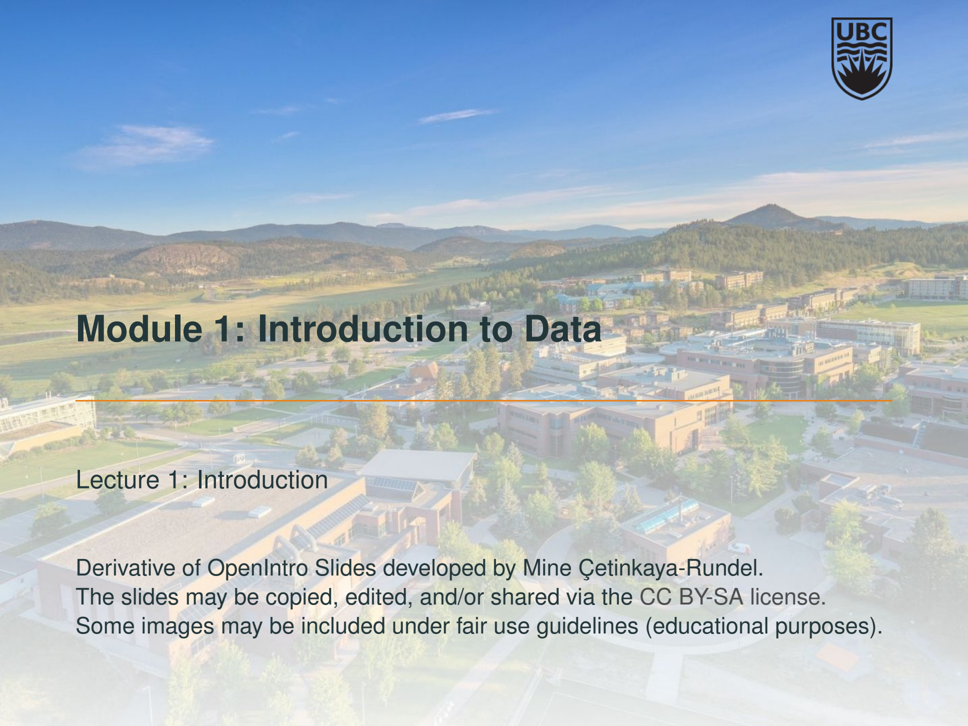
Module 1: Introduction to Data

Lecture 1: Introduction



What is Statistics? (and why you should care)

- Statistics is the science of learning from data.
- Everything dealing with the collection, processing, analysis, and interpretation of data belongs to the domain of statistics.
- Statistics assists you in making decisions in the light of uncertainty
- In general, a base knowledge of statistics is important, or at the very least highly useful, for everyone who might
    - watch the nightly news or read a newspaper
    - play cards, board games, or gamble
    - research the risks of a medical drug
    - ...

Data is abundant and everywhere!

- In 2020, people created 1.7 MB of data every second. [techjury].
- There are many different types of data, corresponding to different types of information.
- In this course, we will most often be concerned with numeric measurements.
- Broadly, data can be qualitative (observations that can be sorted into groups/categories) or quantitative (observations that can be measured or counted).

Types of Qualitative Data

Categorical/Nominal can take on one of a limited (and usually fixed) number of possible values

- Eye colour : blue, green, brown, grey.
- Nationality : Irish, English, Scottish, French, Welsh, Canadian, etc.

Ordinal Exist on an ordinal scale, i.e. data can be sorted by rank

- Dose: less than 10mg; 10 to 20mg; more than 20mg.
- Questionary: Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree
- Rating from 1 to 10

Types of Qualitative Data

Binary Data which can only take on two possible values. This is really just a special case of categorical data (when the number of categories is 2).

- Are you diabetic? yes or no.
- What is the outcome of flipping a coin? heads or tails.

Numerical Data

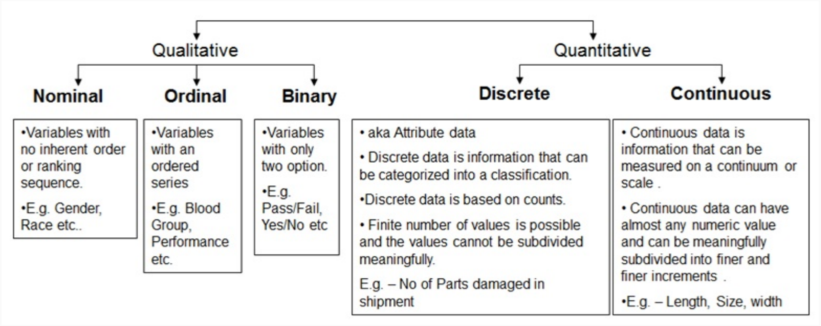
Continuous data arises when all values are possible inside some interval on the real number line (including open-ended intervals).

- volume of water inside a glass
- distance between cities

Discrete data arises when the possible items are countable

- number of hurricanes that hit land each year (e.g. 0, 1, 2, 3,. . .)
- number of flips needed to get 10 heads with a flip of a coin (eg. 10, 11, 12, . . .)

Types of Data



Qualitative
Quantitative
Nominal
Ordinal
Binary
Discrete
Continuous
•Variables with no inherent order or ranking sequence.
•E.g. Gender, Race etc..
•Variables with an ordered series
•E.g. Blood Group, Performance etc.
•Variables with only two option.
•E.g. Pass/Fail, Yes/No etc
• aka Attribute data
• Discrete data is information that can be categorized into a classification.
•Discrete data is based on counts.
• Finite number of values is possible and the values cannot be subdivided meaningfully.
E.g. – No of Parts damaged in shipment
• Continuous data is information that can be measured on a continuum or scale .
• Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
•E.g. – Length, Size, width

Figure 1: Summary of the different types of data. Source for image: here

Some Vocabulary and Notation

- A population is an all-encompassing group of interest; usually unobservable in its entirety for one reason or another (most often the cost of measurement).
  - UBC Okanagan students
  - Rocks on Mars
- A sample consists of a number of observations.
  - The group of you sitting in this classroom is a (biased) sample of UBC students...and each of you individually are an observation.
  - NASA's *Curiosity* rover has been drilling for rock samples on Mars.
- We often denote observations $x_1, x_2, \ldots, x_n$. In this way, x_1 denotes the first observation, x_2 the second, and so on.

Descriptive Statistics

Descriptive Statistics I

- Consider observing n observations : x_1 ,x_2 ,. . .,x_n.
- We call this collection of observed data a sample which we sometimes denote by $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ (set notation).
- The mean, or average, of the sample is denoted $\bar{x}$ and is given by,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

- The standard deviation of this sample s is given by,

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Descriptive Statistics II

- The variance is the standard deviation squared s^2. Informally, it measures how far a set of (random) numbers are spread out from their average value.

- The range of a sample is the distance from the minimum observation to the maximum observation.

  range = maximum - minimum

- The mode is the most frequently occurring observation

- When the observations are ordered from smallest to largest, the median is the ‘middle’ observation.

Descriptive Statistics III

- The pth percentile is a value that has $100p\%$ of the ordered data falling below it, and $100(1-p)\%$ of the ordered data falling above it.
    - eg. the median is the 50th percentile

- We call the 25th percentile the first quartile (Q_1) and the 75th percentile the third quartile (Q_3)

- The interquartile range (IQR) is the size of the gap between the first and the third quartile. It is the ‘distance’ over which the ‘middle half’ of the data is spread.

IQR = Q3 - Q1

Exercises

Consider the following data: 6, 7, 8. What is the mean and standard deviation?

The mean of four numbers is 12. Three of these numbers are 2,11 and 19. What is the other number?

Exercise

Example 1

Consider the following data: 0.13, 0.25, 0.31, 0.44, 0.49, 0.51, 0.55, 0.59, 0.70, 0.81, 12.00. What is the IQR? Given that s=3.48, comment on the difference between s and the IQR.

- The example on the previous slide demonstrates the drastic affect that outliers can have on statistics like standard deviation.
- An outlier is loosely defined as an observation that lies unusually far from the main body of the data.
- We say that the IQR is more *robust* to outliers, in that the presence of a few outliers will not greatly sway it’s value.
- To provide another example, the mean is very susceptible to outliers, while the median is a robust statistic that is more resistant to outliers.

Doing these calculations in R

These calculations are quite tedious to do by hand. In R, these calculations can be done using the following commands:

```
> x <- c(6, 7, 8)
> mean(x)
[1] 7
> sd(x)
[1] 1
```

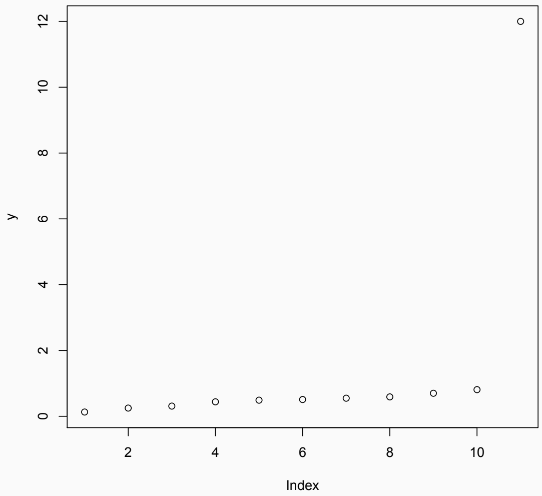
N.B. R haves a function for calculating IQR, but it does so using a slightly different method than we've done on paper.

Graphs

- Looking at data is a very useful first step.
- Often, a picture is the best summary of data.
- We will see various graphs as we go along.

Plot the data for Example 1

For instance, plotting the ordered data from Example 1 makes it very obvious that we have an outlier



Histograms

- A histogram is a graphical display of data which shows the underlying frequency distribution (shape).
- To construct a histogram, we first divide the split of data into intervals (called bins).
- Each bin will have a corresponding bar who's height is equal to the number of data points which falls into that interval of numbers.
- Histograms provide a great visual of the underlying distribution, i.e. is the distribution symmetrical, skewed, does it conatin outliers, etc.

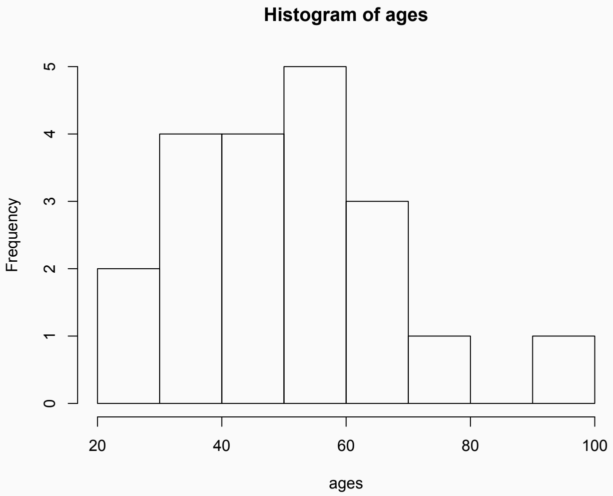
Histogram

36	25	38	46	55	68
67	45	22	48	91	46

Suppose we have data comprised of the 12 ages above. We could split our data into bins of 10-year periods starting at 20 years.

Bin	Frequency	Ages Included in Bin
20-30	2	25,22
30-40	4	36,38,36,38
40-50	4	46,45,48,46
50-60	5	55,55,52,58,55
60-70	3	68,67,61
70-80	1	72
80-90	0	-
90-100	1	91

This histogram would look like this:



Bin	Frequency	Ages Included in Bin
20-30	2	25,22
30-40	4	36,38,36,38
40-50	4	46,45,48,46
50-60	5	55,55,52,58,55
60-70	3	68,67,61
70-80	1	72
80-90	0	-
90-100	1	91

Histogram Shapes

Histograms come in a variety of shapes.

unimodal histogram is one that rises to a single peak and then declines (like the example on previous slide)

bimodal histogram has two different peaks

multimodal histogram with more than two peaks

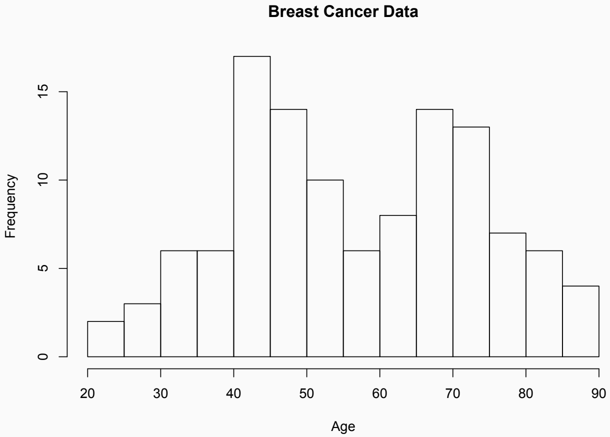
symmetric if the left half is a mirror image of the right half.

positively skewed if the right or upper tail is stretched out compared with the left or lower tail

negatively skewed if the stretching is to the left.

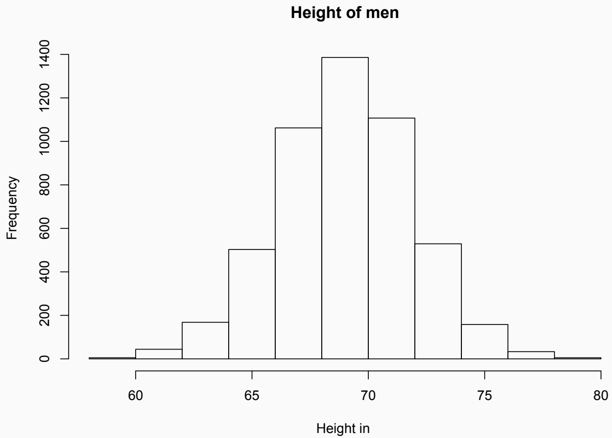
Bimodal Histogram

Consider the age of patients from the Breast Cancer Coimbra Data Set. Since the histogram has two ”humps” we say it is bimodal.



Symmetric Histogram

Consider the height of men from this kaggle data set. Drawing a line down the center, the left roughly mirrors the right; so the histogram is symmetric. Since it has one hump, it is unimodal.



Unimodal vs Bimodal

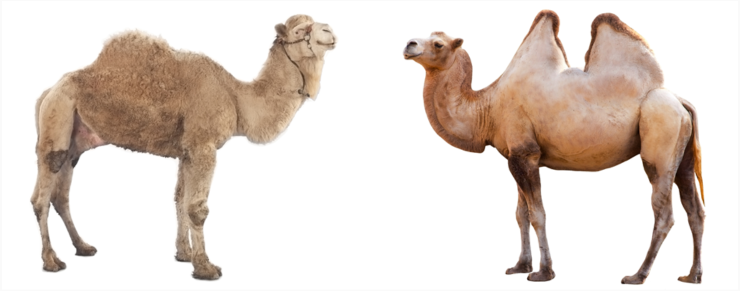
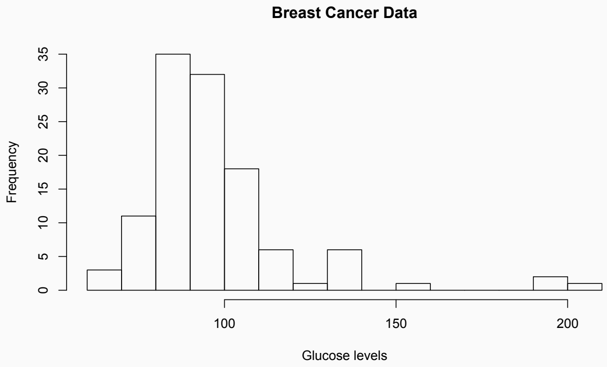


Figure 2: Left: unimodal Right: bimodal

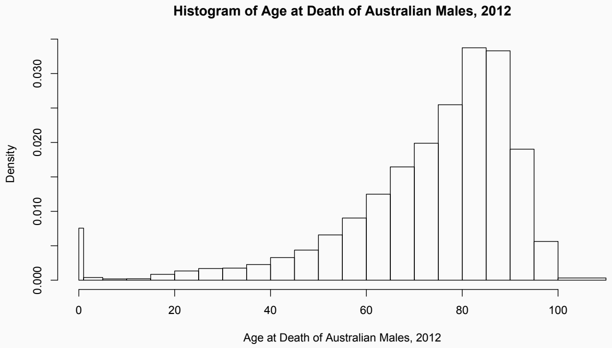
Positively skewed (or right skewed) Histogram

Consider the glucose level of patients from the Breast Cancer Coimbra Data Set. Since the right or upper tail is stretched out compared with the left or lower tail we say it is ”skewed right” or ”positively skewed”.



Negatively skewed (or left skewed) Histogram

Consider the age at death of Australian males from 2012 from the Australian Bureau of Statistics (Table 8). Since the left (ie lower) tail is stretched out compared with the right (i.e. upper) tail we say the histogram is ”skewed left” or ”negatively skewed”.



Boxplots

- Histograms are able to encapsulate the general shape of a dataset, whereas a single summary statistic (eg. mean or standard deviation) focuses on a single aspect of the data (eg. center and spread, resp)
- A boxplot provides a graphical summary of a data set along with the following features:
    - center (median)
    - spread (IQR)
    - symmetry/departure from symmetry (as indicated by the shape of the “box”)
    - outliers (indicated as individual points)

Boxplot outliers

- There are a number of ways one could provide a cut off for determining outliers.
- For boxplots, observations are flagged as outliers if they are either:
  - $\geq Q3 + 1.5*$IQR or
  - $\leq Q1 - 1.5*$IQR
- Not all data will have outliers.
- The ends of the "whiskers" represent the maximum and minimum observation that are not considered outliers.

Boxplot

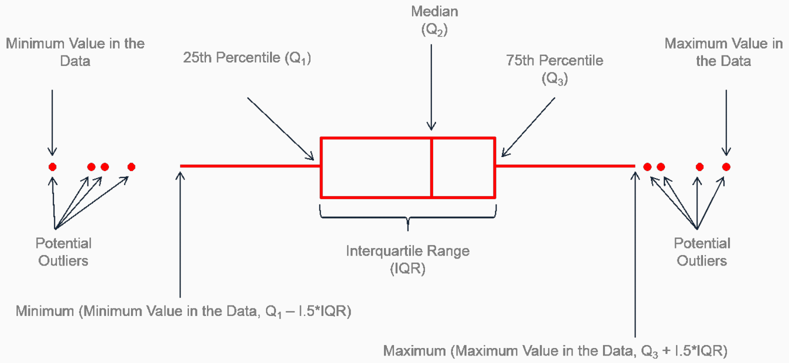
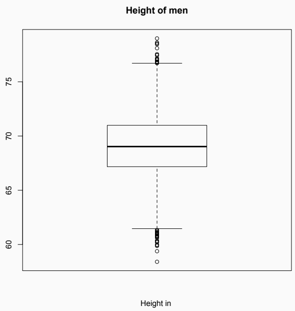
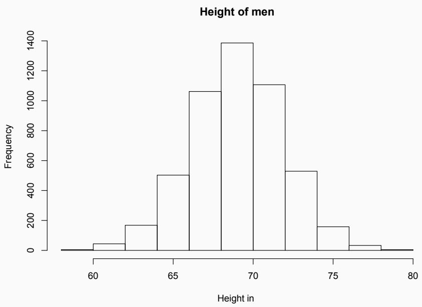

Median (Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, $Q_1 - 1.5*IQR$)
Maximum (Maximum Value in the Data, $Q_3 + 1.5*IQR$)

Image source

Symmetric Boxplot

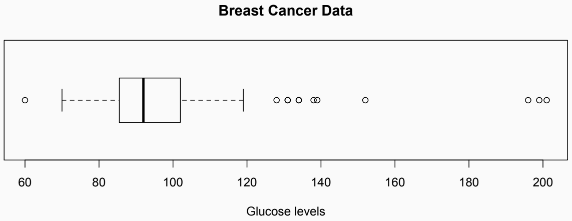
Consider the height of men from the kaggle data set.

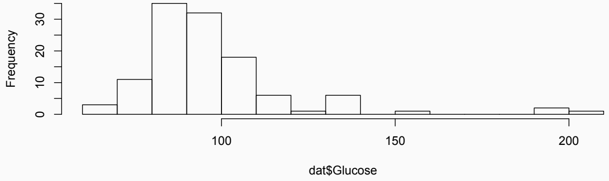




Positively skewed (or right skewed) Boxplot

Consider the glucose level of patients from this data set.

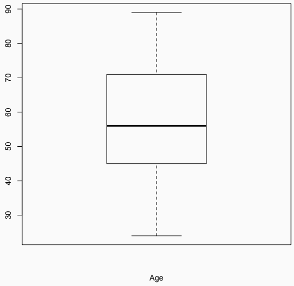


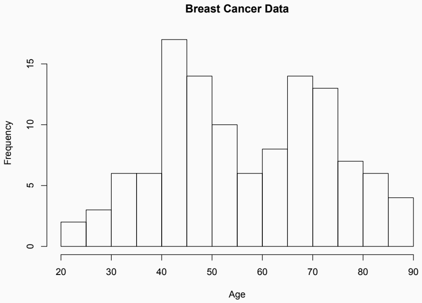


Boxplot of Bimodal data

Notice how bimodal data may not be apparent in boxplots.

Consider the age of patients from this data set.





R: A (very) brief introduction

- And now, a quick introduction to R, . . .
- As mentioned before, you will learn mostly about R in labs
- But I'd like to go few some of the basics here to help you get started.

Download R

- First and foremost you will need to download R
- If you’ve already downloaded R on your computer, make sure you check that you have the latest version
  - R version 4.1.1 (Kick Things) has been released on 2021-08-10.
- Secondly, you’ll want to download Rstudio.
- While you can use R without RStudio it is highly recommended that you use RStudio.

What is R?

R is a free and open source programming language for statistical computing and graphics.

1. One of the most widely used programming languages for statistical analysis.
2. Popular in academia and companies like Microsoft, Google, and Facebook.
3. There are 10’s of thousands of packages (units of reproducible R code) that are available for uses to use for free.
4. R creates high quality graphs and visualizations.

RStudio

RStudio is an integrated development environment (IDE) that uses the R language.

- For this course, this means we will be able to writes and edit code in the RStudio's editor.

It runs on all platforms (Windows, Mac OSX, UNIX, LINUX).

It was started by Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland in 1995.

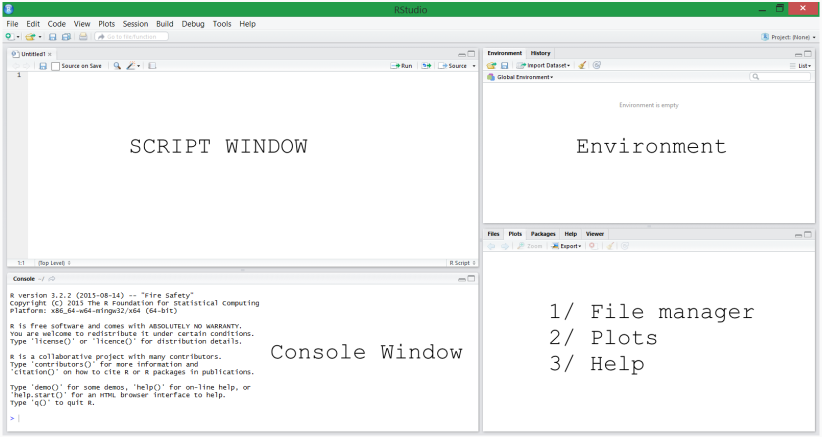
RStudio

It uses a command interface, so it has a steeper “learning curve” than other programs, as there are no menus.

Once you get the hang of using R, it is much easier and is great for doing the same thing a large number of times.

Apart from one-off calculations, you should ALWAYS save a file (or R *script*) with the commands used for any analysis. In this way, R is much better than menus in that you can always reproduce an old analysis.

RStudio Environment



```
R version 3.2.2 (2015-08-14) -- "Fire Safety"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

RStudio IDE

Script Window

Draft and save code. Write a script to run in the console by Cmnd/CTRL (Mac/Windows) +R or Cmd/CTRL+Enter, or pressing the **Run** located at the top of the window.

Console

Where the code goes once run. Shows input (blue), output (black) and any errors or warnings (red). (different “Themes” can be chosen to display different colours.

Environment

Shows saved variables and datasets

File Browser/Plots/Help

Show files in working directory and generated plots, help, open window.

Basics of R

R can be used as a glorified calculator.

Operation	Command
addition	`a+b`
subtraction	`a-b`
multiplication	`a*b`
division	`a/b`
powers	`a^2`
modulos	`a%%2`

N.B.[1]

[1] be careful with copying and pasting minus signs

Basics of R

- To begin we can type calculations directly into the *Console* window.
- There you should see a greater than sign (>), referred to as a *prompt*.
- N.B. if ever you see a plus sign (+) instead of the prompt, it means you haven't completed a command on a previous line (to get rid of the plus sign you will either need to complete the command or press ESC)

Basics of R

Spaces between operators do not matter

```
> 2       +       3
[1] 5
> 2 + 3
[1] 5
> 2+3
[1] 5
> 2 +
+
+ # press ESC to return to prompt
>
```

RStudio

- While we can type code directly into the console, we will usually be working with a script in the *Script Window*
- An **R script** is a file containing commands that you would otherwise be typing in the console (i.e. command line) of R
- R scripts are saved with the extension `.R`
- Files of these type will open in the Script Window (AKA the Editor of RStudio)[2]
- All the R code for todays lecture has been saved to the `01-intro.R` file uploaded to Canvas.

[2]By default this might open in R (rather than Rstudio).

Basics of R

Comments are used by the programmer to document and explain the code.

Comments are not executed by R and should give information about the code to the person reading.

Comment lines begin with #.

There is no multiline commenting in R.

```
> #This is a comment.
```

Basics of R

Variables are assigned with `<-` or = or <- Note: there is some debate on the preferred syntax.

```
> x <- 4
> y <- 10
> x + y
[1] 14
```

To get help type

```
> help(c)
> ?c
```

(or google it–`https://stackoverflow.com` is a great resource)

Note that > indicates a newline in R and `[1]` represents output

R fundamentals

Everything in R is case sensitive:

```
> x <- 4
> x
[1] 4
> X
Error: object 'X' not found
```

R commands may be separated either by a semi-colon or a newline.

```
> 1 + 2; 2+ 3
[1] 3
[1] 5
```

Curly brackets { } are used to group commands together.

Vectors in R

Rather than storing a single value to a variable, we can store multiple. In R we call these vectors.

```
vec <- c(24, 200, 90, 30, 79, 90, 0, 200)
```

- A common mistake is to forget the “c” before the first parenthesis.
- Failing to follow R’s “rules” will result in a syntax error
- Syntax errors always result in an error message. For example:

```
> vec <- (24, 200, 90, 30, 79, 90, 0, 200)
Error: unexpected ’,’ in "vec <- (24,"
```

Vectors in R

Below are some basic operations for vectors.

Operation	Command
Number of elements in x	`length(x)`
Unique elements of x	`unique(x)`
Mean	`mean(x)`
Variance	`var(x)`
Standard Deviation	`sd(x)`
Minimum value	`min(x)`
Maximum value	`max(x)`
Smallest and largest values	`range(x)`
median	`median(x)`

```
> length(vec)
[1] 8
> unique(vec)
[1]  24 200  90  30  79   0
> mean(vec)
[1] 89.125
> var(vec)
[1] 5767.268
> sd(vec)
[1] 75.94253
> min(vec)
[1] 0
> max(vec)
[1] 200
> range(vec)
[1]   0 200
> median(vec)
[1] 84.5
```

Basic Plotting

One of the main reasons data analysts turn to R is for its strong graphic capabilities.

For instance we could create the plot on this slide using:

```
data <- c(0.13, 0.25, 0.31, 0.44, 0.49, 0.51,
0.55, 0.59, 0.70, 0.81, 12.00)
plot(data)
```

We could also create boxplots and histograms using:

```
boxplot(data)
hist(data)
```

A note on rounding

- For this course, I will be working under the assumption that all rounding is only done at the *very end* or multi-steep calculations.
- Round at every step can result in an answer outside of the threshold of a correct answer in a Canvas quiz.
- While I suspect you can all round properly at this point, R does have a rounding function you could use:

```
> round(0.1324234)
[1] 0
> round(0.1324234, digits=4)
[1] 0.1324
> ?round
```

Closing remarks

- This demonstration of R code was meant to highlight some of the basic capabilities of R.
- In labs you will go through R tutorials to help you get a better feel for how R works, and eventually how we can use R to perform the methods we discuss in class.
- R code and output will be tested on both assignments and midterms/exam. So please make sure you keep up with the R components in lab!
- For next class, please ensure that you are ready to use iClicker!