Chapter 2

Collecting Data

Chapter Objectives

- Recognize bias in nonrandom sampling schemes.
- Distinguish between the two fundamental random sampling schemes, simple random sampling and systematic sampling.
- Recognize the two more involved random sampling schemes, stratified sampling and multi-stage sampling.
- State the advantages and disadvantages of each of the four random sampling schemes.
- Distinguish between replication and pseudoreplication.
- Distinguish between snapshot and trajectory studies.
- Distinguish between observational studies and experiments.
- Explain how confounding variables limit the use of observational studies for establishing cause and effect.
- List the three principles of experimental design.
- Explain why impact assessment studies are seldom experiments.
- Know how to carry out before-after, control-impact, and before-after-control-impact studies.

Key Takeaways

- Nonrandom sampling, in particular judgmental sampling, can be biased.
- Simple random sampling and systematic random sampling are unbiased.
- Pseudoreplication results when observations are made too close together spatially or temporally.
- Snapshot studies are ones in which observations are made over a spatial region at a single time point. Trajectory studies are ones in which they're made over several time points at a single spatial location.
- Observational studies can't establish cause and effect because of confounding variables.
- Randomized, controlled experiments can establish cause and effect.
- Impact assessment studies aren't randomized experiments, so using them to establish cause and effect is difficult.
- Both before-after and control-impact studies are vulnerable to confounding variables.
- Before-after-control-impact study designs help to alleviate problems with confounding variables.

2.1 Introduction

Data collection methods for statistical inference usually fall into one of two broad categories, *random sampling* and *designed experiments*. Principles from these two types of data collection methods will provide the framework for designing environmental studies. Other issues related to designing environmental and covered in this chapter include:

1. Replication versus pseudoreplication.
2. Snapshot versus trajectory studies.
3. Observational studies versus experiments.
4. Impact assessment study designs.

In environmental studies, the population of interest is often a spatial region or a period of time, and these scenarios will be the main focus of this chapter. In these cases, sampling involves selecting spatial locations or points in time at which to measure a variable. More elaborate studies in which the variable is measured over space *and* time can use the sampling methods described herein.

2.2 Sampling

2.2.1 Introduction

Sampling procedures that use chance to select from a population's individuals are called ***random sampling*** schemes. Ones that don't use chance are called ***nonrandom*** schemes. We'll look at a few of the most important sampling schemes:

1. Judgmental sampling (nonrandom)
2. Simple random sampling
3. Systematic random sampling
4. Stratified random sampling
5. Two-stage random sampling

A sampling scheme is ***biased*** if it has a systematic tendency to favor certain sub-classes of the population, leading to their *over-representation* in the sample. It's ***unbiased*** if there's no such systematic tendency. Samples selected using unbiased sampling schemes *tend* to be *representative* of the population, but there's no guarantee that any *particular* sample will be representative, particularly if the sample size is small. Small samples can be unrepresentative just by chance even if they're selected using an unbiased scheme.

An example of a biased sampling scheme is ***convenience sampling***, whereby individuals are selected from the population because selecting them is ”convenient”. Selecting sample locations that are close to a road or trail simply because they're easy to access is an example of convenience sampling, and those locations might differ systematically from the rest of the study region.

For drawing inferences about a population, random sampling is preferred because it's unbiased (if done appropriately), and therefore tends to produce representative samples. Furthermore, as we'll see in later chapters, it allows us to make statements about the uncertainty of our inferences using the language of probability.

Of the five sampling schemes listed above, only the first one, judgmental sampling, is nonrandom and therefore can be biased. It's included mainly for expository purposes to demonstrate its potential for bias. The next two, simple random sampling and systematic random sampling, are always unbiased. The last two, stratified random sampling and two-stage random sampling, can be biased if not done carefully, and therefore should only be used with caution.

2.2.2 Replication Versus Pseudoreplication

Replication refers to measuring a variable on *several* individuals (e.g. several soil or water specimens, several quadrats, several time points, etc.). Each additional ***replicate*** increases the sample size by one. True replication requires that each replicate contributes a completely new bit of information about the variable to the sample data.

Measurements made on specimens gathered *too close together spatially* or at *insufficiently separated time points* are *redundant* in the sense that the two values of the variable will tend to be almost the same. Two nitrate measurements made in soil just a few feet apart, for example, will be almost identical, as will two carbon dioxide measurements made at a monitoring station 30 seconds apart. Such measurements are called ***pseudoreplicates***, a term first coined in [10]. In effect, each pseudoreplicate is a duplicate of a measurement already made, and therefore *doesn't* contribute any new information to the sample and *doesn't* increase the sample size.

When measuring a variable over a spatial region or period of time, it's advisable to leave enough separation between sample locations or time points that each measurement contributes a new, independent piece of information to the sample. Not only will the resulting data be more informative, but statistical inference procedures generally require it.

2.2.3 Snapshot Versus Trajectory Studies

When designing environmental studies, one important question is whether replication should be done in space or in time (or both). The terms *snapshot* and *trajectory* are sometimes used to distinguish between the two choices [7]. A ***snapshot study*** is one in which replication is done in space, that is, a variable is measured at several spatial locations (all at roughly the same time). A ***trajectory study*** is one in which replication is done in time, so that the variable is measured at several time points (all at the same location).

If the goal of the study is to make generalizations over a spatial region (the population), a snapshot study is appropriate. But a snapshot study is a sample of size one in time, so it's not amenable to generalizing over time. If the study goal is to make generalizations over a period of time, a trajectory study is appropriate. But a trajectory study is a sample of size one in space, so it's not amenable to generalizing over a spatial region. The two types of studies (snapshot and trajectory) can be combined in a single study – replication can be done over space *and* time.

2.2.4 Judgmental Sampling

Judgmental sampling refers to a type of *nonrandom* sampling in which a small number of the population's individuals are ”hand-picked” to be included in the sample because they're considered to be ”representative” of the population.

Example 2.1: Judgmental Sampling

To estimate the average daily traffic on a certain stretch of road, we could pick a few days that we consider to be ”typical”, count the number of cars on the road for each of those days, and use those counts to estimate the true daily average.

The main drawback of judgmental sampling is that our subjective judgment about which individuals are ”representative” of the population can be wrong. If it is, we (inadvertently) introduce bias into the selection process. For example, we might consider weekdays to be more ”typical” than weekends for estimating average daily traffic. But only including weekdays in our sample would bias the results toward weekday traffic patterns.

In addition, because it's nonrandom, we can't use the language of probability to make statements about the uncertainty of our inferences when judgmental sampling is used.

For these reasons, judgmental sampling is seldom recommended. The one exception is when resources allow for only a very small sample to be taken. In this case, judgmental sampling can ensure that we don't end up with atypical individuals in the sample (such as holidays in Example 2.1) that would distort the study results. The drilling of ice cores in Antarctica, for example, is costly and time consuming, so ice core sample sizes are generally very small, and selecting suitable locations judgmentally (rather than randomly) can help ensure that costly missteps are avoided.

2.2.5 Simple Random Sampling

Sampling people by writing their names on slips of paper and "drawing names from a hat" is an example of *simple random sampling.* Formally, a ***simple random sample*** of ***size n*** is a sample that's selected in such a way that every possible size-n subset of a population's individuals has the same chance of being the subset selected. Thus, for example, each possible group of, say, $n = 10$ names has the same chance of being the group of names "drawn from the hat."

Taking a simple random sample of spatial locations from a study region can be thought of as "throwing n darts" at a map of the region, where each dart is equally likely to land anywhere on the map, regardless of where the previously thrown darts have landed. In practice, the sample would be taken using a computer's random number generator.

Example 2.2: Simple Random Sampling

Suppose we want to estimate the average density of sawgrass plants (plants per 1 m^2) in the Everglades region of Southern Florida, depicted in Fig. 2.1.

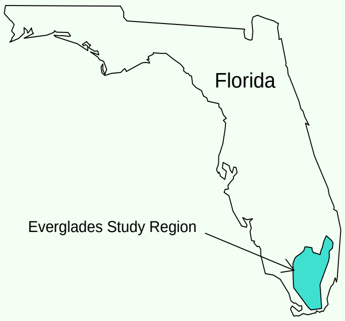


Figure 2.1: Florida and the Everglades study region.

We could take a *simple random sample* of locations, then count the number of sawgrass stems in a 1 m^2 *quadrat* at each location.

One way to take the sample of quadrat locations would be to use a computer random number generator to select their spatial coordinates (e.g. latitudes and longitudes) from within the Everglades study region.

Another way, akin to "drawing names from a hat" but a bit unwieldy, would be to partition the Everglades into 1 m^2 quadrats, assign each one a number, and then use a computer to randomly sample from the list of quadrat numbers. The Everglades covers roughly two million acres, though, which translates to 8,093,712,845 quadrats, so there'd be a lot to choose from!

A computer-generated simple random sample of $n = 50$ locations in the Everglades (using the first method above) is shown below.

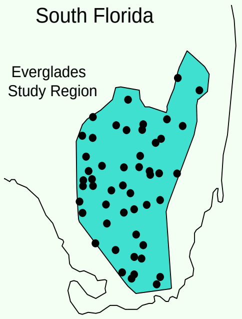


Figure 2.2: A simple random sample of locations of $n = 50$ quadrats from the Everglades study region.

An advantage of simple random sampling is that it's always *unbiased*. Another is that commonly used statistical inference procedures were developed for use with simple random samples, so those procedures would be available for use with the data. A disadvantage, especially when the sample size is small, is that there's no guarantee that the selected locations will be evenly spread over the study region – whether this happens or not is left to chance.

2.2.6 Systematic Random Sampling

In Fig. 2.2, the sampled locations are fairly uniformly (evenly) distributed over the study region. With simple random sampling, there's always a possibility that we'll end up with an "unlucky" set of sample points that lie mostly in one part of the study region or another. For example, in Fig. 2.2, we might have ended up with most of the points lying in the northern part of the Everglades.

To eliminate this possibility, we could instead take a ***systematic random sample***, whereby the sampled locations are taken to lie on a regular grid. To incorporate chance into the sample selection process, the grid is initialized at a single, randomly selected, starting location. As long as the initial grid point is randomly chosen, systematic sampling will be unbiased. Note that a smaller spacing between the grid points leads to a larger sample size, but taking them too close together will lead to pseudoreplication.

Example 2.3: Systematic Random Sampling

To take a systematic random sample of 1 m^2 quadrats from the Everglades, we'd first randomly select a single location in the region, and then take the other locations to lie on a regular grid that includes the original one, as shown in Fig. 2.3.

For the desired sample size of $n = 50$ quadrats, since the Everglades regions is two million acres, we needed a grid point every $2,000,000/50 = 40,000$ acres. The actual sample size ended up being $n = 51$ here due to the irregular shape of the Everglades.

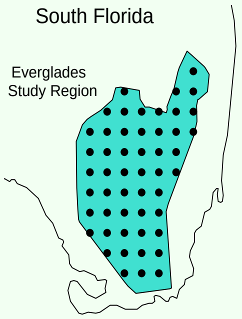


Figure 2.3: Systematic random sample of $n = 51$ quadrat locations from the Everglades study region.

Systematic sampling is the preferred method for sampling spatial locations when the goal is to *map* the spatial distribution of a variable. A number of statistical methods have been developed for analyzing and mapping data collected on a grid. Some of these are covered in Chapter 17.

Systematic random sampling is also the preferred method for sampling time points. In this case, a systematic sample is one for which the sample of time points lie at regular intervals. Data collected at regular time intervals are sometimes called ***time series data***. A wealth of statistical methods have been developed for analyzing time series data. Some of these are covered in Chapter 16. One advantage to collecting data at regular time intervals is that the data are guaranteed to cover the entire time period being studied. The length of the interval separating time points is determined by the desired sample size – a smaller the interval between time points leads to a larger sample size.

Example 2.4: Systematic Random Sampling

For estimating the average daily traffic on a stretch of road, as in Example 2.1, we can avoid having all of our sample of days end up being in summer by taking a systematic random sample of, say, $n = 12$ days over the course of a year.

The study period is 365 days, so we'd sample every 30th day (since $365/12 \approx 30$), that is, about once a month. This would guarantee that our sample would span all four seasons.

To incorporate chance into the sample selection process, we'd start by randomly selecting a single day from among Jan. 1-Jan. 30, and then take every 30th day thereafter.

When choosing the length of the interval separating time points, we need to pay attention to whether there are any repeating, ***cyclical*** patterns of variation in the variable being measured, such as diurnal, weekly, or seasonal patterns. If there are, and the interval coincides with the ***period*** of that cyclical pattern, the sample won't be representative. The next example illustrates.

Example 2.5: Beware of Cyclical Patterns

To estimate the average daily traffic on a stretch of road, a systematic random sample consisting of every seventh day of the year might contain only Sundays, which clearly wouldn't be representative of all the days of the year.

Fig. 2.4 shows a cyclical pattern of variation in a pollutant concentration with a daily period of repeat. The points labeled "A" are pollutant values at time points one day apart, which coincides with the period of the cycle. If the goal is to estimate the average pollutant concentration over the span of the study, represented by the horizontal line in the graph, it's clear that the "A" points would lead to an underestimate. In general, estimates of averages aren't trustworthy if the interval between sampled time points coincides with the period of a cyclical pattern.

The points labeled "B" in Fig. 2.4 are pollutant values for which the sampled points are half a day apart. In this case, the above-average pollutant values in the sample are exactly offset by the below-average ones, leading to an accurate estimate of the time span's true average. In general, estimates of averages will be accurate if the interval between sampled time points is any odd multiple of one half of the period of the cycle.

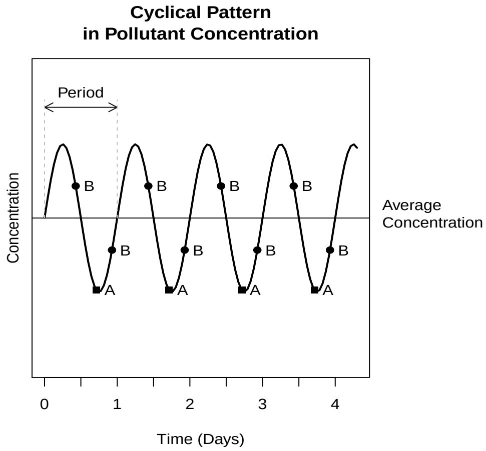


Figure 2.4: Graph of a pollutant concentration following a cyclical (sinusoidal) pattern over time.

Another way to ensure an accurate estimate of the average is to choose the interval between sampled time points such that every portion of the cycle is represented equally in the sample. For example, if the cycle is weekly, all days of the week should be represented in the sample equally.

When the period of a cyclical pattern isn't known, one suggestion is to partition the study's time span into smaller blocks of time, and then take a separate systematic random sample of time points from each block, starting from a randomly selected initial time point in each block [6].

Although less common, cyclical patterns can occur in space too. Repeating patterns brought about by agricultural practices (e.g. growing row crops) is one example. Grid patterns generated by city blocks is another.

In closing, it should be mentioned that systematic random sampling is suitable for making inferences about the population, as long as there are no cyclical patterns in the variable being measured. More specifically, systematic random sampling is unbiased, and most of the statistical inference procedures that have been developed for use with simple random samples are also (at least approximately) valid with systematic samples.

2.2.7 Stratified Random Sampling

Populations are sometimes composed of distinct groups defined by the categories of a categorical variable. For example, a study region might be composed of distinct sub-regions based on soil type, land use type, habitat type, or topography. In the context of sampling, the groups are referred to as ***strata*** (each of which is a ***stratum***).

As mentioned, a simple random sample of spatial locations might end up all being bunched up in one part of the study region by chance. When this happens, one stratum or another may be overrepresented in the sample and others underrepresented. Worse yet, some strata might not be represented at all. *Stratified random sampling* guarantees that all strata will be represented in the sample, and if done carefully, that none will be overrepresented or underrepresented.

A ***stratified random sample*** is taken by dividing the population into strata, taking a separate simple random sample from each stratum, and then combining these separate samples together.

Example 2.6: Stratified Random Sampling

The U.S. Environmental Protection Agency divides the Everglades study region of Fig. 2.1 into the following six sub-regions that differ with respect to their ecosystem characteristics [1]:

Big Cypress National Preserve (BICY)
Everglades National Park (ENP)
Water Conservation Area 1 (WCA1)
Water Conservation Area 2 (WCA2)
Water Conservation Area 3 (WCA3)
Rotenberger Wildlife Area/Everglades Agricultural Area (ROT-EA)

These sub-regions are depicted in the figure below. We'll consider these to be strata for the purpose of illustrating stratified random sampling.

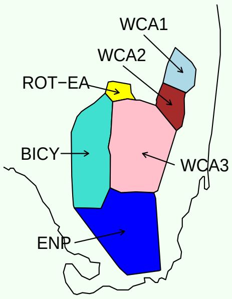


Figure 2.5: Six strata in the Everglades study region.

A careful examination of Fig. 2.2 in Example 2.2 reveals that the simple random sample produced only a single observation from ROT-EA, and only two from WCA1. With so few observations it is difficult to draw any definitive conclusions about these two strata. Taking a stratified random sample would allow us to choose how many observations we end up with in each stratum.

Suppose for example that we want our sample to include $n_1 = 10$ observations from BICY, $n_2 = 11$ from ENP, $n_3 = 7$ from WCA1, $n_4 = 7$ from WCA2, $n_5 = 11$ from WCA3, and $n_6 = 5$ from the ROT-EA. Fig. 2.6 below shows a stratified random sample obtained by taking independent simple random samples of these sizes from the six strata.

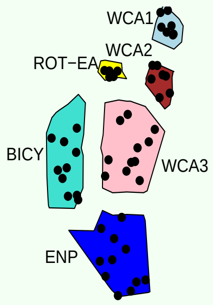


Figure 2.6: Stratified random sample from the Everglades study region. The strata are separated in the figure to emphasize the fact that independent simple random samples have been drawn separately from the six strata.

The advantages of stratified sampling are that it guarantees that all strata will be represented in the sample, and it lets us decide how many sample points to select from each stratum. A disadvantage, though, is that it can be biased if the stratum-specific sample sizes aren't chosen carefully. For example, notice in Fig. 2.6 that there are more sample points per unit area in the smallest strata (ROT-EA, WCA1, and WCA2), so those strata will be *overrepresented* in the overall combined sample, giving biased results. To avoid such bias, we could instead use the areas of the strata to determine the stratum-specific sample sizes. More precisely, we could make each stratum-specific sample size *proportional* to the area of its stratum. This is known as ***proportional allocation***.

Example 2.7: Proportional Allocation

The areas of the six strata in Fig. 2.6 are approximately:

BICY = 466,000 acres
ENP = 965,000 acres
WCA1 = 90,000 acres
WCA2 = 86,000 acres
WCA3 = 374,000 acres
ROT-EA = 19,000 acres

and the total area is two million acres. To take a stratified random sample of, say, $n = 50$ locations using proportional allocation, since BICY represents about 23% of the study region $(466,000/2,000,000 = 0.23)$, 23% of our 50 sample locations should come from this stratum, so its sample size is

$$n_1 = (0.23)(50) \approx 12.$$

Likewise, because ENP represents 48% of the study region $(965,000/2,000,000 = 0.48)$, we want 48% of our sample to come from this stratum, so its sample size is

$$n_2 = (0.48)(50) = 24.$$

The other stratum-specific sample sizes are found in a similar manner giving, in the order listed above,

$$\begin{aligned} n_3 &= (0.05)(50) \approx 2 \\ n_4 &= (0.04)(50) \approx 2 \\ n_5 &= (0.19)(50) \approx 9 \\ n_6 &= (0.01)(50) \approx 1. \end{aligned}$$

Note that the overall sample size is $n = 12 + 24 + 2 + 2 + 9 + 1 = 50$ as desired.

In general, suppose we want to take a stratified random sample of overall size n from a study region whose total area is A. Suppose also that there are L strata whose areas are $A_1, A_2, \ldots, A_L$. The stratum-specific sample sizes $n_1, n_2, \ldots, n_L$, using the *proportional allocation* method, are taken to be

$$n_i = \left(\frac{A_i}{A}\right) n,$$

rounded to the nearest integer, for $i = 1, 2, \ldots, L$.

If instead of sampling locations from a spatial region, we're sampling individuals from a finite population of size N (such as a city with N people), and there are L strata of sizes $N_1, N_2, \ldots, N_L$ (representing, say, L socioeconomic classes), the stratum-specific sample sizes $n_1, n_2, \ldots, n_L$, are given by

$$n_i = \left(\frac{N_i}{N}\right) n,$$

rounded to the nearest integer.

Caution should be used when drawing inferences about a population using a stratified random sample. In particular, most of the formal statistical inference procedures that were developed for use with simple random samples aren't appropriate for use with stratified samples unless we're drawing conclusions separately about each stratum using the stratum-specific samples. If you must use data from a stratified sample draw inferences about a population as a whole, [6], [13] or [14] provide guidance as to which procedures are appropriate.

2.2.8 Cluster Sampling and Two-Stage Random Sampling

Cluster sampling and *two-stage random sampling* and are sometimes used when the population is composed of ***clusters*** (groups) of individuals. For example, the U.S. population is composed of clusters of people – cities and towns. Likewise, some animal species form colonies and some plant species grow in patches. The idea is to first select a simple random sample of *clusters* (e.g. cities), and then, from each of the selected clusters, either take *all* the individuals in that cluster (e.g. the entire city) to be in the sample or take a *simple random sample* of individuals from the cluster (city). The first method, where *all* the individuals in the selected clusters are included in the sample, is called ***cluster sampling***. The second, where only a *sample* of individuals from each cluster are included, is called ***two-stage sampling***.

An advantage of cluster and two-stage sampling is that they can reduce the time, effort, and costs required to collect the data. For example, consider sampling U.S. residents for face-to-face interviews. If a simple random sample of, say, $n = 1,000$ U.S. residents was taken, it's likely that they'd be spread all over the country, and the interviewer would have to spend unreasonable amounts of time and money flying from city to city to conduct the interviews. To reduce those costs, a two-stage sample could be taken by first selecting a simple random sample of, say, $m = 10$ cities and then randomly selecting $\tilde{n} = 100$ residents from each city. In this way, the interviewer would still interview $1,000$ people but would only need to fly to 10 cities.

We'll focus on *two-stage* sampling. In this context, the clusters are called ***first-stage sampling units*** (or ***FSU***s) and the individuals that constitute the clusters are called ***second-stage sampling units*** (or ***SSU***s). Thus two-stage sampling is carried out by, in the first stage, taking a simple random sample of, say, $\boldsymbol{m}$ FSUs, and then, in the second stage, taking a separate simple random sample of SSUs from each of the FSUs that were selected in the first stage. Note that SSUs are individuals of the population, so we end up with a sample of individuals. The process of drawing a second-stage sample from a FSU is sometimes called ***subsampling***. If $\boldsymbol{\tilde{n}}$ SSUs are drawn from each of the m FSUs sampled in the first stage, then the total sample size $\boldsymbol{n}$ will be $n = m\tilde{n}$.

Here are some other examples of two-stage sampling.

Example 2.8: Two-Stage Random Sampling

Forestry researchers are often interested in the total biomass of a forest, for example to help determine the forest's capacity for carbon dioxide sequestration. The total biomass includes tree branches, tree stems, and other foliage in the forest. To estimate the total biomass of just the tree branches, a two-stage sample could be taken by first drawing a simple random sample of m trees, and then, from each tree, taking a simple random sample of $\tilde{n}$ branches.

In this example, the population consists of all the *branches* in the forest, that is, the branches are the individuals. The trees are the FSUs and the branches the SSUs.

Example 2.9: Two-Stage Random Sampling

Two-stage sampling was used in a study of the zinc (Zn) and calcium (Ca) concentrations in soil on a research field in Slovenia [12].

In the first stage, the field was partitioned into subplots, and a simple random sample of m subplots was selected. In the second stage, from each of the selected subplots, a simple random sample of $\tilde{n}$ soil specimens was selected and the Zn and Ca concentrations measured in each specimen.

In this example, the population is the entire field, and *soil specimens* are the individuals. The subplots are the FSUs and the soil specimens within the subplots are the SSUs.

Example 2.10: Two-Stage Random Sampling

Consider again the study to estimate the average sawgrass density in the Everglades (Examples 2.2, 2.3 and 2.6). A two-stage sample could be drawn by first taking a simple random sample of, say, $m = 4$ sites, and then taking a simple random sample of, say, $\tilde{n} = 3$ quadrat locations at each site. Fig. 2.7 shows the result of one such two-stage sample.

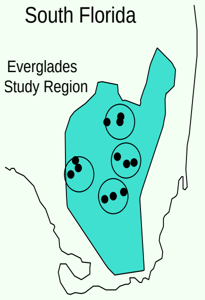


Figure 2.7: Two-stage random sample from the Everglades study region with $m = 4$ FSUs (the large circular plots representing sites) sampled in the first stage, and $\tilde{n} = 3$ SSUs (quadrat locations represented by points) sampled from each FSU in the second stage. The total sample size is $n = m\tilde{n} = 12$ quadrats.

In this example, population (Everglades) consists of all possible quadrat locations (the individuals). Sites are the FSUs, and quadrats within the sites are the SSUs.

Two-stage sampling using equal second-stage sample sizes ($\tilde{n}$), as just described, is unbiased as long as the FSUs in the population are all about the same size (that is, contain equal numbers of SSUs).

In practice, the sizes of the second-stage samples drawn from the FSUs don't all have to be the same. In fact, when they are the same, the results can be *biased* if the FSUs differ considerably in size. In particular, individuals in smaller FSUs will tend to be *overrepresented* in the sample, and if those individuals differ in important ways from individuals in larger FSUs, the results will be biased.

The problem is related to the fact that even though small and large FSUs represent different fractions of the population, they have the same chance of being selected in the first-stage simple random sample of FSUs.

To illustrate, consider a two-stage sample of $n = 1,000$ U.S. residents, where the first stage consists of sampling $m = 10$ U.S. cities (the FSUs), and the second stage consists of sampling $\tilde{n} = 100$ people (the SSUs) from each of the 10 selected cities. Then a small city like Lafayette, Indiana is just as likely to be among the 10 selected cities as a big one like New York. This means that because the second-stage sample sizes would be 100 for both cities, we're just as likely to end up with 100 Lafayette residents in our sample as we are 100 New Yorkers, even though Lafayette residents make up a much smaller fraction of the U.S. population than New York residents do. In other words, residents of small cities like Lafayette would tend to be *overrepresented* in the sample, and residents of large cities like New York would tend to be *underrepresented.*

One way to get around this bias problem is to use unequal sample sizes for the second-stage samples, with larger samples being drawn from the larger FSUs and smaller ones drawn from the smaller FSUs. More precisely, two-stage sampling will be unbiased if we use second-stage sample sizes that are *proportional* to the sizes of their corresponding FSUs. In other words, we still use simple random sampling to select the FSUs in the first stage and also to select the SSUs from each FSU in the second stage, but in the second stage we sample a *fixed fraction* of each FSU's SSUs. For example, instead of choosing 100 residents from Lafayette and 100 from New York, we might choose one person per 100,000 residents from each city so that our second-stage sample sizes are *proportional* to the sizes of the cities.

Note that *cluster sampling* is a special case of two-stage sampling using second-stage sample sizes that are proportional to the FSU sizes, but in this case *all* 100% of each FSU's SSUs are selected in the second-stage samples. Thus cluster sampling is an unbiased sampling method.

Another way to avoid the bias problem that arises when the FSUs are different sizes is to continue to use equal second-stage sample sizes (e.g. 100 residents from Lafayette and 100 from New York), but use what's called ***probability proportional to size sampling*** to select the FSUs in the first-stage rather than simple random sampling. The idea is to give larger FSUs (like New York) a higher probability of being selected than smaller ones (like Lafayette). In particular, each FSU's probability of being selected in the first state is made to be proportional to the FSU's size. If this is done, and equal sample sizes are used for the second-stage samples, then the sampling scheme is unbiased. More information about this method can be found in [6], [13] and [14].

To summarize, two-stage sampling is *unbiased* if any of the following three sets of conditions is met:

1. The sizes of the FSUs that compose the population are all roughly the same (contain roughly the same number of SSUs), simple random sampling is used to select the first-stage sample of FSUs and to select each of the second-stage samples of SSUs, and the second-stage sample sizes ($\tilde{n}$) are all equal.
2. The sizes of the FSUs that compose the population are very different from each other (contain very different numbers of SSUs), simple random sampling is used to select the first-stage sample of FSUs and each of the second-stage samples of SSUs, and second-stage sample sizes are proportional to the FSU sizes.
3. The sizes of the FSUs that compose the population are very different from each other (contain very different numbers of SSUs), probability proportional to size sampling is used to select the first-stage sample of FSUs, simple random sampling is used to select each of the second-stage samples of SSUs, and the second-stage sample sizes ($\tilde{n}$) are all equal.

Cluster sampling satisfies the second set of conditions above, so it's an unbiased sampling method.

As a final cautionary note, care should be taken when using a cluster sample or two-stage sample to draw inferences about a population. Most of the statistical inference procedures that were developed for use with simple random samples aren't appropriate for use with cluster and two-stage samples. If you must use data from a cluster or two-stage sample to draw inferences about a population, [6], [13] or [14] provide guidance as to which procedures are appropriate.

2.2.9 Other Topics in Environmental Sampling

Composite Sampling

Composite sampling is based on the principle that larger sample sizes generally give more reliable results. It's used to improve improve the reliability of results, *without having to analyze more specimens*, when laboratory analysis of soil, water, or biological specimens is expensive or time consuming and resources only allow for a small number of specimens to be analyzed. It's done by physically mixing several randomly selected specimens together to form a single, so-called ***composite specimen*** for lab analysis.

As an example, fish collected from the Columbia River Basin in the northwestern U.S. were composited using a commercial meat grinder and analyzed for heavy metals in a study of fish contamination in [9].

The improved reliability is due to the fact that a *composite specimen* is a kind of *average* of the specimens it's composed of, and averages tend to be more reliable than single observations.

A drawback to compositing specimens is that information about specimen-to-specimen variation is lost. For example, if soil specimens are collected from different locations within a study region, then upon compositing them we lose all information about how much the contaminant (or other variable of interest) fluctuates spatially from one location to the next.

Quadrat Sampling and Transect Sampling

Quadrat sampling and *transect sampling* refer to methods of defining a population's individuals as opposed to methods of selecting the samples.

We've seen (in Example 2.2) that a ***quadrat*** is a small (e.g. 1 m^2), usually square, plot of land. The population consists of all such quadrats within a study region, and thus each quadrat is an individual of the population. ***Quadrat sampling*** refers to sampling from the population of quadrats. It's used in ecological studies to estimate plant or small animal (e.g. insect) population sizes or densities – the plants or animals are counted within each sampled quadrat, so the resulting data are discrete *counts* which are then generalized to the study region. Usually a quadrat-sized square frame made of wood or pvc pipe is used to demarcate each sampled quadrat's boundary prior to making the counts. The sample of quadrat locations is usually selected by simple random sampling or systematic random sampling, but any of the sampling schemes described in Sections 2.2.4-2.2.8 may be used.

A ***transect*** is a line (or strip) (e.g. 100 m long) within a study region and along which a variable is measured. The population (study region) consists of all such transects within the region, so each transect is an individual. ***Transect sampling*** involves selecting a sample of *transects* and measuring a variable along each one. It's is commonly used in ecological studies of plant or animal populations, whereby a person walks along each transect and counts the number of plants or animals seen. Thus, as for quadrat sampling, the resulting data are discrete *counts.* Transect lengths are typically around 100 m but can be up to 1 km or longer for bird surveys. Sometimes the terms ***line transect*** and ***strip transect*** (or ***belt transect***) are used. The first is used when the observer counts plants or animals seen at *any distance* from the line, and the second when he or she only counts those plants or animals that lie within a strip extending a few meters on either side of the line. When line transect counts are made, plants and animals that are closer to the line are easier to see, and therefore more likely to be counted, than those lying farther from it. ***Distance sampling*** refers to measuring the distances from the line at which the plants or animals are seen, for example using a laser rangefinder, and then adjusting the counts to correct for under-counting of plants or animals farther from the line.

To select the sample of transects, first a sample of spatial *locations* is selected, most often using simple random sampling or systematic random sampling. Then the transect *orientations* are either taken to be parallel to each other or determined randomly for each transect, for example by spinning a board-game spinner. Commonly, transects are taken to be parallel and at locations on a grid, giving a systematic random sample of transects.

For large study regions, transect sampling can be done by linear aerial surveys. As an example, aircraft flying along east-west transects spaced 7.5 km apart were used to count muskoxen in Greenland in [2].

Example 2.11: Transect Sampling

Suppose we want to estimate the species richness (number of different species) of plants in the Everglades. One way to do it would be to select a random sample of transects, and then count the number of species seen as we travel along each one. Fig. 2.8 below shows an example of a simple random sample of $n = 5$ (very long) transects with randomly determined orientations.

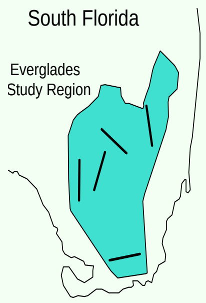


Figure 2.8: A simple random sample of of $n = 5$ transects in the Everglades study region.

Sampling for Hot Spots

Sampling for hot spots refers to using random sampling to try to locate one or more ***hot spots***, or small, concentrated areas of very high pollutant levels such as buried containers of toxic waste.

Example 2.12: Hot Spot Detection

A container of toxic waste is presumed to have been buried somewhere within a cleanup site, but its exact location is unknown. In an attempt to locate it, the ground will be penetrated at a systematic random sample of locations within the cleanup site, as shown below.

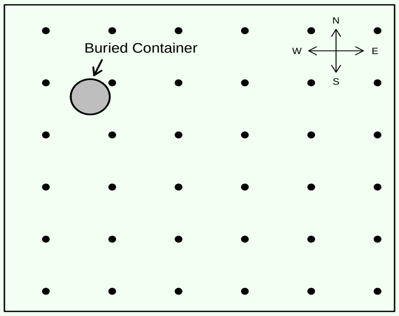


Figure 2.9: Buried container of toxic waste (shaded circle) and systematic random sample of dig locations (dots) within a cleanup site (rectangular area).

When a systematic random sample of locations is used to test for the presence of a hot spot, as in the last example, the chance that one of the sample points will hit the hot spot depends on how big of an area the hot spot spans and how close together the sample points are. For a given hot spot area and spacing between sample points, the chance that one of the points in the sample will hit the hot spot can be obtained from tabulated values given in [6]. Alternatively, the tabulated values can be used to determine how close together the sample points would need to be in order to attain some desired high probability of hitting a hot spot with a given area.

2.3 Designed Studies

2.3.1 Introduction

Often research involves investigating the *relationship* between two variables, one of which is designated the ***explanatory variable*** and the other the ***response variable***. As the names imply, the explanatory variable is presumed to help *explain* variation in values of the *response*.

Example 2.13: Relationships Between Variables

Studies have shown that there's a relationship between ozone levels near large cities and mortality rates in those cities – mortality rates, especially among the elderly, tend to be higher when ozone levels are high.

In such studies, ozone level is the *explanatory variable*, and mortality rate is the *response*.

Example 2.14: Relationships Between Variables

In an *impact assessment study* of the change in a lake's aquatic invertebrate population due to construction of a nuclear power plant on the lake's edge, the period (before or after construction) is the *explanatory variable* and the invertebrate population size is the *response*.

Example 2.15: Relationships Between Variables

A study was carried out to determine the effect of reducing plant species diversity on the overall above ground biomass [8]. In this study, plant diversity is the *explanatory variable* and the above ground biomass is the *response*.

Example 2.16: Relationships Between Variables

A study was carried out to determine if exposure to copper pollution has any effect on the reproductive capabilities of earthworms [17]. Earthworms were exposed to soil containing different concentrations of copper, and their cocoon production was measured under each concentration.

In this study, copper concentration is the *explanatory variable* and cocoon production is the *response*.

2.3.2 Observational Studies Versus Experiments

Studies to investigate the relationship between two variables are of two types: *observational studies* and *experiments*. They differ in terms of:

1. How the data are collected.
2. The conclusions that can be drawn from the data.

In ***observational studies***, the explanatory and response variables are merely *observed* (measured) on individuals, and *no treatments are imposed* on those individuals. In other words, no attempt is made to induce changes in the response by manipulating the value of the explanatory variable. The study of ozone levels and mortality rates in Example 2.13 is an example of an observational study.

In ***experiments***, *treatments are imposed* on individuals in a deliberate attempt to induce a change in their responses. Imposing treatments means manipulating the value of the explanatory variable, which is sometimes called the ***factor*** in the context of experiments. The different levels of the factor are the ***treatments*** that are imposed. The study in which earthworms were exposed to different levels of copper in Example 2.16 is an example of an experiment. The *factor* is copper concentration and the *treatments* are the different concentrations imposed on the worms in the study. Individuals to which the treatments in an experiment are applied, such as earthworms, are called ***experimental units***.

The way the data are collected (observing individuals versus imposing treatments on them) has implications for the types of conclusions that can be drawn from the data. Most importantly, observational studies *cannot*, by themselves, be used to establish cause-and-effect relationships between variables because of the possible presence of so-called *confounding variables.* The following example illustrates.

Example 2.17: Confounding Variables in Observational Studies

As mentioned in Example 2.13, observational studies have shown that mortality rates, especially among elderly, tend to be higher when ozone levels are high. The question then arises as to whether the high ozone levels are *causing* the deaths. This is one plausible explanation, but it's not the only one.

For example, it turns out that the chemical formation of ozone in the atmosphere requires both heat and sunlight, so heightened ozone levels occur on hotter days.

It's also known that mortality rates among the elderly tend to be high on hot days due to heat exposure.

So it may be the heat, not the ozone, that's causing deaths.

In these studies, we call temperature a *confounding variable* and say that its effect on mortality is *confounded* with the effect (if any) of ozone.

In general, a ***confounding variable*** is a variable, often "lurking" in the background of a study (that is, not recorded as part of the study), whose effect on the response variable can't be distinguished from the effect, if any, of the explanatory variable. Because of the possible presence of confounding variables, observational studies, by themselves, *can't* establish cause-and-effect relationships between variables. Establishing cause-and-effect requires conducting a well-designed experiment.

2.3.3 Principles of Designed Experiments

As mentioned at the start of this chapter, principles from both sampling and designed experiments provide the framework for designing environmental studies. Formal concepts of experimental design were introduced in agricultural studies of the 1920's. One of the key ideas was that the effect of the factor (explanatory variable) could often be distinguished from the effects of potentially confounding variables through the use of *randomization* and statistical *control.* Another important idea was that *replication* (using numerous experimental units in each treatment group) improved the *reliability* of the results.

Control

In the context of experiments, ***control*** refers to either:

1. Including in the experiment a ***control group*** that receives no treatment and to which any changes in response to a treatment can be compared. The idea is that any variable whose effect on the response is confounded with the effect of an imposed treatment should influence the control and treatment groups equally, so we can distinguish the effect of the treatment from that of the confounding variable by comparing the changes in responses for the two groups.
2. Holding variables that might influence the response *constant* across treatment and control groups. The idea here is that any variable that would otherwise affect the responses in one group but not the other will now affect both groups equally, so any observed difference in the groups' responses can't be due to that variable.

To illustrate the importance of *control* in designed experiments, in Examples 2.18 and 2.19 we'll first look at *poorly designed* experiments in which confounding is a problem. Then in Example 2.20 we'll look at improved designs in which the confounding variables are *controlled for* and the problem of confounding is eliminated.

Example 2.18: Poorly Designed Experiment

In a *poorly designed* experiment to study the effectiveness of biosolids from a sewage treatment facility for use as fertilizer to increase corn yield, a research farm is split into two fields, one containing eight plots that are treated with biosolids and the other, serving as a *control group*, containing eight plots that are left untreated, as shown in Fig. 2.10.

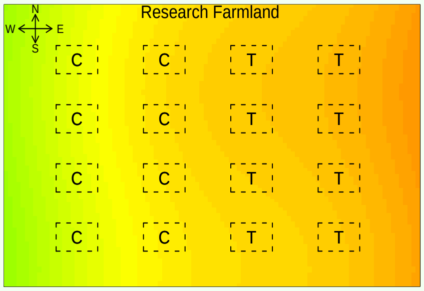


Figure 2.10: Non-random assignment of 16 research plots to treatment (T) and control (C) groups. Gradation in soil quality (depicted by the color shading) is seen to be confounded with the treatment effect.

But if the part of the field that received biosolids had better soil to begin with, as depicted by the darker (brownish) shades in Fig. 2.10, we wouldn't be able to distinguish the effect (if any) of the biosolids from the effect of soil quality – the two effects would be *confounded*.

The experiment was *poorly designed*, in part because we didn't *control* for the effect of soil condition by holding it constant across treatment and control groups.

Example 2.19: Poorly Designed Experiment

Another approach to performing the biosolids experiment would be to grow the corn two years in a row on the *same* field, the first year without using biosolids and the second year using them. In this way the soil condition is held *constant* (it's the same soil both years), so its effect is no longer confounded with that of biosolids.

However, if the weather conditions were more favorable the second year, then we wouldn't be able to distinguish the effect of the biosolids from the effect of weather conditions (the two effects would be *confounded*). So this experiment wasn't designed very well either because we didn't *control* for the effect of weather conditions by holding them constant.

Example 2.20: Controlling for Confounding Variables

To improve upon the previous two experiment designs, we could *control* for the effects of weather *and* soil, thereby eliminating them as possible reasons for a higher corn yield, using either of two designs below.

Study Design #1: We could plant the corn on two different fields two years in a row, the first year without using biosolids on either field, and the second using them on one field but not the other. This way, any increase in yield from year one to year two on the field that didn't receive any biosolids either year (the *control* field) would be due entirely to the change in weather conditions, so if the increase on the other field was even larger, we could conclude that biosolids had an effect.

Study Design #2: We could hold the weather *and* soil conditions constant across the treated and untreated plots by growing the corn with and without biosolids in a climate-controlled greenhouse using the same soil.

Randomization

Even when a control group is included in an experiment, confounding can still be a problem if the experiment doesn't involve ***randomization***, which refers to random assignment of experimental units to the treatment and control groups. The idea is to make the treatment and control groups to as similar as possible with respect to *all* so-called ***extraneous variables***, variables that affect the response but are neither manipulated as treatments nor held constant to control for them. The next example illustrates.

Example 2.21: Randomization

We saw in Fig. 2.10 that if all eight plots treated with biosolids as fertilizer are on one side of the farmland and all eight control plots on the other, the effect of soil condition might be *confounded* with the effect (if any) of the biosolids.

A better experimental design would involve *randomly* assigning the plots to the treatment and control groups. The figure below shows the result of one such *randomization.* Note that soil conditions are fairly balanced across the two groups.

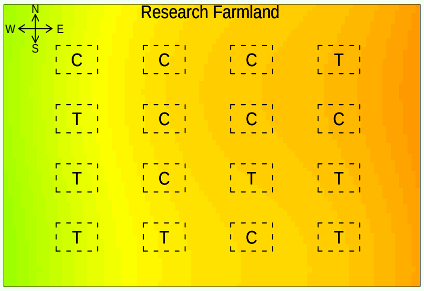


Figure 2.11: Random assignment of 16 research plots to treatment (T) and control (C) groups. The effects of soil conditions and other extraneous variables are balanced out over the two groups.

Because the two groups end up being fairly balanced with respect to soil conditions, randomization can be seen as an alternative to holding soil conditions constant across the groups. In fact, the two groups end up being fairly balanced with respect to *all other extraneous variables* that affect corn yield such as nutrients, drainage properties, sunlight, presence of weeds, and so on. So randomization alleviates the burden of us having to hold *all* these other variables constant.

When *randomization* is used, if the observed difference between yields in treated and control plots was larger than could be explained by chance, we could conclude that the biosolids had an effect. In later chapters we'll see how to use a *hypothesis test* to decide if an observed difference is larger than can be explained by chance.

Replication

The term ***replication*** is used in the context of experiments to mean using *more than one* experimental unit in each treatment group. Replication is used to improve the *reliability* of an experiment's results, meaning the results would about the same if the experiment was repeated independently on a different set of experimental units. More ***replicates*** in an experiment lead to more reliable results.

As an illustration, imagine in Example 2.21 that instead of eight plots in the control and treatment groups, we only used, say, two in each group. Then the randomization might result in two high-quality-soil plots in the treatment group and two poor-quality-soil ones in the control group. But if we repeated the experiment, the second randomization might have the soil-quality conditions for the two groups reversed. In this case, the two experiments might produce opposite results. With eight plots per group, each group is likely to end up with a range of soil qualities that would average each other out, and this would also true if the experiment was repeated. Thus the results of the two experiments would likely be similar.

Three Principles of Experimental Design

We summarize this section with the ***three principles of experimental design*** for establishing a cause-and-effect relationships:

1. **Randomization** - Randomly assign experimental units to treatment and control groups so that the groups will be (roughly) the same with respect to all extraneous variables before any treatments are applied.

2. **Control** - Compare at least two groups in the experiment, even if one of them is a control group that receives no treatment, and hold potentially confounding variables constant across the treatment and control groups, so that the effect of the treatment can be distinguished from the effects of variables that would otherwise be confounded with the treatment.
3. **Replication** - Include more than one experimental unit in the treatment and control groups so that the results of the experiment will be reliable. The more experimental units that are used, the more reliable the results will be.

2.3.4 Impact Assessment Study Designs

Environmental ***impact assessment studies*** are conducted to determine the effect on the environment of some disturbance, usually anthropogenic, such as an oil spill or the construction of a nuclear power plant. They're also used to assess the impacts of changes in management practices, such as those designed to protect sensitive habitats or critical animal species. We'll call the disturbance (or change in management practices) the ***impact event***, and any location or area potentially impacted an ***impact site***. A ***control site*** will be a location or area known to be unaffected by the impact event. Most impact assessment studies can be classified as one of three types:

- A *before-after* study, which answers the question "Did the quality of the environment at the impact site change from the period before the impact event to the period after?"
- A *control-impact* study, which answers the question "Is the quality of the environment at the impact site, after the impact event, different from that at a control site?"
- A *before-after-control-impact* study, which answers the question "Was the change in the quality of the environment at the impact site, from the period before the impact event to the period after, any different from the change at a control site over the same time span?"

  Equivalently, the question could be phrased as "Did the size of the difference between the qualities of the environments at the impact and control sites change from the period before the impact event to the period after?"

In impact assessment studies, we *don't* have the luxury of being able to randomly assign sites to impact and control groups – we can't, for example, randomly assign some lakes to have nuclear power plants built on them and others to serve as controls by remaining power plant-free. Thus impact assessment studies are observational studies, not experiments, so establishing cause and effect is a challenge. In particular, accounting for potentially confounding variables can be very difficult.

Pulse Versus Press Impact Events

Before turning to a discussion of the three types of impact assessment study designs, we note that impact events can be either of two types, *pulse* and *press* [3]. A ***pulse*** event is a relatively short-term anthropogenic or natural disturbance of the environment or an ecosystem. Pulse events are often unforeseen accidents or natural disasters. Examples include oil spills, chemical spills, nuclear plant malfunctions ("meltdowns"), forest fires, and floods. Impacts of pulse events are characterized by sudden, large changes in environmental variables, after which over time the variables return to their pre-disturbance levels. Studies of pulse events are concerned not only with whether the event had an effect, but also with how quickly the environment or ecosystem recovers after the event. The top graph in Fig. 2.12 illustrates a pulse event.

A ***press*** event is a long-term, sustained disturbance, such as a permanent alteration of the environment or ecosystem. Press events are often foreseeable anthropogenic perturbations. Examples include sustained discharge of toxic chemicals or sewage, construction of a permanent structure such as a highway or dam, and sustained overharvesting of a plant or animal species. Impacts of press events are characterized by large, lasting changes in environmental variables. Studies of press events are concerned not only with whether the event had an effect, but also with how large the effect was if there was one. The bottom graph in Fig. 2.12 illustrates a press event.

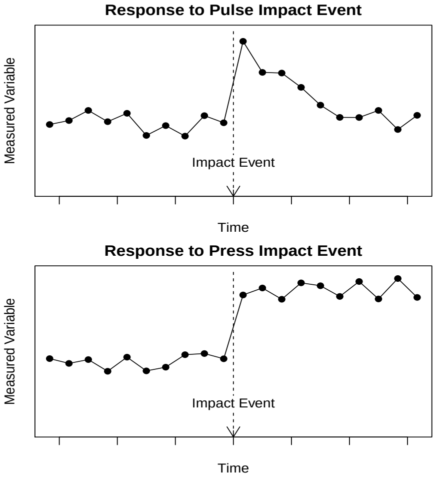


Figure 2.12: The response of a variable to a pulse impact event (top) and to a press impact event (bottom).

Before-After Studies

In ***before-after*** studies, an environmental variable is measured before and after the impact event, but only at the impact site. Typically, a *trajectory* study design is used, whereby a *series* of measurements is made over time at a single location in the impact site before the impact event, and another *series* at that *same* location after the event. Alternatively, a *snapshot* design could be used, whereby the variable is measured at a *random sample* of *locations* within the impact site, all at the same time, before the impact event and then again at those *same* locations, all at the same time, after the event. Finally, a *combination* trajectory/snapshot design could be used, with measurements made over time *and* space both before and after the impact event.

Before-after studies are usually done when no suitable control site is available. They require prior knowledge that the event is going to occur, so they can be used, for example, to assess the impact of new highway construction, but not of unforeseen events like accidental oil spills or floods (unless data happen to have already been collected when the event occurs).

A major drawback of before-after studies is that they don't tell us for sure whether a perceived effect of the impact event is attributable to the actual event or just to naturally occurring changes in the environment over time. Would the observed change have happened regardless of whether or not the impact event had occurred? We can't always tell. The problem is that the effect (if any) of the impact event is *confounded* with the effects of other variables that are changing naturally over time, and would be doing so regardless of whether or not the impact event had occurred. Examples include changing weather conditions, animal immigration and emigration, human population encroachment, etc.

Control-Impact Studies

In ***control-impact*** studies, the environmental variable is measured *after* the impact event at two sites, the impact site and a nearby, unaffected control site. The control site should be chosen to be as similar as possible to the impact site, and not too distant from it, so that the soil, weather, climate, vegetation, etc. are similar at the two sites. A *trajectory* study design could be used, with a *series* of measurements made over time at a *single* location within each site. Whenever possible, the measurements should be made simultaneously at the two sites so that the two time series data sets are synchronized. Alternatively, a *snapshot* design could be used, whereby the variable is measured at a *random sample* of *locations* within the impact site and at another *random sample* of *locations* at the control site. In this design, the measurements at the two sites should be made at the *same* time, or as close to it as possible. A *combination* trajectory/snapshot design could also be used, with measurements made over time *and* space at both the impact site and the control site.

Control-impact studies don't require advance knowledge that the impact event is going to occur, so they can be used to assess the impacts of unforeseen events like oil spills and floods.

Their main drawback, though, is that no matter how hard we might try to choose a control site that's identical to the impact site, there will always be some differences, for example in soil conditions, presence of nearby industries, etc. And if the variables that differ across the two sites are related the one being used to assess the effect of the impact event, we won't know for sure whether a perceived effect is due to the event or to differences across the two sites in those other variables. In other words, the effect (if any) of the impact event will be *confounded* with the effects of those other variables.

Before-After-Control-Impact Studies

In ***before-after-control-impact*** (or ***BACI***) studies, the environmental variable is measured both before *and* after the impact event at both the impact site *and* a control site. As for control-impact studies, the control site should be chosen to be as similar as possible to, and not too far from, the impact site.

There are two ways to look at the results of a BACI study to decide if the impact event had an effect. In the first, the *change* in the measured variable at the impact site, from the period before the impact event to the period after, is compared to the *change* in that variable at the control site over that same time span. If the variable changed more at the impact site than at the control site, it suggests that the event had an effect. The second way is to compare the *difference* in values of the variable between the two sites before the event to the *difference* after. If the difference after the event is larger than it was before, it suggests that the event had an effect.

Typically, a *trajectory* study design is used, whereby a *series* of measurements is made over time at a single location in the impact site, both before *and* after the impact event (same location both periods), and another *series* at a location in the control site, both before *and* after the event (same location both periods). As for before-after studies, the measurements should be made simultaneously at the two sites, if possible, so that the time series data sets are synchronized. Alternatively, a *snapshot* design could be used. Here, the variable is measured at a *random sample* of *locations* within the impact site, all at the same time if possible, before the impact event and then again at those *same* locations, all at the same time, after the event, and similarly the variable is measured at a sample of locations before and after the event at the control site. Finally, a *combination* trajectory/snapshot design could be used, with measurements made over time *and* space both before and after the impact event at both sites.

The tables below reflect these three study designs. An X represents a measurement of the variable, S1, S2, . . . represent locations, and T1, T2, . . . represent time points.

Trajectory Type BACI Study

		Before			After		
		T1	T2	T3	T4	T5	T6
Control	S1	X	X	X	X	X	X
Impact	S2	X	X	X	X	X	X

Snapshot Type BACI Study

		Before	After
		T1	T2
Control	S1	X	X
	S2	X	X
	S3	X	X
Impact	S4	X	X
	S5	X	X
	S6	X	X

Combination Trajectory/Snapshot BACI Study

		Before			After		
		T1	T2	T3	T4	T5	T6
Control	S1	X	X	X	X	X	X
	S2	X	X	X	X	X	X
	S3	X	X	X	X	X	X
Impact	S4	X	X	X	X	X	X
	S5	X	X	X	X	X	X
	S6	X	X	X	X	X	X

Two implicit assumptions in BACI studies are that the control and impact sites were suitably similar to each other prior to the impact event, and that no other events, besides the impact event, affected one site but not the other between the two data collection periods.

BACI impact assessment study designs are preferred to both before-after and control-impact designs. Unlike before-after studies, BACI studies are able to control for external variables that are changing naturally over time, such as weather conditions, human population encroachment, etc., as long as those variables are changing equally at the control and impact sites. And unlike control-impact studies, BACI studies are able to control for variables that differ across the control and impact sites, such as soil conditions, presence of nearby industries, etc., as long as the magnitudes of those differences are the same before and after the impact event.

But like the before-after study designs, BACI designs require advance knowledge that the impact event is going to occur, so they're not feasible for unforeseen events like oil spills and floods.

2.4 Problems

2.1 A sample of households on a certain road is to be taken for the purpose of estimating the total energy usage along that road. Classify each of the following as either a simple random sample or a systematic random sample.

a) A single household from among the first five along the road is randomly selected, and then every fifth household thereafter is also selected.

b) The N households on the road are each assigned a number from 1 to N, and these numbers are written on slips of paper and mixed thoroughly in a box. Then a specified number of households is selected by blindly drawing slips of paper from the box.

2.2 A study is to be carried out to estimate the average arsenic concentration in a region's soil. Classify each of the following as either a simple random sample or a systematic random sample.

a) A computer is used to generate the spatial coordinates of $n = 30$ points in the study region. Each point is generated independently of the other points. Then soil specimens are taken at each point.

b) Starting from a randomly selected point, a regular grid over the study region is established, with grid lines 25 m apart. Then soil specimens are taken at the grid line intersections.

2.3 A sample of a city's households is to be taken for the purpose of estimating the amount of plastic disposed of per week by a typical household. Classify each of the following as either a simple random sample, systematic random sample, stratified random sample, or two-stage sample.

a) The addresses of all the households in the city are written on slips of paper and mixed thoroughly in a box. Then a specified number of households is selected by blindly drawing slips of paper from the box.

b) The city is first partitioned into seven socioeconomic districts, and from each district, a specified number of households are randomly selected.

c) A single household in the city is randomly selected, and ends up being the third house on its block. Then, for every block in the city, the third house on the block is selected for the sample.

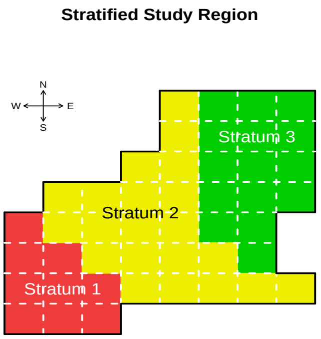
2.4 A sample of quadrats on a reef is to be taken for the purpose of estimating the percent coral cover on the reef. Classify each of the following as either a simple random sample, systematic random sample, stratified random sample, or two-stage sample.

a) The entire reef is partitioned into 3,500 1 m^2 quadrats. The quadrats are numbered from 1 to 3,500 and a computer random number generator is used to randomly choose $n = 25$ of these numbers.

b) First, eight 50 m long transects are randomly selected on reef. Then, along each transect, a simple random sample of five 1 m^2 quadrats is selected, giving a total sample size of $n = 40$ quadrats.

c) Starting from a randomly selected point, a regular grid over the reef is established, with grid lines 10 m apart. Then a 1 m^2 quadrat is selected at each set of grid line intersections.

2.5 The figure below depicts a study region that's been partitioned into three strata.



We want to take a stratified random sample of $n = 60$ locations from within the region. How many locations should be selected from each stratum if proportional allocation is used to determine the stratum-specific sample sizes n_1, n_2, and n_3? **Hint**: The relative areas of the strata can be determined by counting the squares within them.

2.6 A study was carried out to compare the risk to breeding success of herons and egrets due to the presence of contaminants in a nature preserve in Hong Kong [4]. One question of interest was whether contaminant concentrations in eggs would differ for the two species, for example due to different diets and metabolic characteristics.

An egg was taken from each of nine randomly selected egret nests in the Mai Po egretry and nine randomly selected heron nests in the A Chau egretry. The cited paper describes the two sites as follows:

> Mai Po egretry is located in woodland closely adjacent to the Mai Po Marshes and opposite the Mai Po Village, while the A Chau egretry is located in a more remote woodland on an island in Starling Inlet.

The researchers found that the egret eggs from Mai Po had significantly higher concentrations of chlordanes (a contaminant) than heron eggs from A Chau. But they state in the paper:

> Although the above data involve comparisons between different species with different locations of the egretries, it is possible that they indicate a spatial difference in levels of trace organic pollution between the western and eastern waters of Hong Kong.

a) Is the study an observational study or an experiment?

b) Identify the variable, singled out by the researchers, whose effect on contaminant concentrations in eggs might be confounded with the effect, if any, of bird species.

2.7 Two separate studies were carried out to investigate the effects of the herbicide atrazine on enzyme activity in alligators [5].

a) Both studies are described below. Classify each as either an observational study or an experiment.

1. In one study, eggs were collected from alligator nests on Lake Apopka, Florida, a lake heavily polluted by atrazine from agricultural activities, other contaminants from a sewage treatment facility, and a chemical spill at a nearby pesticide manufacturing facility. Eggs were also collected from nests on Lake Woodruff National Wildlife Refuge, also in Florida, an unpolluted lake that served as a control. Enzyme activity was found to be significantly lower in the juvenile alligators hatched from Lake Apopka eggs than in those hatched from Lake Woodruff eggs.
2. In the other study, eggs were collected from nests on Lake Woodruff (the unpolluted lake), and randomly assigned to two treatment groups. In the first group, atrazine was applied to the shells of the eggs during a critical period of embryonic development. In the second group, which served as a control group, no atrazine was applied. The researchers found no significant difference between the enzyme activities of alligators hatched from eggs in the two groups.

b) The result of first study seems to suggest that atrazine lowers enzyme activity in alligators, yet the second study found no such effect. How do you explain the seemingly contradictory study results? In your answer, identify at least one confounding variable in the first study.

2.8 Biodegradation of a substance refers to decomposition by microbes. The rate at which a substance biodegrades in soil can be influenced by several factors. Environmental scientists are interested in knowing which factors influence the biodegradation rates of pesticides. In one study, soil specimens collected from 20 sites were applied with pesticide after first being analyzed for various factors, among them microbial biomass and pH [18]. It was found that the pesticide biodegraded faster in soils with higher pH levels. But soils with higher pH levels were also found to have higher levels of microbial biomass.

a) In this study, what is the response variable?

b) Which two variables are confounded?

c) Explain, in terms of confounding, why we can't conclude from this study that higher soil pH *causes* faster pesticide biodegradation.

d) Another way to mitigate the confounding effect of a variable in an experiment is to *control* for that variable by holding it constant across treatment groups.

Describe how you'd carry out an experiment to decide whether soil pH *causes* faster pesticide biodegradation rates, while controlling for the confounding effect of microbial biomass.

2.9 Among the chemicals used in hydraulic fracturing, an oil and natural gas drilling technique, are some so-called endocrine disrupting chemicals (EDCs), exposure to which can lead to fertility problems, birth defects, and cancer [11], [15].

To determine if hydraulic fracturing can lead to EDC contamination of surface and ground water, researchers collected surface and ground water samples from natural gas drilling-dense areas in Garfield County, Colorado, and compared them to samples collected from control sites in Boone County, Missouri, where little or no drilling had taken place [11]. They found that the EDC levels were higher in the water samples taken from the drilling-dense areas than in those from the control sites.

An oil and gas industry representative responded:

> It [the study] compares Boone County, Mo., with Garfield County, Colo., which is essentially comparing apples to oranges. These are different states, different regions, they have different industries in their areas, completely different terrain. And, the report treats them as if they are exactly the same [15].

a) Is the study a before-after, control-impact, or before-after-control-impact study?

b) Identify at least one variable, singled out by the oil and gas industry representative, whose effect on ECD levels in water might be confounded with the effect, if any, of hydraulic fracturing.

2.10 The Exxon Valdez oil tanker struck a reef in Prince William Sound, Alaska, on March 24, 1989, resulting in the largest marine oil spill in U.S. history.

To determine if the spill resulted in bird mortality or displacement, researchers compared bird abundance data collected along the shores in the three years after the oil spill (1989-1991) to historical abundance data collected before the oil spill (1984-1985) [16]. They found that the abundance of Pigeon Guillemots was significantly lower after the spill than before.

But they cautioned about attributing the decline in bird abundance to the oil spill:

> One of the most challenging aspects of using historical data for impact analyses is sorting out the effects of a perturbation [such as an oil spill] from natural temporal variability that occurs independently within the system.
>
> Natural annual variability often is high for marine birds in south coastal Alaska.
>
> We did not separate oiling effects from natural factors that might have contributed to overall population changes between the two time periods.

a) Is the study a before-after, control-impact, or before-after-control-impact study?

b) Identify the variable, singled out by the researchers, whose effect on bird abundance might be confounded with the effect, if any, of the oil spill.

2.11 Classify each of the following impact assessment studies as before-after, control-impact, or before-after-control-impact.

a) To assess the effect of a new nuclear power plant's hot water discharge on a river's aquatic community, the fish abundance is recorded before the plant begins operations at a site downstream of where the hot water will be discharged, and at another site upstream, where the discharge will have no effect. After the discharge begins, the sampling protocol is repeated, and the change in fish abundance at the downstream site is compared to the change (if any) at the upstream site.

b) To assess the effect of an oil spill on the bird population at a beach on which the oil has washed ashore, bird counts are made at that beach and compared to ones made at a nearby, unaffected beach that has similar characteristics.

c) At a monitoring station near an operating nuclear power plant, the health status of the area's plant and animal communities is monitored. An unforeseen severe malfunction (a "meltdown") suddenly occurs at the power plant. After it's deemed safe for humans to return to the monitoring site, the health status of the same plant and animal communities is compared to what it was prior to the accident.

Bibliography

[1] South Florida ecosystem assessment: Phase I/II (tech. report) - Everglades stressor interactions: Hydropatterns, eutrophication, habitat alteration, and mercury contamination. Technical Report EPA 904-R-01-003, United States Environmental Protection Agency, 2001.

[2] P. Aastrup and A. Mosbech. Transect width and missed observations in counting muskoxen (*Ovibos moschatus*) from fixed-wing aircraft. *Rangifer*, 13(2):99 – 104, 1993.

[3] Edward A. Bender, Ted J. Case, and Michael E. Gilpin. Perturbation experiments in community ecology: Theory and practice. *Ecology*, 65(1):1–13, Feb 1984.

[4] D.W. Connell et al. Risk to breeding success of fish-eating ardeids due to persistent organic contaminants in Hong Kong: Evidence from organochlorine compounds in eggs. *Water Research*, 37:459–467, 2003.

[5] D. Andrew Crain et al. Alterations in steroidogenesis in alligators (*Alligator mississippiensis*) exposed naturally and experimentally to environmental contaminants. *Environmental Health Perspectives*, 105(5):528 – 533, May 1997.

[6] Richard O. Gilbert. *Statistical Methods for Environmental Pollution Monitoring*. John Wiley and Sons, 1987.

[7] Nicholas J. Gotelli and Aron M. Ellison. *A Primer of Ecological Statistics*. Sinauer Associates, Inc., Sunderland, MA, U.S.A., 2004.

[8] A. Hector et al. Plant diversity and productivity experiments in European grasslands. *Science*, 286(1123), 1999.

[9] Jo Ellen Hinck et al. Environmental contaminants and biomarker responses in fish from the Columbia River and its tributaries: Spatial and temporal trends. *Science of the Total Environment*, 366:549 – 578, 2006.

[10] Stuart H. Hurlbert. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2):187 – 211, 1984.

[11] Christopher D. Kassotis, Donald E. Tillitt, J. Wade Davis, Hormann Annette M., and Nagel Susan C. Estrogen and androgen receptor activities of hydraulic fracturing chemicals and surface and ground water in a drilling-dense region. *Endocrinology*, 2013.

[12] K. Kosmelj, A. Cedilnik, and P. Kalan. A comparison of a two-stage sampling design and its composite sample alternative: An application to soil studies. *Environmental and Ecological Statistics*, 8:109–119, 2001.

[13] Paul S. Levy and Stanley Lemeshow. *Sampling of Populations: Methods and Applications*. Wiley, fourth edition, 2008.

[14] Sharon L. Lohr. *Sampling: Design and Analysis.* Brooks/Cole, first edition, 1999.

[15] Lesley McClurg. Hormone disrupting chemicals linked to fracking. Colorado Public Radio, Jan. 16, 2014.

[16] Stephen M. Murphy, Robert H. Day, John A. Wiens, and Keith R. Parker. Effects of the Exxon Valdez oil spill on birds: Comparisons of pre- and post-spill surveys in Prince William Sound, Alaska. *The Condor*, 99:299–313, 1997.

[17] D.J. Spurgeon et al. Effects of cadmium, copper, lead and zinc on growth, reproduction and survival of the earthworm *Eisenia fetida*(savigny): Assessing the environmental impact of point-source metal contamination in terrestrial ecosystems. *Environmental Pollution*, 84:123–130, 1994.

[18] A. Walker, M. Jurado-Exposito, G.D. Bending, and V.J.R. Smith. Spatial variability in the degradation rate of isoproturon in soil. *Environmental Pollution*, 111:407 – 415, 2001.