EXPERIMENTS AND DATA

> by Dr Juan H Klopper

- Research Fellow
- School for Data Science and Computational Thinking
- Stellenbosch University

INTRODUCTION

In this notebook we familiarise ourselves with the terms and defintions that are commonly used in Data Science. As Data Science brings together many fields and can be used in many domains, there are numerous terms. They often overlap, are used in diverse situations with slightly different meaninings. At times their are more than one term used, both with the same definition. Certain terms are then favoured by authors and researchers.

Do not dispair. It is common for spoken languages to have a diveristy of words, definitions, and meaning. As with the spoken language, you will soon be able to *speak Data Science*.

Along with some data-centric definitions, we will also take a look at data capturing principles, research questions, trials, experiments, outcomes, populations, samples, and randomisation. We start our journy, though, by defining data types.

DATA TYPES

From a Data Science point of view, it is important to know the type of data that we are working with, since these types determine what analysis we can do with the data and what types of models we can build.

Data (always pleural) come in a variety of formats. Perhaps very commonly, we have tabular data. With tabular data, we have rows and columns of values in spreadsheet format.

It is typical in a spreadsheet to have individual columns that reflect some measurement. Each row is then the results of that measurment for infividual observations, be they humans, animals, or objects.

Some examples of these columns would be age, weight, colour, temperature, humidity, intensity, and many, many more. We can refer to each of these as variables. As such, a **variable** represents a property. A **random variable** is a function that assigns a value to a measurement for a variable or an outcome of an experiment (see later).

The *measurements* of each the columns is of a certain type. These can be represented by numbers or words. This helps us to create a classification for data types. In one such classification we have numerical and categorical data types.

NUMERICAL DATA

As the term implies, **numerical data** are measured in numbers. This is the most common data type is many settings. Even digital images and audio files are numerical data. A colour digital image is simply an array of pixel intesity values in a grid for each of the three colours red, green, and blue.

In this classification system there are two numerical data subtypes. These are interval numerical data and ratio-type numerical data.

Ratio-type numerical data are values that have a true zero. A true zero indicates absence. A length of 0 is an absence of any length. For this data type a value of say 20 is doble or twice that of a value of 10.

Interval numerical data do not have a true zero. An example would be temperature measured in Celsius or Fahrenheit. On both these scales 0 degrees is not the absence of temperature. We cannot state that 20^oC is twice as hot as 10^oC.

We can also use a different classification system and look at numerical data as either being discrete or continuous. Most numerical data are of the continuous type. **Continuous numerical data** have, by defintion, infinite decimal values. It is only our measurement apparatus that limit the number of decimal values that we can measure. We might also conveniently round certain

values such as monetary values. By definition, though, this is still an example of a continuous data type.

Discrete numerical data come in values that cannot be further divided. The orbital energy levels of electrons in an atom are quatised and therefor discrete.

CATEGORICAL DATA

Categorical data are values other than numbers, although they can be encoded as numbers. These data can be captured as words. There are also two subtypes of categorical data. **Ordinal categorical data** have an inherent order to them. Think of the Likert style selection for a survey question. On this scale, respondents might choose *strongly disagree*, *disagree*, *neither agree nor disagree*, *agree*, and finally, *strongly agree*. This is an example of a five-point Likert scale. There is a natural order to these values. They can even be encoded as numbers, i.e. 1, 2, 3, 4, and 5. Note, though, that these are not true numbers. We cannot state that 4 is twice as much as 2 on this scale.

Another example of ordinal categorical data that we can all imagine is a pain scale. After surgery we are given pain releavers, known as analgesics. A pain scale can be used to measure the amount of pain expeienced. These scales typically vary form 0 indicating no pain to 10, the most severe pain. These are once again not true numbers. For one, there is no fixed difference between consecutive values. We cannot state that the difference between a pains core of 2 and 3 is the same as the difference between 8 and 9. Pain is subjective. We can also not state that a pain score of 6 is twice as much pain as a score of 3. Lastly, we cannot express these results as an average. Such an average over many patients might be 3.7. Such a value has no meaning. Note that we can also refer to the values in this example as being discrete data. They are still not numbers, though.

Nominal catgeorical data have no order. An example might be four types of soil, i.e. sandy, silt, clay, and loamy soil. There is no order to these. Their water content might be expressible in some order, but the type of soil itself is a nominal categorical variable. With respect to this last statement, note that we often use the data type as part of the expression of the variable type.

SAMPLE SPACE

We have seen above that each variable is of a specific type. We can capture the measurements for a variable in a column in a spreadsheet. We refer to these variables as **statistical variables** to

separate them them form other variable that we will come across on this course. These include

Although there are specific defintion for a sampel space, we will think here of a **sample space** of a statistical variable as the collection of all possible values that can be measured or captured for that variable.

In our four soil example, the sample space of a statistical variable named *soil type*, would be the sandy, silt, clay, and loamy. If we include participants in a psychological study, we might restrict them to be between the ages of 25 and 35. All the values in this range are the sample space of a variable named *age*.

TIDY DATA

When working with tabular data (think of a spreadsheet), we have tidy data as a very important concept. Hadley Wickham published a paper on the concept of tidy data. [Wickham H. Tidy data. Journal of Statistical Software. 2014. Vol 59 (10).] The idea here is to have a spreadsheet (or data file) that adheres to certain standards.

Most notably, we require that each subject or observation be assigned a row in the data set and that each column represents a specific statistical variable. Each variable must have a well-defined sample space.

Data integrity is an important aspect of tidy data. Under this umbrella we have common mistakes such as an ill-defined sample space. It is so common to see a spreadsheet file where, as an example, the variable *soil type* will have values captured as *Silt*, *silt*, *silt soil*, and so on. A computer will see these three observations as different. Trailing white space (invisible spaces after a word) is a particular pain.

When designing a spreadsheet file or database, spend some time on it. Make dropdown lists for data captures to select from rather than allowing free-form input. Take special care of dates and times and keep then constant, without using the software program to format them in any way.

Tidy data also refer to avoiding the combination of values. Imagine then a study looking at soil type in an area divided into different section. In one section there might be sandy soil and silt. In yet another there might be sandy soil, loamy soil, and clay. We cannot have a statistical variable (column in the spreadsheet) be named *soil* and have researchers combine the types separated by commas or semi-colons. Instead we need to have individual columns for each of the soil types. We entering data, we would then enter either *yes* or *no* or 1 or 2 into each column.

Next we mention special coding for unknown data values. Some researhers (especially using expensive commercial statistical software) have codes such as 999 to indicate missing data. This is also frustrating when analysing the data. If a measure is unknown, leave it blank. If there are different reasons why data might be missing, consider individual

We have the actual column header values. The column headers in a spreadsheet is the first row that are the names of the different statistical variables. While we used the variable same *soil type*, this is actually not acceptable. For ease of analysis in Python do not use illegal characters such as spaces. A better name would be *soil_type* (called snake-case) or *soilType* (called camelCase). Make the statistcal variables names expressive of the data.

Lastly, we have spreadsheet formatting. This can be driven by the user, allowing empty rows, colouring, and all manner of calculations and plots in a spreadsheet. Individual columns can also be formatted to express percentages when fractions are entered or monetary units when money values are entered. These serve no purpose when analysing data and can lead to frustration. Save files in the universal comma separated values (CSV) format instead of the program proprietary format at all times. This removes all sorts of unwanted formatting and makes the file size smaller and more accessible.

RESEARCH QUESTION

It is common for some inexperienced researhcers to ill-define the aim of their research. Research must have very well defined research questions.

A well-defined research question must be very specific in its design by allowing individual statistical variable to answer the question.

In keeping with our four soil example, a poor research question might be *I want to compare the soil in different regions*. We have to be much more specific. An examples would include *Proportion differences of soil type in the various regions*. For a simplified two soil example, this question we can gather data for the statistical variables *region*, *region_size*, *sandy*, *silt*, *sandy_size*, *silt_size*. Our data table might look like the one below.

region	region_size	sandy	silt	sandy_size	silt_size
A	245	Y	Y	100	145
B	280	N	Y		280
C	320	Y	N	320	
D	120	Y	Y	80	40

TRIALS, EXPERIMENTS, AND OUTCOMES

These three terms can be confusing and used differently.

A **trial** is a sort of action leading to a measurement. Sticking to our soil example, a trial is the actual identification of a sample of soil. We repeat this trial many times over as we continue sampling. In some setting the term trail might mean a whole reseacrh project such as a randomised trial to investigate a new vaccine.

An **experiment** combines a number of trials into a research project. It is sometimes used as a synonymn for trail, though.

An **outcome** is the value of the measurement of a trial. The results is a value from the sample space of a statistical variable.

POPULATIONS AND SAMPLES

The term **population** refers to all the members of a set. We could view all human being on earth as the population of a research project. This number can be much smaller if our population are all the people on earth with a rare disease. In yet another example, we might view all the cultures for a specific organism in our lab as the population for that organism.

A *sample* refers to a select subset taken from a population. In most cases, we want to select the sample individuals from the population at random. When this selection is done by intriducing some pattern, the result is a **biased sample**.

A biased sample might contained confounders. **Confounders** are statistical variable that determine the actual result or has an influence on the final results, but are not recognised as such.

The results obtained from analysing a sample of the population can be inferred back to the population. It is usally impossible to investigate a whole popultion. We commonly select a random, unbiased sample from the population. The results can then be used on individuals in the population.

A **parameter** is some summary of the population. All human beings on earth are of certain length and if we knew this at any one time, we could calculate the average height of all human.

This average would be a parameter.

A **statistic** is a summary of a sample. If our population is all students at a University, we might slect a random sample of students and measure their height. The average height of this sample would be a statistic. Since the population is the students at the Univeristy, we could infer this statistic to the whole population of students if the sample contained no bias.

RANDOMISATION

To avoid bias in a sample, we often employ various randomisation techniques when selecting individuals or subjects from a population.

SIMPLE RANDOMISATION

A technique is used to create a random value, to which we assign a subject to a group. Each subject that enters the research project is randomly assigned to a group by this process, one at a time.

While simple, this method can introduce imbalance in smaller projects. Even the flip of a fair coin can land heads or tails up quite a few times in row or one side can land facing up at least more often that the other in a smaller set.

BLOCK RANDOMISATION

To reduce the risk of unequal group sizes, block randomisation divides the total number of planned subjects into equally sized subsets, i.e. six. If the aim is to assign a subject to one of two groups, three in the block will be randomized to one group and three to the other. All six participants are assigned before a new block is started.

Bias can be introduced with this method as those involved in the assignment will know what the last one, two, or even three randomisations would be. If the first three were randomly assigned to one group, all of the following three will be assigned to the other group. Even if the assignment was equally distributed for the first five cases, the last one will be known. If those involved with the project were not blinded to the *content* of the groups, bias can be introduced by postponing a subject's randomisation until the situation arises that a *sixth* assignment scenario occur as described.

STRATIFIED RANDOMISATION

In this method other variables that may influence the outcome of the project are monitored. As an example, if age could cause such an influence, randomisation will follow a pattern to ensure that an equal or near-equal number of participants are randomised to each group.

Using stratified randomization, subjects are assigned to blocks such that there is an even spread of these variables within a block. Simple randomization then occurs within each block.

This type of assignment is not always possible to do. This is the case in projects during which participants are enrolled as a matter of emergency.

ADAPTIVE RANDOMISATION

Adaptive randomisation aims to correct for imbalances in variables that may affect the outcome. It might become apparent that one group contains significantly more subjects with one or more of those variable values. When new subjects are recruited they are screened for these variables and a weighted randomisation can be used in this case.

GROUP RANDOMISATION

It is not always required to randomise every subject. Instead, if the research project is distributed among more than one physical site, these sites or geographic areas can be randomised. All subjects in a single such setting receives the same intervention.

CONCLUSION

We have covered many topics and defitions common to Data Science projects. These will stand us in good stead as we continue our journey of discovery.

```
1
```

✓ 0s completed at 13:47