Topic 2: Data Analysis

Before making any decisions…
Take a look at your data

Diversity of Biological Data

The main reason to analyze collected data is to reach conclusions or some tip of information that allows us to **make decisions**, **apply certain policies**, or develop new theories from which to continue this marvelous path of knowledge called **science**.

Unfortunately, data analysis is not an easy task, and the user needs to know some basic and not so basic **statistics,** apart from the specific knowledge required in the field from which the data comes. Even if the data is treated by professional statisticians, the user needs to formulate **adequate questions** to obtain adequate answers, and this requires a certain mastery of the techniques through which data can be analyzed.

But first, it is very important to get a feeling of your data, to familiarize yourself with it. By doing so, we make it less likely to reach the wrong conclusions in the analysis, and it will be easier to understand the outputs of it.

TYPES OF BIOLOGICAL DATA:

- **Data on a Ratio Scale:** Data coming from assigning numerical values to natural properties. What defines a ratio scale measurement is that the intervals or differences can be compared (i.e., the difference between 30 and 40 cm is higher than 45 to 50 cm. 1 cm is twice 0,5 cm, and so on). Ratio scale possesses a zero point or character of origin which any other type of data possesses.
- **Data on an Interval Scale**: Some measurement scales have a constant interval size but there is no true zero. Temperature is a good example of it. Even though 10 is two times 5, 10°C is not two times 5°C (consider the same temperature in Fahrenheit, it would be 50°F and 41 °F respectively).
- **Data on an Ordinal Scale**: In some cases, variable differences are relatively measured, rather than quantitatively (i.e. larger vs. smaller fish).
- **Data on a Nominal Scale**: Sometimes, the variable to study is classified by some quality rather than a numerical measurement (for example, colour, names, etc.).

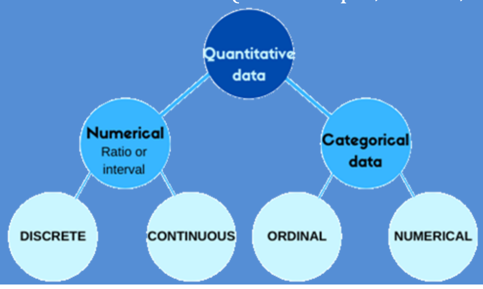

Figure 1. Typologies of data

Types of variables:

Within a dataset, there may be different measurements that refer to characteristics that vary from one entity to another, those characteristics are called **variables**. The amount of phosphate in the soil, the expression of a gene in different tissues, the number of individuals that carry a certain illness or the temperature of different organisms, are examples of variables, and they present essential differences with which a standard classification can be done (box „Types of Biological Data“). When aiming to relate causes and consequences of the change of any variable, scientists differentiate between **predictor variables**, which are the ones driving a change of a specific variable, and **response variables,** which are the variables responding to a change in the predictor variables.

Data coming from a Ratio or an interval scale will always have associated values and units in it, and they can either be **continuous**, in which the difference between one value and the next one is infinitesimal (i.e. length, or weight of an organism), or **discrete**, in which no decimals are considered (i.e. the number of rabbits in a field).

RESEARCH QUESTION:

Before starting any kind of analysis, it is of utmost importance to think of the questions that the user wants to answer during the analytic process. The questions normally arise from an initial hypothesis, which is required to statistically test for significant differences with our data. The Hypothesis must be falsifiable. When looking for deer in a garden, the hypothesis could be either:

1. There is a deer in my garden. ✓
2. There are no deer in my garden. ✗

The first statement is not falsifiable. It could always be that the deer hides every time that the garden is studied, or that the sample area is avoided by the deer. It is not possible to prove that the deer is really not in the garden, since the fact that we do not find it cannot prove that it is not there. However, the second statement turns false at the very moment when a deer is found in the park. **The only way to build scientific knowledge is by proving wrong falsifiable statements.**

Matematical approaches to reality are always simplifications of the reality. Predictor variables can become response variables in a certain context, and the same is true for response variables. In the end, the world is a complex system of interconnected variables that have many effects on many scales. Scientists try to simplify this complex net of interactions into small individual pieces.

REFERENCES

1. Zat, Jerrold H. *Biostatistical Analysis.* Upper Saddle River, N. J: Prentice Hall, 1996. Print.

ABOUT THIS POLICY BRIEF

This Policy Brief is part of a series aiming to inform policy-makers on the key results of the NeDiT research project and provide recommendations to policy-makers. The series of NeDiT Policy Briefs can be found at http://nedit.net/downloads/. This publication was commissioned, supervised and produced by NeDiT project partners.

IMPRINT

Authors:

Cescc Gordo[1], Hauke Kegler[1], Daudi Msangameno[2], Christopher Muhando[2] and Hauke Reuter[1]

The authors work at, or are affiliated with, [1]Leibniz Center for Tropical Marine Ecology (ZMT) and [2]Institute of Marine Sciences Zanzibar (IMS).

You can find more information about the project here.

The NeDiT project was funded as part of the MeerWissen Initiative by GIZ, you can find more information here.