Introduction to Biological Data Analysis and Statistics

Steps in the process of understanding data:
1. Collecting the data
2. Summarizing the data
3. Analyzing the data
4. Interpreting the results and reporting them

Note that before carrying out any of the above, there is presumably some underlying question or hypothesis you have formulated, which you wish to use the data to address. There are a few key types of approaches through which we can address scientific questions:
(i) observation (natural history - see what occurs where and interpret the results based upon differences in the locations or history);
(ii) experiment (vary aspects of the environment in order to tease apart how the biological components respond); and
(iii) theory (make assumptions about the natural world and analyze the implications of those assumptions using verbal, graphical, and mathematical arguments).
Each of these approaches involves quantitative approaches, and an objective of this course is to provide you with an understanding of some of the methods needed.

Step 1 above involves the area of "design of experiments" in which the process by which the data are to be collected is determined based upon the objectives of the study and the limitations imposed (e.g. cost, time, available personnel, accessibility of the study area, etc.). Design implies that the scientist considers alternative methods to collect the data as well as the manner in which the factors deemed to affect the data collection are manipulated. Examples would be deciding where and when to put out traps to collect animals in the field, how many replicates of an evaluation test to use in determining the efficacy of a new drug, how to many different levels of growth medium with what nutrient constituents to use in evaluating the impact of a new antibiotic on bacterial population growth in the lab, or in determining the response of an organism's respiration rate to temperature, how many different temperature treatments are applied, in what order and for how long.

Step 2 in the process is typically called "descriptive statistics" in which the objective is to abstract out certain properties of the data in order to better interpret them. The assumption here is that the data are too complex to understand well by simply looking at them as lists or tables. The simplest example of this is the computation of an "average" value of the data. Many of us obtain a better grasp of a data set by having some summary of the data available, particularly in graphical form, rather than simply a tabular elaboration of the data. Note that whatever methods are utilized here, there is a loss of information associated with the description provided - the description (e.g. the average value of the data) does not include the full amount of information in the complete data set. An objective in descriptive statistics is to choose the appropriate level of description between complete enumeration of the data, and a coarse simple summary (such as the average value), so as to be able to address the questions you posed in the first place. As an example, consider the height of all students in this course. Having these displayed as a long list would not be readily useful, whereas if we state that the average height of students is 65 cm, this gives you a simple way to possibly compare this characteristic of the students to the students in another course. More information would be provided by a histogram (bar chart) of the heights of the students in the course, but even then there would be some loss in information since we could not develop from the histogram the full list of heights of all students in the course.

Step 3 in the process typically involves the area of inferential statistics - parameter estimation and hypothesis testing. These refer to using the data to

determine estimates of values of particular interest (respiration rate, photosynthetic rate, hemoglobin level, etc.) from the observations, a process called parameter estimation. One might then use the data to evaluate hypotheses (respiration rate increases with temperature, the hemoglobin content of two species differs) in which one compares a "null hypothesis" (respiration rate is independent of temperature) to an "alternative hypothesis" (respiration rate increases with temperature).

Step 4 uses the results of the inferential statistics developed to evaluate the results of the observations and provide an interpretation of the results (there is a significant effect of temperature on respiration, and this implies that the species has limited lattitudinal range due to the effects of temperature; two species differ significantly in their photosynthetic rate and you expect one species to outcompete the other under certain environmental conditions).

This course will focus on the descriptive statistics aspects of the above process. All life scientists are well served by being exposed to a formal statistics course that includes aspects of experimental design and hypothesis testing however, so you are encouraged to enhance your training in this area beyond the limited coverage included here. The emphasis on descriptive statistics here arises due to regular comments by life science practitioners that an extremely important aspect of quantitative training for their colleagues is the ability to interpret graphs, and utilize diverse graphical approaches to explain and interpret experiments.