

February 22, 2009

## CHAPTER 1: INTRODUCTION

**Statistics** is the science of data collection and analysis. Most of our information in scientific endeavors comes in the form of data which are often large files containing numbers. Staring at a large collection of numbers typically will not provide much insight into the problems at hand. In order to make sense of the data, methods are needed to organize the data in meaningful ways using descriptive and visual aids. Using the power of probability theory, statistical inferences can be made about the population or the model from which the data originated. The field of statistics is certainly not static – new statistical methodologies continue to be developed based on mathematics and high speed computing to address ever emerging problems in data analysis. However, the primary topics covered in this class deal with classical statistical analysis topics that have been in use for decades. Therefore, this is an *applied* statistics course where we will apply existing statistical methodologies to solve problems related to biostatistics.

Statistics deal with data analysis. However, before analyzing the data, the data needs to be collected. Thus, statistics also encompasses the science of collecting data. The branch of statistics called **sampling design** focuses on collecting observational data and the branch called **experimental design** focuses on collecting data through experiments, like in a laboratory (see below).

**Biostatistics** is the branch of statistics that deals with problems arising in medical sciences and in biological problems. This branch of statistics has seen enormous growth in recent years primarily due to the continued research in medical sciences. In order to properly evaluate the results of any clinical trial, statistical methodologies need to be applied to the resulting data.

Ideally students taking this class will be able to perform some basic statistical analysis of data when the class is finished. It is fairly easy to use a statistical software package to perform a basic statistical analysis. However, the hard part is knowing what the output from a statistical software package means and how to properly interpret the output. It is also very important to know what sort of statistical analysis is called for when analyzing a particular data set or if it is reasonable to use a particular statistical methodology on a given data set.

Of course, we cannot hope to turn students into expert applied statisticians in a single introductory biostatistics course. One of the primary goals of this course is to educate students on how to interpret statistical output that is reported in studies. If you are reading a research article pertaining to a topic of interest, you will very likely see statistical results reported using statistical terms and language. In order to completely understand the reported research, the reader must be familiar with basic statistical principals and terminology. It is not unusual to see an expression like, “an ANOVA  $F$ -test yielded a  $p$ -value of 0.0012” in a scientific report. A basic background in statistics is needed in order to understand such statements.

One of the powerful aspects of statistics is that proper statistical tools allow us to extract all the relevant information that is in a large collection of data in an economical manner. Data from an observational study or an experiment often comes in the form of a file with columns of numbers. It is difficult to understand what information the data has to offer by staring at the numbers in the columns. The power of statistics is to be able to extract the relevant information from the data file in an efficient manner and draw inferences. For instance, it is often possible to represent all the information in a large collection of data by two relatively simple *statistics*: (1) the mean (or average) which measures where the data values are centered and (2) the standard deviation, which measures the “spread” of the data values (we will talk about this in more detail later).

In many problems, interest lies in understanding a particular population, say a population of a species of animals. Data can be obtained by *observing* these animals. This type of data would be called **observational data**. In other cases, we may want to conduct an experiment and note how specific *factors* affect some response of interest. For instance, we could look at the growth of a plant (response = height of the plant), and see how the height is affected by **factors** such as the amount of fertilizer and amount of water etc. Data obtained from experiments like these are called **experimental data**. In the plant example, the response variable (height) is known as a *dependent* variable because it depends on the *independent* variables which are amount of fertilizer and amount of water.

**Question:** *How many observations should one collect?* Think of data as evidence. How much evidence is needed to answer a particular scientific question? The more data we have, the more information we have. However, collecting data is often expensive and time consuming. Thus, how much data is needed to obtain a confident answer to a question? For instance, on average, how long does it take for a new pain medicine to begin working? Does the new pain medicine begin working quicker than an older pain medicine? Can we answer this question by performing an experiment with say four or five patients? Or do we need to do an experiment with thousands of patients? If the data has too few observations, then there will not be enough *power* to detect if the new medication is better than the old medication. If we conduct a huge study with thousands of subjects, we may be able to confidently answer the question, but we may have wasted a lot of money and time if a smaller sample would also have provided an answer to the question. Questions of sample size determination will be covered in later chapters.

In the pain medication example, will the time it takes for the medication to take effect be the same for all patients? The answer is very likely no! The time till pain relief will vary for patients and hence the issue of *variability* is extremely important in statistics. Also, if we randomly choose a patient, how likely is it that this patient will get pain relief say one minute after taking the medication, or two minutes? *Probability* theory provides answers to the likelihood of certain events happening. Probability is therefore the foundation upon which statistics is founded. Chapter 3 will cover the basics of probability theory.

**Statistical Software.** As mentioned earlier, statistics is the science of data collection and analysis. Once the data is collected, how do we analyze it? Recall that data is often a large collection of numbers. In the old days (before computers), the analysis had to be done by hand which severely limited the statistical analysis. Of course we now have very powerful computers which can be used to “crunch” the numbers. Many statistical software programs are available for doing statistical analysis. Statistical software is essential for analyzing data. We shall use the statistical software package **SAS** to do the data analysis for this course. SAS is a widely used statistical software package. Later chapters will illustrate how to write SAS programs and how to interpret the statistical output from the programs.

**Populations and Samples.** The primary goal of statistics is to infer something about a population based on a sample obtained from the **population**. Typically populations may be very large (or even hypothetically infinite in size). Thus it is often impossible or impractical to observe every object in the population. However, with the power of statistics, we can make very accurate statements about the population based on a **sample** from the population even if the sample is a tiny fraction of the whole population.

**Example: Opinion Polls.** Everyone is familiar with opinion polls. The population of interest in an opinion poll may be the set of all adults in the U.S.A. The sample may consist of 1000 adults selected at random. The data may consist of a yes or no response to a question posed to the people being sampled.

The data can be recorded as a series of zeros and ones (for yes and no answers respectively). Provided the people polled are representative of the general population, the results of the opinion poll are very accurate even though a sample of 1000 adults is a very tiny fraction of the millions of adults in the U.S.A. A representative sample is practically guaranteed provided adults are chosen at random. There is an entire branch of statistics referred to as *Sampling Design* that deals with the issue of obtaining observational data in efficient and representative ways, whether we are sampling humans or doing a wildlife study. Of course, it is very easy to obtain a sample that is not representative of the population of interest and therefore great care must be taken to ensure that the sample is indeed representative of the population.

Consider an example – suppose we want to estimate the proportion of U.S. adults that support a particular candidate for president. If we let  $p$  denote this proportion, then  $p$  is known as a *parameter* of the population (U.S. adults). Our goal is to try and figure out the value of  $p$ . Of course, the true value of  $p$  will not be known because it is almost impossible to ask every single adult their opinion. We can estimate the value of  $p$  using statistics. Suppose 1000 adults are selected at random and asked if they support a particular candidate. We can summarize all the information in the sample of 1000 responses by a single number, the sample proportion, i.e. the proportion of the 1000 polled individuals who support the candidate. Typically, the sample proportion will be denoted using the notation  $\hat{p}$  (“ $p$ -hat”).  $\hat{p}$  is an example of a *statistic*. This statistic is an *estimator* of the parameter  $p$ . Because the 1000 polled individuals were selected at random, the value obtained for  $\hat{p}$  will depend on which individuals happened to have been selected for the poll. That is, the value of  $\hat{p}$  varies depending on the sample obtained and its value is random because a random sample of individuals was obtained. Thus,  $\hat{p}$  is also an example of a *random variable*. Using probability theory, we can determine how  $\hat{p}$  will behave and that in turn tells us how reliable  $\hat{p}$  is as an estimator of  $p$ , the true population proportion.

**Statistical Models.** The scientific question that one hopes to investigate using statistical analysis can usually be formalized using statistical models. We shall use the following example to illustrate a couple simple models.

**Example.** Data was collected on the thickness of Anacapa pelican eggs (in millimeters) (Risebrough 1972). There were  $n = 65$  birds examined. Thus, the sample size is 65 and we usually use  $n$  to denote sample size, i.e. the number of observations. We can define a **variable**  $Y$  to denote the thickness of a pelican egg. Each bird has a different egg thickness, so the value of  $Y$  varies from bird to bird. If the birds for this study are selected at random from the entire population of Anacapa pelicans, then the values observed for  $Y$  are also random. Therefore, we call  $Y$  a *random variable*. Let  $Y_1$  denote the egg thickness for the first bird, and  $Y_2$  denote the thickness for the second bird, on up to  $Y_{65}$ . Once the data is collected, we will have fixed observed values for  $Y_1, Y_2, \dots, Y_n$ . Typically we shall use lower-case letters to denote the realized value of a random variable. Therefore, from the table below, we let  $y_1 = 0.14, y_2 = 0.19$  etc. Another *variable* that was measured was the PCB (polychlorinated biphenyl) concentration for each bird (in parts per million). Let  $X$  to denote the PCB concentration. Here is the observed data for the pairs  $(x, y)$  for each of the  $n = 65$  birds, where the data has been ordered in ascending order for the values of  $y$ :

### Pelican Eggshell Data

452	0.14	204	0.28	261	0.34	199	0.46
184	0.19	89	0.28	150	0.34	216	0.46
315	0.20	320	0.28	143	0.35	236	0.47
115	0.20	138	0.29	229	0.35	206	0.49
139	0.21	198	0.29	132	0.36	237	0.49
177	0.22	191	0.29	175	0.36		

214	0.22	193	0.29	173	0.36
356	0.22	316	0.29	220	0.37
246	0.23	265	0.29	212	0.37
177	0.23	122	0.30	236	0.37
289	0.23	305	0.30	119	0.39
166	0.23	203	0.30	144	0.39
175	0.24	396	0.30	147	0.39
296	0.25	250	0.30	171	0.40
205	0.25	230	0.30	232	0.41
260	0.26	214	0.30	216	0.41
188	0.26	256	0.31	164	0.42
208	0.26	46	0.31	185	0.42
324	0.26	204	0.32	216	0.42
109	0.27	218	0.34	87	0.44

Considering only the eggshell thickness, a very simple model for this data can be expressed as

$$y_i = \mu + \epsilon_i,$$

where  $\mu$  (“mu”) is an overall average or mean value for the eggshell thickness.  $\mu$  is another example of a *parameter* that is used to describe the entire population of pelican eggs. We can use this data to estimate  $\mu$  by simply averaging the 65 eggshell thicknesses.

Because the eggs are not all of the same thickness, there is **variability**. That is, the eggshell thicknesses vary from egg to egg. The error term  $\epsilon_i$  is needed in the model because of the variability in eggshell thickness from bird to bird. We shall give a formal definition for a population variance later – the population variance is usually denoted by the Greek letter  $\sigma^2$  (“sigma-squared”). Low values of  $\sigma^2$  indicate that the eggshell thicknesses are tightly clustered about the mean value  $\mu$  and large values of  $\sigma^2$  indicate that the eggshell thickness values are spread out considerably.

**Key Point:** In practice, the value of a parameter that describes an entire population is *unknown* and cannot possibly be known exactly unless the entire population is sampled which is either impossible or highly impractical. One of the primary goals of statistics is to *estimate* the unknown parameters that define a population and determine the precision of the estimate. We don’t expect that our estimates of the parameters will equal the exact value of the parameter BUT we can show mathematically the following properties of a good estimator:

1. A good estimator will not systematically under or overestimate the population parameter of interest (in statistics we say such an estimator is **Unbiased**).
2. With high probability, a good estimator will be close in value to the true parameter value (*precision or reliability*).
3. The value of a good estimator will become arbitrarily close to the true parameter value as the size of the sample becomes large (*consistency*).

**Statistic:** A statistic is a number that is computed from the data in a sample. That is, a statistic is computed by plugging numbers (i.e. the data) into a formula that does not require knowing the true parameter values of the population. One of the simplest examples of a statistic is to simply average the data values to compute the mean. Thus, in the pelican example, we can estimate the average eggshell thickness for the population by taking the average of the 65 eggshell thicknesses in the sample – this will be called the sample mean and is an example of a statistic. To compute the sample mean does not require

that we know the average eggshell thickness  $\mu$  for the whole population. The sample mean of a variable  $y$  will be denoted by  $\bar{y}$ .

In order to model the relation between eggshell thickness  $y$  and PCB concentration  $x$ , we need a more complicated model than before. A very general model would be

$$y = f(x)$$

for some function  $f(x)$ . However this model will not suffice because there of the variability in the eggshell thicknesses. Once again we need to incorporate an error term into the model:

$$y_i = f(x_i) + \epsilon_i,$$

where the index  $i$  corresponds to the different observations. In the pelican eggshell example,  $i = 1, 2, \dots, n$  where the sample size  $n = 65$ . The statistical problem now becomes estimating the function  $f(x)$  so that we can determine what sort of affect PCB's have on the pelican eggshell hardness. This can be a difficult problem because the function  $f$  is not known ahead of time. However, it turns out that often we can approximate  $f$  by a fairly simple function and end up with a model that explains the data quite nicely. For instance, (using Taylor's theorem from calculus), the function  $f$  can often be approximated by a polynomial in  $x$ . The simplest type is to consider a linear function:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

This model is known as a **simple linear regression model**.  $\beta_0$  and  $\beta_1$  are parameters of this model and are unknown.  $\beta_0$  is known as the  $y$ -intercept because it is equal to the value of  $y$  when  $x = 0$ .  $\beta_1$  is the slope of the regression line and is typically the parameter of primary importance in a regression analysis. The slope measures the change in  $y$  for a unit change in  $x$ . Statistical methods from *regression analysis* are used to estimate these two parameters. Note that if  $\beta_1 = 0$ , then there is no effect of PCB on eggshell thickness. Thus, a question of interest that can be answered using statistical *hypothesis testing* is whether or not  $\beta_1$  equals zero.

In the regression example we had two variables  $x$  and  $y$ :  $x$  is known as an *independent* variable (or predictor or regressor variable) and  $y$  is the *dependent* variable. The eggshell thickness  $y$  depends on the level of PCB  $x$  and that is why we call  $y$  the dependent variable.

The two statistical models considered here are very simple. We can consider other statistical models in order to solve more complicated problems.

This concludes our brief introduction to statistics. The following chapters will provide more details on specifics.