TYPES OF DATA

BEST FOR AFRICA

RAISA RAMDEEN
INTRODUCTION

- Discerning between different types of data from an experiment or observational study is important.
- The type of data being collected determines the method used in collection, the way the data is managed and analysed.
- Simply put, there are two main types of data, qualitative data and quantitative data.
- We will see how these classifications can be further stratified.
- Then we will see how these classifications compare to the data type designations assigned by R.
- Please note that throughout this presentation we refer to “classification of data” to which we are referring to the classification of variables within a dataset into different categories of data types.
TYPES OF DATA

Qualitative data
- Nominal
- Ordinal

Quantitative data
- Discrete
- Continuous
Some variables can be classified by qualitative means or quantitative means. Let’s look at two examples, and their corresponding qualitative data.

For our first example, we consider leaf characteristics, in this case we look at leaf shape. Leaf shape is qualitative and **nominal**. It gets the term nominal, because there is no clear order between each category.

For our second example we consider temperature. A qualitative way of assessing temperature, are by the categories: cold, cool, warm and hot. This variable would then be qualitative and **ordinal**, since there is a specific order to the category options.

In some instances qualitative data can also be numerical. However, qualitative data cannot be used for further calculations. For example, finding the sum or average of a postal code would not generate useful information (Nolan & Lang 2005).
Following up on the examples in the previous slide, let’s look at how the given examples of leaf characteristics and temperature can be classified quantitatively.

For leaf characteristics, if we wanted to know the number of leaves per plant, this could be counted for each plant. Leaf number would be quantitative and discrete. The term discrete is used for this type of data because there are a limited number of alternative values (Surfstat n.d.).

Finally, temperature can be classified quantitatively using a thermometer, and attaining readings in degrees Celsius. This type of data is there quantitative and continuous since the values can be any number within a given range (Surfstat n.d.).
DATA PRESENTATION

• Often times qualitative data and quantitative data can be confused with one another. This can be due to a misunderstanding of terms, an error in data labelling or simply because there is an apparent overlap in its classification.

• For example, if we wanted to investigate the favourite colour for 30 individuals in a classroom, it should be clear that the variable “colour” would be qualitative. So even though the method of data collection— a tally - and the presentation of data below- a count – are numeric, the data still is considered qualitative.

• Consider the following graphical representation:
• In the previous example, we looked at how qualitative data could be tallied in a count and presented numerically in a graph.

• Now we can now look at an instance where quantitative data can be summarised into qualitative brackets for simplicity.

• In this example, using the same classroom example of 30 individuals, if we asked for the annual household income we can plot a distribution. But, the graph may be more visually comprehensive if each income is grouped into brackets as illustrated below.
DATA MANAGEMENT IN EXCEL

• Since Excel is commonly used to collect and store data from a single or multiple experiments, it is important to keep track of your data type and data structure where necessary, before moving forward to using your data in a statistical analysis software.

• Although some of the information on data types may seem obvious to you now, when returning to your Excel file in months or years, some of the small details may not be as easy to recall from memory.

• The practice of keeping a separate worksheet for “Metadata” is an efficient way of keeping track of the small details of the variables in your data sheets for your future self or other people.

• The column headings can be transposed into the first column of a new worksheet called “Metadata”.

• Details such as the type of data/structure, meaning of an abbreviated name or acronym, the range of the variable or to explain the calculation performed, are all useful and should be recorded.
TYPES OF DATA IN R

• Data in R is classified in two different ways. When working in R, it is likely that you will use textbooks and/or online guides to help you figure out problems when you are confused. It is noteworthy to remember the distinction between the two ways data is classified in R since not every guide will communicate this clearly.

• We will distinguish between the two classifications by referring to the first as “R Data Types” and the second as “R Data Structures”.

• These notes are intended to help you understand how your data will be classified by default when it is imported and how to reorganize your data so that it is recognized to accurately reflect the content of your data.

• We will look at what commands can be used to check what classifications R has given to each variable and what commands can be used to re-classify your variables.
R DATA TYPES

• This type of data classification more easily relates to the general types of data previously discussed: qualitative data and quantitative data.
• After importing data, R automatically recognizes and categorizes data into classes which may or may not be correct.
• There are five ways in which data in R is recognized: numeric, integer, complex, logical and character (Yau, C 2009).
  • **Numeric** values are those which contain decimal places (Yau, C 2009), any variable containing decimal values will be viewed as numeric by R.
  • **Integers** are also recognized by R from an imported dataset as variables containing integer values, or integers can be created using the `as.integer` function (Yau, C 2009).
  • **Complex** values are those containing the imaginary unit, $i$ (Yau, C 2009).
  • **Logical** values occur as either TRUE or FALSE, as a result of a comparison between variables or under some determined rule (Yau, C 2009).
  • **Character** objects are members of a sequence of values or string (Yau, C 2009).
• So now that we understand the groups that the data can be classified differently, how do we know if R has recognized our data accurately?
R DATA STRUCTURES

• This type of data classification is more related to the **structure** of the data rather than the **type** of data.
• This topic is mentioned because there is some mention of these data structures as data types in some statistics guides which may cause some confusion, and it is noteworthy to mention the difference between its
• Common data structures used are vectors, matrices and arrays, factors, lists and data frames (Wickham, H n.d.)
• When data has been imported from Excel it will be recognized as a **data frame** by default.
• Numerics, characters and logical data types are all have **vector** structures (Wickham, H n.d.)
• An imported dataset can be used to construct all of these forms of data structures.
• As we will see in the next slide, it may be necessary to convert one or more variables to a different data structure to reflect its true classification.
Here we will cover the commands which can be used to determine the classification of variables in your imported data; then we will look at ways to re-classify your data for a more accurate representation.

The command `class()` can be used to check what data type R has given a specific variable.

Alternatively, the command `str()` provides both the data structure and the data type for each variable in the dataset. This command is therefore particularly useful to get an overview of the classifications of data types all of your variables after you have made multiple changes.

The `as.` function can be used to re-assign a variable to a different data type.

For example, in an imported dataset called “Experiment2015” which includes a column listing the replicate number is included named “Rep”, R recognizes replicate as numeric variable, when it should be a factor; in order to re-classify this variable the `as.` is used:

```
Experiment2015$Repf <- as.factor(Experiment2015$Rep)
```

In the above example, a new variable is added to the dataset, as “Repf”; this ensures that the original variable “Rep” is kept and remains unchanged.
BASIC DATA MANAGEMENT IN R (2)

- The following examples are provided for a dataset called “DEMO”
- For recalling the exact names given to each variable used:
  \[ \text{names(DEMO)} \]
- To determine the data classification of the variable “postcode” in a dataset called “DEMO” is:
  \[ \text{class(DEMO$postcode)} \]
- If R has incorrectly recognized the qualitative variable “postcode” as a numerical variable (quantitative), it would need to be re-labelled as a “factor”
- Here we can use the \text{assign} command, and the \text{as.factor} command:
  \[ \text{DEMO$pc <- as.factor(DEMO$postcode)} \]
- As seen before, the above command keeps the original variable unchanged and adds a new column to the dataset, “DATA”, with the new variable called “pcfactor” with the same values as the original variable “postcode”
- For determining the data structure for this dataset, and the data type for all the variables, use:
  \[ \text{str(DEMO)} \]
CONCLUSION

• Within this presentation we looked at how data is classified and presented as either qualitative or quantitative

• Next we looked at how R reads and classifies imported data and how to organize the variables with correct data type and data structure classifications

• Remember to have a look at the listed references for more information

• More details on each command can be found within R by using the “?” command (for example if you wanted to know more about vectors, enter “?vectors” and some explanatory notes will be appear in the Help box within your R window)
REFERENCES

Types of Data:


Types of Data in R:


Data Structure in R:

