# **TOPIC ONE: EXPLORING AND COLLECTING DATA**

## **DEFINITIONS & BACKGROUND**

“Why should I learn Statistics?” you might ask. “After all, I don’t plan to do this kind of work. In fact, I’m going to hire people to do all of this for me.” That’s fine. But the decisions you make based on data are too important to delegate. You’ll want to be able to interpret the data that surrounds you and to come to your own conclusions. And you’ll find that studying Statistics is much more important and (perhaps) enjoyable than you thought.

**SO WHAT IS STATISTICS?**

The term statistics has two definitions.

**Definition one**: Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.

**Definition two**: Statistics (plural) are quantities calculated from sample data- mean, median, standard deviation, e.t.c. Usually a sample is used because it is at times not possible to get information about the entire population- due to constraints such as cost and time. If it were possible to get information from the entire population, the corresponding quantities calculated are called *parameters*.

**WHAT ARE DATA?** The term data is the plural form of datum- this is the reason we use ‘are’ and not ‘is’. So, what are data? Data are values along with their context.

**PURPOSE OF THE COURSE**

If we want to analyse student perceptions of business ethics, should we administer a survey to every single university student in Kenya—or, for that matter, in the world? Well, that wouldn’t be very practical or cost effective. What should we do instead? Give up and abandon the survey? Maybe we should try to obtain survey responses from a smaller, representative group of students. Statistics can help us make the leap from the data we have at hand to an understanding of the world at large. We discuss about the specifics of sampling later, and the theme of generalizing from the specific to the general is one that we will revisit throughout this course. This course will empower *you* to draw conclusions from data and make valid business decisions in response to such questions as:

* Do university students from different parts of the world perceive business ethics differently?
* What is the effect of advertising on sales?
* Is there a seasonal cycle in your firm’s revenues and profits?
* What is the relationship between fertilizer input and cereal yields?
* How reliable are the annual forecasts for your farm production and sales?
* Are there common characteristics about your customers and why they choose your products

Our ability to answer questions such as these and draw conclusions from data depends largely on our ability to understand *variation*. That may not be the term you expected to find at the end of that sentence, but it is the essence of Statistics. The key to learning from data is understanding the variation that is all around us.

**DATA**

As mentioned earlier, data are values together with their context. Businesses have always relied on data for planning and to improve efficiency and quality. Now, more than ever before, businesses rely on the information in data to compete in the global marketplace. Most modern businesses collect information on virtually every transaction performed by the organization, including every item bought or sold. These data are recorded and stored electronically, in vast digital storehouses called **data warehouses**. Companies use data to make decisions about many aspects of their business. By studying the past behaviour of customers and predicting their responses, they hope to better serve their customers and to compete more effectively. This process of using data, especially of **transactional** data (data collected for recording the companies’ transactions) to make other decisions and predictions, is sometimes called **data mining** or *predictive analytics*. The more general term **business** **analytics** (or sometimes simply analytics) describes *any* use of statistical analysis to drive business decisions from data whether the purpose is predictive or simply descriptive.

For data to have meaning, it must have context- the set of facts or circumstances that surround an event. The table below shows data without context. It is impossible to tell whether the towns cited refer to their birth place, place of residence, or destination (in case they are travellers). The other data are pretty meaningless too. It is hard to make any conclusions from this dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| James | 100 | 00100 | 54 | 12 | Nairobi |
| Irene | 250 | 40500 | 63 | 16 | Kisumu |
| Paul | 300 | 20300 | 30 | 20 | Wajir |
| Sheila | 20 | 10101 | 22 | 13 | Mombasa |
| Purity | 465 | 20100 | 19 | 12 | Nakuru |

Table 1: Data without context

Try to guess what these data represent. Why is that hard? Because these data have no *context*. Whether the data are numerical (consisting only of numbers), alphabetic (consisting only of letters), or alphanumerical (mixed numbers and letters), they are useless unless we know what they represent. Newspaper journalists know that the lead paragraph of a good story should establish the “Five W’s”: *who, what,* *when, where,* and (if possible) *why*. Often, we add *how* to the list as well. Answering these questions can provide a **context** for data values and make them meaningful. The answers to the first two questions are essential. If you can’t answer *who* and *what*, you don’t have data, and you don’t have any useful information.

We can make the meaning clear if we add the context of *who* the data are about and *what* was measured and organize the values into a **data table** such as this one.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NAME | POSTAL ADDRESS | POSTAL CODE | AGE | YEARS OF EDUCATION | CITY OF BIRTH |
| James | 100 | 00100 | 54 | 12 | Nairobi |
| Irene | 250 | 40500 | 63 | 16 | Kisumu |
| Paul | 300 | 20300 | 30 | 20 | Wajir |
| Sheila | 20 | 10101 | 22 | 13 | Mombasa |
| Purity | 465 | 20100 | 19 | 12 | Nakuru |

Table 2: Data with Context

In general, the rows of a data table correspond to individual **cases** about which we’ve recorded some characteristics called **variables**. Cases go by different names, depending on the situation. Individuals who answer a survey are referred to as **respondents**. People on whom we experiment are **subjects** or (in an attempt to acknowledge the importance of their role in the experiment) **participants**, but animals, plants, websites, and other inanimate subjects are often called **experimental units**. Often we call cases just what they are: for example, *NAME, AGE*, or YEARS OF EDUCATION. In a database, rows are called **records**—in this example, purchase records. Perhaps the most generic term is **cases**.

If you collect the data yourself, you’ll know what the cases are and how the variables are defined. But, often, you’ll be looking at data that someone else collected. The information about the data, called the **metadata**, might have to come from the company’s database administrator or from the *information technology* department of a company. **Metadata** typically contains information about *how,* *when*, and *where* (and possibly *why*) the data were collected; *who* each case represents; and the definitions of all the variables.

**VARIABLE TYPES**

A variable is a quantity that can assume any set of values. In table 2 above, NAME is a variable, so is AGE, YEARS OF EDUCATION and so on. In a data table, usually the topmost row contains the names of the various variables. Variables play different roles, and knowing the variable’s type is crucial to knowing what to do with it and what it can tell us. When a variable names categories and answers questions about how cases fall into those categories, we call it a **categorical**, or **qualitative**, variable. When a variable has measured ***numerical values*** with ***units*** and the variable tells us about the quantity of what is measured, we call it a **quantitative variable**. Classifying a variable into categorical or quantitative can help us decide what to do with a variable, but doing so is often more about what we hope to learn from a variable than about the variable itself. It’s the questions we ask of a variable (the why of our analysis) that shape how we think about it and how we treat it.

Descriptive responses to questions are often categories. For example, the responses to the questions “What type of mutual fund do you invest in?” or “What variety of maize seed do you plant on your farm?” yield categorical values. An important special case of categorical variables is one that has only two possible responses (usually“yes” or “no”), which arise naturally from questions like “Do you invest in the stock market?” or “Do you make online purchases from this website?” If the variable has values that are not numbers, it’s clearly categorical (or needs to be recoded). However, if the values are numbers, you need to be careful. It may be considered quantitative if the values actually measure a quantity of something. Otherwise, it’s categorical. For example, POSTAL CODES are numbers, but the numerical values of the codes don’t have numerical meaning. The numbers assigned by the postal codes are codes that *categorize* the phone number into a geographical area. So, we treat area code as a categorical variable. For quantitative variables, the **units** tell how each value has been measured. Even more important, units such as shilling, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement, so we know how far apart two values are. Without units, the values of a measured variable have no meaning. It does little good to be promised a raise of 5000 a year if you don’t know whether it will be paid in euros, dollars, yen, or Estonian krooni. An essential part of a quantitative variable is its units. Sometimes the type of the variable is clear. But some variables can answer both kinds of questions and how they are classified depends on their use. For example, the variable *Age* would be considered quantitative if the responses were numerical and they had units. A doctor would certainly need *Age* to be quantitative. The units could be years, but for infants, the doctor would want even more precise units, like months, or even days. On the other hand, if a researcher asked your *Age*, it might lump together the values into categories like “Child (12 years or less),” “Teen (13 to 19),”“Adult (20 to 64),” or “Senior (65 or over).” In this case, *Age* has made a categorical variable.

**IDENTIFIERS**

What’s your student registration number? Or national ID number? It may be numerical, but is it a quantitative variable? No, it doesn’t have units. Is it categorical? Yes, but a special kind. Look at how many categories there are and at how many individuals there are in each category. There are exactly as many categories as individuals and only one individual in each category. While it is easy to count the totals for each category, it’s not very interesting. This is an **identifier variable**. Karatina University wants to know who you are when you sign in to register for courses and doesn’t want to confuse you with some other customer (who may even bear a name similar to yours). So they assign you a unique identifier.

**OTHER DATA TYPES**

A survey might ask:

“How satisfied were you with the service you received?”

1) Not satisfied; 2) Somewhat satisfied; 3) Moderately satisfied; or 4) Extremely satisfied.

Is this variable categorical or quantitative? There is certainly an *order* of perceived worth; higher numbers indicate higher perceived worth. An employee whose customer responses average around 4 seems to be doing a better job than one whose averages are around 2, but are they *twice* as good? Because the values are not strictly numbers, we can’t really say and so we should be careful about treating *Customer* *Satisfaction* as purely quantitative. When, as in this example, the values of a categorical value have an intrinsic order, we can say that the categorical variable is **ordinal**. By contrast, a categorical variable that names categories that don’t have order is sometimes called **nominal**. Values can be individually ordered (e.g., the ranks of employees based on the number of days they’ve worked for the company) or ordered in classes (e.g., First year, Second year; Baby, Middle, Top). Ordering is not absolute; how the values are ordered depends on the purpose of the ordering. For example, are the categories Infant, Youth, Teen, Adult, and Senior ordinal? Well, if we are ordering on age, they surely are and how to order the categories is clear. But if we are ordering on purchase volume, it is likely that either Teen or Adult will be the top group.

**Cross-Sectional, Time Series and Panel Data**

A **time series** is a single variable measured at regular intervals over time. Time series are common in business. Typical measuring points are months, quarters, or years, but virtually any consistently-spaced time interval is possible. Variables collected over time hold special challenges for statistical analysis.

By contrast, most other methods are better suited for **cross sectional data**, where several variables are measured at the same time point. On theother hand, if we collect data on sales revenue, number of customers, and expensesfor last month at *each* Naivas Supermarket (more than 100 locations) at onepoint in time, this would be cross-sectional data. Cross-sectional data may containsome time information (such as dates), but it isn’t a time series because it isn’t measuredat regular intervals. Because different methods are used to analyze these differenttypes of data, it is important to be able to identify both time series andcross-sectional data sets. **Panel data** have the dimensions of both time series and cross-sections, e.g. the average prices of a number of crops (maize, beans, coffee, and tea) produce over five years. The estimation of panel regressions is an interesting and developing area. Fortunately, virtually all of the standard techniques and analysis in econometrics are equally valid for time series and cross-sectional data.

**Continuous and discrete data**

As well as classifying data as being of the time series or cross-sectional type, we could also distinguish them as being either continuous or discrete, exactly as their labels would suggest. Continuous data can take on any value and are not confined to take specific numbers; their values are limited only by precision. For example, the return on a farm could be 6.2%, 6.24% or 6.238%, and so on per annum. On the other hand, discrete data can only take on certain values, which are usually integers (whole numbers), and are often defined to be count numbers. For instance, the number of students at Karatina University or the number of shares traded during a day.