

Overview: Introduction to Probability and Random Variables

In this orientation week, we will be doing a quick revision on basic concepts of probability, types of variables and the definition of random variables so that we will be ready to start with the content which will be covered in Week 1.

For the practicals, this course will be using R as the statistical computing software to run statistical analysis tasks. In general, there are also other available alternatives to run these tasks, such as Python, SAS or SPSS. However, R is widely used amongst the statistics (and data science) community partly due to its rich resource of packages and capabilities for statistical computing. If you have not used R before, the last part of this week's content gives an introduction to get you started.

0.1 Concept of Probability and Statistics

As data science revolves around the use of data, and often massive amount of data, it is important to be able to use the right method or technique to extract meaning from these data. In data science, statistical methods provide a set of useful and essential tools to elicit patterns and relationships from the data which we have. More importantly, statistical methods also give us an insight into unique characteristics of data and the know-how to identify and deal with these characteristics. For example, numeric and non-numeric data are measured and modelled using different statistical models.

Statistical learning should not be viewed as a series of black boxes.

No single approach will perform well in all possible applications.

Without understanding of all the cogs inside the box, or the interaction between those cogs, it is impossible to select the best box. (James, Witten, Hastie and Tibshirani, 2021, p. 8)

One of the primary interests in data science is statistical learning, which refers to a set of approaches for estimating the function f between output variable(s) and a set

of input variables. The two main reasons for estimating f are prediction and inference.

- **Prediction** revolves around predicting the value of the output variable Y when a set of inputs X are available. For example, suppose that X_1, X_2, \dots, X_p are characteristics of a house and Y is the sale price. We can attempt to predict Y using X , so that we can then pitch the sale price in the best interest of the seller and/or purchaser. We typically do not expect the prediction to be perfect hence it will also be of interest to understand the errors associated with it.
- **Inference** problems are interested in understanding the association between the output variable Y and a set of input variables X_1, X_2, \dots, X_p . For example, we may ask "Which predictors are associated with the sale price of a house?" or "What is the relationship between the sale price and each predictor?"

There are many linear and non-linear approaches for estimating the function f , which will be further investigated in the subsequent weeks in this course.

Prevalent in the field of statistical studies is the concept of probability, which is used to quantify uncertainty in various scenarios ranging from weather forecasting to stock price prediction. In statistics, probability values are between 0 and 1, inclusive.

The following reading list gives an overview on statistics in general and also an introduction to probability. Going through the chapter on probability will prepare you for next week's topic on probability axioms.

Ware, W.B., Ferron, J.M., & Miller, B.M. (2012). Introductory Statistics: A Conceptual Approach Using R. Taylor and Francis.

<https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=cat01176a&AN=thol.5304207&scope=site&custid=s7320401>

- Chapter 1: Introduction and Background (pp. 14-15).

Mann, P.S. (2021). Introductory Statistics. Wiley.

<https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=cat01176a&AN=thol.5304207&scope=site&custid=s7320401>

- Chapter 1: Introduction (pp. 1-7).
- Chapter 4: Probability (pp. 141-147).

0.2 Random Variables

Why is it important to classify variables? Different statistical tools and methods are required to analyse different types of variables. For example, using mean or average to describe a qualitative variable is inappropriate whereas the mode is often not informative when it is used to describe quantitative variables. Watch the following video for a quick overview on random variables.

<https://www.youtube.com/watch?v=WJLEnu3rgSM>

Source: (Selina YC Low, 2023)

Problems with a quantitative response (or output variable) is referred to as regression problems, whereas problems involving a qualitative response variable are often known as classification problems. However, do note that there are exceptions. For example, logistic regression is a method used with a qualitative response hence it is a classification method despite its name bearing the word "regression". Some approaches can be used with both quantitative and qualitative responses.

Our decision in selecting statistical learning methods is usually made based on the type of the response variable. Linear regression is a well-known option for quantitative response when the underlying assumptions are met. Logistic regression has been discussed in the preceding paragraph.

0.3 Introduction to R

In this course, all statistical programming will be done using the R programming language. If you have not used R before, watch the following video for a quick

introduction to R. Feel free to skip this video if you are already familiar with the R environment.

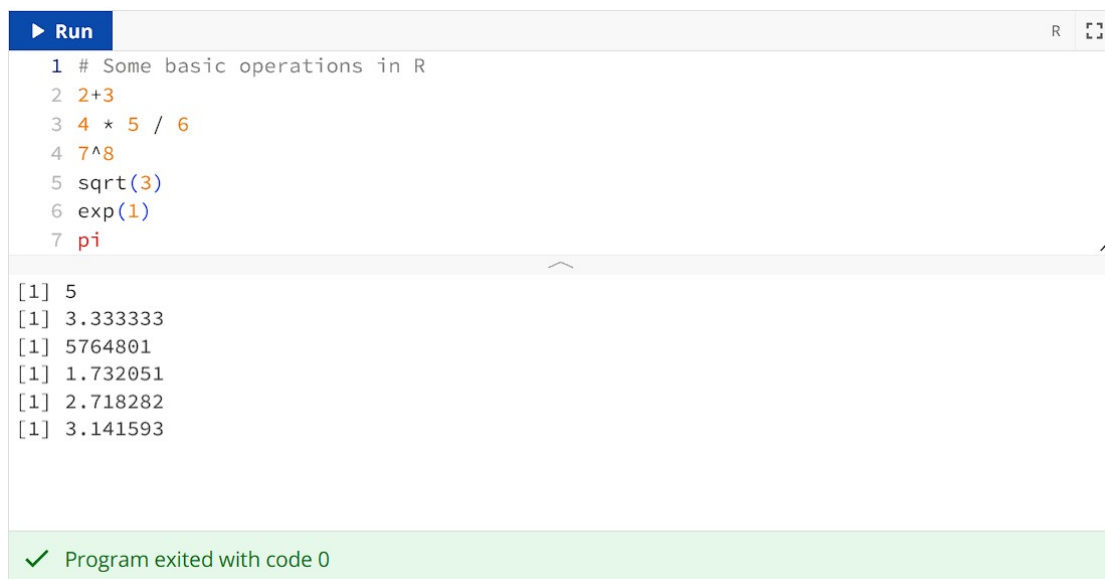
Intro to R

<https://www.youtube.com/watch?v=gD5JX3Zs7iQ>

Source: ([Selina YC Low](#), 2023)

Basic Operations

R can perform basic arithmetic operations. The commenting operator in R is the hash (#) key. Anything typed after a # symbol is ignored by R. Run the following R code and examine the output.



```
▶ Run R [ ]
1 # Some basic operations in R
2 2+3
3 4 * 5 / 6
4 7^8
5 sqrt(3)
6 exp(1)
7 pi

[1] 5
[1] 3.333333
[1] 5764801
[1] 1.732051
[1] 2.718282
[1] 3.141593

✓ Program exited with code 0
```

In R, we can add two or more vectors of the same length. The length of a vector can be checked using the `length()` function. The number of rows and columns in a matrix can be examined using the `dim()` function. Run the following code and associate the outputs to their respective functions.

```
► Run R [ ]
1 x <- c(74, 35, 98, 51, 76, 34, 13, 56, 94)
2 y <- c(1, 2, 3, 5, 6, 10, 20, 30, 40)
3 A <- matrix(1:16, 4, 4)
4 length(x)
5 len_y <- length(y)
6 x + y
7 dim(A)

[1] 9
[1] 75 37 101 56 82 44 33 86 134
[1] 4 4

✓ Program exited with code 0
```

Loading Data into R

In most cases, the first step of an analysis is to load a data set into R. Before attempting to load a data set, we must make sure that R is searching in the correct directory. The approach to ensure this depends on the operating system that is being used.

R can read file types of the following: `.txt`, `.csv`, `.tsv`, `.xsl`.

The following functions are used for reading data into R:

- `load()`: all of the R objects saved in the file are loaded into R. The names given to these objects when they were originally saved will be given to them when they are loaded.
- `read.table()`: read delimited data files (space delimited, tab delimited, etc)
- `read.csv()`: read comma-delimited files

In this course, three data sets are being used for learning purposes.

- `AmesHouseNormal.csv` data set
- `campaign-success.csv` data set
- `bikesharing2011.csv` data set

The following is an example of an R code to load and examine the `AmesHouseNormal.csv` data set.

```
1 AmesHouse <- read.csv("AmesHouseNormal.csv", header=T, stringsAsFactors = T)
2 View(AmesHouse) # to view the data set in a spreadsheet like window
3 head(AmesHouse) # to view the first few rows of the data
4 head(AmesHouse[1:6], n=10)
5 dim(AmesHouse)
6 AmesHouse <- na.omit(AmesHouse) # remove rows containing missing observations
```

Variables

In an earlier video we have discussed the types of variables, namely qualitative and quantitative variables. Quantitative variables can be further classified as either discrete or continuous variables. Let us look at some examples of variables in the datasets package that comes with base R.

Example 1: The U.S. Geological Survey recorded the lengths (in miles) of several rivers in North America. They are stored in the vector rivers in the datasets package.

```
► Run
1 str(rivers)

num [1:141] 735 320 325 392 524 ...

✓ Program exited with code 0
```

The output says that rivers is a numeric vector of length 141, and the first few values are 735, 320, 325 and so on. These data are quantitative and measurements have been rounded to the nearest mile. Therefore, it can be considered as discrete data. However, in some of the statistical procedures it may be more convenient to take data like these as continuous.