Introduction to Data Management, Life History, and Demography

Data are the backbone of scientific research and exploration, so learning how to collect and process data efficiently is a critical skill for scientific professionals. Most people are not born understanding how to collect, record, enter, and analyze data, but with guidance and practice you can learn how to handle information and create world-class datasets.

Scientific organizations, especially large ones, spend a lot of time and effort determining the best ways to process data. The National Ecological Observatory Network (NEON) is one such organization that has made efficient data collection and processing a priority. NEON was designed to collect long-term ecological data on a continental-scale to help researchers address questions related to climate change, land-use, and invasive species. Data are collected at field sites called domains using standardized protocols, which allow for comparisons across large geographic ranges. Data on dozens of different variables and species will be collected every year for 30 years, yielding a comprehensive look at ecological processes across the entire United States.

As you watch this introduction to the NEON program, think about how much data (and what types) are being collected and the challenges associated with managing these data.



NEON Intro

Given the scale and scope of the project, NEON will create literally terabytes of data every year, so the information needs to be very well organized to be useful. In this module, you'll become familiar with data management practices and then use these skills to collect data, and then construct a life table and survivorship curve for humans.

Learning Objectives

At completion of this learning module, you will be able to:

- explain the importance of data management in the scientific process
- identify elements of a useful, organized data sheet

- create a spreadsheet data table
- manipulate data in a spreadsheet program
- calculate vital statistics of a population using life tables
- collect, manage and analyze data to test hypotheses about human demography

Data Management

What is data management?

Essentially it is organized, efficient data collection. A field or lab notebook may look like this:

but ultimately you need to turn it into a format more like this:



for ease of analysis and sharing with others. Data management encompasses everything about how data are collected, processed, stored, and archived. Data management procedures and practices should be an integral part of any research project planning and should be in place prior to any data being collected.

Why do it?

Data are essential to the process of science. Even after an initial analysis is done and published, the data are essential to keep track of and archive.



Think about why it is important to archive/back up your data. There are a lot of physical threats to data storage, even if it's in the cloud.



And storage formats do change over time....



And finally, it is important that the research (sampling and recording data) can be repeated regardless of changes in the project or staff.



For any project, scientists need to think about:

- what data will be recorded
- how the data will be recorded
- safety (people, organisms, data)

It is also important to note that field collection doesn't have to be limited to the "field" in the way that we often think about being outside, it could mean data collected in the lab, or gathered from other papers or even data repositories.

Best Practices:

How is it done?

Data Files

When you think about what the data sheet for a project should look like in order to make it readable and accessible to as many people as possible, think about the content of the data and how it should be presented. For example:

1) Columns (variables) & Rows (data)

- Single row of descriptive headers avoid spaces or starting headers with numbers
- Data disaggregation make sure there is only one cell per variable (e.g., toe length & tail length in separate columns)
- Each cell has one type of data Cell should only contain numbers or letters.
- Use plain text

2) Used standardized formats whenever possible (e.g. date and time)

- Date: YYYY-MM-DD (Year-Month-Day)
- Time: hh:mm:ss (use 24-hour time)

To put these together, separate the date and time with the letter "T": therefore, 6:18pm on December 7, 1941 would be written as 1941-12-07T18:18:00. **Practice:** how would you write your birthday as a combined date & time?

• Time Zones: At the start of a project you must decide if you will record you data in local time (easier if working in a single area) or in UTC (Coordinated Universal Time, aka Greenwich Mean Time; better if making comparisons across large spatial scales). You

can signify a date/time is in UTC by adding a Z at the end of the date/time: 1941-12-07T18:18:00Z. You must include your time keeping methods in the metadata (see section on metadata below).

3) Use full taxonomic names - Genus and Genus species

4) Retain separate files of raw and analyzed data. Then you always have a "fresh" dataset to work with.

5) Use easily transferrable file formats & hardware

- .csv format, not .xls
- Internet/cloud storage & backup
- Non-proprietary formats

6) Descriptive file names (no spaces)

7) Long-term data storage/archiving

Metadata

Metadata are data about data. They describe the content, quality, condition, and other characteristics of a dataset.

Metadata record answers questions such as:

- Why was the dataset created?
- What processes were used to create the dataset?
- What geographic projection are spatial data in?
- When was the dataset last updated?
- Who created the data?
- What scale was used?
- What fields are in the table?
- What do the values in those fields mean?
- Who do I contact about getting more information about the data?
- How do I obtain a copy of the data?
- Do the data cost anything?
- Are there any limitations to the data?

Thus, metadata are valuable tools. Metadata records preserve the usefulness of data over time by detailing methods for data collection and dataset creation. Metadata greatly minimize duplication of effort in the collection of expensive data and fosters the sharing of digital data resources.

You may be thinking, why should I make such a fuss about my data files?

Have you ever thought to yourself during a class laboratory exercise that there is no way you will forget what the methods were, only to go home to find that you don't remember the details when writing the lab report?

Or have you had to compile data with other students for a group project, and you can't make sense of something one of your group members recorded/contributed?

Now imagine you are doing research for an independent study, and when you graduate, the dataset you leave behind is not clear to your professor or lab mates. That can be a huge problem for scientists who must combine data with collaborators, justify findings to a grant agency, relay findings to the general public and publish the work.

Some projects are very simple, and data management can be a fairly easy task. However, the more complex the project, the more important management becomes.

Example:

Scientific organizations, especially large ones, spend a lot of time and effort determining the best ways to process data. The National Ecological Observatory Network (NEON) described in the introduction to this module is an ecological observatory funded by the National Science Foundation that has made efficient data collection and processing a priority. Recall that NEON was designed to collect long-term ecological data on a continental-scale to help researchers address questions related to climate change, land-use, and invasive species. Data are collected at field sites called domains using standardized protocols, which allow for comparisons across large geographic ranges. Data on dozens of different variables and species will be collected every year for 30 years, yielding a comprehensive look at ecological processes across the entire United States.

NEON provides:

- •Free and open data on the drivers of and responses to ecological change
- •A standardized and reliable framework for research and experiments

•Data interoperability for integration with other national and international network science projects

An example of a NEON project is small mammal trapping. Small mammals were chosen by NEON to be bioindicators because they are present across the country in a wide variety of habitats. Their small size and short lifespan make them sensitive to environmental changes, and they are responsible for spreading or maintaining a wide diversity of diseases in an environment. They are also easy to safely collect as live specimens using arrays of traps. Live trapping has the advantage of being able to return the animal to their habitat without having to destructively sample. In just a few minutes, a lot of information can be collected from an individual animal. Because researchers want to reduce the stress on the animal while it is captured, it's important to have an efficient framework for recording that data.

Please view the following video. See if you can identify what variables need to be recorded in each of the "column" categories on a datasheet. Also think about why data management is a concern with a project like this one.



NEON Small Mammal Monitoring

Hopefully you were able to pick up on some of the following types of information being collected as part of this project:

- Plot ID
- Date of capture
- Species
- Individual ID
- Sex
- Weight
- Etc.

Think about how you might build a data sheet to include all of this information.

In reality, the datasheets for this project are VERY complex. Does the data sheet below look anything like what you imagined? Why or why not?

Think for a few minutes about the data that will be collected in these sheets by all types of observers. Do you think the NEON data sheets are well designed for all researchers to use? To transfer the information from the field to an electronic data file? What challenges do you see? What could make the process easier?

Also, what if you were responsible for entering data from hundreds of data sheets. How would you make sure you were not making mistakes? What types of checks could you do to make sure you were correctly transferring the data?

Let's try to keep these data management ideas in mind as we learn about life history and prepare to analyze data to answer some questions about human demography.



Then Provide Background Material on Life History and Demography

I also gave them a life table practice activity before the Demography Assignment.

I also made it mandatory that they complete the discussion before they were allowed to see the assignment to avoid seeing any possible answers to the discussion questions.

Human Demography Discussion:

For this discussion, let's turn our attention to a population that has grown at an incredibly rapid rate over the last 400 years - humans. The study of human demography is important, but this year everyone is extremely aware of the potential for human population change. First, 2020 is a census year. You or your parents should have received instructions for completing the US 2020 census in the mail. The census is how the number of people in this country is determined, along with information about the age and sex of every individual in every household. The census is only conducted once every 10 years.

And now 2020 is also the year of the COVID-19 pandemic. Every day the news is filled with statistics of numbers of cases and deaths due to the disease, both in the US and around the world.

Although humans are considered a global population, there are differences in how changes in human populations occur. These population dynamics can differ depending on **where** the humans live (spatially) and **when** the humans live(d) (temporally).

1) What specific spatial and/or temporal factors could cause differences in population dynamics (list **at least two**)?

2) Choose one factor you listed in #1. What kind of data would you need to collect to determine whether that factor affects population dynamics, and where might you find those data?

3) Do you think COVID-19 will have a measurable effect on human population dynamics? Why or why not?

4) Respond to at least two classmates - provide thoughts on their responses to question 2 or 3 (or both).

Human Demography Assignment:

As you've worked through this module, it should now be clear that life tables and survivorship curves are useful in helping to understand the interaction between an organism and its environment. We've addressed how species vary in schedules of mortality and reproduction. For example, oysters experience high mortality early in life but produce numerous offspring annually, whereas elephants have a high probability of survival after birth but females produce only one calf at a time.

The life table you filled out provided many different columns of information about a rabbit population, including information related to mortality and reproduction.

We've also learned about survivorship curves, graphical representations of the numbers (or fractions) of individuals born at the same time (a cohort) that die at a given age. There are species that have (a) low mortality at a young age, (b) constant mortality throughout life, or (c) high mortality at a young age (Types I, II, and III, respectively). These curves can help us determine possible causes of population limitation—periods of heaviest mortality that have the greatest impact on population growth.

Knowing patterns of death and reproduction can be important in managing plant and animal populations. For example, fisheries managers might adjust the allowed catch after good and poor years of reproduction; managers trying to eliminate invasive plants might be able to target specific ages of plants for removal; and managers seeking to protect rare species might know what ages of individuals most need protection.

Now, how do we determine survivorship curves and calculate life tables for humans?

We already discussed that human population dynamics vary in space and time, and you had to come up with some reasons for why that might be. You also had to think about what kind of data you could collect, and how to collect it.

If you said you could use census data, great job! That is definitely one way to get data to answer questions about differences in human populations dynamics. Another way is to visit a cemetery (or cemeteries) to collect data. This is not always practical, and certainly during social distancing we should not be out and about visiting cemeteries. Fortunately, the amount of data publicly available online is incredible, along with the capacity for virtual searches using sites like <u>findagrave.com</u>.

For the first part of this assignment, you will have an opportunity to collect your own demography data from this **"virtual cemetery"** site to make your own data sheet and begin to explore how you would compare survivorship between two populations. For example, you could compare males vs. females, people living at different times, people living in different places, or, perhaps, people of different ethnicities or socio-economic backgrounds. There are many possible interesting comparisons.

For the remainder of the assignment, you will use a dataset I collected to construct survivorship curves and life tables to compare two populations.

Download the Word and Excel files to begin the assignment.