Notes

1 Introduction

MTH 3240 Environmental Statistics

Spring 2020

Topics

Objectives

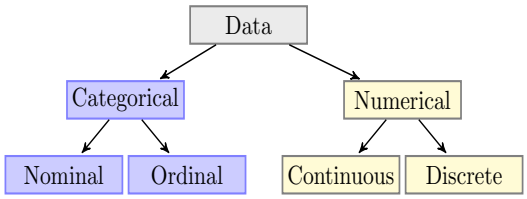
Objectives:

- Distinguish between categorical and numerical variables.
- Distinguish between discrete numerical variables and continuous ones.
- For a given study, identify the population and its individuals.
- For a given study, identify the population parameter of interest.

Variables and Data

- ***Statistics*** is the science of collecting, organizing, analyzing, and drawing conclusions from *data*.
- ***Data*** are measurements, or observed values, of variables.
- A ***variable*** is any characteristic that varies from one ***individual*** to the next. The ”individuals” can be **spatial locations** or **time points**.

  Examples: Lead concentrations in soil specimens, hourly nitrogen oxide emissions from vehicles at a particular intersection.



Notes

- Variables can be ***categorical*** or ***numerical***.
- A ***categorical*** variable takes values in a set of categories.

  Examples: Soil specimens classified by soil type (clay, silt, sand, loam, etc.), plots of land classified by habitat type.
- A ***numerical*** variable takes numerical values.

  Examples: Lead concentrations in soil specimens, densities of rock specimens.

Notes

- Numerical variables are ***discrete*** or ***continuous***.
- A variable is ***discrete*** if the possible values are isolated numbers (such as integers) with gaps between them.

  Example: The numbers of eggs laid by fish during a given spawning season.
- A variable is ***continuous*** if its set of possible values forms an entire continuous interval.

  Examples: Weights of fish, concentrations of a pollutant in water.

Notes

- Two additional types of data are measurements of
  - ***Spatial location*** variables (e.g. latitude and longitude).
  - ***Time*** variables (e.g. date).

Populations and Samples

- ***Statistical inference*** means drawing conclusions about a *population* using a *sample*.
- A ***population*** is a large group of individuals or items about which we seek information.
- A ***sample*** is a subset of the population's individuals that's selected in some prescribed manner.
- ***Randomly selected*** samples tend to be representative of the population.

Notes

- A ***statistic*** is any numerical value computed from a **random sample**.
- Statistical inference involves using **statistics** to ***estimate*** or ***test hypotheses*** about numerical characteristics of the population called ***population parameters***.

Notes

Example

The South Florida Ecosystem Assessment is a long-term monitoring project in the Florida Everglades initiated by the EPA.

Environmental and ecological variables were recorded at each of a sample of 757 sites in the Everglades.

Notes

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

Notes

Below is 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)

Notes

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.

Notes

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.

Notes

Example

The NY Department of Environmental Conservation measured total precipitation-deposited Hg at a site in the Bronx weekly from January 2008 through September 2010.

Here, because Hg is measured at weekly time points, the **population** is the time period from January 2008 through September 2010.

The **individuals** in the population are time points making up that period.

One population **parameter** of interest might be the mean weekly wet Hg deposition over the period.

The **variable**, Hg concentration, is **numerical**, and it's ***continuous***.

Notes