

# Chapter 1

## Introduction

Regardless of the area of research, all investigations have at the core the scientific method (Figure 1.1). Observations lead to the development of a research question, which is then formulated into a formal statement of the research hypothesis. Through experimentation, the form of which will vary depending on the nature of the research question, data are collected and analyzed to evaluate the validity of the hypothesis. Evaluation of the hypothesis usually will lead to more hypotheses, which is why science is cyclical in nature. One of the goals of statistics is to allow the researcher to objectively separate biological differences from natural variation. Statistics are only a tool; they can't fix flawed experimental designs or poorly collected data. However, statistics are powerful tools that need to be understood before they can be used correctly.

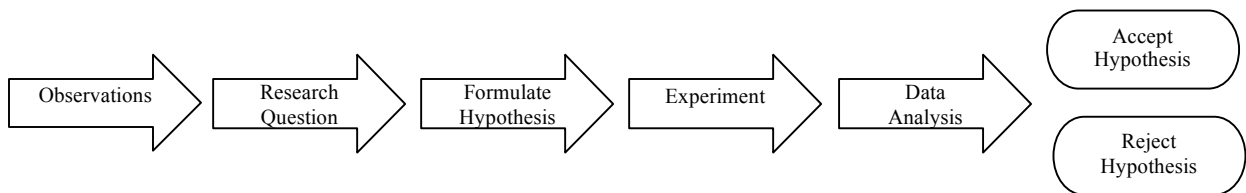


Figure 1.1. Schematic diagram of the scientific method leading to evaluation of the research hypothesis.

Before you can begin to understand statistics, you must first understand the nature of data as different types of data necessitate particular statistical procedures. There are five main classes of data: **nominal**, **discrete**, **ranked**, and **continuous**. Nominal data refers to data that describes classes, categories, or attributes and is generally textual in nature. For instance, if a researcher were to investigate bee visitation to different colored flowers, he or she may come up with different categories of flower color (e.g. red, white, and yellow). In this case the flower color class is nominal data. Discrete data, in contrast, are numerical and fixed in nature. Often discrete data are integers; for example, in the previous scenario the number of visits to each category of flower would be discrete data since it would be impossible to have a fraction of a visit. Ranked data are also integers, but differ from discrete data in that they reflect hierarchy in classification (e.g. birth order). Continuous data assume any number of values between two points (e.g. height). In some cases continuous data is derived in nature, as is the case with rates, ratios and differences. Derived data requires mathematical transformation to be biologically relevant.

It is important to know the nature of data, prior to collection so that you can be sure that it is appropriate to answer the research question. For example, in testing the effectiveness of an antibiotic on a particular type of bacteria there are several data sets that can be collected: presence/absence, number of colonies, or area of inhibition. Each one is a different type of data (nominal, discrete and continuous, respectively) and each one is appropriate for a slightly different research question. The first (presence/absence) addresses *whether* a particular antibiotic is effective and the last two data sets (number of colonies and area of inhibition) address *how* effective the antibiotic is on bacteria growth. Secondly, although the number of colonies and area of inhibition both address the same general research question, they would need to be analyzed using very different statistical methods. Thus, in addition to

planning how to conduct an experiment, it is also important to understand the nature of the data being collected prior to beginning a research project.

Just as there are different types of data, there are different types of scientific investigations. The most familiar types of science are empirical, in which controlled observations are made in order to test a particular hypothesis. Within the category of empirical science, studies can be descriptive, manipulative, comparative, perturbation/response, or correlational. Most initial investigations within a particular field are descriptive in which there is a comparison of measured variables under different conditions. Descriptive studies rarely address cause-effect relationships directly, but they are useful in generating such hypotheses that can be addressed by further research. Manipulative experiments are the most familiar application of the scientific method. In a manipulative experiment the researcher imposes one or more treatments and observes the response of some variable of interest to the treatment(s). You can describe a manipulative experiment in terms of independent and dependent variables: the **independent variable** is the treatment or condition that is manipulated, and the **dependent variable** is the variable you measure (also called the **response variable**). The limitations of manipulative experiments is that they are often limited to small numbers and short time scales and not all areas can be investigated using manipulative experiments for ethical reasons. Comparative studies are often used in ecology and physiology, and are similar to manipulative experiments except that treatments are determined by natural conditions (e.g. comparison of photosynthetic rates of northern and southern populations of a particular plant species). A perturbation/response study also takes advantage of natural conditions, but in this case it is a study of recovery that is conducted following a large-scale disturbance. One can think of a perturbation/response study as a special type of descriptive study in that they have the same limitations.

Correlational studies differ from the other types of empirical studies in that a broad research question is often investigated rather than a discrete hypothesis. A correlational study examines a survey of variables for potential relationships. This type of study is often used when manipulative experiments are impractical or unethical, yet they are limited because they can only suggest causal links not definitive cause-effect relationships. Many medical and environmental studies fall into this category of empirical science.

Another category of scientific investigation is deductive science, or modeling. Scientific models specify values for variables or conditions, and produce a predicted outcome using logical or mathematical relationships. The values of the input variables often come from empirical studies, so one can think of modeling as a higher level of investigation (but not necessarily better or more important!). Measured data from empirical studies is often compared to model predictions as a means of validating the model and pinpointing gaps in our knowledge, which frequently leads to further empirical studies.

## Presenting Data

The primary rule that applies to all types of data is that raw data are *never* presented in a scientific report; instead one presents the summarized data (discussed further in Chapter 3). Typically data is presented in tables or figures. A table is an arrangement of data into rows and columns, and a figure is any other type of graphical representation (i.e. graphs,

photos, maps, etc.). Some instructors and most journals require that the information presented in tables and figures not be the same, but any formal report will usually have both. For instance, it can be very useful to highlight the most important data (that which is crucial to your interpretation) into a figure and then place other summarized data or the statistical parameters into a table.

A table design should always facilitate the readers' understanding of its contents. All columns should be properly labeled with units presented whenever appropriate. Each group should be clearly identified and the data for each group clearly organized within a single row. While tables should not have titles, all tables must have a caption that describes the table contents. A caption differs from a title in that it is written as a sentence. Table captions always appear *above* the table, and tables are numbered numerically as they are referred to in the text (e.g. "Table 1"). Tables that are inserted into the text of a report should be located as close to the initial reference as possible, but not in such a way that a paragraph is broken up.

Table 1.1. Mean tissue nitrogen content (mg N g<sup>-1</sup>) of fertilized or unfertilized bean seedlings (n = 12).

Sample	Nitrogen Content (mg N g <sup>-1</sup> )	SE	95% Confidence Limits		Range	
			Low	High	Low	High
Fertilized	32.27	1.26	29.03	35.51	28.2	36.2
Unfertilized	30.33	0.72	28.49	32.18	27.3	32.3

Like tables, figures need to be numbered and presented in consecutive order. Most of the figures that appear in scientific reports are charts depicting the data graphically, more commonly referred to as graphs. There are a few rules to constructing graphs that need to be followed for an effective presentation. First of all, the dependent variable always goes on the Y-axis and the independent variable on the X-axis. Both axes need to be properly labeled with units. Make sure to adjust the scales of each axis so that your data cover most of the area – white space is distracting to the eye. If you have more than one symbol or line, make sure to identify each either within the figure or in the caption. When graphing means, you must always provide error bars (you'll learn why later on). All figures, like tables, need a caption that describes the content of the figure; however, a figure caption is placed *below* the figure.

While there are many different types of graphs, not all of them are appropriate for biological data and the type of data dictates the type of graph. The two most common types of graphs for biological data are bar graphs and scatter plots. Bar graphs (Figure 1.2) are appropriate only for frequencies or percentages. Scatter plots (Figure 1.3) are appropriate when the response variable or dependent variable is continuous. Scatter plots are appropriate either when the independent variable is nominal and means (with error bars) are presented for the dependent variable (Figure 1.3) or when both the independent and dependent variable are continuous, X-Y pairs (Figure 1.4). When data are X-Y pairs the individual points should not be connected unless points are related, as is the case with a time sequence.

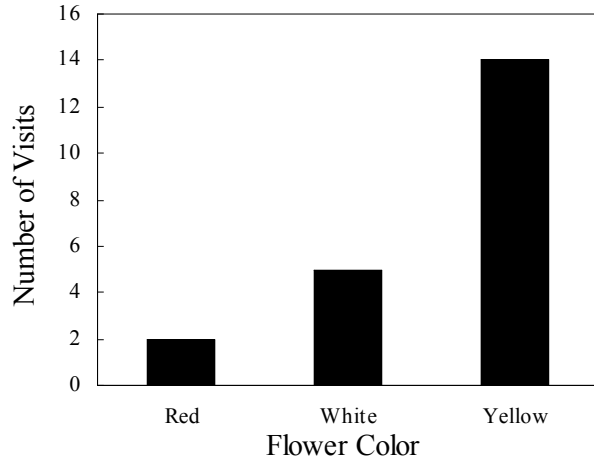


Figure 1.2. Frequency of bee visitations to different colored flowers.

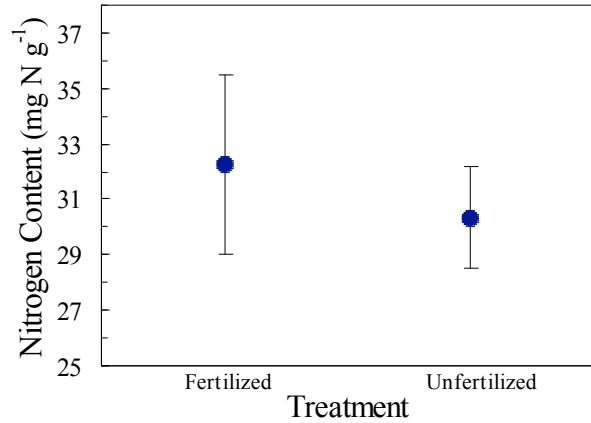


Figure 1.3. Mean nitrogen contents of fertilized and unfertilized bean plants. Bars represent 95% confidence intervals.

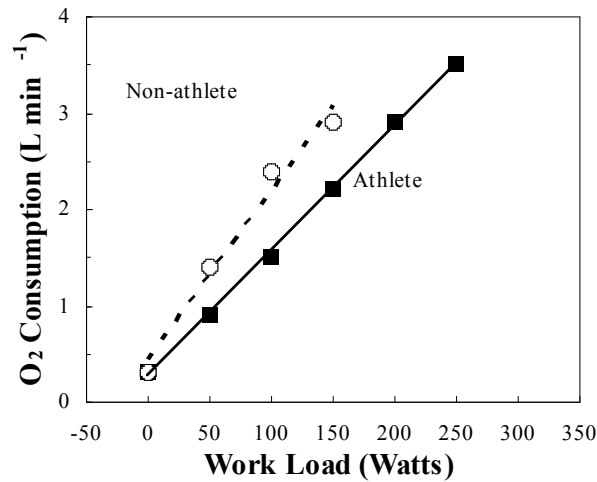


Figure 1.4. Oxygen consumption (L/min) at increasing levels of work (Watts) for athletes (filled symbols and solid line) and non-athletes (open symbols and dashed line). The lines represent the line of best fit for each data set.

Data analysis and presentation can be facilitated by careful consideration of the data prior to experimentation. It is important to consider what type of data is going to be collected to ensure that it is appropriate for the research question. Understanding the nature of the data is necessary to choose the appropriate statistical procedure and presentation style.