1 Introduction to statistics and simple descriptive statistics

This chapter discusses several topics, from why statistics is important in anthropological research to statistical notation. The first section (statistics and scientific enquiry) defines basic scientific terms and explains the role of statistics in anthropological research. The second section (basic definitions) reviews the vocabulary we need for the rest of the book. The third section (statistical notation) explains the fundamentals of statistical notation.

1.1 Statistics and scientific enquiry

If you are an anthropologist, you have probably been asked what anthropology is, and what it is good for. Many of us are at a loss to explain the obvious: how else could we look at the world, but with a cross-cultural and evolutionary perspective? What you may not be quite convinced about is the need for you to include a statistical aspect to your anthropological data analysis. In this section, I hope to explain why statistics are an integral part of a scientific approach to anthropological enquiry.

The word **statistics** is part of popular culture and every-day jargon. For that reason, I wish to clarify our working definition of the term in this book. Let us agree that statistics are the figures which summarize and describe a data set (**descriptive statistics**), and the methods used to arrive at those figures. In addition, statistical analysis allows us to make predictions about the wider universe from which a data set was obtained (**inferential statistics**). The reason statistics are an integral part of the anthropology curriculum is that statistical methods allow us to approach our subject of study in a manner which lets us test hypotheses about the subject. Or as G. Jenkins says: "The preoccupation of some statisticians with mathematical problems of dubious relevance to real-world problems has been based on the mistaken notion that statistics is a branch of mathematics – In contrast to the more sensible notion that it is part of the mainstream of the methodology of science" (Jenkins 1979). Indeed, some students enter my classroom fearful because they say they are not good at mathematics. However, the statistics class offered by my department is not a mathematics class; it is a class on methodology useful for testing hypotheses in anthropology.

There are other terms (such as hypothesis) which are used with different meanings in popular and scientific parlance. Within science, different disciplines have different

definitions for them. For that reason I wish to define some key terms at the start of the book.

I would first like to define the term **fact**. It is a fact that you are reading this book. Such fact is easily verified by yourself and others. People who are visually impaired can verify this by using their sense of touch and by asking other people to confirm this. Therefore, a fact is something that is verified usually, though not always by human senses. The existence of sub-atomic particles is verified by computers. Although a human might need her senses to read or feel the output generated by the computers, it is the latter which was able to detect the sub-atomic particles. Thus, verification of facts is not limited by human senses (otherwise, how could we detect sound used by other animal species which we humans are unable to detect?). Science would have hardly advanced if we had limited observation to what we humans are able to detect with our senses. Thus, facts are verifiable truths, where the verification is not limited to human senses.

A **hypothesis** is an explanation of facts. What makes a hypothesis scientific is that it can be tested and rejected by empirical evidence. A hypothesis that cannot be tested is not scientific. Scientific hypotheses explain observed facts in testable ways. For example, it is a well-known fact that the frequency of different colors changed in peppered moths in London during the height of the industrial revolution. Prior to the onset of heavy pollution in London, the majority of the moth population was lightly pigmented because light-colored moths were well camouflaged on light tree bark from their predators (birds). However, due to heavy pollution, tree bark became progressively darker. As a result, dark moths became better camouflaged, and the frequency of light moths decreased and dark ones increased. The frequency of dark peppered moths increased from less than 1% in 1848 to 95% in 1898. With the control of pollution in the 1900s, tree barks became lighter, light moths had the survival advantage, and the frequency of dark moths decreased. These are the observed and verifiable facts. Various hypotheses can be proposed to explain these facts. For example, it could be proposed that such changes in moth coloration are the result of a supernatural Being testing our faith. Or it could be hypothesized that birds acted as a natural selection agent, and that the change in moth coloration was an evolutionary change experienced by the moth population. Both propositions are explanations of the facts, but only the second one can be empirically tested and therefore be considered scientific. Please note that a hypothesis cannot be proven to be true. It can, however, be rejected. A hypothesis that has not been rejected after many studies is more likely to be correct than one that has been supported by only a single study or none at all.

There is quite a difference in the meaning of **theory** in popular culture and in science. In the former, the word theory is sometimes used dismissively, as if it were something with no factual base. This is certainly not how a theory is understood to be in science. We define a **theory** as a set of unified hypotheses, none of which has been rejected. For example, the theory of plate tectonics encompasses several hypotheses which explain several facts: the shape of Africa and South America "fit," the stratigraphy of both continents also correspond with each other, the shape of Madagascar "fits" with Africa, etc. The currently accepted (not proven) hypotheses to explain these facts is that there is continental movement, that Africa and South America were once a single land

mass, and that Madagascar split from Africa. Therefore, a theory is able to explain facts with hypotheses driven by the theory. These hypotheses are tested and accepted (for the moment). Should any of these hypotheses be rejected later (perhaps because better observation is possible as a result of new equipment), then the theory encompassing the rejected hypothesis must be revisited. But the entire theory does not fall apart.

The example above illustrates what is (in my opinion) the most distinctive trait of science as a form of human knowledge different from other forms of knowledge: science is by definition a changing field. Hypotheses which have been accepted for decades could very well be rejected any day, and the theory which drove those hypotheses revisited. As Futuyma so clearly puts it: "... good scientists *never* say they have found absolute 'truth' (emphasis in text) (Futuyma 1995).

Statistical methods are of fundamental importance for the testing of hypotheses in science. Researchers need an objective and widely recognized method to decide if a hypothesis should be accepted or rejected. Statistical tests give us this method. Otherwise, if each of us were to decide on what criteria to use to accept or reject hypotheses, we would probably never allow ourselves the opportunity to accept a hypothesis we want to reject or vice versa. As a tool to test hypotheses and advance theories, statistics are an integral part of scientific studies.

There is one more reason why statistics are so important for hypothesis-driven anthropological research: if we quantify results, we are able to compare them. The need for comparing results is due to the fact that scientific results should be replicable. As you must have learned in high school science classes, different people following the exact same procedures in a scientific experiment should be able to obtain the same results. The problem in anthropology, of course, is that it is impossible to replicate a historical event such as a migration. But if we use statistics to summarize data, we can compare our results with the results of other researchers. For example, I am interested in determining if migrant communities have increased body mass index (BMI, computed as $\text{BMI} = \frac{kg}{m^2}$) and hypertension rates in comparison with non-migrant communities. Although I cannot replicate a migration event, I can compare data on BMI and hypertension in several migrant communities and determine if the migrant communities do or do not differ from non-migrant communities. By using statistical methods, we were able to show that whereas migrant groups experienced an increase in BMI, they did not always experience an increase in hypertension rates (Madrigal *et al*. 2011). We should remember that anthropology is by definition a comparative science. A cross-cultural view of anything human is intrinsic to the field. With statistics anthropologists are able to compare their results with the results of others.

1.2 Basic definitions

1.2.1 Variables and constants

One sure way to favorably impress your instructor is to refer to data in the plural and not singular. You can say, for example, that your data are a collection of measures on a group

of 100 women from a village. Your data include the following information: woman's age, religion, height and weight and number of children produced. The **unit of analysis** in this data set is the woman, from each of whom you obtained the above information, which includes both variables and constants. The fact that all of your subjects are women and the fact that they all live in the same village means that gender and village are both constants. **Constants** are observations recorded on the subjects which do not vary in the sample. In contrast, **variables** are observations which do vary from subject to subject. In this example the number of children produced, the age, the height, and the weight all vary. They are therefore variables. A singular observation in a subject (age 25, for example) may be referred to as an **observation**, an individual **variate** or as the **datum** recorded in the subject.

1.2.2 Scales of measurement

A cursory look at statistics textbooks will indicate that different authors favor different terms to refer to the same concept regarding the scale of measure of different variables. Please do not be surprised if the terms I use here are different from those you learned before.

1.2.2.1 Qualitative variables

Qualitative variables classify subjects according to the kind or quality of their attributes. These variables are also referred to as attributes, categorical or nominal variables. An example of such variables is the religion affiliation of the women. If an investigator works with qualitative variables, he may code the different variates with numbers. For example, he could assign a number 1 to the first religion, 2 to the second, etc. However, simply because the data have been coded with numbers, they cannot be analyzed with just any statistical method. For example, it is possible to report the most frequent religion in our sample, but it is not possible to compute the mean religion. I have always preferred to enter qualitative data into spreadsheets by typing the **characters** (Christian, Muslim, etc.) instead of using numbers as codes. However, not all computer packages allow you to enter data in this manner.

Another important point about research with qualitative variables concerns the coding system, which should consist of mutually exclusive and exhaustive categories. Thus, each observation should be placed in one and only one category (**mutual exclusiveness**), and all observations should be categorized (**exhaustiveness**). Qualitative variables will help us group subjects so that we can find out if the groups differ in another variable. For example, we could ask if the women divided by religion differ significantly in the number of children they produced, or in their height or weight. Qualitative variables themselves will be the focus of our analysis when we ask questions about their frequency in chapter eight.

1.2.2.2 Ranked or ordered variables

Ranked or ordered variables are those whose observations can be ordered from a lower rank to a higher rank. However, the distance or interval between the observations is not

fixed or set. For example, we can rank individuals who finish a race from first to last, but we do not imply that the difference between the first and second arrivals is the same as that between the second and the third. It is possible that the difference between the first and second place is only 2 seconds, while the difference between the second and third place is 3 minutes. A more appropriate anthropological example would be a situation in which we do not have the actual age of the women we interviewed because they do not use the Western calendar. In this case, we could still rank the women from youngest to oldest following certain biological measures and interviews to confirm who was born before whom. We will use ranked variables quite a bit in our non-parametric tests chapter (chapter seven).

1.2.2.3 Numeric or quantitative variables

Numeric or quantitative variables measure the magnitude or quantity of the subjects' attributes. Numeric variables are usually divided into discontinuous/discrete and continuous variables.

Discontinuous numerical variables have discrete values, with no intermediate values between them, while the distance between any two values is the same (as opposed to ranked variables). In the research project mentioned above, the number of children born to a woman is an example of discontinuous variables because it can only be whole numbers. Also, the difference between one and two children is the same as the difference between 11 and 12 children: the difference is one child. As we well know, discontinuous numeric variables are amenable to statistical analysis which may produce counterintuitive results. If we compute the mean number of children of two women, one of whom produced one and one of whom produced two children, the result will be 1.5 children. Therefore, it is possible to compute the mean of discrete numerical variables, whereas it is not possible to compute the mean of qualitative variables such as religious membership, *even if the latter are coded with numbers*.

Continuous numeric variables are numeric data which do allow (at least theoretically) an infinite number of values between two data points. In the research project mentioned above, the weight and height of the women are continuous numeric variables. In practice, investigators working with continuous variables assign observations to an interval which contains several measurements. For example, if an anthropologist is measuring her subjects' height, and a subject measures 156.113 cm and another measures 155.995, the researcher will probably assign both subjects to one category, namely, 156 cm. White (1991) discusses the issue of measurement precision in osteological research. He specifically focuses on the appropriate procedure to follow when slightly different measures of the same tooth or bone are obtained. The problems associated with measurement in osteology show that the measurement of continuous variables is approximate, and that the true value of a variate may be unknowable (White 1991). You will sometimes see a distinction between **interval** and **ratio** continuous numeric variables. In an interval scale a value of 0 does not mean total absence of the item measured by the scale. For example, a 0 value for temperature measured in the Fahrenheit or Celsius scales does not mean absence of temperature. In contrast, a 0 value in a ratio

variable does mean total absence, such as 0 kilos. In terms of statistical manipulation, the difference between ratio and interval scales is not important, so they are both treated in the same manner in this book. The analysis of continuous numeric data is the main purpose of this book.

1.2.3 Accuracy and precision

An **accurate measurement** is one that is close to the true value of that which is measured. When doing research, we should strive to obtain accurate data, but this is not as easy as it sounds for some variables. If we are working with easily observable discrete numeric variables (let's say the number of people in a household at the time when we visit it) then it's easy to say that there are three, six, or ten people. If the variable we wish to measure is not so easily observable (let's say the number of children produced by the woman) we might not be able to determine its true value accurately. It is possible that the woman had a baby when she was young and gave it up for adoption, and nobody in her household knows about it. She is not going to tell you about this baby when you interview her. In this case the accurate (true) number of children produced by this woman is the number of children she declares plus one (assuming that the number of children she declares is accurate). The problems associated with obtaining accurate measures of continuous numeric variables are different, and I already alluded to them in section 1.2.2.3. The better the instrument for measuring height in living subjects, or length of a bone, the more accurate the measurement. If we can determine with a tape measure that a subject's height is 156 cm but with a laser beam that she is 156.000789 cm, the latter measure is more accurate than the former. A **precise measure** is one that yields consistent results. Thus, if we obtain the same value while measuring the height of our subject, then our measure is precise. Although a non-precise measure is obviously non-accurate, a precise measure may not be accurate. For example, if you interview the woman about how many children she had, she may consciously give you the same response, knowing that she is concealing from you and her family that one baby she had when she was very young and gave up for adoption. Since she gives you the same response, the answer is precise; but it is not accurate.

1.2.4 Independent and dependent variables

The independent variable is the variable that is manipulated by, or is under the control of the researcher. The **independent variable** is said to be under the control of the experimenter because she can set it a different level. The **dependent variable** is the one of interest to the researcher, and it is not manipulated. Instead, she wishes to see how the independent variable affects the dependent variable, but she does not interfere or manipulate the latter. In a laboratory setting, it is easier to manipulate an independent variable to see its effects on the dependent one. Many readers have seen films of the Harry Harlow experiments on the effects of isolation on the behavior of young monkeys, in which the independent variable was degree of isolation, and the dependent variable was the behavior of the animals. For example, Harlow raised some monkeys with their

mothers, while he raised others in the company of other young monkeys and he raised others alone. By varying the degree of isolation, Harlow manipulated the independent variable, and observed its effects on the behavior of the monkeys.

In a non-laboratory setting it is much more difficult to have such tight control over an independent variable. However, according to the definition above, the independent variable is under the control of the researcher. Thus, we can separate the women in our research project by religion (independent variable) and ask if the two groups of women differ in their mean number of children (dependent variable).

In this book we will denote independent variables by an X and dependent variables by a Y. This distinction will be important in our regression chapters only. However, since it is our wish to understand the behavior of the dependent variable, we will usually refer to a variable with the letter Y. If we are discussing more than one variable we will use other letters, such as X, Z, etc.

1.2.5 Control and experimental groups

Let us go back to the research project in which you have data on a group of women from a village, from whom you collected each woman's age, religion, number of children produced, height, and weight. Let us say that you are working for an NGO which seeks to give some kind of employment and therefore better economic prospects to the women. Let us say that you recruit 50 women into the new program for an entire year and keep 50 out of the program. The following year you measure the 50 women in the program and the 50 not in the program for all the same variables. It would be more proper to refer to these two groups as the experimental group (the women in the program) and the control group (the women not in the program). Therefore, the **experimental group** receives a **treatment**, while the **control group** remains undisturbed, and serves as a comparison point. Assuming that participation in this program is beneficial to the women, we could predict that we would see a difference in the two groups of women. Specifically, we could predict that women in the program will have a healthier weight after a year when compared with women not in the program. It is by having a control group that we are able to show that a change does or does not have an effect on our subjects. Please note that we could express this example using independent/dependent variables terminology: in this example the independent variable is participation in the program (yes or no, under the control of the investigator) and the dependent variable is weight (which we do not manipulate).

If subjects are to be divided into experimental and control groups, the statistical decision derived from the experiment rests on the assumption that the assignment to groups was done randomly. That is, the researcher must be assured that no uncontrolled factors are influencing the results of the statistical test. For example, if you assign to the new program only women of one religion, and use as a control group the women of the other religion, and the mean weight between the groups differs, you do not know if you are seeing the effects of religion, the program, or both combined, on weight. If subjects are randomly assigned to treatment or control groups and the treatment does not have

an effect, then the results of the experiment will be determined entirely by chance and not by the treatment (Fisher 1993).

1.2.6 Samples and statistics, populations and parameters. Descriptive and inferential statistics. A few words about sampling

A statistical population is the entire group of individuals the researcher wants to study. Although statistical populations can be finite (all living children age 7 in one particular day) or infinite (all human beings when they were 7 years old), they tend to be incompletely observable (how could all children age 7 in the world be studied in one day?). A **parameter** is a measure (such as the mean) that characterizes a population, and is denoted with Greek letters (for example, the population mean and standard deviation are designated with the Greek letters μ -*mu*- and σ -*sigma*- respectively). But since populations are usually incompletely observable, the value of the population parameter is usually unknown.

A **sample** is a subset of the population, and generally provides the data for research. Some students upon taking their first statistics class feel that there is something wrong about the fact that we don't work with populations but rather with samples. Please do not let this bother you. Most research in realistic situations must take place with samples. You should however be concerned with obtaining a representative sample (this is discussed below). A **statistic** is a measure that characterizes a sample. Thus, if a sample of children age 7 is obtained, its average height could easily be computed. Statistics are designated with Latin letters, such as $\bar{Y}$ (Y-bar) for the sample mean and s for the sample standard deviation. This difference in notation is very important because it provides clear information as to how the mean or standard deviation were obtained. It should also be noted that population size is denoted with an uppercase N, whereas sample size is denoted with a low case n. In this book we will differentiate the **parametric** from the **sample notation**.

Descriptive statistics describe the sample by summarizing raw data. They include measures of central tendency (the value around which much of the sample is distributed) and dispersion (how the sample is distributed around the central tendency value) such as the sample mean and standard deviation respectively. Descriptive statistics are of extreme importance whether or not a research project lends itself to more complex statistical manipulations.

Inferential statistics are statistical techniques which use sample data, but make inferences about the population from which the sample was drawn. Most of this book is devoted to inferential statistics. Describing a sample is of essential importance, but scientists are interested in making statements about the entire population. Inferential statistics do precisely this.

Sampling. Since this book is about statistical analysis and not about research design, I will not discuss the different types of sampling procedures available to researchers. Moreover, the research design and sampling of anthropologists can vary a lot, whether they are doing paleoanthropology, primatology, door-to-door interviews, or archaeological excavations. However, I would like to discuss two issues.

(**A) Samples must be representative and obtained with a random procedure**. In the research project we have been discussing in this chapter, our data set included women from both religious groups. This was done because if we had only measured women of one group, our village sample would have been biased, or it would not accurately represent the entire village. A **representative sample** is usually defined as having been obtained through a procedure which gave every member of the population an equal chance of being sampled. This may be easier said than done in anthropology. An anthropologist in a particular community needs to understand the nuances and culture of the population, to make sure that an equal chance of being sampled was given to each and every member of the population. In many instances, *common sense is the most important ingredient to a good sampling procedure*. If you know that the two religious groups are segregated by geography, you need to obtain your sample in both areas of the village so that you have members of both groups (sample is representative).

There are some situations in anthropological research in which random sampling can hardly be attempted. For example, paleoanthropologists investigating populations of early hominids would hope to have a random sample of the entire population. But these researchers can only work with the animals that were fossilized. There is really no sampling procedure which could help them obtain a more representative sample than the existing fossil record. In this situation, the data are analyzed with the acknowledgment that they were obtained through a sampling procedure that cannot be known to be random.

(B) Samples must be of adequate size. The larger the sample, the more similar it is to the entire population. But what exactly is large? This is not an easy question, especially because in anthropology it is sometimes impossible to increase a sample size. Paleoanthropologists keep hoping that more early hominids will be unearthed, but can only work with what already exists. However, if a research project involves more easily accessible data sets, you should consider that most statistical tests work well (are **robust**) with samples of at least 30 individuals. Indeed, there is a whole suite of non-parametric statistical tests specifically designed for (among other situations) cases in which the sample size is small (discussed in chapter seven). As I mentioned above, sometimes the most important aspect of research design is common sense. When you are designing your project and you are trying to determine your ideal sample size, you should talk to experts in the field and consult the literature to determine what previous researchers have done. In addition, you might be able to perform a power analysis to help you determine your ideal sample size, although not everyone has the necessary information to do this. Power analysis is discussed in chapter four.

1.3 Statistical notation

Variables are denoted with capital letters such as X, Y, and Z while individual variates will be denoted with lower case letters such as x, y, and z. If more than one variable is measured in one individual, then we will differentiate the variables by using different letters. For example, we might refer to height as Y, and to weight as X. Distinct observations can be differentiated through the use of subscripts. For example, y_1 is the observation recorded

in the first individual, y_2 is the observation recorded in the second one, and y_n is the last observation, where n is the sample size.

Sigma ($\sum$) is a **summation sign** which stands for the sum of the values that immediately follow it. For example, if Y stands for the variable height, and the sum of all the individuals' heights is desired, the operation can be denoted by writing $\sum Y$. If only certain values should be added, say, the first ten, an index in the lower part of sigma indicates the value at which summation will start, and an index in the upper part of sigma indicates where summation will end. Thus: $\sum_1^{10} Y = y_1 + y_2 + \ldots + y_{10}$. If it is clear that the summation is across all observations, no subscripts are needed. Below are two brief examples of the use of sigma:

X	Y
6	7
8	9
5	2
3	10
9	1
10	3
$\sum X = 41$	$\sum Y = 32$

A frequently used statistic is $(\sum Y)^2$ which is the sum of the numbers, squared. In our example: $(\sum X)^2 = 41 = 1681$ and $(\sum Y)^2 = 32 = 1024$.

Another frequently used statistic is $\sum Y^2$ which is the sum of the squared numbers (and is called the uncorrected sums of squares by SAS). Using our previous examples, we can square the numbers, and sum them as follows:

X	X^2	Y	Y^2
6	36	7	49
8	64	9	81
5	25	2	4
3	9	10	100
9	81	1	1
10	100	3	9
$\sum X^2 = 315$		$\sum Y^2 = 244$	

The reader should not confuse $\sum Y^2$ with $(\sum Y)^2$. The former refers to the sum of squared numbers, whereas the latter refers to the square of the sum of the numbers. These two quantities are used in virtually every statistical test covered in this book.