

March 28, 2005

Chapter 1: Introduction

ES714 is an Environmental statistics course for students who have had at least one previous course in statistics that covers the basics of statistical principals such as probability distributions, hypothesis testing, confidence intervals, t -tests, analysis of variance and regression analysis.

From this foundation, we shall cover more advanced statistical topics that are commonly encountered in the environmental sciences. The field of statistics is vast and the statistical applications pertinent to environmental statistics is also vast. It will be impossible to cover in great detail all the topics that we shall encounter in this course. In fact, entire courses can be devoted to each of the topics we will cover this quarter. These topics include:

- Sampling Designs
- Multiple Regression topics
- Logistic Regression and Generalized Linear Models
- Time Series Analysis
- Spatial Statistics

Sampling Designs. Environmental studies require data. Often the data is obtained in field studies where the investigator must obtain samples. In order to make valid statistical inferences from the data, the data must be representative of the population from which it is obtained. This typically involves randomly sampling the units in the study in some fashion. There are a wide variety of methods available. If one does not obtain a sample in a statistically valid fashion, then the resulting data may fail to be of much use due to inadvertent selection biases. Additionally, a prudent choice of a sampling design may lead to more efficient estimation of population parameters and save time and money in collecting data.

Multiple Regression Topics. Regression is one of the most used statistical tools. The reason for its high use is that regression allows us to model complicated relations between variables. We shall quickly review the fundamentals of multiple regression and then follow that up with topics involving regression with indicator variables, polynomial regression, and nonlinear regression models.

Generalized Linear Models. Regression models can be generalized to handle response variables that are not normally distributed. The most commonly used type of a generalized linear model is *logistic regression*. In logistic regression, the response

is a Bernoulli 0 or 1 variable indicated success or failure. The estimated regression function gives the probability of success based on a covariate.

Time Series. Time series are very important in environmental studies. Any time data is collected over time (such as temperatures), the resulting data is a time series. Time series analysis allows us to study seasonal trends and overall trends in the data as well as correlations between successive measurements.

Spatial Statistics. Environmental studies often involve data collected spatially such as soil samples obtained in a large field. It is of interest in such studies to investigate if clusters exist or if the data is spatially correlated.

Some Basics of Experimental and Sampling Design

Experimental design is extremely important when embarking on any environmental study. Experimental design deals with designing experiments. Scientific results come from the analysis of data from such experiments. Here we introduce some basic terminology and ideas.

An **experiment** is when researchers control the allocation of treatments to the **experimental units**. Thus, in an experiment, the researcher has direct control over treatments received. The *experimental unit* is the smallest unit to which treatment combinations are applied. The *observational unit* is the unit upon which measurements are taken. Experimental units can be people, animals, beakers of liquids, plots of land etc. Sometimes the experimental and observational units are one in the same, and other times they differ. For example, a beaker of water may contain several organisms and the implementation of the experiment is to expose the organisms to a particular chemical. If the chemical is added to the water, then the experimental unit is the beaker. If responses are recorded for individual organisms within the beaker, then the organisms are the observational units.

In observational designs, the researchers do not have control on the allocation of treatments, but instead *observe* a population of interest. The **population** is the collection of all possible sampling units. When identifying the population, the investigator must determine to which group will the conclusions of the statistical inference be applied.

What constitutes a sampling unit may not always be well-defined. For instance, if one is performing an observational study of the emerald ash borer in Ohio, the population may be defined as all ash trees in Ohio with trees being the sampling units. Alternatively, sampling units could be defined as plots of land in this example. Sometimes the population will be hypothetical. For instance, in the emerald ash borer example, the population may be all trees infested with the borer in Ohio. If the borer has not entered Ohio yet, then the population is hypothetical. Nonetheless, experiments could be conducted by infesting trees with the borer in order to study different treatments for eradicating the borer.

If the goal of an experiment is to study the effect of some treatment, say an environmental toxin, on a population of interest, then good experimental design requires the

elimination of all other variables that may effect the outcome to the greatest extent possible. For example, if we want to study the effect of a toxin on a particular organism, then one would want to control to the greatest extent possible the effect of any other variables on the organism. If an effect is discovered when analyzing the data, the experimenter would like to be able to attribute that effect to the toxin. However, this will not be possible if there are other variables that were not controlled that could have caused the effect. This is sometimes referred to as **local control**. Experimental units should be as similar to each other as possible in order to obtain local control. If there is a high degree of variability among the experimental units, then it may be difficult to detect effects due to the increased variability. On the other hand, if high levels of variability exist among experimental units, then it may not be possible to differentiate any observed effects based on experimental conditions or differences in experimental units. If we want to observe the effect of a toxin on an organism, then we want to factor out **confounding** factors due to differences in temperature, differences in lab conditions and lab technicians, solution preparations, etc.

Blinding: Another aspect of local control is the notion of blinding. There are many examples of studies whose results came out wrong because the experimenter had reason to believe the experiment would come out in a particular way even before the experiment was even conducted. This phenomenon occurs everyday: if someone believes that something is a particular way, they find evidence that supports their view while discarding or ignoring evidence that does not support their view. Often this happens on a subconscious level. Unfortunately, scientists are not exempt from this sort of bias. For this reason, it is important for investigators to protect their research by *blinding* or **masking**. For instance, if one wants to study the effect of an environmental toxin on an organism in an experiment using different levels of toxin exposure, then the investigator should not know which experimental units received which treatment. A single-blind study is one where the subjects do not know which treatment they are receiving. This may not be very relevant in a study on fish say, but it can have a big impact in clinical trials involving humans. A double-blind study is one where neither the subjects nor the experimenter know which subjects are getting which treatments.

Controls: Another fundamental aspect of experimental design is that of a **control group**. If a study of the effect of a toxin on an organism shows an effect, how can the experimenter know the effect is more than what could occur by chance alone? A control group in this situation may be a group of organisms that are as similar as possible to the organisms that are exposed to the toxin except the control group is not subject to exposure. Therefore, if a difference is observed between the treatment group and control group that is too big to have occurred by chance alone, then it may be reasonable to attribute the difference to the toxin. In studies of anti-depressants, the control group is often a group of subjects who take a placebo. It is important that these studies be double-blind. It is well-known that there is a high placebo-response rate in studies of depression. In other words, many depressed subjects will report some improvement by simply knowing that they are taking something that is supposed to make them feel better, even if that something does not actually work. If a group of subjects receiving the actual anti-depressant treatment do improve on average, the experimenters need to determine if this improvement is greater than

what would be observed from the placebo effect alone.

Another way of incorporating controls into an experiment besides a placebo control, is to have the control group receive a well-established standard treatment which will then be compared to the treatment of interest.

When implementing an experiment there will be a variable of interest (or perhaps several variables). We may record something as simple as whether or not an organism survives, or we may record how long an organism survives. The variable of interest may be a quantitative measurement such as the size of a plant or animal (length, width, etc). The experiment will then control **factors** that are thought to effect the variable. For instance, in the toxicity study on fish, the dose of the toxin would be the factor and we may perform the experiment at several different **levels** of the factor, i.e. different doses. Often times an experiment will incorporate other factors as well, for instance water temperature, pH level of the water, etc. A **blocking** factor is one whose effect on the variable is not of interest but is known to effect the outcome and therefore it must be controlled for. A blocking factor then is used to factor out extraneous sources of variability in order that a more focused study of the effect of the treatment can be made. For instance, suppose interest lies in accessing the effect of pesticide runoff on the thickness of turtle eggshells. Suppose also that a turtle's diet has a very big effect on eggshell thickness. If a study is undertaken to study the effect of the pesticide on the eggshell thickness but there is no control over the diet of the turtles, then it could be the case that most of the variability in the observed eggshell thicknesses is due to the difference in diets. If this variability is too great, then it may not be possible to detect a difference in mean eggshell thicknesses due to the pesticide. However, if diet is used as a blocking factor, then this source of variability can be controlled leading to a more powerful statistical analysis, i.e. an analysis that is more likely to find an effect due to pesticide if such an effect actually exists.

Randomization: We have discussed issues related to factoring out variability due to factors that are not of concern in an experiment. It must be noted however that there will always be some intrinsic variability between sampling units. When studying animals, no two animals are exactly the same, even after considering factors such as age, size, etc. In order to balance out uncontrolled systematic effects in an experiment, units are assigned to treatment combinations using **randomization**. Randomization is the process of assigning treatments to experimental units at random. If there exist systematic effects that we have not (or could not) control for, then the use of randomization will hopefully wash out differences between treatments that are not due to the treatments themselves. Suppose in the fish toxicity example that some fish have a genetic predisposition to be immune to the harmful effects of the toxin and the experimenters are unaware of this genetic effect. If the fish are randomized to different treatments based on varying doses of the toxin, then randomization will help make it possible with a high degree of probability that the fish with and without this genetic predisposition will balance out somewhat between the treatments. Random number generators can be used to implement randomization in practice. Most statistical software packages have random number generators.

Non-random allocation of treatment to experimental units once again puts the ex-

periment at risk of bias, even when it is unintentional.

Finally, remember the most sophisticated statistical analysis available cannot salvage a poorly designed or implemented experiment.