NRES 798 — Statistical Methods for Ecologists
Part II: Designing Experiments

Oscar García

March 7, 2013

Contents

This Part, chapters 6, 7, and 8, discusses general aspects of project planning and management. The textbook is a "Primer" supposed to be a first introduction to Statistics, with your research experience you should already be familiar with much of this. Maybe not, and it contains generally good advice, read it.

Not much to add, mostly comments on what we will need for later. The material on experimental design in Chapter 7 may make more sense together with the models and analysis, we will come back to it.

1 Chapter 6, designing field studies

Clarifying a statement on page 138: "(1) many ecological hypothesis do not generate simple, falsifiable predictions; and (2) even when a hypothesis does generate predictions, they are not unique". This must refer to a faulty design or insufficient data, a hypothesis whose effects cannot be detected is not scientific. Hypotheses should be falsifiable.

As pointed out in the book, in the ecological literature experiments (or *designed experiments*) are sometimes called "manipulative experiments", and observational studies are incorrectly called "natural experiments". This terminology is best avoided.

"Trajectory" studies are commonly called longitudinal, repeated observations, time series, or dynamic. Common names for "snapshot" are cross-sectional, transversal, or static. In ecology, cross-sectional data that is used to infer a time trend is called a *chronosequence*. Although literally that means "sequence over time", exactly the opposite.

The "rule of 10" for the number of replicates may be appropriate for the type of studies that the authors are interested in, but that tends to be problem-dependent. In many experimental designs, for instance in agriculture, it is typical to use only 2 replicates, using available resources for covering more treatment combinations.

When designing a large and expensive study, it is now feasible and a good idea to simulate it first. Generate random data like what you expect to observe, with reasonable guesses for standard deviations, etc. Then carry out the proposed analysis and see what you get. Repeat several times.

The purpose of ensuring independence and randomizing is to get a sample that complies with model assumptions. The observations should have the same distribution as the population (in simple random sampling). Randomization balances-out ignored factors, avoiding confounding and making an analysis more manageable.

2 Chapter 7, sampling and experimental design

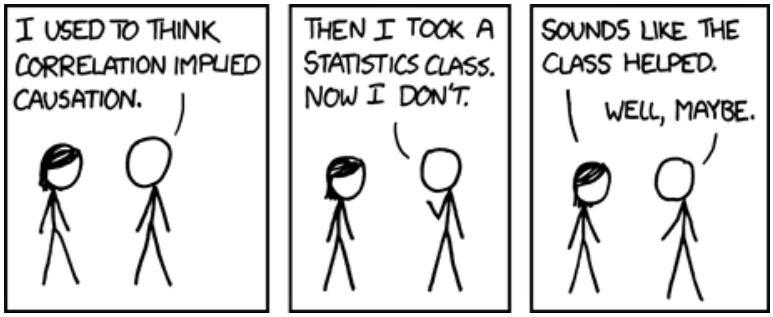
We will use only simple random sampling, where the distribution of the observations and of the population are (supposed to be) the same. There are more efficient strategies, studied in Survey Sampling and in Resource Inventory. For instance, stratification chooses random samples of given sizes from population subsets, not simply from the whole population. Variable probability sampling concentrates effort on the more important units. And/or one can measure expensive variables in a sub-sample, and use ratio or regression estimators. Possibly at several hierarchical levels. The analysis has to take into account all that. It is a large and specialized subject, if you ever need it you could try Schreuder et al. *Statistical techniques for sampling and*

monitoring natural resources from `http://www.fs.fed.us/rm/pubs/rmrs_gtr126.html`; they will mail you a free copy if you ask.

The other topic is experimental design. Many of the points made apply also to observational studies.

Before, we modelled a simple random variable Y. Now we are interested in relationships $Y = f(x_1, ..., x_p)$, between Y and one or more other variables. The analysis includes problems of hypothesis testing, and of estimation. In some situations the x_i are random variables X_i, so that there is a multivariate joint distribution and one has "proper" correlations. More often the x_i are taken as ordinary variables that can be chosen or fixed at given values, or the analysis is based on the conditional distribution of Y given $X_i = x_i$ (regression).

Y is called the dependent variable, response, or simply y-variable. The x_i are independent variables, predictors, or x-variables. There is usually a hypothesized cause-effect relationship, with the x_i representing causes that affect or determine Y. The statistical analysis, however, can only establish an association or correlation among the variables. "Correlation does not imply causation".



`http://xkcd.com/552/`

Variables x_i and Y can be of two types: (a) Numerical, called *continuous* although they may actually be discrete, taking only integer values. (b) *Categorical*, taking only a small number of possible values called *levels*; they can be modelled as discrete RVs. In *R*, categorical variables are called *factors*, although factor has a somewhat wider meaning in general statistical usage. Levels are often identified by names, but integers can also be used. Sometimes it is important to distinguish between *ordered* categorical variables or factors (like low, medium, high, aka ordinal variables), and unordered categorical variables or factors (e.g., sex).

The experimental designs table in the textbook can be applied to models, design, and analysis. A slightly modified version is:

	Independent(s)		
Dependent	Continuous	Categorical	Both
Continuous	Regression	ANOVA	ANCOVA
Categorical	Log.reg, GLM, etc.	Cont. tables	

I have included the possibility of more than one independent variable, in which case there can be a mixture of continuous and categorical variables. Bernouilli categorical responses can be handled through a logistic transformation (logistic regression); binomial, Poisson and other RV's may be analyzed with generalized linear models (GLM) or other methods. Categorical-categorical data is common in the social sciences, contingency tables being an important analysis tool.

Only the case of a single Y is treated here. If Y is a vector one uses *multivariate methods*.

ANOVA traditionally works with multi-way tables. Like this two-way one for the ants example near the end of the notes for Chapter 5:

	Habitat	
Site	Field	Forest
1	13	7
2	8	3
3	5	7
4	14	7
5	3	10
6	7	3

It contains ant nest counts for each combination of the categorical variables *Site* (6 levels) and *Habitat* (2 levels). One might have several observations for each entry, called *replicates*. The data is *balanced* if the number of observations for all combinations are the same. The levels are sometimes referred to as *treatments*, especially when the variable is the subject of interest, as it may be for *habitat* here. Others may be called *blocks*, when they are introduced mainly to reduce variability in the comparisons, like *site* in the example.

Using i for habitat and j for site, the two-way ANOVA model can be written as

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \ ,$$

where the observation Y_{ij} consists of an over-all mean μ, a habitat effect α_i, a site effect β_j, and a residual error ϵ_{ij}. *R* and other modern statistical packages specify this with a formula notation proposed by Wilkinson & Rogers in 1973:

```
nests ~ habitat + site
```

(the mean is implicit), with the data given in a data frame

```
habitat  site  nests
field    1     13
forest   1     7
field    2     8
forest   2     3
...
```

The *R* functions `stack`, `unstack` and `reshape` convert between the two data formats.

3 Chapter 8, managing data

Important aspects are data entry, organization, storage, and documentation. Besides this chapter, a good resource on the management of research data is `http://libraries.mit.edu/guides/subjects/data-management/index.html`.

Other preliminary work includes the checking of data for gross errors and outliers, and exploratory data analysis (EDA).