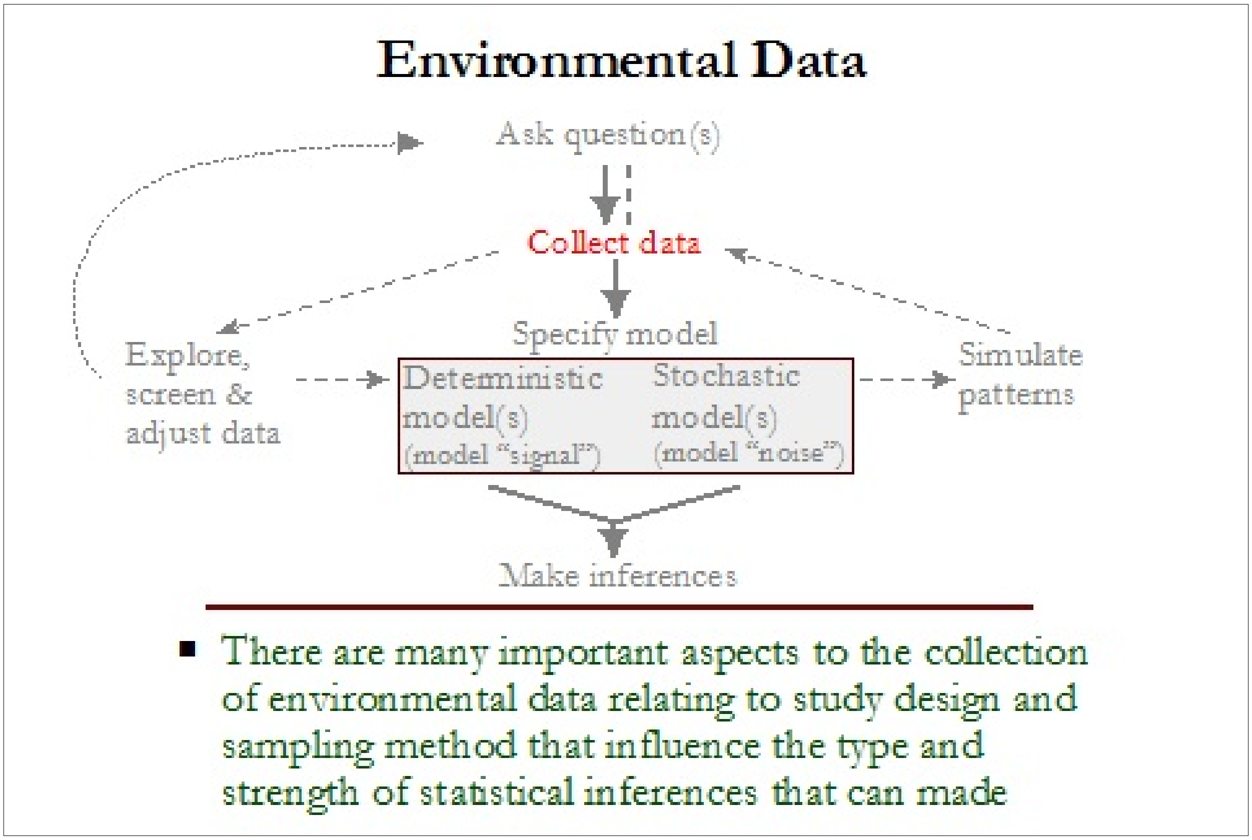
Analysis of Environmental Data

Chapter 2. Conceptual Foundations:
Environmental Data



Environmental Data

Ask question(s)

Collect data

Specify model

Explore, screen & adjust data

Deterministic model(s) (model "signal")

Stochastic model(s) (model "noise")

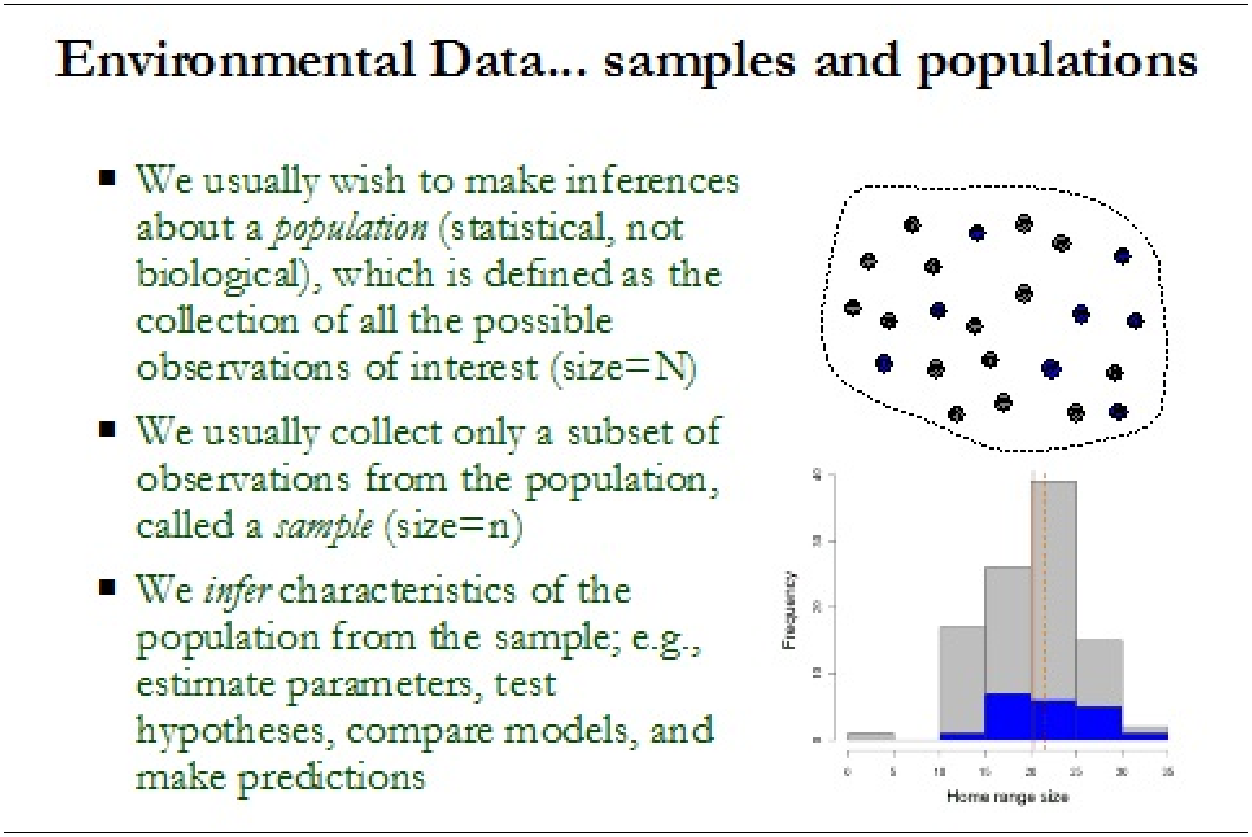
Simulate patterns

Make inferences

- There are many important aspects to the collection of environmental data relating to study design and sampling method that influence the type and strength of statistical inferences that can made

1. Purpose of data collection

Ideally, once the environmental question has been identified, the study is designed and the data is collected in a manner that will result in strong inferences. There are many important aspects to the collection of environmental data relating to study design and sampling method that will influence the type and strength of statistical inferences that can made: identifying the desired scope of inference, choosing appropriate observational/experimental (sampling) units, choosing the types of data to collect, and establishing a robust sampling scheme (i.e., spatial and/or temporal distribution of units and method(s) of collecting the data) to ensure accurate and precise inferences. These issues will be discussed in the Research Concepts course. Here, we will briefly distinguish samples from populations, describe the major different types of environmental data and some of the properties of each type, the types of variables and their relationships, and preview some of the important issues to consider in study design that will be discussed further in the Research Concepts course.



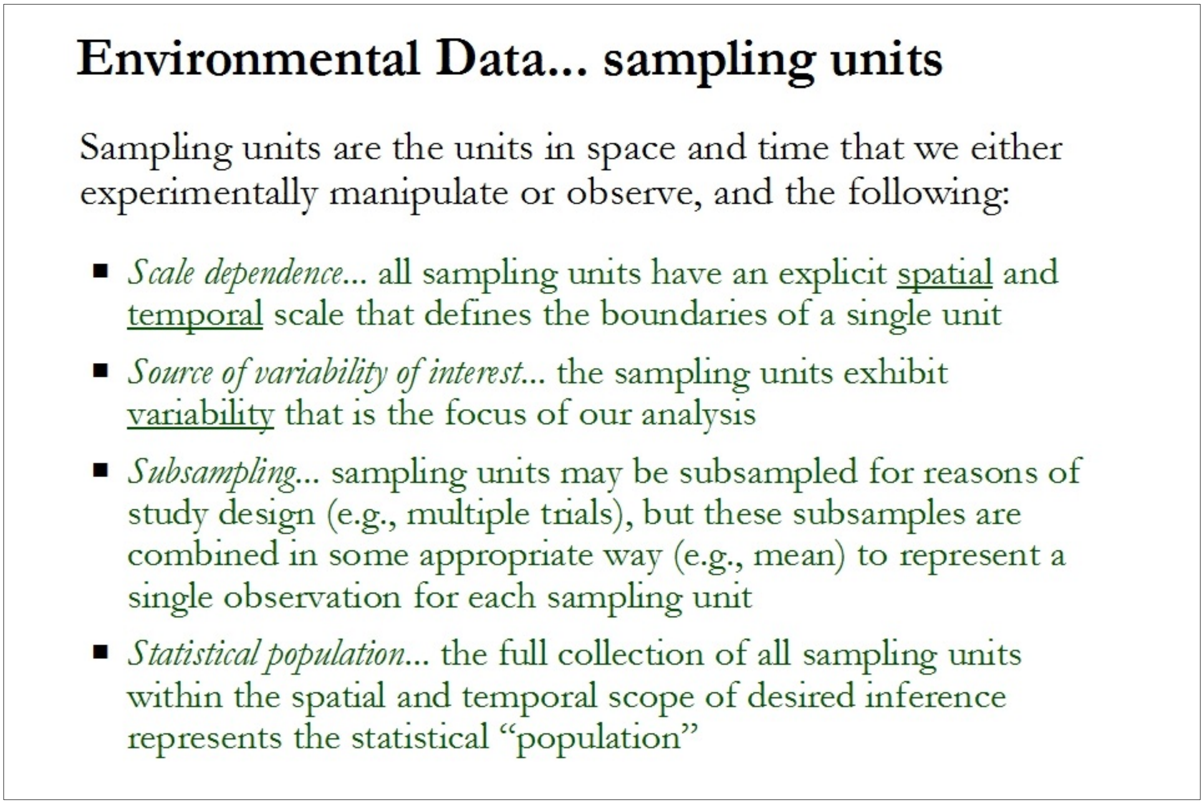
Environmental Data... samples and populations

- We usually wish to make inferences about a *population* (statistical, not biological), which is defined as the collection of all the possible observations of interest (size=N)
- We usually collect only a subset of observations from the population, called a *sample* (size=n)
- We *infer* characteristics of the population from the sample; e.g., estimate parameters, test hypotheses, compare models, and make predictions

2. Samples and populations

We usually wish to make inferences about a population (statistical, not biological), which is defined as the collection of all the possible observations of interest. A biological population under consideration may or may not constitute the statistical population if, for example, the functional population extends over a broader geographic extent than the study area. We usually represent the size of the statistical population in statistical formulae as upper case N. For lots of practical reasons, we usually collect only a subset of observations from the population, and we represent the size of the sample as lower case n.

Importantly, we *infer* characteristics of the population from the sample; e.g., estimate parameters, test hypotheses, compare models, and make predictions. Thus, the entire realm of inferential statistics applies when we seek to draw conclusions from a sample about the underlying population. Otherwise, we may be interested in or forced to merely describe the patterns in the sample without explicit inference to the population – the realm of descriptive statistics. Note, in rare cases, we may actually observe every possible entity of interest – the population – in which case simple descriptive statistics suffice to draw conclusions from about the population since we know with exactness (to the precision of our measurement system) the characteristics of the population we are studying.



Environmental Data... sampling units

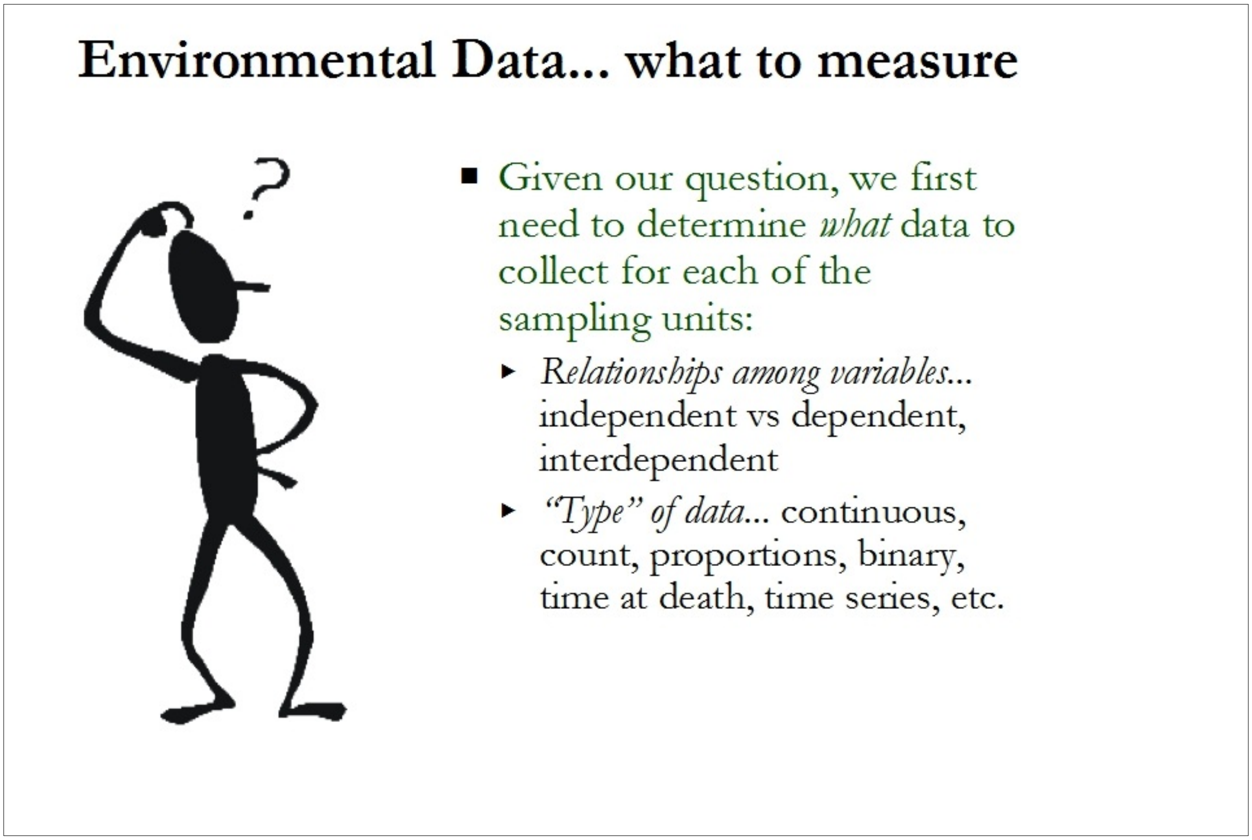
Sampling units are the units in space and time that we either experimentally manipulate or observe, and the following:

- *Scale dependence*... all sampling units have an explicit spatial and temporal scale that defines the boundaries of a single unit
- *Source of variability of interest*... the sampling units exhibit variability that is the focus of our analysis
- *Subsampling*... sampling units may be subsampled for reasons of study design (e.g., multiple trials), but these subsamples are combined in some appropriate way (e.g., mean) to represent a single observation for each sampling unit
- *Statistical population*... the full collection of all sampling units within the spatial and temporal scope of desired inference represents the statistical "population"

3. Sampling units

A sample is a collection of sampling units from the underlying statistical population. These sampling units are the units in both space and time that we either experimentally manipulate or observe (and thus also referred to as experimental and observational units, respectively) and for which we measure the attributes of interest and relevant to the question or objective under consideration. Sampling units can be anything so long as they are defined in a manner and at a scale that is relevant to the question under consideration. There are several important considerations:

- All sampling units should be defined with an explicit spatial and temporal scale. For example, does the sampling unit represent, e.g., a 0.1-hectare plot, 1-hectare plot, or 100-hectare plot, or similarly, does it represent, e.g., a plant cell, single leaf, single branch, single tree, single forest stand, or single forested landscape? Similarly, does the sampling unit represent, e.g., a second, a day, a year, or a decade in the system under investigation?
- Sampling units represent the source of variability in the system that is the focus of the investigation, and thus each sampling unit should, at least potentially, vary with respect to one or more measured variables.
- Sampling units may be subsampled for reasons of study design (e.g., when multiple trials are required to determine a proportional response). Subsamples are measurements/observations made below the scale of the focal sampling unit; as such, these subsamples are combined in some appropriate way (e.g., mean) to represent a single observation for each sampling unit.
- The full collection of all potential sampling units (whether they can be enumerated or not) within the spatial and temporal scope of desired inference represents the statistical "population".



Environmental Data... what to measure

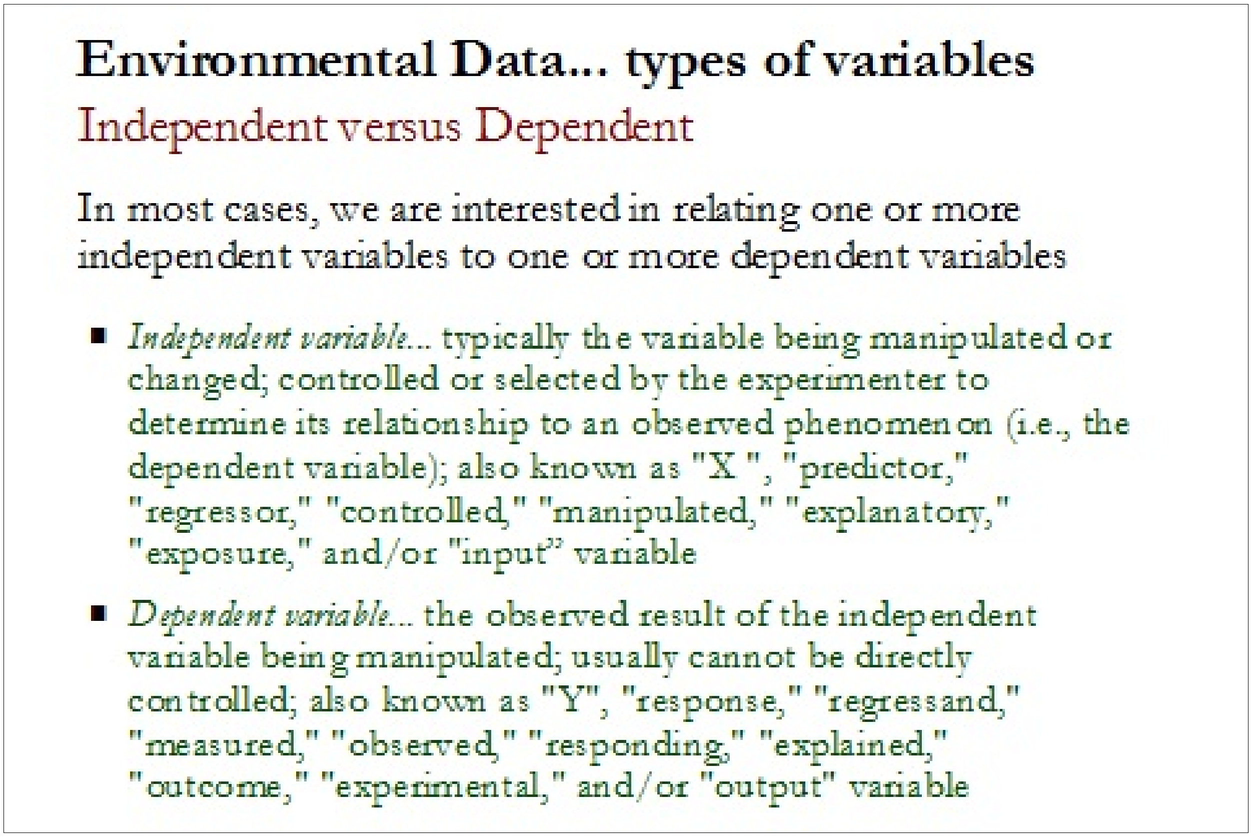
- Given our question, we first need to determine *what* data to collect for each of the sampling units:
  - *Relationships among variables...* independent vs dependent, interdependent
  - *"Type" of data...* continuous, count, proportions, binary, time at death, time series, etc.

4. What to measure

Once we have identified our environmental question, the first thing we need to do is determine what data to collect. This is one of the most important steps in the entire modeling process, because if we collect the wrong type of data, no statistical model of any kind will allow us to answer our environmental question. While there are many important considerations to this step, we need to carefully consider the "type" of data and the relationships among variables.

There are at least three major types of variables based on their relationships to each other: 1) independent variables, 2) dependent variables, and 3) interdependent variables.

In environmental studies, there are several major types of data: 1) continuous data, 2) counts, 3) proportions, 4) binary data, 5) time at death, 6) time series, and 7) circular data. Importantly, here we are principally referring to the response data or dependent variable when a distinction is made between dependent and independent variables. This is noteworthy, because it is typically the response variable type that determines the appropriate statistical model or class of statistical methods.



Environmental Data... types of variables

Independent versus Dependent

In most cases, we are interested in relating one or more independent variables to one or more dependent variables

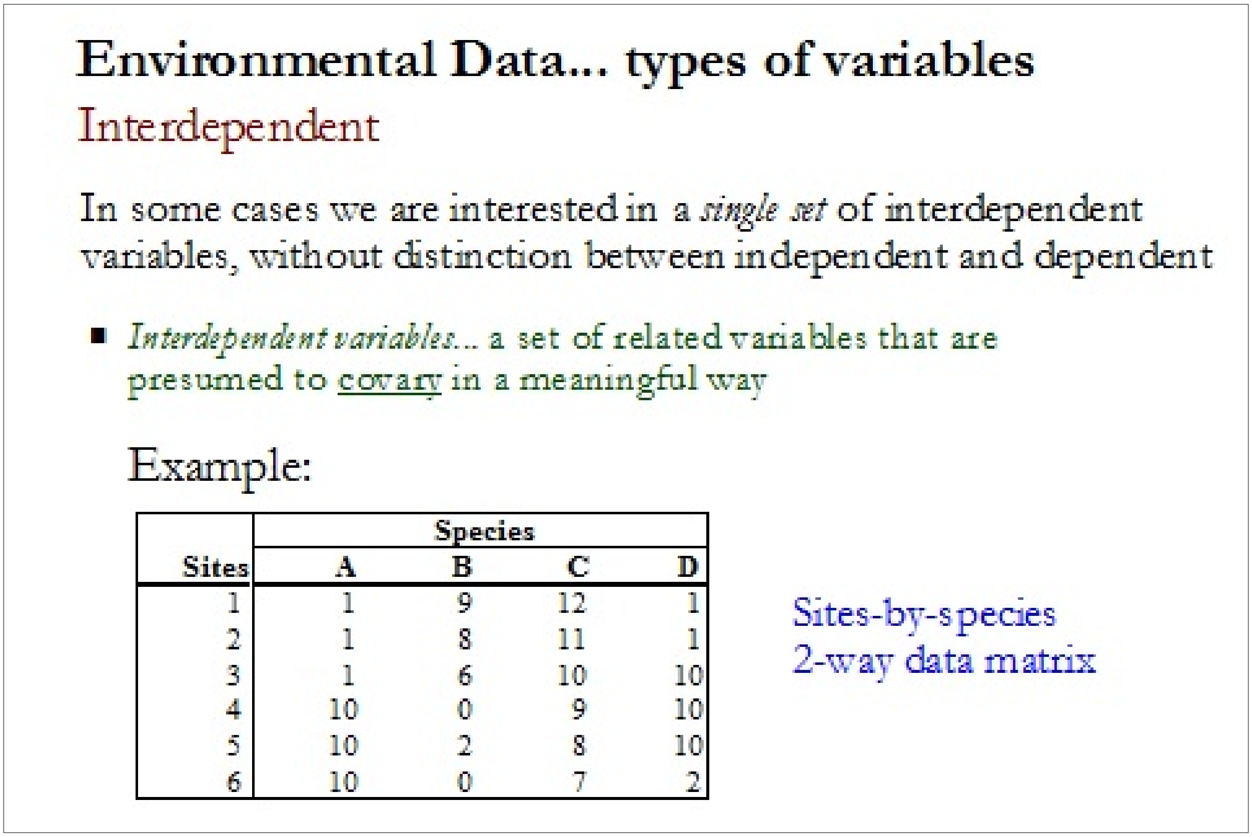
- *Independent variable*... typically the variable being manipulated or changed; controlled or selected by the experimenter to determine its relationship to an observed phenomenon (i.e., the dependent variable); also known as "X ", "predictor," "regressor," "controlled," "manipulated," "explanatory," "exposure," and/or "input" variable
- *Dependent variable*... the observed result of the independent variable being manipulated; usually cannot be directly controlled; also known as "Y", "response," "regressand," "measured," "observed," "responding," "explained," "outcome," "experimental," and/or "output" variable

5. Types of variables

In most, but not all, studies, our environmental question requires that we collect data on two or more variables in which one or more variables are considered as “independent” variables and one or more are considered as “dependent” variables. This distinction is critical to most statistical models, but note that variables are not intrinsically “dependent” or “independent”, rather this distinction is one of context as defined by the researcher.

Independent variable... typically the variable(s) being manipulated or changed, or the variable(s) controlled or selected by the experimenter to determine its relationship to an observed phenomenon (i.e., the dependent variable). In observational studies, the independent variable(s) is not explicitly manipulated or controlled through experimentation, but rather observed in its naturally occurring variation, yet it is presumed determine or influence the value of the dependent variable. The independent variable is also known as "x ", "explanatory," "predictor," "regressor," "controlled," "manipulated," "exposure," and/or "input” variable.

Dependent variable... the observed result of the independent variable(s) being manipulated, and it usually cannot be directly controlled. The dependent variable is generally the phenomenon whose behavior we are interested in understanding. The dependent variable is also known as "y", "response," "regressand," "measured," "observed," "responding," "explained," "outcome," "experimental," and/or "output" variable.



Environmental Data... types of variables

Interdependent

In some cases we are interested in a *single set* of interdependent variables, without distinction between independent and dependent

- *Interdependent variables...* a set of related variables that are presumed to covary in a meaningful way

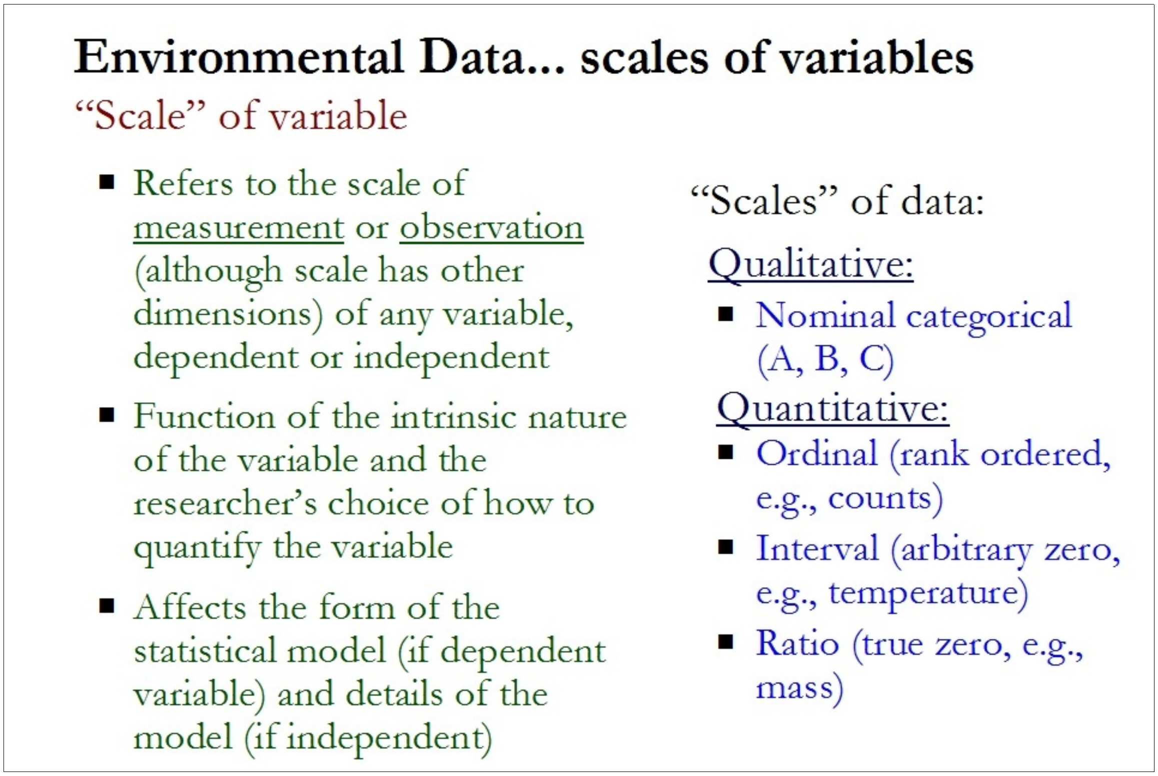
Example:

	Species			
Sites	A	B	C	D
1	1	9	12	1
2	1	8	11	1
3	1	6	10	10
4	10	0	9	10
5	10	2	8	10
6	10	0	7	2

Sites-by-species
2-way data matrix

In some cases we are interested in a *single set* of interdependent variables, without distinction between independent and dependent

Interdependent variables... a set of related variables that are presumed to covary in a meaningful way. A common example is a community data set consisting of *n* sites by *p* species abundances, arranged in a two-way data matrix in which the rows represent the sites and the columns represent the species. In this case, the species are the variables and there is no distinction of independent and dependent. In fact, they are all presumed to be interdependent on each other since they presumably covary in meaningful ways. Moreover, they are generally considered to be inter-*dependent* variables because they are presumed to respond to other perhaps unmeasured independent variables that are not part of this variable set.



Environmental Data... scales of variables

"Scale" of variable

- Refers to the scale of measurement or observation (although scale has other dimensions) of any variable, dependent or independent
- Function of the intrinsic nature of the variable and the researcher's choice of how to quantify the variable
- Affects the form of the statistical model (if dependent variable) and details of the model (if independent)

"Scales" of data:

Qualitative:

- Nominal categorical (A, B, C)

Quantitative:

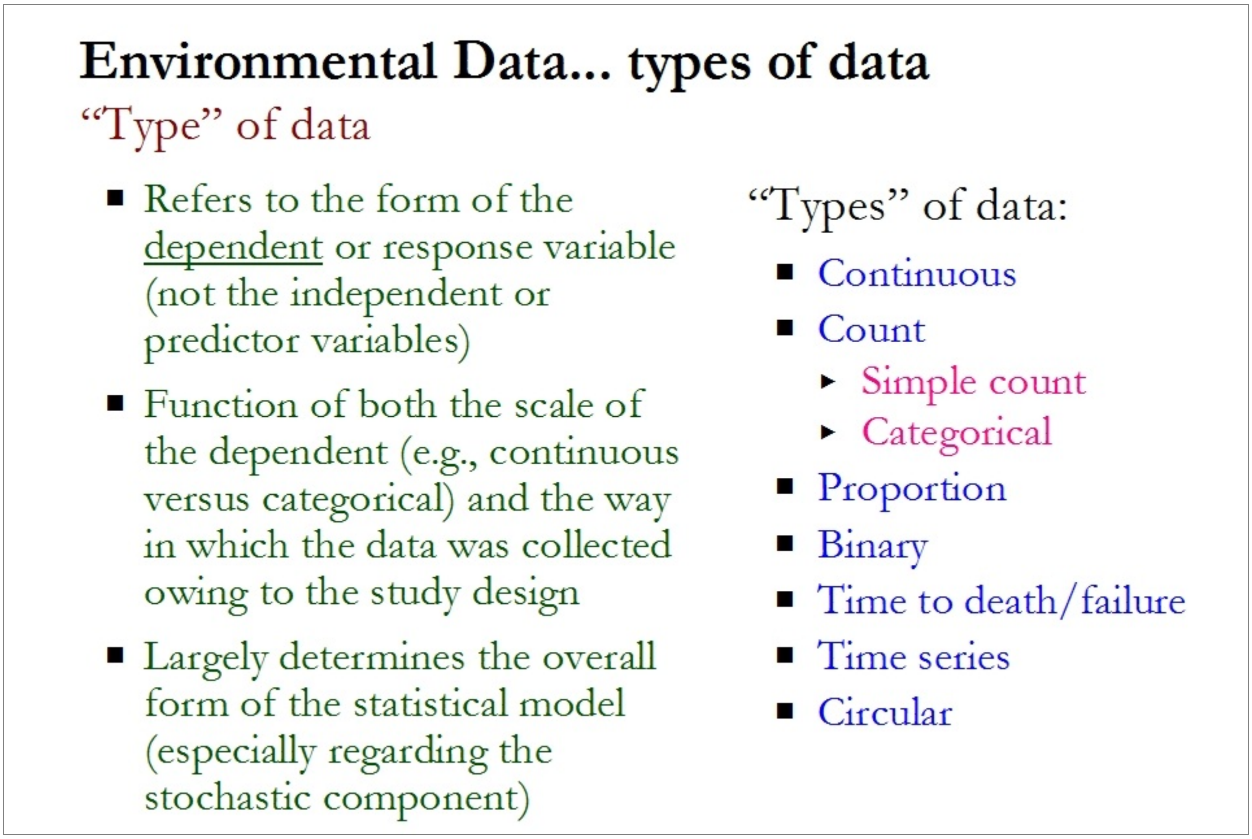
- Ordinal (rank ordered, e.g., counts)
- Interval (arbitrary zero, e.g., temperature)
- Ratio (true zero, e.g., mass)

6. Scales of variables

The "scale" of a variable refers to the scale of measurement or observation, although note that there are other dimensions of scale, and this applies to all variables regardless of whether they are being treated as dependent or independent variables. Importantly, the scale of the variable is a function of both the intrinsic nature of the variable (e.g., a nominal-scaled qualitative variable cannot be coerced into a ratio-scaled quantitative variable) and the researcher's choice of how to quantify the variable. For example, an intrinsically ratio-scaled quantitative variable (e.g., tree height) can easily be measured in height classes (e.g., short, medium, tall) and treated as an ordinal-scale variable. Lastly, the scale of the variable affects the form of the statistical model, and thus helps determine the "type" of data if it is the dependent variable and details of the model and model fitting procedures if it is an independent variable.

Although there is no agreed upon classification system for defining the "scale" of a variable, one convenient system includes four basic scales: nominal, ordinal, interval and ratio. A nominal-scale variable is categorical or discrete and fundamentally *qualitative* in nature; i.e., there is no quantitative information present in the scale. A good example is the variable species. The remaining scales are all *quantitative* in the sense that they convey quantitative information about the observational/ experimental units. An ordinal-scale variable is also discrete, in that units cannot be infinitely subdivided, and expresses rank order information. A good example is age class: young, middle-aged, old. Also, counts (# of eggs) are usually treated as ordinal scale and discrete. Interval and ratio scale differ in whether there is a true zero (ratio) or not (interval), and in other subtle ways, but most statistical methods treat these scales equivalently and thus we often just refer to them as "continuous" variables. Importantly, these so-called "continuous" variables are such that they can take on any value along a number continuum; i.e., they can be measured to any level of precision limited only by the precision of the measuring device. For example, temperature can take on any value along a number continuum, and thus the measured values are limited only by the precision of the measuring device. For example, we could measure temperature to the nearest degree, tenth of a degree, hundredth of a degree, and so on, depending on the precision of the thermometer.

Note, the measurement/observation scale of the dependent variable is of utmost importance in determining the "type" of data and thus the form of the statistical model with respect to the stochastic (or error) component of the model, as discussed in later chapters. Discrete dependent variables, including both nominal and ordinal scales, generally warrant the use of discrete probability distributions for the stochastic component of the model (i.e., to describe the error). Likewise, continuous dependent variables, including both interval and ratio scales, necessitate the use of continuous probability distributions for the stochastic component of the model. Thus, one cannot form an appropriate statistical model without knowing the scale of the dependent variable. On the other hand, the scale of the independent variable is generally of minor consequence as it generally only influences subtle details of how the model is formulated mathematically and the mechanics of the model fitting procedures.



Environmental Data... types of data

"Type" of data

- Refers to the form of the dependent or response variable (not the independent or predictor variables)
- Function of both the scale of the dependent (e.g., continuous versus categorical) and the way in which the data was collected owing to the study design
- Largely determines the overall form of the statistical model (especially regarding the stochastic component)

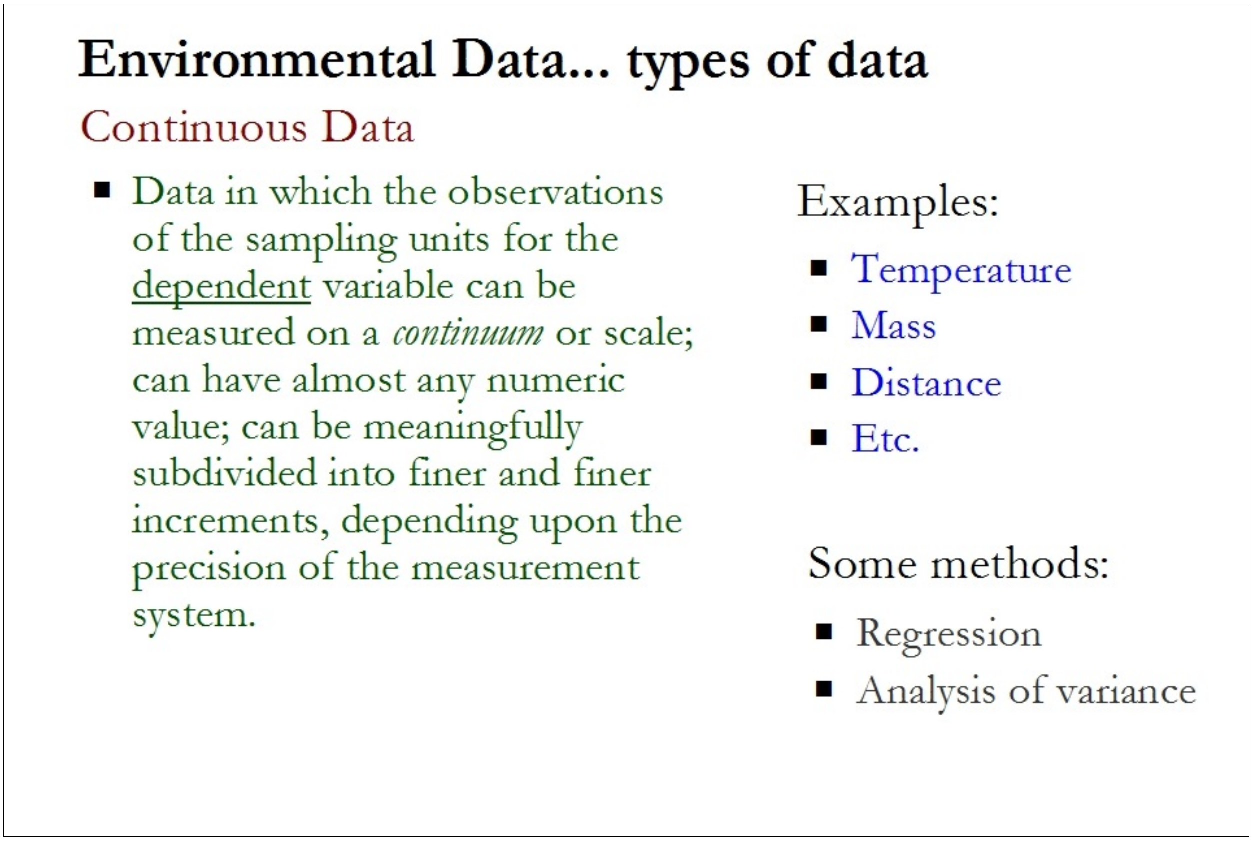
"Types" of data:

- Continuous
- Count
  - Simple count
  - Categorical
- Proportion
- Binary
- Time to death/failure
- Time series
- Circular

7. Types of data

The "type" of data refers to the form of the dependent or response variable, not the independent or predictor variables, when a distinction is made between dependent and independent variables. And in cases when there is no distinction being made, for example when only a single variable is measured, the variable of interest is often assumed to be or at least treated as if it were the dependent variable. Importantly, the "type" of data is a function of both the scale of measurement or observation (e.g., discrete or continuous) and the way in which the data was collected owing to the study design, as described in the material that follows. Lastly, the "type" of data is critical to determine because it largely determines the overall form of the statistical model, especially with regards to the choice of the stochastic component of the model, as discussed in later chapters.

As noted above, in environmental studies there are several major types of data commonly encountered:: 1) continuous data, 2) counts, 3) proportions, 4) binary data, 5) time at death, 6) time series, and 7) circular data. Each of these types will be described separately below.



Environmental Data... types of data

Continuous Data

- Data in which the observations of the sampling units for the dependent variable can be measured on a *continuum* or scale; can have almost any numeric value; can be meaningfully subdivided into finer and finer increments, depending upon the precision of the measurement system.

Examples:

- Temperature
- Mass
- Distance
- Etc.

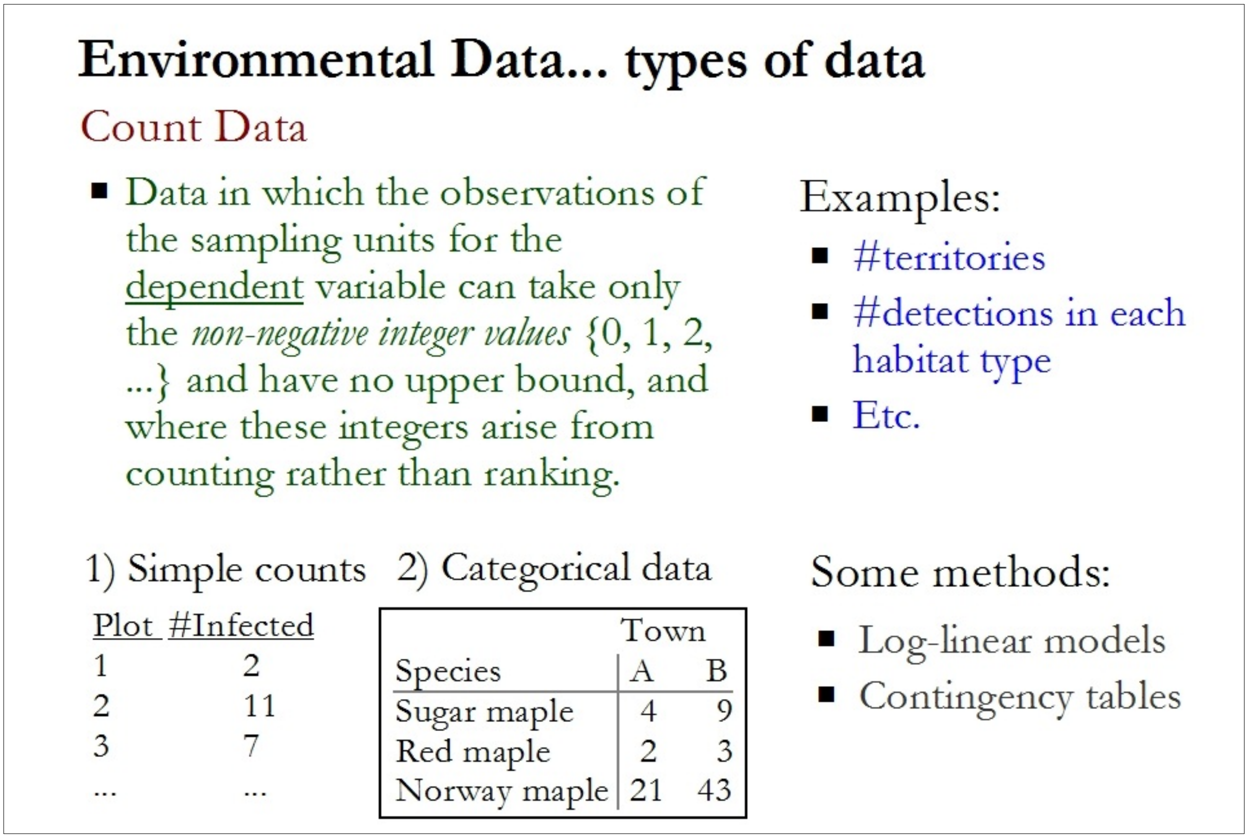
Some methods:

- Regression
- Analysis of variance

7.1 Continuous data

Continuous data is data in which the observations of the dependent variable can be measured on a *continuum*; can have almost any numeric value; and can be meaningfully subdivided into finer and finer increments, depending upon the precision of the measurement system. There are lots of examples of continuous data: temperature, mass, distance, etc. This is the most common type of environmental data collected and there are lots of statistical methods designed to work with this type of data, such as regression and analysis of variance.

The key distinction here is that the dependent variable is continuously scaled; the scale of the independent variable(s) does not matter.



Environmental Data... types of data

Count Data

- Data in which the observations of the sampling units for the dependent variable can take only the *non-negative integer values* {0, 1, 2, ...} and have no upper bound, and where these integers arise from counting rather than ranking.

Examples:

- #territories
- #detections in each habitat type
- Etc.

1) Simple counts

Plot	#Infected
1	2
2	11
3	7
...	...

2) Categorical data

	Town	
Species	A	B
Sugar maple	4	9
Red maple	2	3
Norway maple	21	43

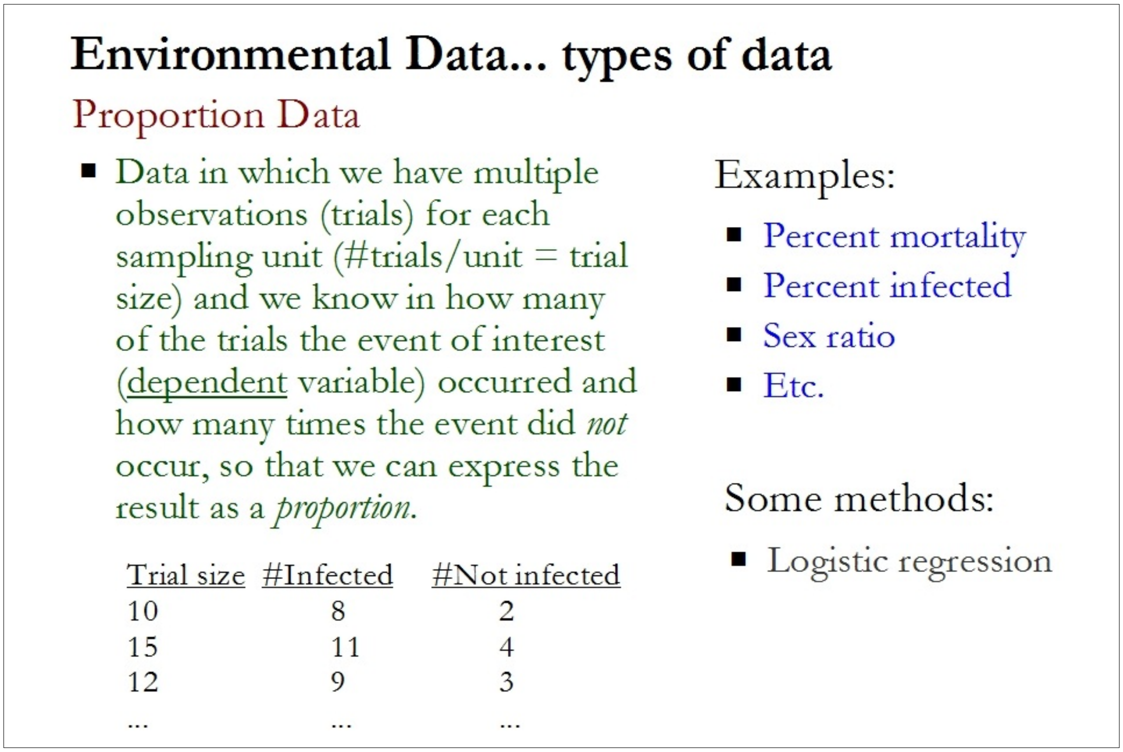
Some methods:

- Log-linear models
- Contingency tables

7.2 Count data

Count data is a form of discrete ordinal scale data in which the observations of the dependent variable can take only the *non-negative integer values* {0, 1, 2, ...}, and where these integers arise from counting rather than ranking. Count data is usually of one of two forms: 1) simple counts, e.g., the number of plants infected by a disease on a plot, the number of eggs in a nest, etc., and 2) categorical (nominal) data, in which the counts are tallied for one or more categorical explanatory variables, e.g., the number of infected plants classified into tree species and town. With simple counts, a count of something (number of infected trees) is made for each sampling unit (e.g., fixed-area plot); in this example the plot is the unit of observation at which variability will be modeled. Note, with simple counts we don't know how many 'somethings' we don't have, so we can't express the result as a proportion.

With categorical data, each sampling unit (e.g., infected tree) is placed in one mutually exclusive category on the basis of one or more categorical factors (e.g., species and town); in this example, the infected tree is the sampling unit at which variability will be modeled. With simple counts, the goal is usually to explain or predict the counts based on one or independent or explanatory variables, and the method of generalized linear modeling is used for this purpose. With categorical data, the goal is usually to determine whether the distribution of counts among categories differs from expected, and the method of contingency table analysis employing log-linear modeling is often used for this purpose.



Environmental Data... types of data

Proportion Data

- Data in which we have multiple observations (trials) for each sampling unit (#trials/unit = trial size) and we know in how many of the trials the event of interest (dependent variable) occurred and how many times the event did *not* occur, so that we can express the result as a *proportion*.

Trial size	#Infected	#Not infected
10	8	2
15	11	4
12	9	3
...	...	...

Examples:

- Percent mortality
- Percent infected
- Sex ratio
- Etc.

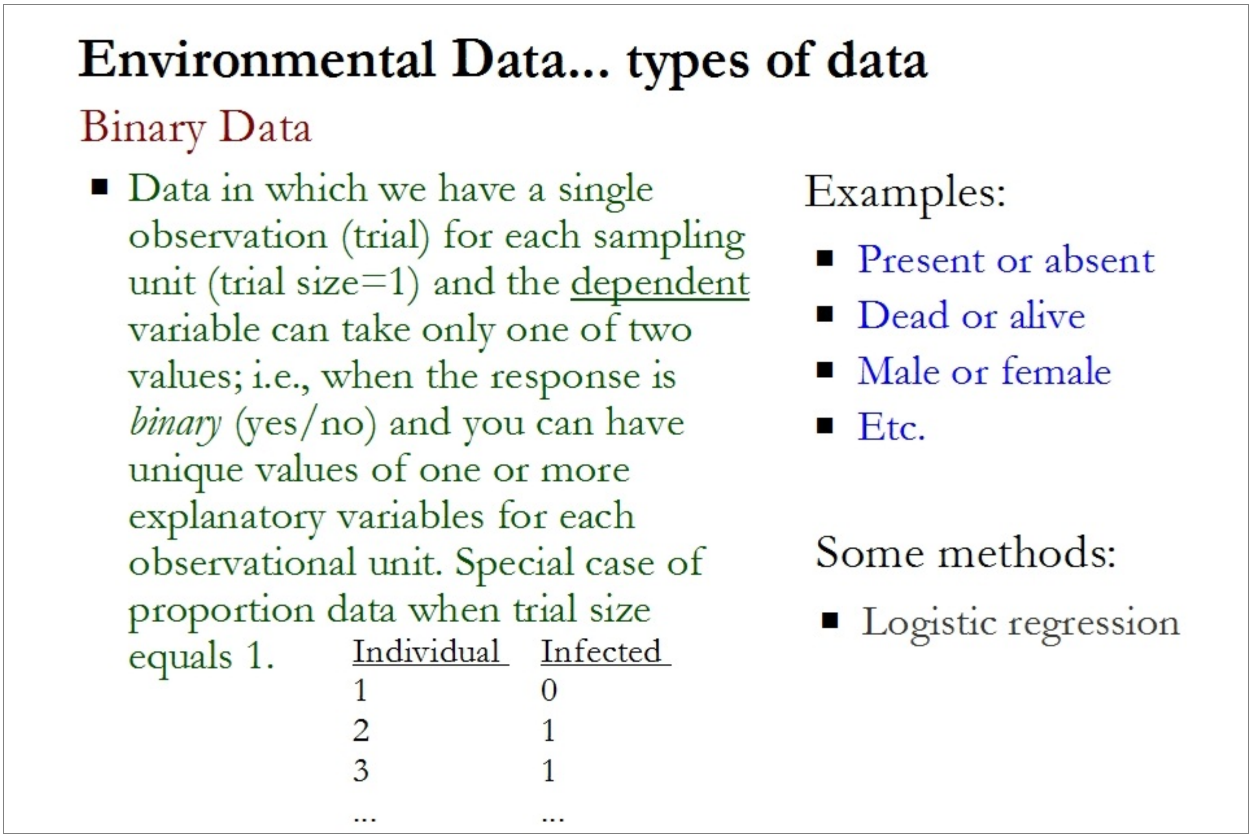
Some methods:

- Logistic regression

7.3 Proportion data

Proportion data is another form of discrete data in which we know how many of the observations of the dependent variable are in one category (i.e., an event occurred) and we also know how many are in each other category (i.e., how many times the event did *not* occur). This is an important distinction, since it allows the data to be represented as proportions instead of frequencies, as with count data. There are lots of environmental examples of proportion data: percent mortality, percent infected, sex ratio, etc.. The key distinction of proportion data is that the frequency of the event, e.g., individual died, is known as well as the total number of events, e.g., total number of individuals. Each event is considered a "trial" and there are one or more trials conducted for each sampling unit. Thus, for each sampling unit, the response (dependent variable) can be expressed as the proportion of trials that were successful. Moreover, the trial size (i.e., number of trials) can vary among sampling units to accommodate various sampling designs. With proportion data, the goal is typically to explain or predict the proportional response based on one or more explanatory variables, and the method of logistic regression is designed for this purpose. Note, here the explanatory variables are measured for each sampling unit, as opposed to each individual trial. This is an important distinction between proportion data and binary data (discussed next).

The key distinction here is that the dependent variable is represented as a proportion; in other words, there are multiple measurements or observations (i.e., trials) made on each sampling unit so that the response can be represented as a proportion. There must be two or more trials or subsamples (i.e., trial size ≥ 2) per sampling unit in order to be able to express the result as a proportion. And the unit of variability of interest is (as always) the sampling unit, not the individual trial.



Environmental Data... types of data

Binary Data

- Data in which we have a single observation (trial) for each sampling unit (trial size=1) and the dependent variable can take only one of two values; i.e., when the response is *binary* (yes/no) and you can have unique values of one or more explanatory variables for each observational unit. Special case of proportion data when trial size equals 1.

Individual	Infected
1	0
2	1
3	1
...	...

Examples:

- Present or absent
- Dead or alive
- Male or female
- Etc.

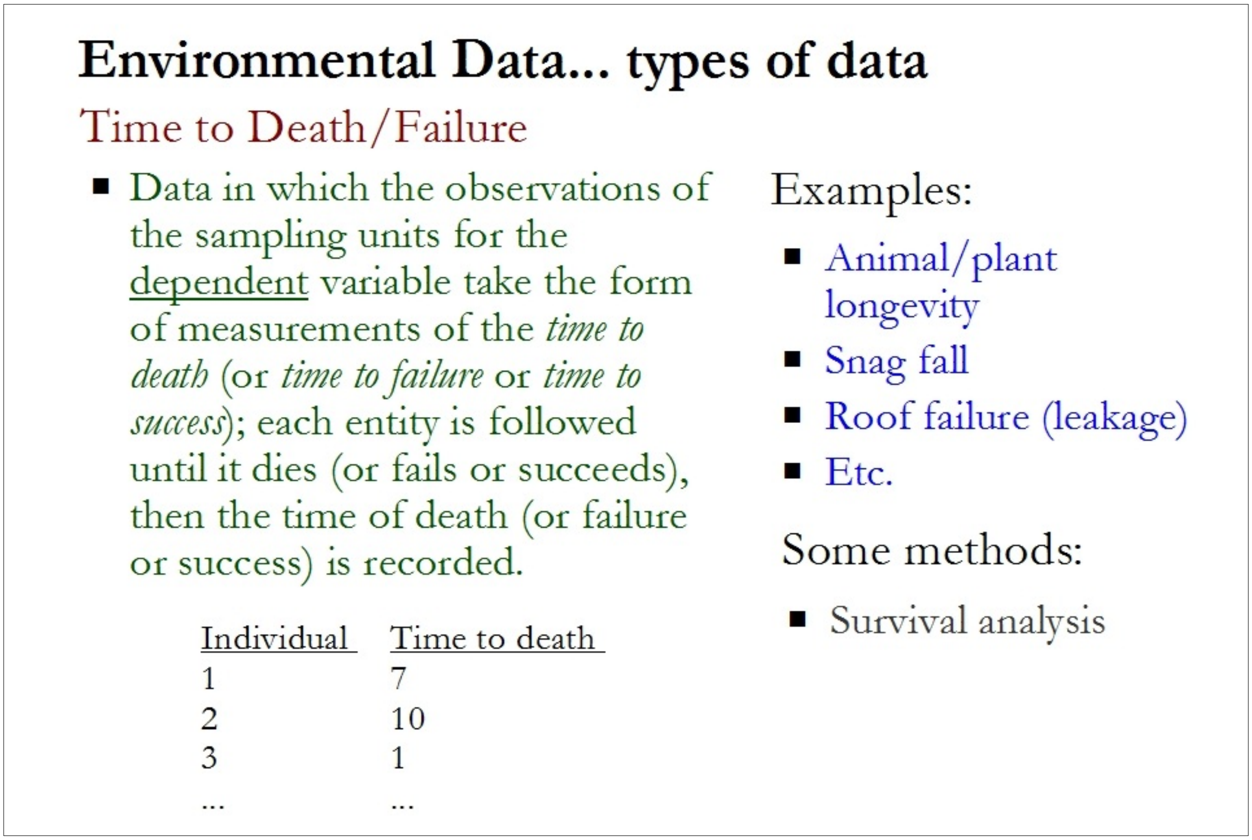
Some methods:

- Logistic regression

7.4 Binary data

Binary data is discrete data in which the observations of the dependent variable can take only one of two values, for example, alive or dead, present or absent, male or female, etc.. Binary data is useful when you have unique values of one or more explanatory variables for each and every "trial", such that each trial is treated as a separate sampling unit. This is an important distinction from proportional data in which the sampling unit consists of multiple trials. Note, in both cases the explanatory data is still collected at the level of the sampling unit; they differ in whether the sampling unit consists of multiple trials (proportion data) or a single trial (binary data). Binary data is typically analyzed with the method of logistic regression, like proportion data.

The key distinction here is that the dependent variable is dichotomous (i.e., can take on only one of two values, e.g., yes or no) and the dichotomous outcome is recorded for each sampling unit. Note, binary data is the special case of proportional data when the trial size is fixed at one for all sampling units.



Environmental Data... types of data

Time to Death/Failure

- Data in which the observations of the sampling units for the dependent variable take the form of measurements of the *time to death* (or *time to failure* or *time to success*); each entity is followed until it dies (or fails or succeeds), then the time of death (or failure or success) is recorded.

Individual	Time to death
1	7
2	10
3	1
...	...

Examples:

- Animal/plant longevity
- Snag fall
- Roof failure (leakage)
- Etc.

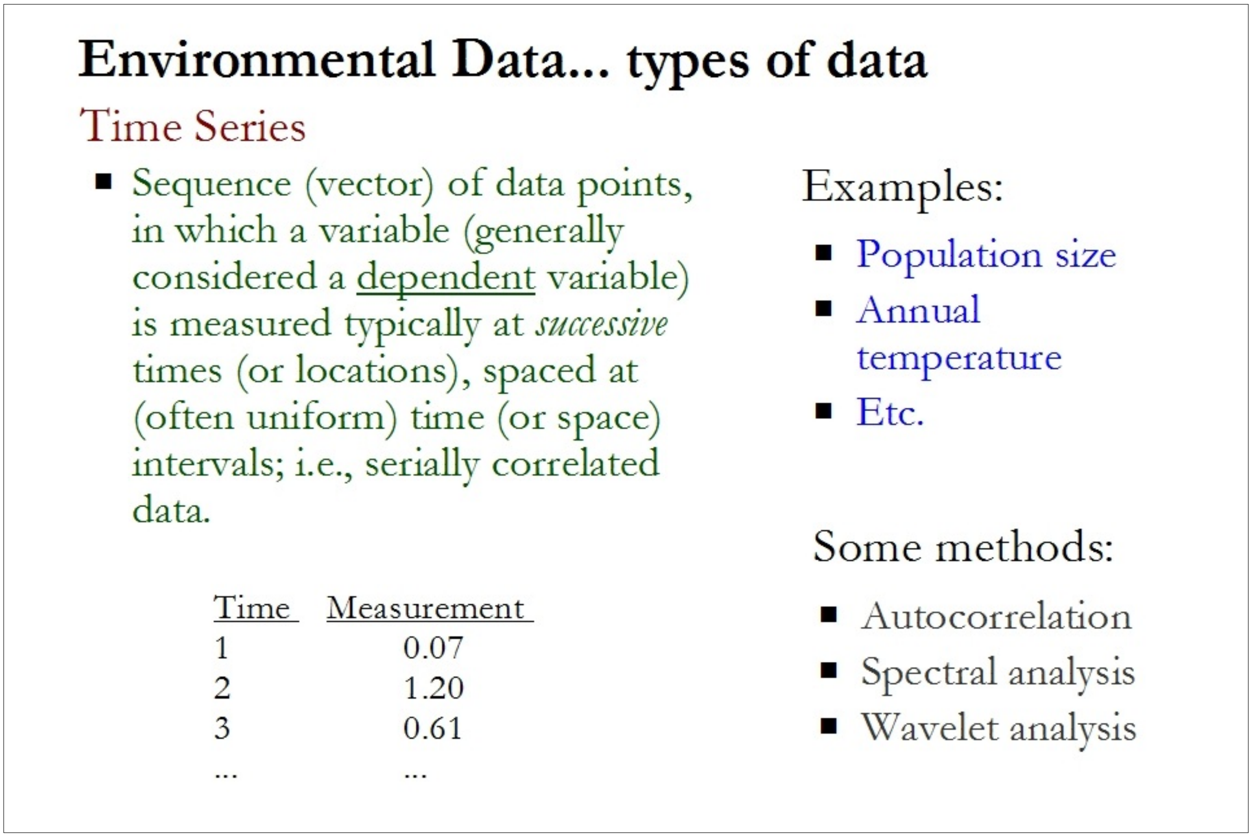
Some methods:

- Survival analysis

7.5 Time to death/failure data

Time to death/failure data is data (dependent variable) that take the form of measurements of the *time to death* (or the *time to failure* or, conversely, the *time to success* of a component); each individual is followed until it dies (or fails or succeeds), then the time of death (or failure or success) is recorded. Time to death/failure data is not limited to plant and animal longevity studies; it applies to any situation in which the time to completion of a process is relevant (e.g., the time it takes juveniles to disperse out of the study area, or the time it takes a snag to fall). Time to death/failure data is analyzed by the special method of survival analysis, which is highly complex and rapidly evolving to account for all sorts of variations in sampling designs.

The key distinction here is that the dependent variable represents the time to death (or failure or success) for each sampling unit. Note, time to death/failure can be measured and treated such that time is considered as a either a discrete or continuous variable, and the form of the stochastic component of the statistical model will vary depending on which scale is chosen.



Environmental Data... types of data

Time Series

- Sequence (vector) of data points, in which a variable (generally considered a <u>dependent</u> variable) is measured typically at *successive* times (or locations), spaced at (often uniform) time (or space) intervals; i.e., serially correlated data.

Time	Measurement
1	0.07
2	1.20
3	0.61
...	...

Examples:

- Population size
- Annual temperature
- Etc.

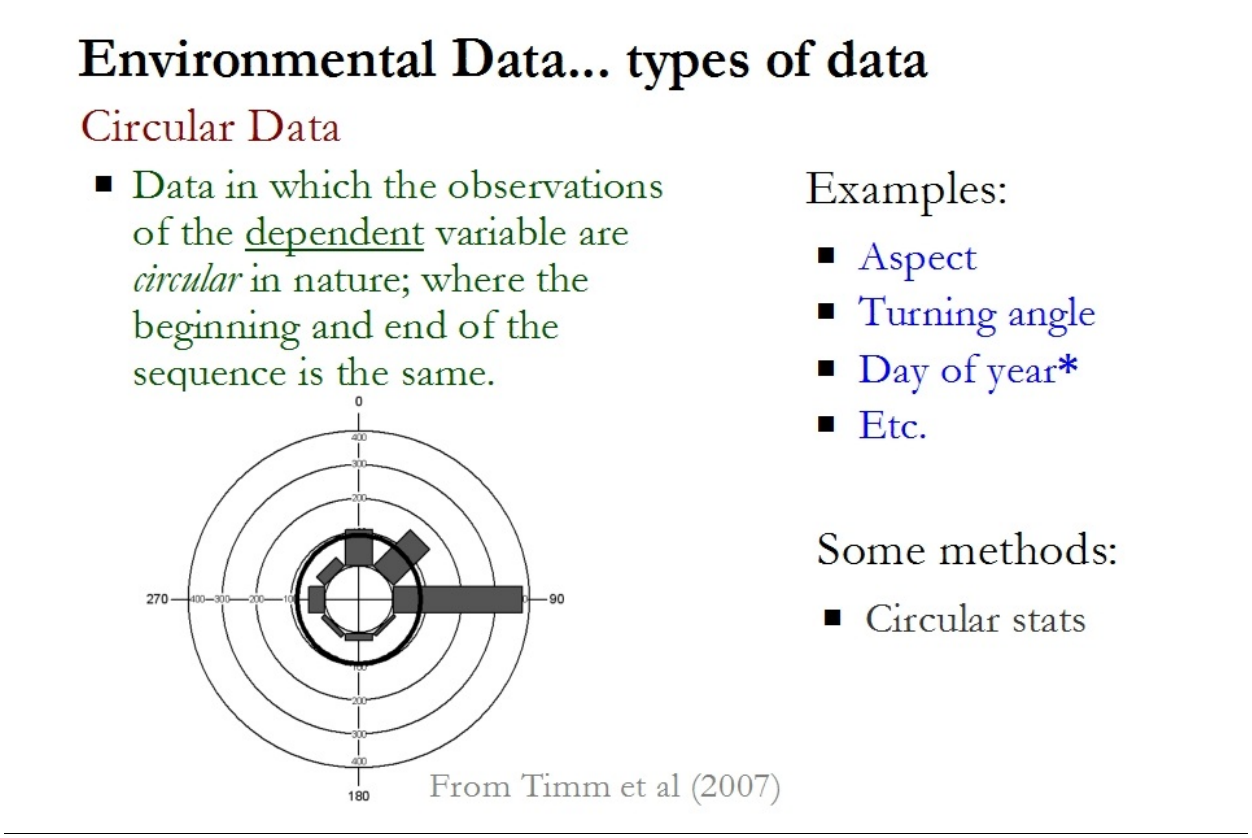
Some methods:

- Autocorrelation
- Spectral analysis
- Wavelet analysis

7.6 Time series data

Time series data involves a sequence (vector) of data points in which the variable of interest (generally considered the dependent variable) is measured typically at successive times (or locations), spaced at (often uniform) time (or space) intervals. Usually time series data contains repeated patterns of variation, and identifying and quantifying the scale(s) of the repeated pattern is often the focus of the analysis. There are many examples of time series data in environmental science: population size measured annually, temperature data measured at fixed intervals, river discharge measured over time, etc. And let's not forget that time series data also includes spatial data that is serially correlated in space rather than time, such as variables measured at intervals along transects, e.g., plant cover, soil chemistry, water depth, etc.. There several specialized analytical methods for time series data, include autocorrelation analysis, spectral analysis, and wavelet analysis to name just a few.

The key distinction here is that the variable of interest (often implied to be the "dependent" variable even if no other "independent" variables are measured) represents a series of measures in time or space and it is the patterns of variability in time or space (i.e., the patterns of rises and falls in the measure) that is of primary interest. Note, is not to be confused with repeated measures of sampling units; i.e., studies in which the sampling units are repeatedly measured/observed over time resulting in a series of observations for each sampling unit. Here the focus is generally to relate one or more independent variables measured for each sampling unit each time it is sampled (i.e., each repeated measure) to the dependent variable, but account for the fact that the observations are grouped by sampling unit, so as to account for the serial correlation of observations within each unit. Note, here the time series is simply a way to observe/measure the system that accounts for changes over time, but the focus is still on the independent-dependent variable relationship, not in identifying the magnitude and scales of repeated patterns of variation in space or time as in true time series analysis. Thus, the distinction between true times series data and repeated measures data is a subtle but important one in terms of study design and objectives, and in the analytical methods employed. However, the phrase "time series" is often used (or misused) in practice to loosely refer to both of these approaches.



Environmental Data... types of data

Circular Data

- Data in which the observations of the <u>dependent</u> variable are *circular* in nature; where the beginning and end of the sequence is the same.

0
400
300
200
100
90
180
270

From Timm et al (2007)

Examples:

- Aspect
- Turning angle
- Day of year*
- Etc.

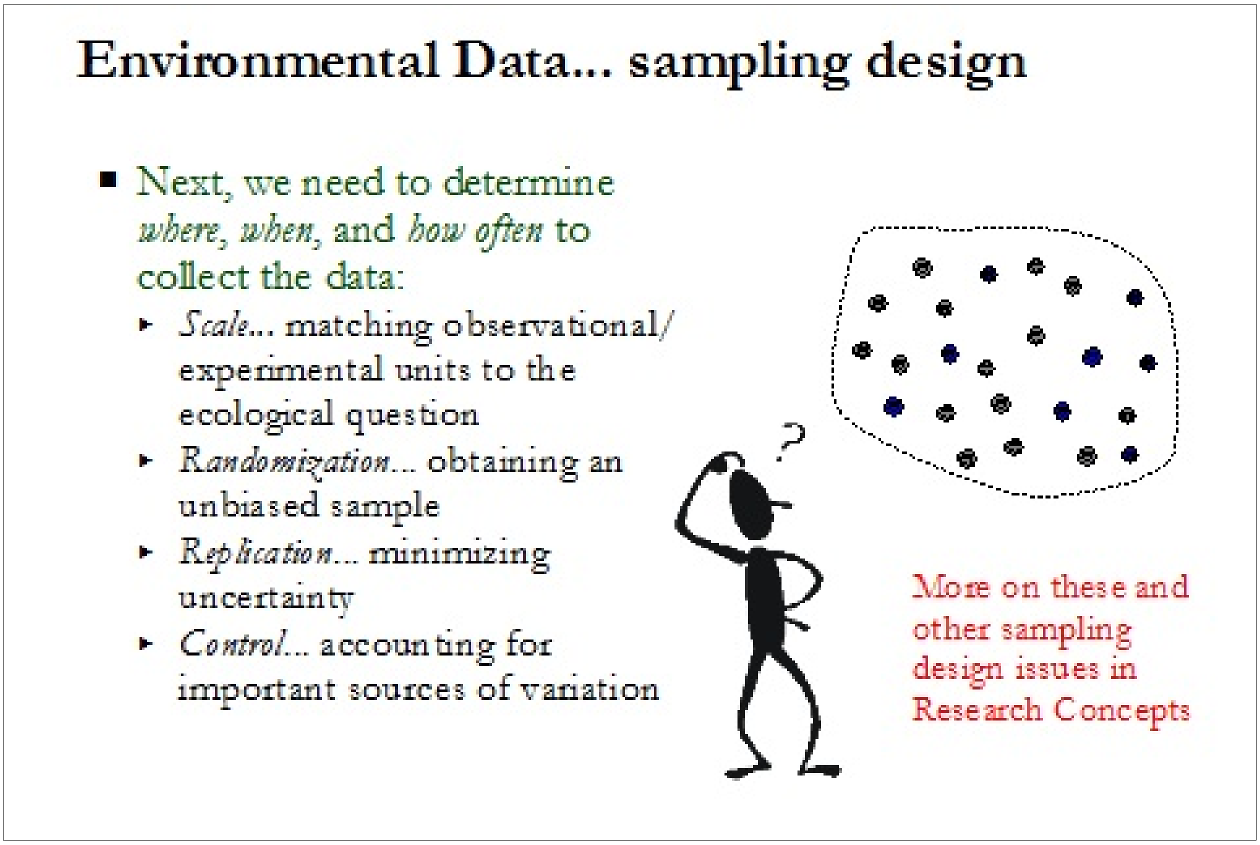
Some methods:

- Circular stats

7.7 Circular data

Circular data is, not surprisingly, data in which the observations of the dependent variable are circular in nature; where the beginning and end of the sequence is the same. Classic examples of circular data are topographic aspect, day of the year, and orientation of movement. The figure shown here is a circular histogram depicting red-spotted newt (*Notophthalmus viridescens*) departure from a uniform distribution for emigrating juveniles leaving a natal pond in western Massachusetts based on eight directional bins. In the histogram, each arm depicts one of the eight directional bins, concentric circles represent a given raw number of individuals, and the bold circle delineates the expected bin value given a uniform distribution. Circular data is typically analyzed with specialized methods that employ special probability distributions, such as the wrapped Cauchy and von Mises distributions, that are designed for circular data.

The key distinction here is that if the dependent variable is circular then circular statistics should be used. Do not confuse this with circular independent variables. For example, aspect is a circular variable and it is often used as an independent variable to explain the distribution and/or abundance of organisms. In this context, as an independent variable, aspect should be transformed into a linearly scaled variable using a cosine transformation (see later chapter) for proper use in the modeling, and circular statistics are not warranted.



Environmental Data... sampling design

- Next, we need to determine *where, when,* and *how often* to collect the data:
  - *Scale...* matching observational/ experimental units to the ecological question
  - *Randomization...* obtaining an unbiased sample
  - *Replication...* minimizing uncertainty
  - *Control...* accounting for important sources of variation

More on these and other sampling design issues in Research Concepts

8. Sampling Design

Once we have determined what data to collect to answer our environmental question, the next thing we need to is determine *where*, *when*, and *how often* to collect the data. This is the complicated arena of sampling design and there are many critical issues to consider, such as:

- *Scale...* matching observational/ experimental units to the environmental question
- *Randomization...* obtaining an unbiased sample
- *Replication...* minimizing uncertainty
- *Control...* accounting for important sources of variation

Each of these issues will be discussed in more detail along with other important study design issues in the Research Concepts course. For now, let's assume the simplest case in which our sampling units are scaled perfectly to match our environmental question, we have designed a simple random sampling scheme in which observations are drawn at random from the population to guarantee unbiased parameter estimates, we have ensured a large sample size to minimize uncertainty in our parameter estimates, and we have measured all important sources of variation (i.e., independent causes of variation in the dependent variable) to minimize the unexplained variation in the model. Now, go out and collect the data!