo, 13% thought film B was the placebo, and 71% had no opinion."

The HIV Transmission Study

Volunteer samples of sex-workers were recruited from 3 clinics in Asia (Thailand) and 3 in Africa (Benin, Côte d'Ivoire and South Africa). Two gel treatments were assigned randomly to the women, one containing Nonoxynol-9, believed to reduce the likelihood of HIV-1, and the other a placebo.
Neither subjects nor researchers knew who was getting which gel ... double blinding
Each clinic had a control group
Each clinic had balanced (i.e., roughly equal sized) treatment and control groups
Subjects were blocked (i.e., grouped) within each clinic

	Nonoxynol-9		Placebo	
Clinic	n	No. Infected	n	No. 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

The design reduced *potential bias* via a (i) control group, (ii) randomization, and (iii) double-blinding, and *sampling error* via (i) replication (multiple independent subjects received treatment/placebo), (ii) balance, and (iii) blocking

Unfortunately ...

Matching

Matching for Quasi-Experimental Designs

Quasi-Experiments can benefit from matching ... essentially a regression-based approach to creating roughly equal treatment and control groups
How? Equal in the sense that all possible confounding variables are used to create similar groups
Logic: For every unit you only see one outcome (Y_0 or Y_1) but you want to know whether the treatment had any effect (i.e., $Y_1 - Y_0$), and this involves the counterfactual – what would have happened if unit i had received the placebo instead of the treatment?
How does it work? ... see the Lalonde example that follows

Example

"We estimate the effect of total nitrogen on macroinvertebrate taxon richness in streams in the western United States, from the U.S. EPA's western EMAP dataset. We use propensity scores to control for five potentially confounding covariates: catchment area, sediment, agricultural land use, annual precipitation, and chloride. ... defining strata becomes increasingly difficult as the number of covariates grows. ... Stratification by propensity score solves this problem via a balancing approach, in which a single metric (the propensity score) combines the effects of multiple original covariates."

Matching in R

Example

Labor economists have long wondered whether job training programs help the unemployed and the underemployed. Towards this end the typical study involves gathering a large sample of individuals who were exposed to job training and others who were not. The outcomes are analyzed via regression-based techniques that try to control for various confounding variables. These techniques are not as powerful as what you can get from Matching because people are not randomly assigned to job training programs but instead end up self-selecting to a large extent.

... switch to R

Choosing Needed Sample Size

How Large a Sample Do I Need?

- When looking to reject H_0 two objectives drive the choice of sample size:
  1. We are looking to achieve a specific degree of precision ... i.e., end up with a 95% or 99% confidence interval that is as small as possible
  2. We are looking to achieve a specific power for the test ... i.e., be able to reject H_0 when H_0 is not true in at least 80% of the trials[1]
- Both objectives are complicated because they vary according to the type of test we are looking to carry out
- Online calculators and free apps are available (for example, G*power)
- Here we will look at a few specific testing scenarios ...

[1] Convention sets this 80% rule on the basis of suggestions by Cohen (1977, 1988) that $\beta = 0.20$ ... i.e., power = $1 - \beta = 1 - 0.20 = 0.80$ .. is as high as one should go.

Precision

- Precision is all about the confidence intervals we end up with... how close do we want to be the truth? The narrower the interval, the closer we are.
- Assume we want to test whether the means of two groups really differ. Assume we also decide to have equally sized samples ... $n_1 = n_2$
- Recall that CI = $\bar{Y_1} - \bar{Y_2} \pm t_{\alpha/2;df}\left(SE_{\bar{Y_1}-\bar{Y_2}}\right)$ ... which is $\bar{Y_1} - \bar{Y_2} \pm$ margin of error
- Assume we want the margin of error $= 1$ ..., for our estimated difference to be within 1 of the true difference, i.e., $\bar{Y_1} - \bar{Y_2} \pm 1$.
- We haven't sampled yet so we don't know n_1, n_2, s_1, s_2, df_1, df_2, or t
- Instead we have to come up with some estimate of σ_1, σ_2 and use z instead of t.
- What estimate of σ_1, σ_2 would be good? ... The values of s_1, s_2 from previous studies or then the values derived from a pilot study. Let us assume variances to be equal in the two groups
- What about z? Well that is just 1.96 for a 95% confidence interval

The Calculations

margin of error $= 1.96 \times \sqrt{s^2_{pooled}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ where $s^2_{pooled} = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}$

Squaring both sides to get rid of the square-root ...

$(\text{margin of error})^2 = (1.96)^2\left(\sigma^2_{pooled}\left(\frac{1}{n}+\frac{1}{n}\right)\right)$ [... since we set $n_1 = n_2$]

$(\text{margin of error})^2 = (1.96)^2\left(\frac{2\sigma^2_{pooled}}{n}\right)$

Solving for n yields $n = \frac{(1.96)^2(2\sigma^2_{pooled})}{(\text{margin of error})^2}$

If desired margin of error $= 1$, and $\sigma^2_{pooled} = 2$ then $n = \frac{(1.96)^2(2\times 2)}{(1)^2} \approx 8$

If desired margin of error $= 1$, and $\sigma^2_{pooled} = 4$ then $n = \frac{(1.96)^2(2\times 4)}{(1)^2} \approx 31$

If desired margin of error $= 0.1$, and $\sigma^2_{pooled} = 2$ then $n = \frac{(1.96)^2(2\times 2)}{(0.1)^2} \approx 1,537$

What about for a Proportion?

Want to test whether toads are equally right-handed and left-handed

We know that the 95% CI is given by $\hat{p} \pm z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

$\text{margin of error} = z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

$(\text{margin of error})^2 = z^2\left(\frac{p(1-p)}{n}\right)$

$n = z^2\left(\frac{p(1-p)}{(\text{margin of error})^2}\right)$

Let us say we want the 95% CI to be within 0.1 of the true proportion; i.e., margin of error=0.1 and $z = 1.96$

If we can guess what p might be we can use that value; else just set $p = 0.5$

Then $n = (1.96)^2\left(\frac{0.5(1-0.5)}{(0.1)^2}\right) = (1.96)^2\left(\frac{0.25}{0.01}\right) = (1.96)^2(25) = 96.04 \approx 97$

What if we want to be within 0.05?

$n = (1.96)^2\left(\frac{0.5(1-0.5)}{(0.05)^2}\right) = (1.96)^2\left(\frac{0.25}{0.0025}\right) = (1.96)^2(100) = 384.16 \approx 385$

Note ... if $p > 0.5$ or $p < 0.5$ then the needed sample size shrinks a bit

In the table below we have set margin of error $= 0.1$ and $z = 1.96$ for calculating the needed sample size and then rounded n

p	$1-p$	$p*(1-p)$	n
0.10	0.90	0.09	35
0.20	0.80	0.16	61
0.30	0.70	0.21	81
0.40	0.60	0.24	92
0.50	0.50	0.25	96
0.60	0.40	0.24	92
0.70	0.30	0.21	81
0.80	0.20	0.16	61
0.90	0.10	0.09	35

Section 14.9 in the text has quick formulas for various testing situations but be careful

Planning for Power

Power of a Test

Recall that $\alpha = P(\text{Type I Error})$ ... aka, rejecting $H_0 | H_0$ is True
We also have $\beta = P(\text{Type II Error})$ ... aka, not rejecting $H_0 | H_0$ is False
Power of a test, for a specific H_a, is the probability of rejecting H_0 and measured as $1 - \beta$
In general the following quantities are linked so that given any 3 we can solve for the 4^{th}

1. n ... the sample size
2. α ... probability of finding an effect that is not real
3. $power = 1 - \beta$ ... probability of finding an effect that is real; typically set to at least 0.80
4. d ... the effect size

See Lenth's Power calculator and at some point read his memo Two Bad Habits