Topic 2: Data Analysis

Before making any decisions…
Take a look at your data II

Plotting is understanding

When the user is already familiarized with the type of data that needs to be analyzed, it is time to step into the data itself: to understand how it behaves and to be able to describe it. Most importantly, to realize how and with what frequency our data vary. If one wished to study the population of the green turtle (*Chelonia Mydas*) on an island, the first step would be to decide which kind of data should be recorded. In this case, an example would be the number of sea turtles spotted per day. After realizing that this is a **discrete numerical variable**, one could graph the turtles observed in a day during every day through a **scatter plot** (Figure 1), but another option could be to plot the "number of turtles per day" in the x-axis, and "days" in the y-axis. In doing so, one would obtain a frequency distribution, which would show how often each specific ratio "turtles per day" is observed (Figure 2). If this was done infinite times, one would obtain the infinite limit of the **frequency distribution**, which is the probability of getting a specific frequency, so the **probability distribution** of that specific variable (Figure 3). The probability distribution of a variable reflects exactly how this variable behaves in nature, which is what the researcher is interested in.

TYPES OF BIOLOGICAL DATA:

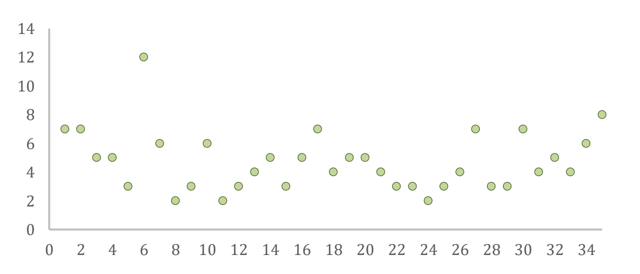

Figure 1. Daily turtles record in the island. X axis accounts for days.

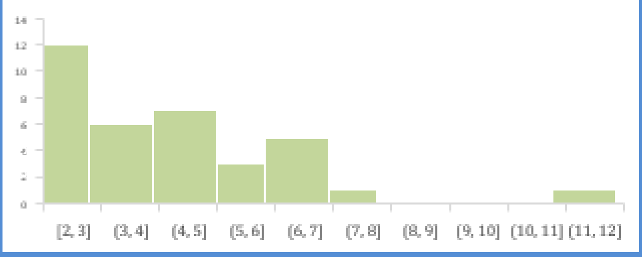

Figure 2. Turtle daily records frequency distribution during one year. X axis represents the number of turtles per day.

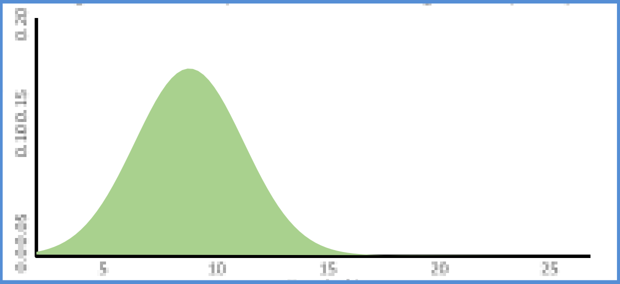

Figure 3. Probability distribution of a specific daily record. X axis

In a similar way, the probability of a certain number of events occurring in a certain amount of time follows a Poisson distribution, and there are some other distributions for other types of data (binomial, negative binomial,...). From here, it is plausible to assume that data comes from some of these probability distributions (the scientist has to find out from which!), and it is possible to explore the reasons for which the data diverges from it.

The statistical tests are the method to calculate the probability of the data belonging to specific probability distribution. If the data is different enough from this distribution, it is assumed to be some explanation (maybe the change of a variable) causing these differences. Then, we assume that the variable under study is the driver of this difference. However, the study has to be perfectly designed to control all the other variables, and the selection of the initial probability distribution has to be cautiously considered!This kind of relationships is what scientists try to establish to understand natural functioning.

Describing the data, describing the world

It is not possible to count all the turtles that come to each beach, every day. Therefore, it is not possible to refer to all the population of turtles on the beaches. However, we can count the turtles at some random beaches, some random days, and from these data extrapolate conclusions to the rest of the population. This process is called **sampling**, and the data obtained through it is the **sample**, which is only a part of the whole **population.** Once a sample is obtained, information of the whole population can be obtained. The characteristics of a population are called **parameters**. The **mean** or the **median** are parameters that try to capture the central tendency of the population while the range or the variance tries to capture the dispersion of the population; how much a specific variable varies across a population. Since the populations are normally not reachable, the parameters are seldom measured, but estimated from the sample. The parameters of the sample are called **statistics**. The mean of the sample, for example, is a good estimator for the mean of the population.

Looking for trends and correlations

Plotting the data across time is sometimes the most direct approach to visualize how the data behave, but the data can also be plotted against any other variable, such as temperature, amount of light, pH, or whichever variable from which the researcher has data from (strictly speaking, time is only a reflection of another process or variable changing since time alone does not explain change). Plotting the data is a very useful way of looking for **correlations**, meaning common ways of variation among variables. If the number of turtles on the beach increase day after day, it could be argued that the turtles increase as the time increases (since the beginning of the sampling). The reason for which more organisms are observed is not known, but the quantity of turtles seems to be **positively correlated** with time. Another case would be that the number of turtles decreased as the temperature increased, which would be a **negative correlation.** When variables are correlated with time, the term “trend” is used to designate the way of the variation.

Remember that correlation and causality are completely different things:

The number of flowers on an island and the number of sea turtles can correlate if both decline with a temperature increase. However, none of them will ever directly cause the decline of the other, neither the increase.

Probability

REFERENCES

1. Zar, 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 ZMT research projects and provide recommendations to policy-makers based on research results. The series of ZMT Policy Briefs can be found at www.leibniz-zmt.de/policy_briefs.html. This publication was commissioned, supervised and produced by ZMT. DOI: 10.21244/zmt.2016.001

DISCLAIMER

The policy recommendations made do not necessarily reflect the views of the ZMT or its

IMPRINT

Authors:

The authors work at, or are affiliated with, a Leibniz Center for Tropical Marine Ecology (ZMT).

You can find more information about the project here.

Published by the Leibniz Center for Tropical Marine Ecology
Fahrenheitstr. 6, D-28359 Bremen, Germany
Editor: Bevis Fedder E-Mail: bevis.fedder@leibniz-zmt.de
Phone: +49 421 23800 -67 Homepage: http://www.leibniz-zmt.de