Chapter 1

Introduction to Environmental Statistics

Chapter Objectives

- Distinguish between categorical and numerical variables.
- Distinguish between discrete numerical variables and continuous ones.
- For a given study, identify the population and the individuals that comprise it.
- For a given study, identify the population parameter of interest, and distinguish it from a sample statistic.

Key Takeaways

- Variables can be any of several types: categorical, numerical (discrete or continuous), spatial location, or date/time.
- Statistics is concerned with drawing conclusions (inferences) about a population using a sample.
- Statistical inference usually involves using a statistic to draw conclusions about a parameter.

1.1 Variables and Data

Lead concentrations measured in soil vary spatially from one soil specimen to the next. A vehicle's hourly nitrogen oxide emissions vary from one hour to the next. Any feature or quantity that varies from one specimen or instance to the next is called a ***variable***, and measured or just observed values of a variable are called ***data***. ***Statistics*** is the science of collecting, organizing, analyzing, and drawing conclusions from data.

Variables are usually one of two types, *categorical* and *numerical.* A ***categorical*** variable is one that takes values in a set of categories, for example, soil type (clay, silt, sand, loam, etc.) and habitat type (forest, shrubland, grassland, etc.). A ***numerical*** variable is one that takes numerical values, for example, the lead concentration or nitrogen oxide emission. Numerical variables can be further split into two types, *discrete* and *continuous.* A ***discrete*** numerical variable is one whose possible values are isolated numbers (such as integers) separated by gaps. Any variable that's a *count*, such as the number of eggs laid by a fish or number of cones on a pine tree, is discrete because it can only take the values $0, 1, 2, \ldots$. A ***continuous*** numerical variable is one that can take *any* value over an entire continuum, or interval. The weight of a fish, for example, or concentration of lead are both continuous because each can take *any* value over an entire continuum.

Two additional types of data are measurements of ***spatial location*** variables (e.g. latitude and longitude) and ***date/time*** variables (e.g. month, day, year).

1.2 Populations and Samples

Statistical inference refers to drawing conclusions about a *population* using data in a *sample* from that population. A ***population*** is a large group of ***individuals*** about which we seek information. The population could consist of people, animals, or items, but in environmental studies, it's often a spatial region or a period of time. In the first case, the population's individuals are the small, location-specific portions of soil, water, or air that make up the region. In the second, they're the time points that make up the period. A ***sample*** is a subset of the population, usually selected ***randomly*** so that the it will tend to be representative of the population, rendering inferences valid. The population and its individuals should be identified as precisely as possible *before* the sample is drawn so that a proper sampling protocol can be developed and followed.

1.3 Parameters and Statistics

Once the sample of individuals has been drawn and the variable measured on them, statistical inferences about the population can be made. Usually, this involves making a statement about the value of a numerical characteristic of the population, such as its average soil lead concentration or its percentage of trees that are infested by bark beetles.

Any numerical characteristic of a *population* is called a ***parameter***. That same characteristic computed from just a *sample* is called a ***statistic***. So, for example, average soil lead concentration over an entire region (the population) is a *parameter*, but the average in a *sample* of soil specimens is a *statistic*.

Example 1.1: Population, Individuals, and Parameters

The South Florida Ecosystem Assessment is a long-term monitoring project in the Florida Everglades initiated by the U.S. Environmental Protection Agency [1]. Environmental and ecological variables were recorded at each of a sample of 757 sites in the Everglades. A few rows of the resulting data set are below.

Florida Everglades Data

STATID	DECLAT	DECLONG	DATE	WEATHER	SOILTKNSFT	SOILTYPE	CO2SDF
M496	26.6	-80.4	36292	CLEAR	13.0	Peat	4.6
M497	26.6	-80.4	36292	CLEAR	12.6	Peat	4.5
M498	26.6	-80.4	36291	OVERCAST	8.5	Peat	2.2
M499	26.5	-80.4	36291	OVERCAST	8.0	Peat	3.3
M500	26.5	-80.2	36291	OVERCAST	1.3	Peat	1.8
M501	26.5	-80.3	36291	OVERCAST	14.0	Peat	3.4
M502	26.5	-80.3	36291	OVERCAST	9.6	Peat	2.7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
M746	25.3	-80.6	36425	CLEAR	0.3	Marl	4.5
M747	25.3	-80.8	36425	CLEAR	0.5	Marl	4.3

Here's a description of the variables in the data set:

Variable	Description
STATID	Sampling station name
DECLAT	Latitude (decimal degrees)
DECLONG	Longitude (decimal degrees)
DATE	Sample collection date
WEATHER	Weather conditions
SOILTKNSFT	Soil thickness (ft)
SOILTYPE	Description of soil type
CO2SDF	Carbon Dioxide in soil (log of μmole/g dry weight)

Here the population is the entire Everglades region and the individuals are spatial locations (sites) within the region. Also:

DECLAT and DECLONG are spatial location variables.

DATE is a time variable.

WEATHER and SOILTYPE are categorical variables.

SOILTKNSFT and CO2SDF are continuous numerical variables.

Three population parameters of interest might be:

The percentage of soil in the Everglades region that's sandy.

The mean soil thickness over the Everglades region.

The mean carbon dioxide concentration in soil over the Everglades region.

Example 1.2: Population, Individuals, and Parameters

To establish a baseline for assessing the effect of a planned mercury (Hg) emissions reduction program, the New York Department of Environmental Conservation measured total precipitation-deposited Hg at a site in the Bronx weekly from January 2008 through September 2010 [3].

Here, the variable, Hg, is a continuous numerical variable, and because it's measured at weekly time points, the population is the time period January 2008 through September 2010. The individuals that make up the population are time points over that period. One population parameter of interest might be the average rate of Hg deposition over the study period.

Statistics used to *estimates* (unknown) population parameter values are called ***estimators***, and ones used to *test of hypotheses* (claims) about their values are called ***test statistics***. A more common use of statistics, though, is to merely *summarize* the information that's contained in a data set. Statistics used for this purpose are called a ***summary statistics***.

Comment: A population may be ***concrete*** (tangible) or ***conceptual*** (abstract). A forest is a concrete population. So is the surface soil over a spatial region. The following example illustrates a conceptual population.

Example 1.3: Concrete versus Conceptual Populations

The accuracy of a radon detector is assessed by exposing it repeatedly to a known, certified concentration of radon and comparing the readings to the true value. If it's exposed to 100 pCi/L of radon 25 times, the readings are unlikely to equal 100 exactly, even the detector is properly calibrated, and are likely to vary from one to the next due to *measurement error* caused by small, unpredictable disturbances (such as breezes, dust particle interference, vibrations, humidity and temperature fluctuations, and so on). The 25 readings can be thought of as a random sample from a *conceptual* population consisting of the (infinitely many) *potential* instances of exposing the detector to the radon.

1.4 Environmental Statistics

Environmental statistics is the branch of statistics concerned with collecting and drawing conclusions from environmental data. It's distinguished from the field of statistics in general by the unique data collection and analysis challenges environmental scientists face, including:

1. The spatial and temporal nature of the data collection process. The target populations from which samples are drawn are often spatial regions, periods of time, or both. This has implications for designing studies and also for analyzing the resulting data
2. Data exhibiting *right skewed* distributions, meaning most of the data are close to zero, but a small fraction of them are very large. For such data, commonly used statistical procedures, developed for use with bell-shaped (*normal*) data, aren't valid.
3. Data that contain *nondetects*, which result when trace amounts of a chemical are measured using instruments that can't reliably distinguish them from zero.
4. Specialized types of studies, including:
   - *Impact assessment studies* of the effects of anthropogenic disturbances on the environment.
   - *Environmental monitoring studies* to track environmental changes over time or to provide early warnings signaling violations of environmental regulatory standards.
   - *Hot spot detection studies* for identifying small areas of unusually high pollution levels known as *hot spots*, often corresponding to pollution point sources.
   - *Laboratory quality assurance studies* to assess the accuracy and reliability of laboratory results.

In the chapters ahead, we'll look at some common statistical procedures used across a variety of disciplines as well as some more specialized ones used in environmental science.

1.5 Problems

1.1 The Iowa Department of Natural Resources selected a random sample of private wells from across the state of Iowa and tested the water in each well for the presence of coliform [4]. The goal was to estimate the percentage of private wells in Iowa that are contaminated by coliform.

a) Identify the population and the individuals that comprise it.

b) Identify the variable observed for each of the sampled individuals.

c) State whether the variable is categorical or numerical, and if it's numerical, whether it's discrete or continuous.

d) Identify the population parameter of interest.

1.2 In a study of the food availability for bald eagles nesting near the shores of Lake Superior, a small sample of nests from the region was observed for a day, and for each nest, the food delivery rate (number of food deliveries from dawn to dusk by adult eagles to nestlings) recorded [2]. The goal was to estimate the average food delivery rate for nests in the region.

a) Identify the population and the individuals that comprise it.

b) Identify the variable observed for each of the sampled individuals.

c) State whether the variable is categorical or numerical, and if it's numerical, whether it's discrete or continuous.

d) Identify the population parameter of interest.

1.3 In a study of heavy metal contamination due to industrialization in the Manali area of Chennai, India, arsenic (As) was measured in soil at a sample of locations in the area [6]. One goal of the study was to estimate the average As concentration in the area's soil.

a) Identify the population and the individuals that comprise it.

b) Identify the variable measured for each of the sampled individuals.

c) State whether the variable is categorical or numerical, and if it's numerical, whether it's discrete or continuous.

d) Identify the population parameter of interest.

1.4 The biological oxygen demand (BOD) in water is the amount of dissolved oxygen needed to break down the organic material present and is used as an indicator of organic pollution (from sewage, plant and animal waste, etc.). To establish a baseline for assessing the effect on water quality of a planned dam on the Tafna River, Algeria, the river's BOD was measured monthly from October 1996 through July 1997, prior to the dam's construction, at a site downstream from its proposed location [7]. A goal of the study was to estimate the average BOD over that study period.

a) Identify the population and the individuals that comprise it.

b) Identify the variable measured for each of the sampled individuals.

c) State whether the variable is categorical or numerical, and if it's numerical, whether it's discrete or continuous.

d) Identify the population parameter of interest.

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] Cheryl R. Dykstra et al. Low reproductive rates of Lake Superior bald eagles: Low food delivery rates or environmental contaminants? *Journal of Great Lakes Research*, 24(1):32–44, 1998. International Association of Great Lakes Research.

[3] Dirk Felton and Kevin Civerolo. Air monitoring plan for establishing an ambient mercury baseline for New York State. Technical report, New York State Department of Environmental Conservation, Apr 2011. Local-Scale Air Toxics Ambient Monitoring Project In Response to EPA RFA NO: OAR-EMAD-05-16.

[4] G.R. Hallberg et al. The Iowa state-wide rural well-water survey: Site and well characteristics and water quality. Technical report, Iowa Department of Natural Resources, Geological Survey Bureau, 1992. Technical Information Series 23, 43 p.

[5] Dale Hess, Marie-Collette van Lieshout, Bill Payne, and Alfred Stein. A review of spatio-temporal modelling of quadrat count data with application to striga occurrence in a pearl millet field. *International Journal of Applied Earth Observation and Geoinformation*, 3(2), 2001.

[6] A. K. Krishna and P. K. Govil. Assessment of heavy metal contamination in soils around Manali Industrial Area, Chennai, Southern India. *Environmental Geology*, 54(7):1465–1472, 2008.

[7] Amina Taleb, Nouria Belaidi, and James Gagneur. Water quality before and after dam building on a heavily polluted river in semi-arid Algeria. *River Research and Applications*, 20:943–956, 2004.