Systems Immunology Workshops

Center for Computational Biomedicine

1/18/23

Table of contents

Systems Immunology Workshops

We will be working from this workbook for our first 4 workshop sessions.

Before the first session, be sure to complete all pre-work steps.

Note: These materials are still a little unrefined, I mainly plan to add reccomended additional readings a resources, as well as a thorough proofreading

Part I

Pre-Work

Install R and RStudio

Before the first session, please install R and RStudio following the instructions Chapter 1.

If you already have R and RStudio installed, make sure that you have the latest versions installed (you can do this by simply following the installation instructions). While this will likely not cause an issue for the first few sessions, it will in later sessions when we use more advanced packages and software.

If you encounter issues installing R or RStudio, please reach out to `christopher_magnano@hms.harvard.edu` or one of the TAs. If we are unable to resolve your issue via email, we ask that you come 30 minutes early to the first session.

Familiarize yourself with RStudio

If you have never used RStudio or are completely new to programming, please review Chapter 2. This material will introduce you to the RStudio interface and how to assign values to variables in R.

1 Installing R and RStudio

1.1 Mac Users

1.1.1 To install R

1. Open an internet browser and go to www.r-project.org.
2. Click the "download R" link in the middle of the page under "Getting Started."
3. Select a CRAN location (a mirror site) and click the corresponding link.
4. Click on the "Download R for (Mac) OS X" link at the top of the page.
5. Click on the file containing the latest version of R under "Files."
6. Save the .pkg file, double-click it to open, and follow the installation instructions.
7. Now that R is installed, you need to download and install RStudio.

1.1.2 To install RStudio

1. Go to www.rstudio.com and click on the "Download RStudio" button.
2. Click on "DOWNLOAD" in the upper right corner.
3. Download the Free version of RStudio Desktop.
4. Save the .dmg file on your computer, double-click it to open, and then drag and drop it to your applications folder.

1.2 Windows Users

1.2.1 To install R

1. Open an internet browser and go to www.r-project.org.
2. Click the "download R" link in the middle of the page under "Getting Started."
3. Select a CRAN location (a mirror site) and click the corresponding link.
4. Click on the "Download R for Windows" link at the top of the page.
5. Click on the "install R for the first time" link at the top of the page.
6. Click "Download R for Windows" and save the executable file somewhere on your computer. Run the .exe file and follow the installation instructions.
7. Now that R is installed, you need to download and install RStudio.

1.2.2 To install RStudio

1. Go to www.rstudio.com and click on the "Download RStudio" button.
2. Click on "DOWNLOAD" in the upper right corner.
3. Download the Free version of RStudio Desktop.
4. Save the executable file. Run the .exe file and follow the installation instructions.

1.3 Reference

Instructions adapted from guide developed by HMS Research computing

2 Introduction to RStudio

2.1 Learning Objectives

- Describe what R and RStudio are.
- Interact with R using RStudio.
- Familiarize various components of RStudio.

2.2 What is RStudio?

RStudio is freely available open-source Integrated Development Environment (IDE). RStudio provides an environment with many features to make using R easier and is a great alternative to working on R in the terminal.

- Graphical user interface, not just a command prompt
- Great learning tool
- Free for academic use
- Platform agnostic
- Open source

2.3 Creating a new project directory in RStudio

Let's create a new project directory for Systems Immunology.

1. Open RStudio
2. Go to the `File` menu and select `New Project`.
3. In the `New Project` window, choose `New Directory`. Then, choose `New Project`. Name your new directory whatever you want and then "Create the project as subdirectory of:" the Desktop (or location of your choice).
4. Click on `Create Project`.
5. After your project is completed, if the project does not automatically open in RStudio, then go to the `File` menu, select `Open Project`, and choose `[your project name].Rproj`.
6. When RStudio opens, you will see three panels in the window.

7. Go to the `File` menu and select `New File`, and select `R Script`. The RStudio interface should now look like the screenshot below.

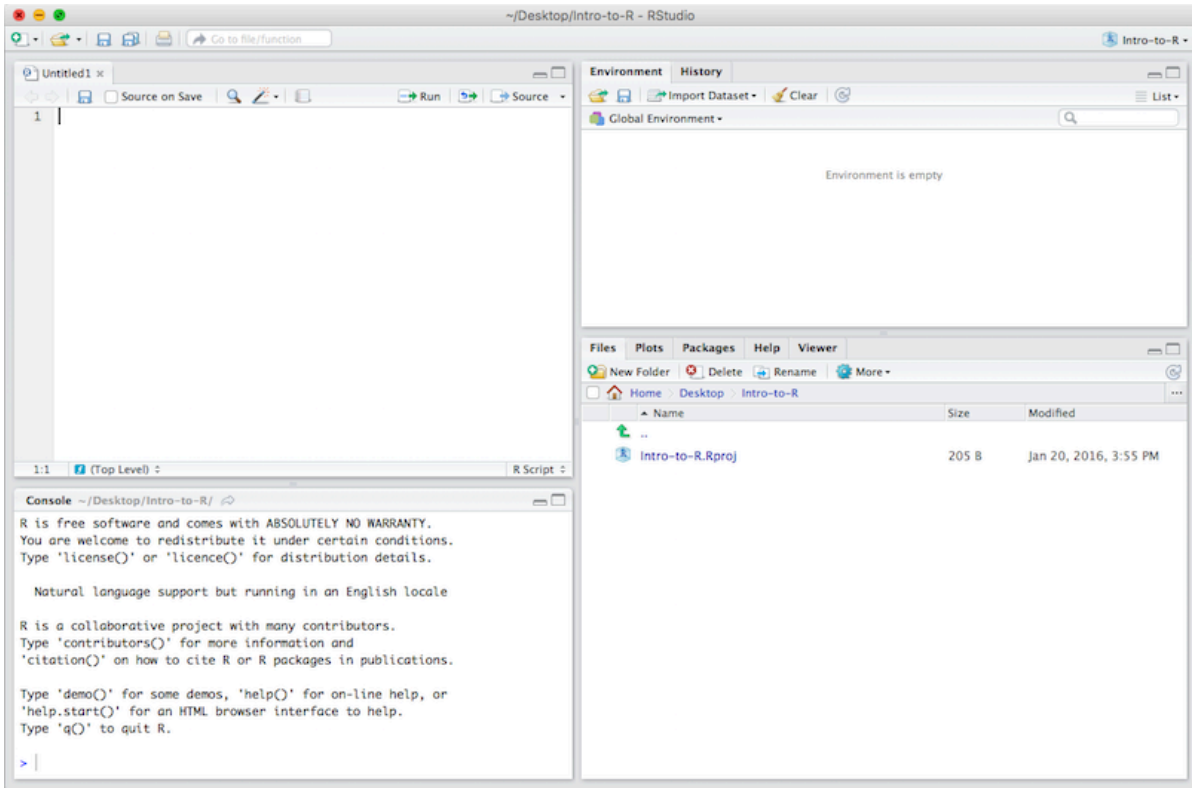


Figure 2.1: RStudio interface

2.3.1 What is a project in RStudio?

It is simply a directory that contains everything related your analyses for a specific project. RStudio projects are useful when you are working on context- specific analyses and you wish to keep them separate. When creating a project in RStudio you associate it with a working directory of your choice (either an existing one, or a new one). A `. RProj file` is created within that directory and that keeps track of your command history and variables in the environment. The `. RProj file` can be used to open the project in its current state but at a later date.

When a project is **(re) opened** within RStudio the following actions are taken:

- A new R session (process) is started
- The .RData file in the project's main directory is loaded, populating the environment with any objects that were present when the project was closed.

- The .Rhistory file in the project's main directory is loaded into the RStudio History pane (and used for Console Up/Down arrow command history).
- The current working directory is set to the project directory.
- Previously edited source documents are restored into editor tabs
- Other RStudio settings (e.g. active tabs, splitter positions, etc.) are restored to where they were the last time the project was closed.

Information adapted from RStudio Support Site

2.4 RStudio Interface

The RStudio interface has four main panels:

1. **Console**: where you can type commands and see output. *The console is all you would see if you ran R in the command line without RStudio.*
2. **Script editor**: where you can type out commands and save to file. You can also submit the commands to run in the console.
3. **Environment/History**: environment shows all active objects and history keeps track of all commands run in console
4. **Files/Plots/Packages/Help**

2.5 Organizing your working directory & setting up

2.5.1 Viewing your working directory

Before we organize our working directory, let's check to see where our current working directory is located by typing into the console:

```
getwd()
```

Your working directory should be the `Intro-to-R` folder constructed when you created the project. The working directory is where RStudio will automatically look for any files you bring in and where it will automatically save any files you create, unless otherwise specified.

You can visualize your working directory by selecting the `Files` tab from the **Files/Plots/Packages/Help** window.

If you wanted to choose a different directory to be your working directory, you could navigate to a different folder in the `Files` tab, then, click on the `More` dropdown menu and select `Set As Working Directory`.

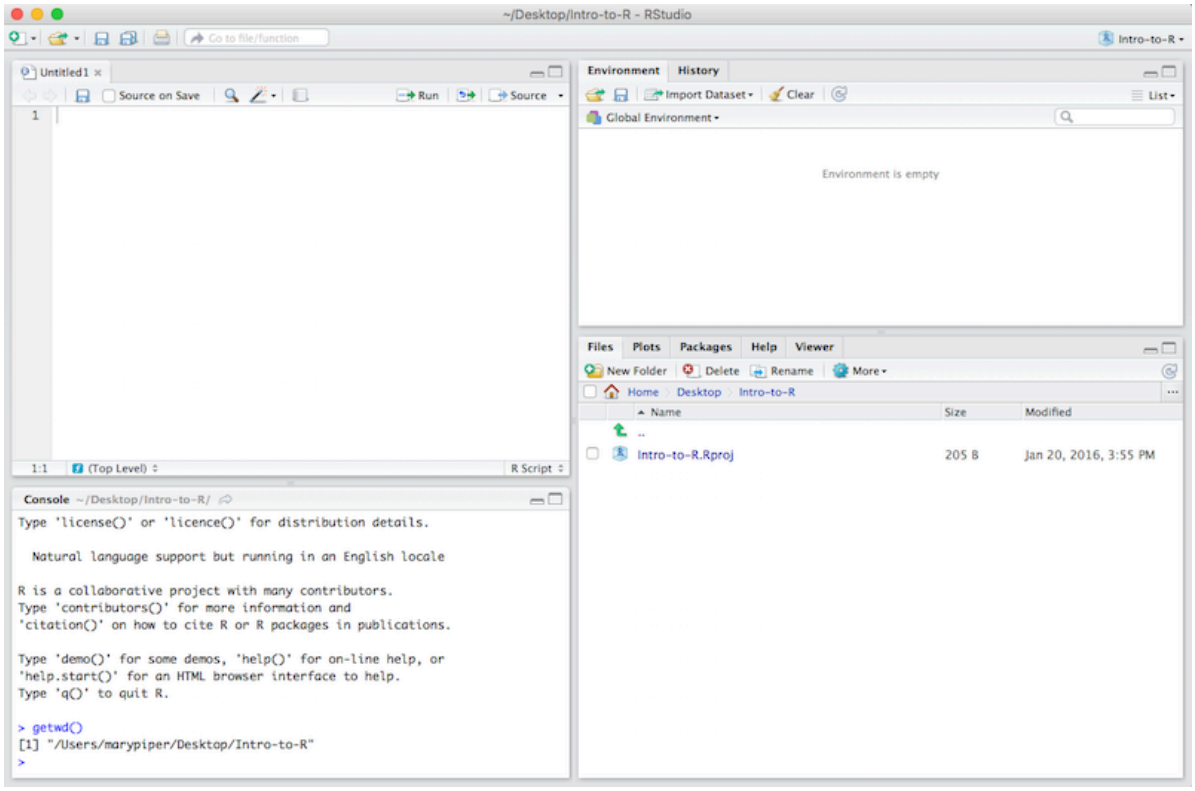


Figure 2.2: Viewing your working directory

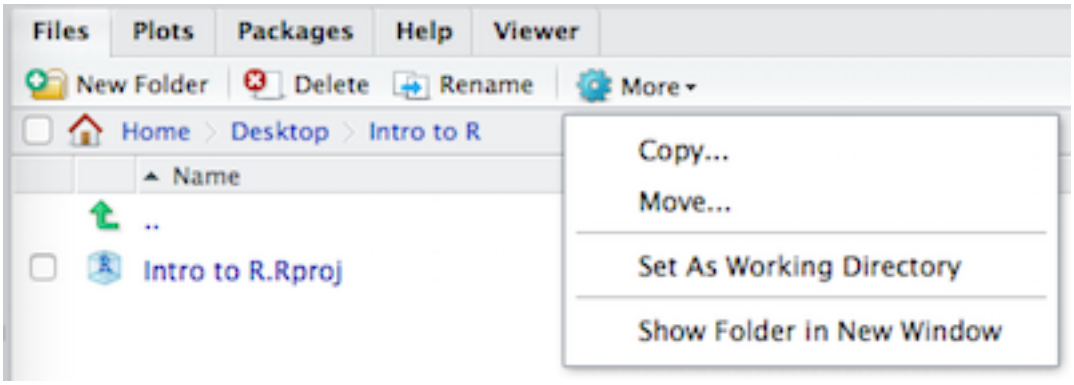


Figure 2.3: Setting your working directory

2.5.2 Structuring your working directory

To organize your working directory for a particular analysis, you typically want to separate the original data (raw data) from intermediate datasets. For instance, you may want to create a `data/` directory within your working directory that stores the raw data, and have a `results/` directory for intermediate datasets and a `figures/` directory for the plots you will generate.

Let's create these three directories within your working directory by clicking on `New Folder` within the `Files` tab.

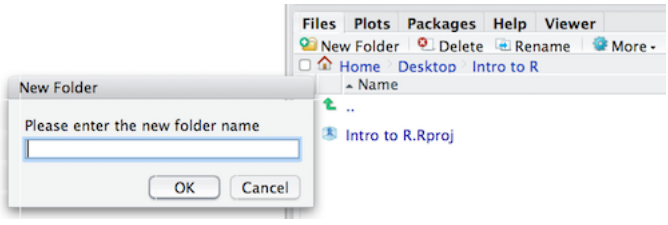


Figure 2.4: Structuring your working directory

When finished, your working directory should look like:

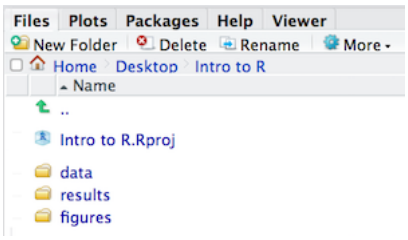


Figure 2.5: Your organized working directory

2.5.3 Setting up

This is more of a housekeeping task. We will be writing long lines of code in our script editor and want to make sure that the lines "wrap" and you don't have to scroll back and forth to look at your long line of code.

Click on "Tools" at the top of your RStudio screen and click on "Global Options" in the pull down menu.

On the left, select "Code" and put a check against "Soft-wrap R source files". Make sure you click the "Apply" button at the bottom of the Window before saying "OK".

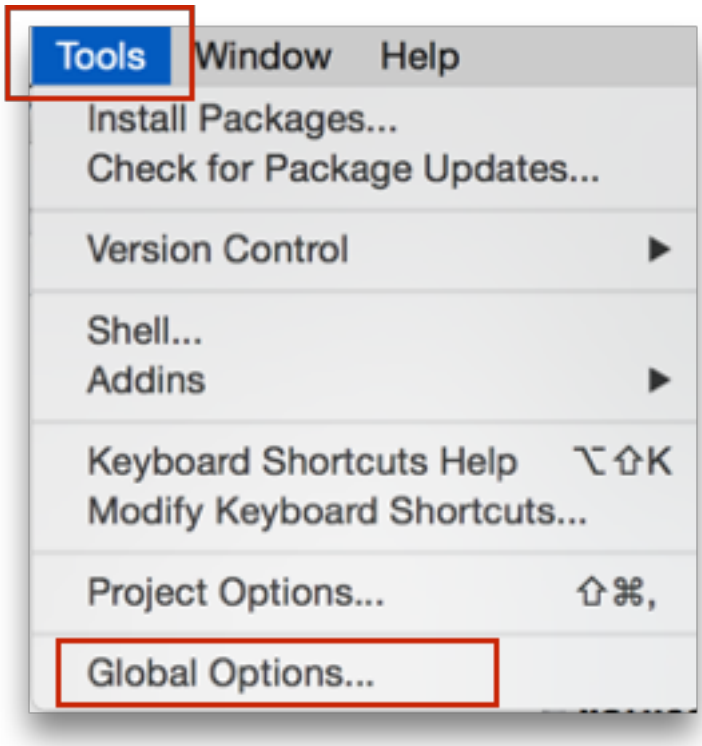


Figure 2.6: options

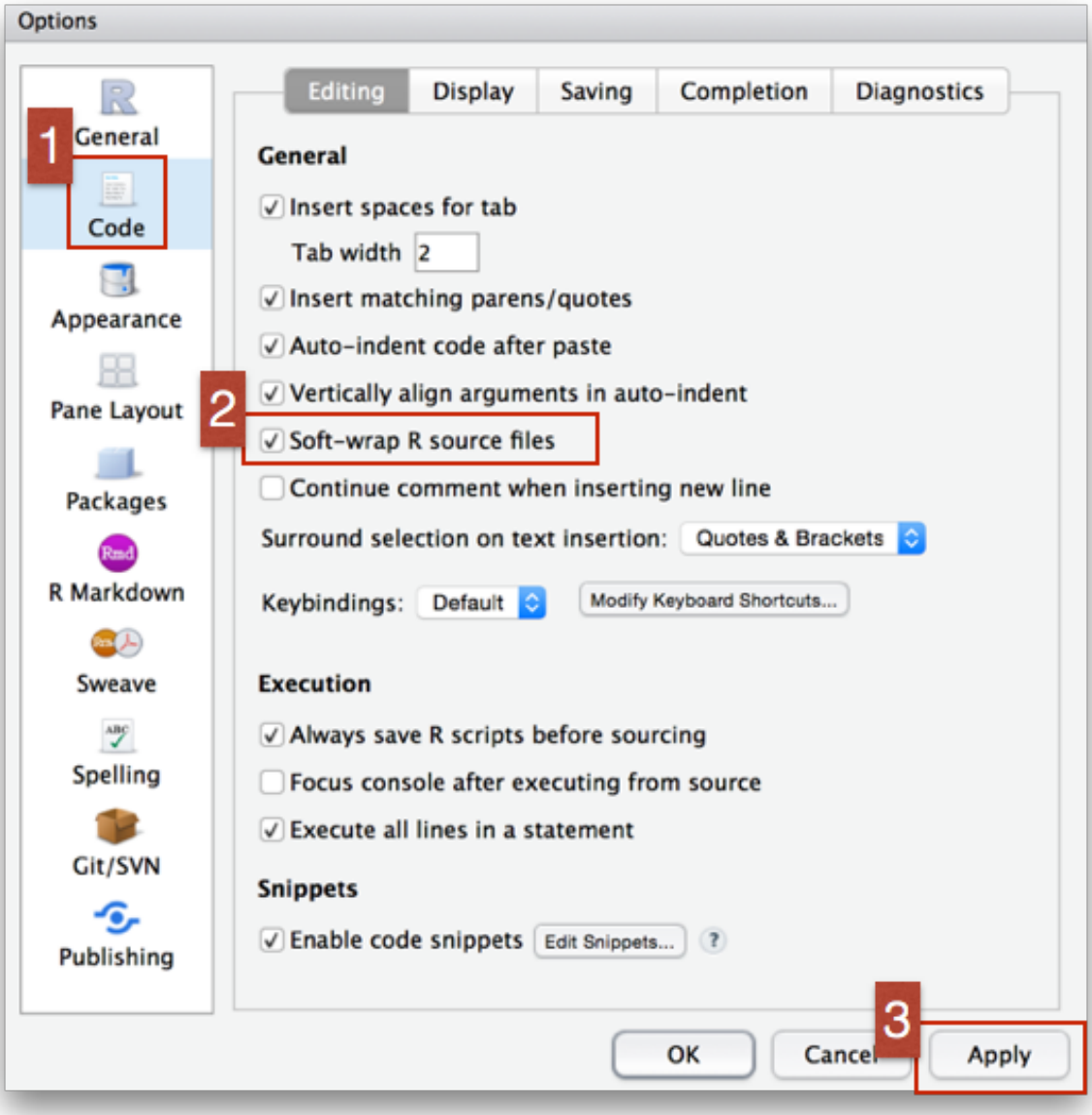


Figure 2.7: wrap_options

2.6 Interacting with R

Now that we have our interface and directory structure set up, let's start playing with R! There are **two main ways** of interacting with R in RStudio: using the **console** or by using **script editor** (plain text files that contain your code).

2.6.1 Console window

The **console window** (in RStudio, the bottom left panel) is the place where R is waiting for you to tell it what to do, and where it will show the results of a command. You can type commands directly into the console, but they will be forgotten when you close the session.

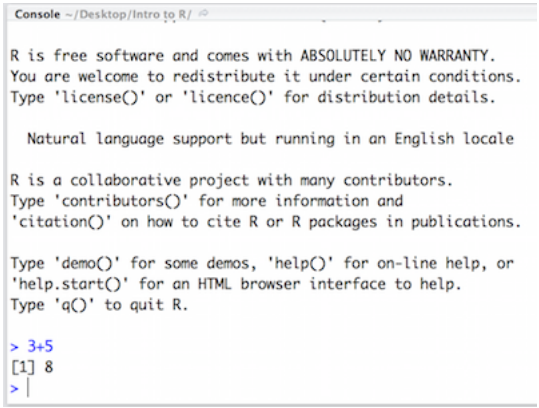


Figure 2.8: Running in the console

2.6.2 Script editor

Best practice is to enter the commands in the **script editor**, and save the script. You are encouraged to comment liberally to describe the commands you are running using `#`. This way, you have a complete record of what you did, you can easily show others how you did it and you can do it again later on if needed.

The Rstudio script editor allows you to 'send' the current line or the currently highlighted text to the R console by clicking on the `Run` button in the upper-right hand corner of the script editor. Alternatively, you can run by simply pressing the `Ctrl` and `Enter` keys at the same time as a shortcut.

Now let's try entering commands to the **script editor** and using the comments character `#` to add descriptions and highlighting the text to run:

```
# Session 1
# Feb 3, 2023

# Interacting with R

# I am adding 3 and 5.
3+5
```

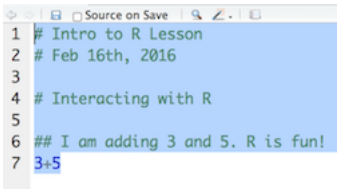


Figure 2.9: Running in the script editor

You should see the command run in the console and output the result.

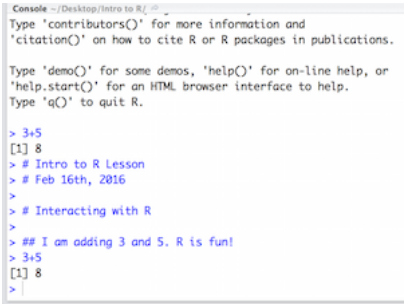


Figure 2.10: Script editor output

What happens if we do that same command without the comment symbol #? Re-run the command after removing the # sign in the front:

```
I am adding 3 and 5. R is fun!
3+5
```

Now R is trying to run that sentence as a command, and it doesn't work. We get an error in the console *"Error: unexpected symbol in"I am" means that the R interpreter did not know what to do with that command."*

2.6.3 Console command prompt

Interpreting the command prompt can help understand when R is ready to accept commands. Below lists the different states of the command prompt and how you can exit a command:

Console is ready to accept commands: `>`.

If R is ready to accept commands, the R console shows a `>` prompt.

When the console receives a command (by directly typing into the console or running from the script editor (`Ctrl-Enter`), R will try to execute it.

After running, the console will show the results and come back with a new `>` prompt to wait for new commands.

Console is waiting for you to enter more data: `+`.

If R is still waiting for you to enter more data because it isn't complete yet, the console will show a `+` prompt. It means that you haven't finished entering a complete command. Often this can be due to you having not 'closed' a parenthesis or quotation.

Escaping a command and getting a new prompt: `esc`

If you're in Rstudio and you can't figure out why your command isn't running, you can click inside the console window and press `esc` to escape the command and bring back a new prompt `>`.

2.6.4 Keyboard shortcuts in RStudio

In addition to some of the shortcuts described earlier in this lesson, we have listed a few more that can be helpful as you work in RStudio.

key	action
Ctrl+Enter	Run command from script editor in console
ESC	Escape the current command to return to the command prompt
Ctrl+1	Move cursor from console to script editor
Ctrl+2	Move cursor from script editor to console
Tab	Use this key to complete a file path
Ctrl+Shift+C	Comment the block of highlighted text

2.7 R syntax

Now that we know how to talk with R via the script editor or the console, we want to use R for something more than adding numbers. To do this, we need to know more about the R syntax.

The main "parts of speech" in R (syntax) include:

- the **comments** `#` and how they are used to document function and its content
- **variables** and **functions**
- the **assignment operator** `<-`
- the `=` for **arguments** in functions

NOTE: indentation and consistency in spacing is used to improve clarity and legibility

We will go through each of these "parts of speech" in more detail, starting with the assignment operator.

2.8 Assignment operator

To do useful and interesting things in R, we need to assign *values* to *variables* using the assignment operator, `<-`. For example, we can use the assignment operator to assign the value of `3` to `x` by executing:

```
x <- 3
```

The assignment operator (`<-`) assigns **values on the right** to **variables on the left**.

In RStudio, typing `Alt + -` *(push* `Alt` *at the same time as the* `-` *key, on Mac type* `option + -`*) will write* `<-` *in a single keystroke.*

2.9 Variables

A variable is a symbolic name for (or reference to) information. Variables in computer programming are analogous to "buckets", where information can be maintained and referenced. On the outside of the bucket is a name. When referring to the bucket, we use the name of the bucket, not the data stored in the bucket.

In the example above, we created a variable or a 'bucket' called `x`. Inside we put a value, `3`.

Let's create another variable called `y` and give it a value of 5.

```
y <- 5
```

When assigning a value to an variable, R does not print anything to the console. You can force to print the value by using parentheses or by typing the variable name.

```
y
```

You can also view information on the variable by looking in your `Environment` window in the upper right-hand corner of the RStudio interface.

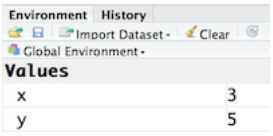


Figure 2.11: Viewing your environment

Now we can reference these buckets by name to perform mathematical operations on the values contained within. What do you get in the console for the following operation:

```
x + y
```

Try assigning the results of this operation to another variable called `number`.

```
number <- x + y
```

2.9.1 Tips on variable names

Variables can be given almost any name, such as `x`, `current_temperature`, or `subject_id`. However, there are some rules / suggestions you should keep in mind:

- Make your names explicit and not too long.
- Avoid names starting with a number (`2x` is not valid but `x2` is)
- Avoid names of fundamental functions in R (e.g., `if`, `else`, `for`, see here for a complete list). In general, even if it's allowed, it's best to not use other function names (e.g., `c`, `T`, `mean`, `data`) as variable names. When in doubt check the help to see if the name is already in use.
- Avoid dots (`.`) within a variable name as in `my.dataset`. There are many functions in R with dots in their names for historical reasons, but because dots have a special meaning in R (for methods) and other programming languages, it's best to avoid them.
- Use nouns for object names and verbs for function names
- Keep in mind that **R is case sensitive** (e.g., `genome_length` is different from `Genome_length`)
- Be consistent with the styling of your code (where you put spaces, how you name variable, etc.). In R, two popular style guides are Hadley Wickham's style guide and Google's.

2.10 Best practices

Before we move on to more complex concepts and getting familiar with the language, we want to point out a few things about best practices when working with R which will help you stay organized in the long run:

- Code and workflow are more reproducible if we can document everything that we do. Our end goal is not just to "do stuff", but to do it in a way that anyone can easily and exactly replicate our workflow and results. **All code should be written in the script editor and saved to file, rather than working in the console.**
- The **R console** should be mainly used to inspect objects, test a function or get help.
- Use `#` signs to comment. **Comment liberally** in your R scripts. This will help future you and other collaborators know what each line of code (or code block) was meant to do. Anything to the right of a `#` is ignored by R. A shortcut for this is Ctrl+Shift+C if you want to comment an entire chunk of text.

Part II

Session 1

Learning Objectives

- Explore how probability distributions inform the mathematical form of statistical tests.
- Explore different types of hypothesis tests and when they should be used.
- Apply hypothesis tests commonly used in biological systems analyses.
- Install and manage packages from CRAN and Bioconductor.
- Identify and use different data types in R.

3 R Syntax and Data Structures

3.1 Basic Data Types

Variables can contain values of specific types within R. The six **data types** that R uses include:

- `"numeric"` for any numerical value, including whole numbers and decimals. This is the most common data type for performing mathematical operations.
- `"character"` for text values, denoted by using quotes (" ") around value. For instance, while 5 is a numeric value, if you were to put quotation marks around it, it would turn into a character value, and you could no longer use it for mathematical operations. Single or double quotes both work, as long as the same type is used at the beginning and end of the character value.
- `"integer"` for whole numbers (e.g., `2L`, the `L` indicates to R that it's an integer). It behaves similar to the `numeric` data type for most tasks or functions; however, it takes up less storage space than numeric data, so often tools will output integers if the data is known to be comprised of whole numbers. Just know that integers behave similarly to numeric values. If you wanted to create your own, you could do so by providing the whole number, followed by an upper-case L.
- `"logical"` for `TRUE` and `FALSE` (the Boolean data type). The `logical` data type can be specified using four values, `TRUE` in all capital letters, `FALSE` in all capital letters, a single capital `T` or a single capital `F`.
- `"complex"` to represent complex numbers with real and imaginary parts (e.g., `1+4i`) and that's all we're going to say about them
- `"raw"` that we won't discuss further

The table below provides examples of each of the commonly used data types:

Data Type	Examples
Numeric:	1, 1.5, 20, pi
Character:	"anytext", "5", "TRUE"
Integer:	2L, 500L, -17L
Logical:	TRUE, FALSE, T, F

The type of data will determine what you can do with it. For example, if you want to perform mathematical operations, then your data type cannot be character or logical. Whereas if you want to search for a word or pattern in your data, then you data should be of the character data type. The task or function being performed on the data will determine what type of data can be used.

3.2 Data Structures

We know that variables are like buckets, and so far we have seen that bucket filled with a single value. Even when `number` was created, the result of the mathematical operation was a single value. **Variables can store more than just a single value, they can store a multitude of different data structures.** These include, but are not limited to, vectors (`c`), factors (`factor`), matrices (`matrix`), data frames (`data.frame`) and lists (`list`).

3.2.1 Vectors

A vector is the most common and basic data structure in R, and is pretty much the workhorse of R. It's basically just a collection of values, mainly either numbers,

Figure 3.1: numeric vector

or characters,

Figure 3.2: character vector

or logical values,

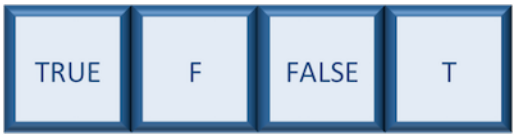


Figure 3.3: logical vector

Note that all values in a vector must be of the same data type. If you try to create a vector with more than a single data type, R will try to coerce it into a single data type.

For example, if you were to try to create the following vector:

Figure 3.4: mixed vector

R will coerce it into:

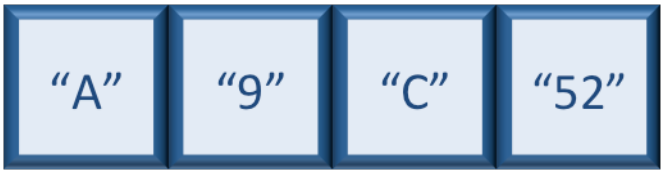


Figure 3.5: transformed vector

The analogy for a vector is that your bucket now has different compartments; these compartments in a vector are called *elements*.

Each **element** contains a single value, and there is no limit to how many elements you can have. A vector is assigned to a single variable, because regardless of how many elements it contains, in the end it is still a single entity (bucket).

Let's create a vector of genome lengths and assign it to a variable called `glengths`.

Each element of this vector contains a single numeric value, and three values will be combined together into a vector using `c()` (the combine function). All of the values are put within the parentheses and separated with a comma.

```
# Create a numeric vector and store the vector as a variable called 'glengths'
glengths <- c(4.6, 3000, 50000)
glengths
```

```
[1]     4.6  3000.0 50000.0
```

Note your environment shows the ***`glengths`*** *variable is numeric (num) and tells you the* ***`glengths`*** *vector starts at element 1 and ends at element 3 (i.e. your vector contains 3 values) as denoted by the [1:3].*

A vector can also contain characters. Create another vector called `species` with three elements, where each element corresponds with the genome sizes vector (in Mb).

```
# Create a character vector and store the vector as a variable called 'species'
species <- c("ecoli", "human", "corn")
species
```

```
[1] "ecoli" "human" "corn"
```

What do you think would happen if we forgot to put quotations around one of the values? Let's test it out with corn.

```
# Forget to put quotes around corn
species <- c("ecoli", "human", corn)
```

Note that RStudio is quite helpful in color-coding the various data types. We can see that our numeric values are blue, the character values are green, and if we forget to surround corn with quotes, it's black. What does this mean? Let's try to run this code.

When we try to run this code we get an error specifying that object 'corn' is not found. What this means is that R is looking for an object or variable in my Environment called 'corn', and when it doesn't find it, it returns an error. If we had a character vector called 'corn' in our Environment, then it would combine the contents of the 'corn' vector with the values "ecoli" and "human".

Since we only want to add the value "corn" to our vector, we need to re-run the code with the quotation marks surrounding corn. A quick way to add quotes to both ends of a word in RStudio is to highlight the word, then press the quote key.

```
# Create a character vector and store the vector as a variable called 'species'
species <- c("ecoli", "human", "corn")
```

Exercise

Try to create a vector of numeric and character values by *combining* the two vectors that we just created (`glengths` and `species`). Assign this combined vector to a new variable called `combined`. *Hint: you will need to use the combine* `c()` *function to do this.* Print the `combined` vector in the console, what looks different compared to the original vectors?

3.2.2 Factors

A **factor** is a special type of vector that is used to **store categorical data**. Each unique category is referred to as a **factor level** (i.e. category = level). Factors are built on top of integer vectors such that each **factor level** is assigned an **integer value**, creating value-label pairs.

For instance, if we have four animals and the first animal is female, the second and third are male, and the fourth is female, we could create a factor that appears like a vector, but has integer values stored under-the-hood. The integer value assigned is a one for females and a two for males. The numbers are assigned in alphabetical order, so because the f- in females comes before the m- in males in the alphabet, females get assigned a one and males a two. In later lessons we will show you how you could change these assignments.

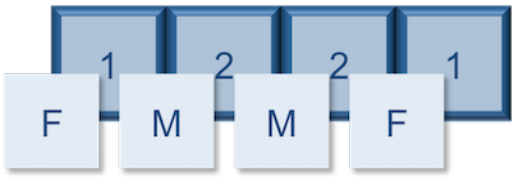


Figure 3.6: factors

Let's create a factor vector and explore a bit more. We'll start by creating a character vector describing three different levels of expression. Perhaps the first value represents expression in mouse1, the second value represents expression in mouse2, and so on and so forth:

```
# Create a character vector and store the vector as a variable called 'expression'
expression <- c("low", "high", "medium", "high", "low", "medium", "high")
```

Now we can convert this character vector into a *factor* using the `factor()` function:

```
# Turn 'expression' vector into a factor
expression <- factor(expression)
```

So, what exactly happened when we applied the `factor()` function?

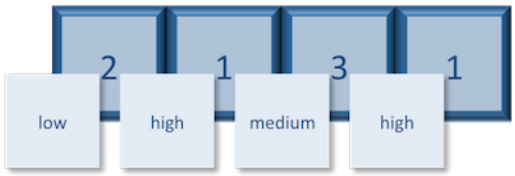


Figure 3.7: factor_new

The expression vector is categorical, in that all the values in the vector belong to a set of categories; in this case, the categories are `low`, `medium`, and `high`. By turning the expression vector into a factor, the **categories are assigned integers alphabetically**, with high=1, low=2, medium=3. This in effect assigns the different factor levels. You can view the newly created factor variable and the levels in the **Environment** window.

Figure 3.8: Factor variables in environment

So now that we have an idea of what factors are, when would you ever want to use them?

Factors are extremely valuable for many operations often performed in R. For instance, factors can give order to values with no intrinsic order. In the previous 'expression' vector, if I wanted the low category to be less than the medium category, then we could do this using factors. Also, factors are necessary for many statistical methods. For example, descriptive statistics can be obtained for character vectors if you have the categorical information stored as a factor. Also, if you want to denote which category is your base level for a statistical comparison, then you would need to have your category variable stored as a factor with the base level assigned to 1. Anytime that it is helpful to have the categories thought of as groups in an analysis, the factor function makes this possible. For instance, if you want to color your plots by treatment type, then you would need the treatment variable to be a factor.

Exercises

Let's say that in our experimental analyses, we are working with three different sets of cells: normal, cells knocked out for geneA (a very exciting gene), and cells overexpressing geneA. We have three replicates for each celltype.

1. Create a vector named `samplegroup` with nine elements: 3 control ("CTL") values, 3 knock-out ("KO") values, and 3 over-expressing ("OE") values.
2. Turn `samplegroup` into a factor data structure.

3.2.3 Matrix

A `matrix` in R is a collection of vectors of **same length and identical datatype**. Vectors can be combined as columns in the matrix or by row, to create a 2-dimensional structure.

Matrices are used commonly as part of the mathematical machinery of statistics. They are usually of numeric datatype and used in computational algorithms to serve as a checkpoint.

Figure 3.9: matrix

For example, if input data is not of identical data type (numeric, character, etc.), the `matrix()` function will throw an error and stop any downstream code execution.

3.2.4 Data Frame

A `data.frame` is the *de facto* data structure for most tabular data and what we use for statistics and plotting. A `data.frame` is similar to a matrix in that it's a collection of vectors of the **same length** and each vector represents a column. However, in a dataframe **each vector can be of a different data type** (e.g., characters, integers, factors). In the data frame pictured below, the first column is character, the second column is numeric, the third is character, and the fourth is logical.

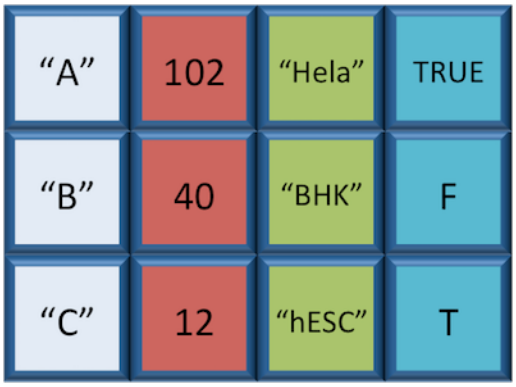


Figure 3.10: dataframe

A data frame is the most common way of storing data in R, and if used systematically makes data analysis easier.

We can create a dataframe by bringing **vectors** together to **form the columns**. We do this using the `data.frame()` function, and giving the function the different vectors we would like to bind together. *This function will only work for vectors of the same length.*

```
# Create a data frame and store it as a variable called 'df'
df <- data.frame(species, glengths)
```

We can see that a new variable called `df` has been created in our `Environment` within a new section called `Data`. In the `Environment`, it specifies that `df` has 3 observations of 2 variables. What does that mean? In R, rows always come first, so it means that `df` has 3 rows and 2 columns. We can get additional information if we click on the blue circle with the white triangle in the middle next to `df`. It will display information about each of the columns in the data frame, giving information about what the data type is of each of the columns and the first few values of those columns.

Another handy feature in RStudio is that if we hover the cursor over the variable name in the `Environment`, `df`, it will turn into a pointing finger. If you click on `df`, it will open the data frame as it's own tab next to the script editor. We can explore the table interactively within this window. To close, just click on the X on the tab.

As with any variable, we can print the values stored inside to the console if we type the variable's name and run.

```
df
```

```
  species glengths
1   ecoli      4.6
2   human   3000.0
3    corn  50000.0
```

3.2.5 Lists

Lists are a data structure in R that can be perhaps a bit daunting at first, but soon become amazingly useful. A list is a data structure that can hold any number of any types of other data structures.

If you have variables of different data structures you wish to combine, you can put all of those into one list object by using the `list()` function and placing all the items you wish to combine within parentheses:

```
list1 <- list(species, df, expression)
```

We see `list1` appear within the Data section of our environment as a list of 3 components or variables. If we click on the blue circle with a triangle in the middle, it's not quite as interpretable as it was for data frames.

Essentially, each component is preceded by a colon. The first colon give the `species` vector, the second colon precedes the `df` data frame, with the dollar signs indicating the different columns, the last colon gives the single value, `number`.

If I click on `list1`, it opens a tab where you can explore the contents a bit more, but it's still not super intuitive. The easiest way to view small lists is to print to the console.

Let's type list1 and print to the console by running it.

```
list1
```

```
[[1]]
[1] "ecoli" "human" "corn"

[[2]]
  species glengths
1   ecoli      4.6
2   human   3000.0
3    corn  50000.0

[[3]]
[1] low    high   medium high   low    medium high
Levels: high low medium
```

There are three components corresponding to the three different variables we passed in, and what you see is that structure of each is retained. Each component of a list is referenced based on the number position.

---

4 Probability Primer

4.1 Defining Probability

Informally, we usually think of probability as a number that describes the likelihood of some event occurring, which ranges from zero (impossibility) to one (certainty).

To formalize probability theory, we first need to define a few terms:

- An **experiment** is any activity that produces or observes an outcome. Examples are flipping a coin, rolling a 6-sided die, or trying a new route to work to see if it's faster than the old route.
- The **sample space** is the set of possible outcomes for an experiment. We represent these by listing them within a set of squiggly brackets. For a coin flip, the sample space is {heads, tails}. For a six-sided die, the sample space is each of the possible numbers that can appear: {1,2,3,4,5,6}. For the amount of time it takes to get to work, the sample space is all possible real numbers greater than zero (since it can't take a negative amount of time to get somewhere, at least not yet).
- An **event** is a subset of the sample space. In principle it could be one or more of possible outcomes in the sample space, but here we will focus primarily on *elementary events* which consist of exactly one possible outcome. For example, this could be obtaining heads in a single coin flip, rolling a 4 on a throw of the die, or taking 21 minutes to get home by the new route.

Let's say that we have a sample space defined by N independent events, $E_1, E_2, ..., E_N$, and X is a random variable denoting which of the events has occurred. $P(X = E_i)$ is the probability of event i:

- Probability cannot be negative: $P(X = E_i) \geq 0$
- The total probability of all outcomes in the sample space is 1; that is, if the , if we take the probability of each Ei and add them up, they must sum to 1. We can express this using the summation symbol $\sum$:

$$\sum_{i=1}^{N} P(X = E_i) = P(X = E_1) + P(X = E_2) + ... + P(X = E_N) = 1$$

  This is interpreted as saying "Take all of the N elementary events, which we have labeled from 1 to N, and add up their probabilities. These must sum to one."

- The probability of any individual event cannot be greater than one: $P(X = E_i) \leq 1$. This is implied by the previous point; since they must sum to one, and they can't be negative, then any particular probability cannot exceed one.

4.1.1 Conditional probability

These definitions allow us to examine simple probabilities - that is, the probability of a single event or combination of events.
However, we often wish to determine the probability of some event given that some other event has occurred, which are known as *conditional probabilities*.

To compute the conditional probability of A given B (which we write as $P(A|B)$, "probability of A, given B"), we need to know the *joint probability* (that is, the probability of both A and B occurring) as well as the overall probability of B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

That is, we want to know the probability that both things are true, given that the one being conditioned upon is true.

4.1.2 Independence

The term "independent" has a very specific meaning in statistics, which is somewhat different from the common usage of the term. Statistical independence between two variables means that knowing the value of one variable doesn't tell us anything about the value of the other. This can be expressed as:

$$P(A|B) = P(A)$$

That is, the probability of A given some value of B is just the same as the overall probability of A.

4.2 Probability distributions

A *probability distribution* describes the probability of all of the possible outcomes in an experiment. To help understand distributions and how they can be used, let's look at a few discrete probability distributions, meaning distributions which can only output integers.

4.2.1 Binomial success counts

Tossing a coin has two possible outcomes. This simple experiment, called a **Bernoulli trial**, is modeled using a so-called Bernoulli random variable.

R has special functions tailored to generate outcomes for each type of distribution. They all start with the letter `r`, followed by a specification of the model, here `rbinom`, where `binom` is the abbreviation used for binomial.

Suppose we want to simulate a sequence of 15 fair coin tosses. To get the outcome of 15 Bernoulli trials with a probability of success equal to 0.5 (a fair coin), we write:

```
rbinom(15, prob = 0.5, size = 1)
```

```
 [1] 0 1 1 0 1 0 1 0 1 1 1 0 0 0 1
```

We use the `rbinom` function with a specific set of **parameters** (called **arguments** in programming): the first parameter is the number of trials we want to observe; here we chose 15. We designate by `prob` the probability of success. By `size=1` we declare that each individual trial consists of just one single coin toss.

For binary events such as heads or tails, success or failure, CpG or non-CpG, M or F, Y = pyrimidine or R = purine, diseased or healthy, true or false, etc. we only need the probability p of one of the events (which we, often arbitrarily, will label "success") because "failure" (the complementary event) will occur with probability $1-p$. We can then simply count the number of successes for a certain number of trials:

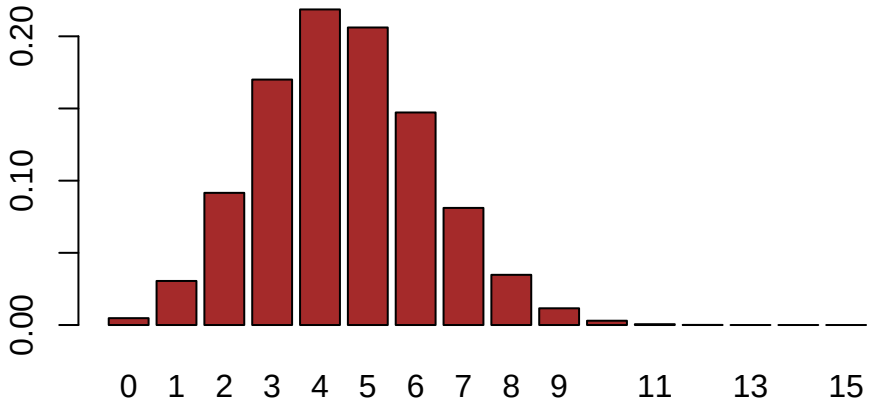
```
rbinom(1, prob = 0.3, size = 15)
```

```
[1] 5
```

This gives us the number of successes for 15 trials where the probability of success was 0.3. We would call this number a **binomial random variable** or a random variable that follows the $B(15, 0.3)$ distribution.

We can plot the **probability mass distribution** using `dbinom`:

```
probabilities <- dbinom(0:15, prob = 0.3, size = 15)
barplot(probabilities, names.arg = 0:15, col = "brown")
```



For X distributed as a binomial distribution with parameters (n, p), written $X\ B(n, p)$ the probability of seeing $X = k$ sucesses is:

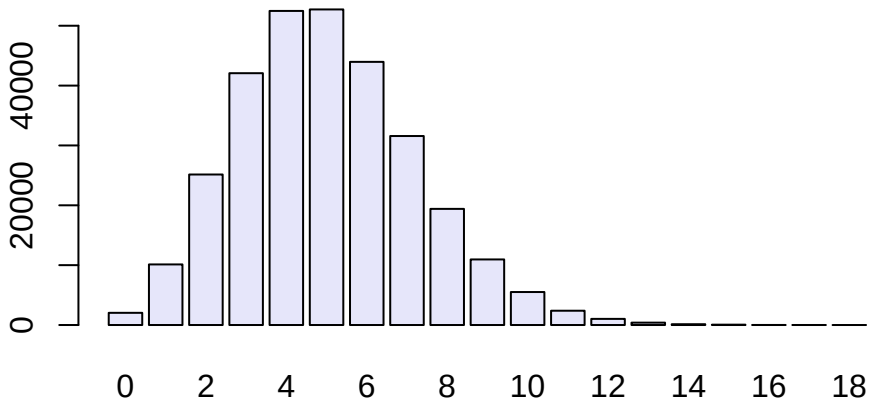
$$P(k; n, p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

4.2.2 Poisson distributions

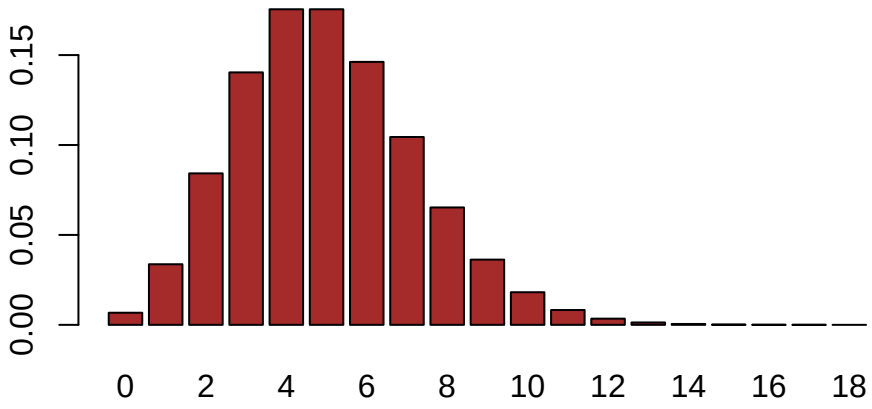
When the probability of success p is small and the number of trials n large, the binomial distribution $B(n, p)$ can be faithfully approximated by a simpler distribution, the Poisson distribution with rate parameter $\lambda = np$

The Poisson distribution comes up often in biology as we often are naturally dealing very low probability events and large numbers of trials, such as mutations in a genome.

```
simulations = rbinom(n = 300000, prob = 5e-4, size = 10000)
barplot(table(simulations), col = "lavender")
```




```r
probabilities <- dpois(0:18, lambda=(10000 * 5e-4))
barplot(probabilities, names.arg = 0:18, col = "brown")
```



4.2.3 Multinomial distributions

When modeling four possible outcomes, for instance when studying counts of the four nucleotides [A,C,G] and [T], we need to extend the binomial model.

We won't go into detail on the formulation, but we can examine probabilities of observations using a vector of counts for each observed outcome, and a vector of probabilities for each outcome (which must sum to 1).

```
counts <- c(4,2,0,0)
probs <- c(0.25,0.25,0.25,0.25)
dmultinom(counts, prob = probs)
```

```
[1] 0.003662109
```

---

5 Distributions to Hypothesis Tests

5.1 Calculating the chance of an event

When testing certain pharmaceutical compounds, it is important to detect proteins that provoke an allergic reaction. The molecular sites that are responsible for such reactions are called epitopes.

Epitope: A specific portion of a macromolecular antigen to which an antibody binds. In the case of a protein antigen recognized by a T-cell, the epitope or determinant is the peptide portion or site that binds to a Major Histocompatibility Complex (MHC) molecule for recognition by the T cell receptor (TCR).

Enzyme-Linked ImmunoSorbent Assays (ELISA) are used to detect specific epitopes at different positions along a protein. Suppose the following facts hold for an ELISA array we are using:

- The baseline noise level per position, or more precisely the **false positive rate**, is 1%. This is the probability of declaring a hit – we think we have an epitope – when there is none. We write this $P(declareepitope|noepitope)$
- The protein is tested at 100 different positions, supposed to be independent.
- We are going to examine a collection of 50 patient samples.

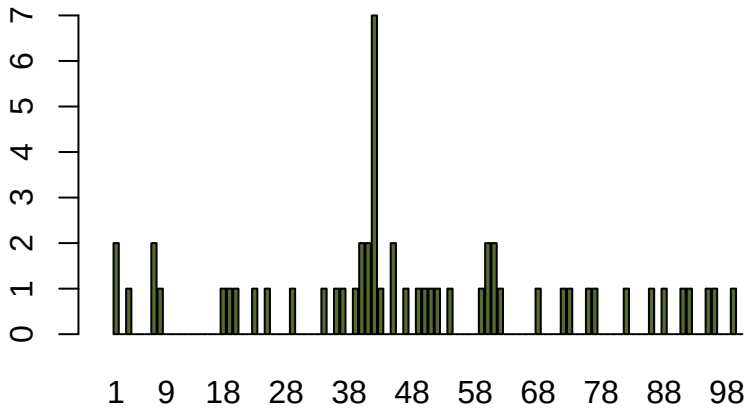
The data for one patient's assay look like this:

```
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

where the 1 signifies a hit (and thus the potential for an allergic reaction), and the zeros signify no reaction at that position.

We're going to study the data for all 50 patients tallied at each of the 100 positions. If there are no allergic reactions, the false positive rate means that for one patient, each individual position has a probability of 1 in 100 of being a 1. So, after tallying 50 patients, we expect at any given position the sum of the 50 observed $(0, 1)$ variables to have a Poisson distribution with parameter 0.5.

```
load("data/e100.RData")
barplot(e100, ylim = c(0, 7), width = 0.7, xlim = c(-0.5, 100.5),
  names.arg = seq(along = e100), col = "darkolivegreen")
```



The spike is striking. What are the chances of seeing a value as large as 7, if no epitope is present? If we look for the probability of seeing a number as big as 7 (or larger) when considering one $Poisson(0.5)$ random variable, the answer can be calculated in closed form as

$$P(X \geq 7) = \sum_{k=7}^{\infty} P(X = k)$$

This is, of course, the same as $1 - P(X \leq 6)$. The probability is the so-called **cumulative distribution** function at 6, and R has the function `ppois` for computing it, which we can use in either of the following two ways:

```
1 - ppois(6, 0.5)
```

```
[1] 1.00238e-06
```

```r
ppois(6, 0.5, lower.tail = FALSE)
```

```
[1] 1.00238e-06
```

You can use the command `?ppois` to see the argument definitions for the function.

We denote this number, our chance of seeing such an extreme result, as ϵ. However, in this case it would be the incorrect calculation.

Instead of asking what the chances are of seeing a `Poisson(0.5)` as large as 7, we need to instead ask, what are the chances that the *maximum of 100 `Poisson(0.5)` trials is as large as 7*? We order the data values $x_1, x_2, ..., x_{100}$ and rename them $x_{(1)}, x_{(2)}, ..., x_{(100)}$, so that denotes $x_{(1)}$ the smallest and $x_{(100)}$ the largest of the counts over the 100 positions. Together, are called the **rank statistic** of this sample of 100 values.

The maximum value being as large as 7 is the **complementary event** of having all 100 counts be smaller than or equal to 6. Two complementary events have probabilities that sum to 1. *Because the positions are supposed to be independent*, we can now do the computation:

$$P(x_{(100)} \geq 7) = \prod_{i=1}^{100} P(x_i \leq 6) = (P(x_i \leq 6))^{100}$$

which, using our notation, is $(1-\epsilon)^{100}$ and is approximately 10^{-4}. This is a very small chance, so we would determine it is most likely that we did detect real epitopes.

5.2 Computing probabilities with simulations

In the case we just saw, the theoretical probability calculation was quite simple and we could figure out the result by an explicit calculation. In practice, things tend to be more complicated, and we are better to compute our probabilities using the **Monte Carlo** method: a computer simulation based on our generative model that finds the probabilities of the events we're interested in. Below, we generate 100,000 instances of picking the maximum from 100 Poisson distributed numbers.

```r
maxes = replicate(100000, {
  max(rpois(100, 0.5))
})
table(maxes)
```

```
maxes
    1     2     3     4     5     6     7     8
    8 23462 60399 14434  1547   139    10     1
```

So we can approximate the probability of seeing a 7 as:

```
mean( maxes >= 7 )
```

```
[1] 0.00011
```

We arrive at a similarly small number, and in both cases would determine that there are real epitopes in the dataset.

5.3 An example: coin tossing

Let's look a simpler example: flipping a coin to see if it is fair. We flip the coin 100 times and each time record whether it came up heads or tails. So, we have a record that could look something like `HHTTHTHTT...`

Let's simulate the experiment in R, using a biased coin:

```
set.seed(0xdada)
numFlips = 100
probHead = 0.6
coinFlips = sample(c("H", "T"), size = numFlips,
  replace = TRUE, prob = c(probHead, 1 - probHead))
head(coinFlips)
```

```
[1] "T" "T" "H" "T" "H" "H"
```

Now, if the coin were fair, we would expect half of the time to get heads. Let's see.

```
table(coinFlips)
```

```
coinFlips
 H  T
59 41
```

That is different from 50/50. However, does the data deviates strong enough to conclude that this coin isn't fair? We know that the total number of heads seen in 100 coin tosses for a fair coin follows $B(100, 0.5)$, making it a suitable test statistic.

To decide, let's look at the sampling distribution of our test statistic – the total number of heads seen in 100 coin tosses – for a fair coin. As we learned, we can do this with the binomial distribution. Let's plot a fair coin and mark our observation with a blue line:

```
library("dplyr")
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

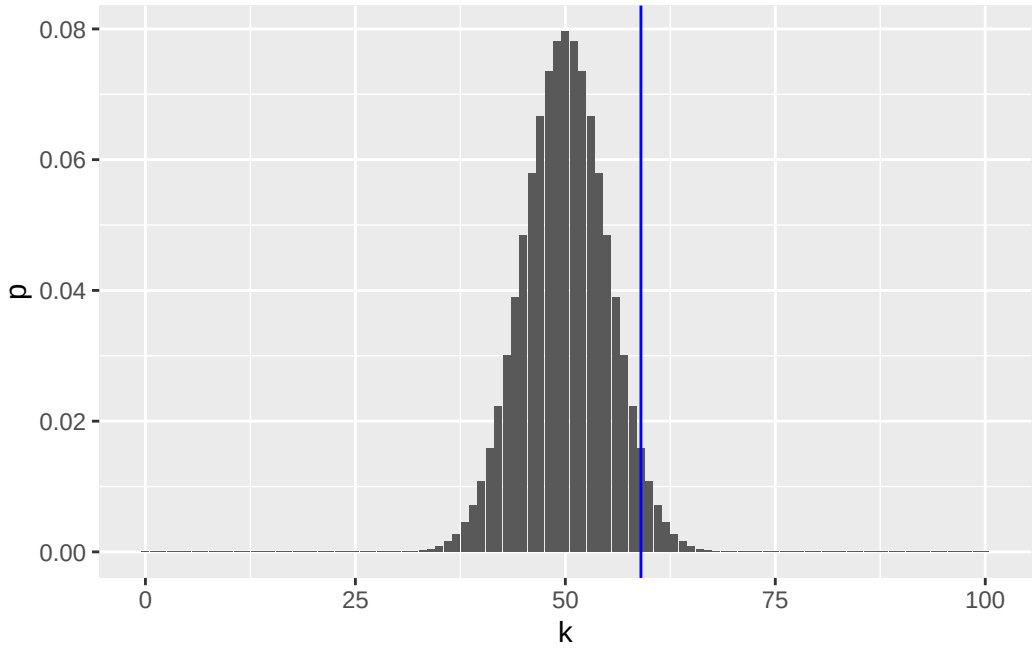
```
library("ggplot2")
```

```
Warning: package 'ggplot2' was built under R version 4.2.2
```

```
k <- 0:numFlips
numHeads <- sum(coinFlips == "H")
p <- dbinom(k, size = numFlips, prob = 0.5)
binomDensity <- data.frame(k = k, p = p)
head(binomDensity)
```

```
  k            p
1 0 7.888609e-31
2 1 7.888609e-29
3 2 3.904861e-27
4 3 1.275588e-25
5 4 3.093301e-24
6 5 5.939138e-23
```

```
ggplot(binomDensity) +
  geom_bar(aes(x = k, y = p), stat = "identity") +
  geom_vline(xintercept = numHeads, col = "blue")
```



How do we quantify whether the observed value is among those values that we are likely to see from a fair coin, or whether its deviation from the expected value is already large enough for us to conclude with enough confidence that the coin is biased?

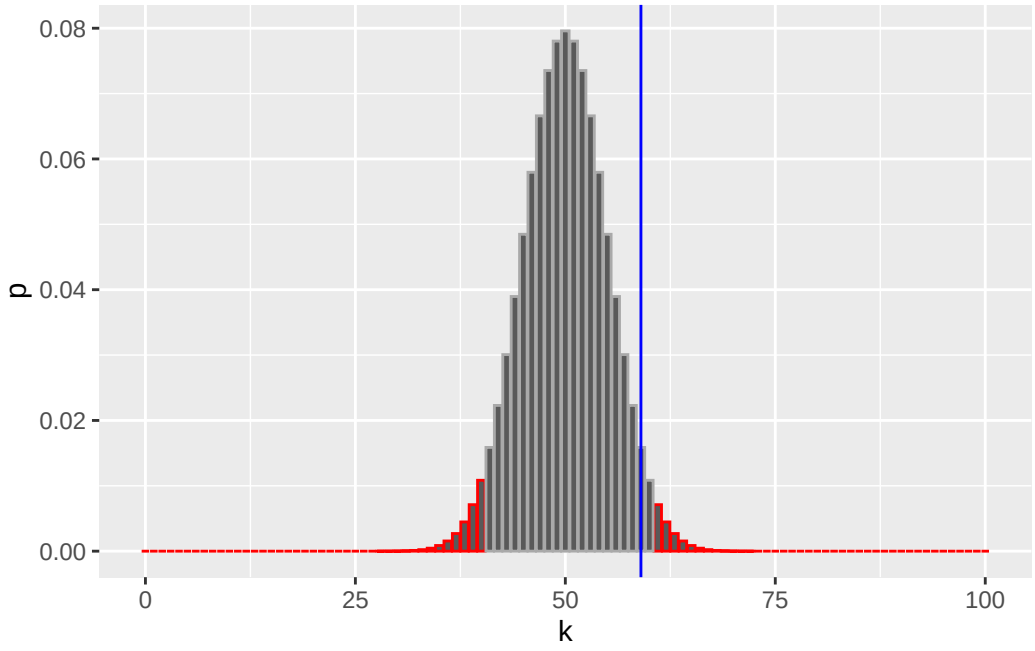
We divide the set of all possible $k(0-100)$ in two complementary subsets, the **rejection region** and the region of no rejection. We want to make the rejection region as large as possible while keeping their total probability, assuming the null hypothesis, below some threshold α(say, 0.05).

```
alpha <- 0.05
binomDensity <- binomDensity[order(p),]
binomDensity$reject <- cumsum(binomDensity$p) <= alpha
head(binomDensity)
```

```
      k            p reject
1     0 7.888609e-31   TRUE
101 100 7.888609e-31   TRUE
```

```
2     1 7.888609e-29   TRUE
100  99 7.888609e-29   TRUE
3     2 3.904861e-27   TRUE
99   98 3.904861e-27   TRUE
```

```
ggplot(binomDensity) +
  geom_bar(aes(x = k, y = p, col = reject), stat = "identity") +
  scale_colour_manual(
    values = c(`TRUE` = "red", `FALSE` = "darkgrey")) +
  geom_vline(xintercept = numHeads, col = "blue") +
  theme(legend.position = "none")
```



We sorted the p-values from lowest to highest (`order`), and added a column `reject` by computing the cumulative sum (`cumsum`) of the p-values and thresholding it against `alpha`.

The logical column `reject` therefore marks with `TRUE` a set of ks whose total probability is less than α.

The rejection region is marked in red, containing both very large and very small values of k, which can be considered unlikely under the null hypothesis.

`R` provides not only functions for the densities (e.g., `dbinom`) but also for the cumulative distribution functions (`pbinom`). Those are more precise and faster than `cumsum` over the

probabilities.

The (cumulative) *distribution function* is defined as the probability that a random variable X will take a value less than or equal to x.

$$F(x) = P(X \leq x)$$

We have just gone through the steps of a **binomial test**. This is a frequently used test and therefore available in `R` as a single function.

We have just gone through the steps of a binomial test. In fact, this is such a frequent activity in R that it has been wrapped into a single function, and we can compare its output to our results.

```
binom.test(x = numHeads, n = numFlips, p = 0.5)
```

```
    Exact binomial test

data:  numHeads and numFlips
number of successes = 59, number of trials = 100, p-value = 0.08863
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4871442 0.6873800
sample estimates:
probability of success
                  0.59
```

5.4 Hypothesis Tests

We can summarize what we just did with a series of steps:

1. Decide on the effect that you are interested in, design a suitable experiment or study, pick a data summary function and test statistic.
2. Set up a null hypothesis, which is a simple, computationally tractable model of reality that lets you compute the null distribution, i.e., the possible outcomes of the test statistic and their probabilities under the assumption that the null hypothesis is true.
3. Decide on the rejection region, i.e., a subset of possible outcomes whose total probability is small.
4. Do the experiment and collect the data; compute the test statistic.
5. Make a decision: reject the null hypothesis if the test statistic is in the rejection region.

5.5 Types of Error

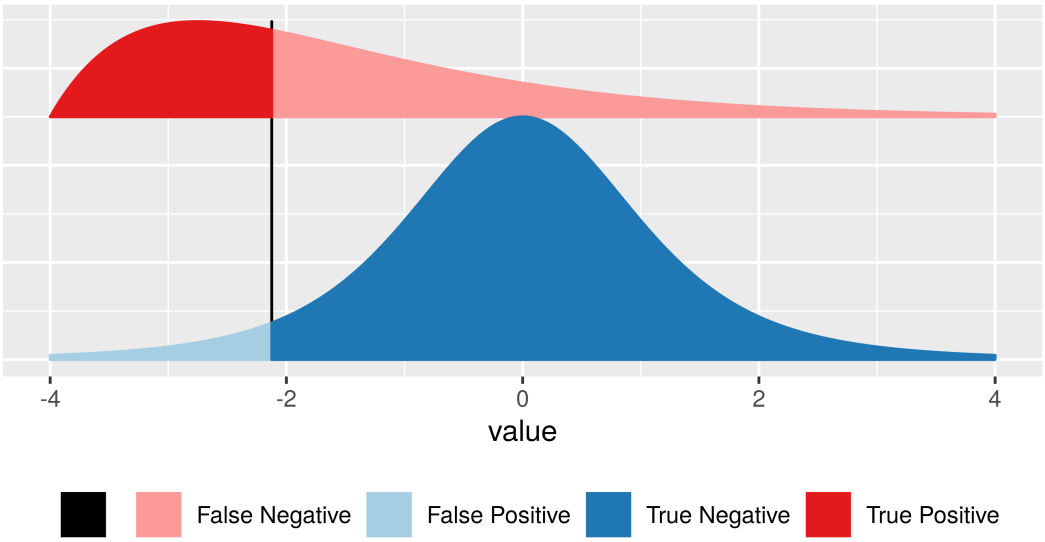


Figure 5.1: From "Modern Statistics for Modern Biology"

Having set out the mechanics of testing, we can assess how well we are doing. The following table, called a **confusion matrix**, compares reality (whether or not the null hypothesis is in fact true) with our decision whether or not to reject the null hypothesis after we have seen the data.

Test vs reality	Null is true	Null is false
Reject null	Type I error (false positive)	True postitive
Do not reject null	True negative	Type II error (false negative)

It is always possible to reduce one of the two error types at the cost of increasing the other one. The real challenge is to find an acceptable trade-off between both of them. We can always decrease the **false positive rate** (FPR) by shifting the threshold to the right. We can become more "conservative". But this happens at the price of higher **false negative rate** (FNR). Analogously, we can decrease the FNR by shifting the threshold to the left. But then again, this happens at the price of higher FPR. T he FPR is the same as the probability α that we mentioned above. $1-\alpha$ is also called the **specificity** of a test. The FNR is sometimes also called β, and $1-\beta$ the **power**, **sensitivity** or **true positive rate** of a test. The power of a test can be understood as the likelihood of it "catching" a true positive, or correctly rejecting the null hypothesis.

Generally, there are three factors that can affect statistical power:

- Sample size: Larger samples provide greater statistical power
- Effect size: A given design will always have greater power to find a large effect than a small effect (because finding large effects is easier)
- Type I error rate: There is a relationship between Type I error and power such that (all else being equal) decreasing Type I error will also decrease power.

In a future session, we will also see how hypothesis tests can be seen as types of **linear models**.

---

6 Categorical Data in R

6.1 Factors

Since factors are special vectors, the same rules for selecting values using indices apply.

```
expression <- c("high","low","low","medium","high","medium","medium","low","low","low")
```

The elements of this expression factor created previously has following categories or levels: low, medium, and high.

Let's extract the values of the factor with high expression, and let's using nesting here:

```
expression[expression == "high"]    ## This will only return those elements in the factor
```

```
[1] "high" "high"
```

> **Nesting note**:
>
> The piece of code above was more efficient with nesting; we used a single step instead of two steps as shown below:
>
> Step1 (no nesting): `idx <- expression == "high"`
>
> Step2 (no nesting): `expression[idx]`

6.2 Releveling factors

We have briefly talked about factors, but this data type only becomes more intuitive once you've had a chance to work with it. Let's take a slight detour and learn about how to **relevel categories within a factor**.

To view the integer assignments under the hood you can use `str()`:

```
expression
```

```
[1] "high"   "low"    "low"    "medium" "high"   "medium" "medium" "low"
[9] "low"    "low"
```

The categories are referred to as "factor levels". As we learned earlier, the levels in the `expression` factor were assigned integers alphabetically, with high=1, low=2, medium=3. However, it makes more sense for us if low=1, medium=2 and high=3, i.e. it makes sense for us to "relevel" the categories in this factor.

To relevel the categories, you can add the `levels` argument to the `factor()` function, and give it a vector with the categories listed in the required order:

```
expression <- factor(expression, levels=c("low", "medium", "high"))     # you can re-facto
```

Now we have a releveled factor with low as the lowest or first category, medium as the second and high as the third. This is reflected in the way they are listed in the output of `str()`, as well as in the numbering of which category is where in the factor.

> Note: Releveling becomes necessary when you need a specific category in a factor to be the "base" category, i.e. category that is equal to 1. One example would be if you need the "control" to be the "base" in a given RNA-seq experiment.

7 Performing and choosing hypothesis tests

There are many factors which can go into choosing an appropriate hypothesis test for a particular problem. As we've seen if we know or can reasonably assume a model for how our data was generated, we can directly calculate a p-value using a chosen distribution. Additionally, if our data is structured in a way which makes classical hypothesis tests difficult to apply, we can also use strategies involving randomization such as the Monte Carlo method or another strategy called **permutation testing**, where we randomize one of our variables to create null samples.

If we consider the steps of a hypothesis test again we can identify a few factors:

1. Decide on the **effect** that you are interested in, design a suitable **experiment** or study, pick a data summary function and test statistic.
2. Set up a **null hypothesis**
3. Decide on the **rejection region**
4. Do the experiment and collect the data; compute the test statistic.
5. Make a decision: reject the null hypothesis if the test statistic is in the rejection region.

Note that this is **not** meant to be a definitive guide. Instead, we aim to highlight some of the most common tests and factors which need to be considered.

7.1 Performing a Hypothesis Test

Many experimental measurements are reported as rational numbers, and the simplest comparison we can make is between two groups, say, cells treated with a substance compared to cells that are not. The basic test for such situations is the t-test. The test statistic is defined as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means of the two groups, S_1^2 and S_2^2 are the estimated variances of the groups, and n_1 and n_2 are the sizes of the two groups. Because the variance of a difference between two independent variables is the sum of the variances of each individual variable ($var(A-B) = var(A) + var(B)$), we add the variances for each group divided by their sample

sizes in order to compute the standard error of the difference. Thus, one can view the the t statistic as a way of quantifying how large the difference between groups is in relation to the sampling variability of the difference between means.

Let's try this out with the `PlantGrowth` data from R's **datasets** package.

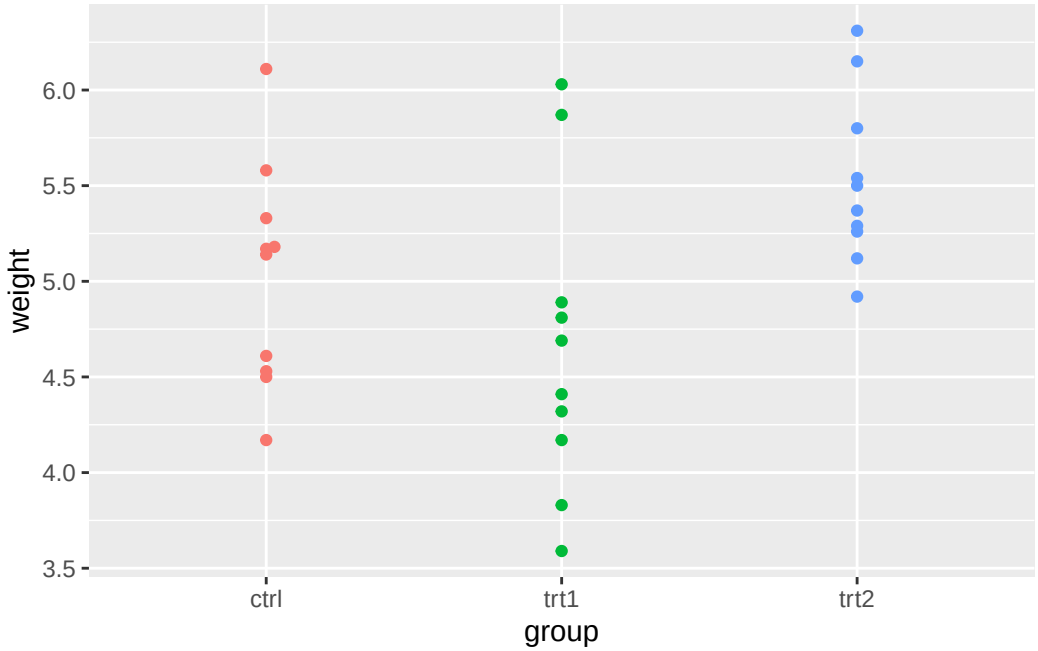
```
library("ggbeeswarm")
```

```
Warning: package 'ggbeeswarm' was built under R version 4.2.2
```

```
Loading required package: ggplot2
```

```
Warning: package 'ggplot2' was built under R version 4.2.2
```

```
data("PlantGrowth")
ggplot(PlantGrowth, aes(y = weight, x = group, col = group)) +
  geom_beeswarm() + theme(legend.position = "none")
```




```
tt1 = t.test(PlantGrowth$weight[PlantGrowth$group =="ctrl"],
       PlantGrowth$weight[PlantGrowth$group =="trt1"],
```

```
      var.equal = TRUE)
tt2 = t.test(PlantGrowth$weight[PlantGrowth$group =="ctrl"],
      PlantGrowth$weight[PlantGrowth$group =="trt2"],
      var.equal = TRUE)
tt1
```

```
    Two Sample t-test

data:  PlantGrowth$weight[PlantGrowth$group == "ctrl"] and PlantGrowth$weight[PlantGrowth$gr
t = 1.1913, df = 18, p-value = 0.249
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2833003  1.0253003
sample estimates:
mean of x mean of y
    5.032     4.661
```

```
tt2
```

```
    Two Sample t-test

data:  PlantGrowth$weight[PlantGrowth$group == "ctrl"] and PlantGrowth$weight[PlantGrowth$gr
t = -2.134, df = 18, p-value = 0.04685
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.980338117 -0.007661883
sample estimates:
mean of x mean of y
    5.032     5.526
```

To compute the p-value, the `t.test` function uses the asymptotic theory for the t-statistic. This theory states that under the null hypothesis of equal means in both groups, the statistic follows a known, mathematical distribution, the so-called t-distribution with $n_1 + n_2 - 2$ degrees of freedom. The theory uses additional technical assumptions, namely that the data are independent and come from a normal distribution with the same standard deviation.

In fact, most of the tests we will look at assume that the data come from a normal distribution. That the normal distribution comes up so often is largly explained by the central limit theorem in statistics. The Central Limit Theorem tells us that as sample sizes get larger, the sampling

distribution of the mean will become normally distributed, *even if the data within each sample are not normally distributed.*

The normal distribution is also known as the *Gaussian* distribution. The normal distribution is described in terms of two parameters: the mean (which you can think of as the location of the peak), and the standard deviation (which specifies the width of the distribution).
The bell-like shape of the distribution never changes, only its location and width.

An important note about the central limit theorem is that it is asymptotic, meaning that it is true as the size of our dataset approaches infinity. For very small sample sizes, even if we are taking the mean of our samples the data might not follow the normal distribution closely enough for tests which assume it to make sense.

The independence assumption

Now let's try something peculiar: duplicate the data.

```
with(rbind(PlantGrowth, PlantGrowth),
       t.test(weight[group == "ctrl"],
              weight[group == "trt2"],
              var.equal = TRUE))
```

```
    Two Sample t-test

data:  weight[group == "ctrl"] and weight[group == "trt2"]
t = -3.1007, df = 38, p-value = 0.003629
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8165284 -0.1714716
sample estimates:
mean of x mean of y
    5.032     5.526
```

Note that estimates of the group means (and thus the difference) are unchanged, but the p-value is now much smaller!

7.2 Choosing the Right Test

7.2.1 Variable Types (Effect)

The types of our variables need to be considered. We will go through some choices if our variables are quantitative (continuous; a number or qualitative (discrete; a category or factor).

However, note that other tests exist for some specific properties like proportions.

If we wish to consider the relationship between **two quantitative variables**, we need to perform a correlation analysis. The Pearson correlation directly analyses the numbers (is parametric) while Spearman's rank correlation considers ranks (and is nonparametric).

For **two qualitative variables**, we typically will use a Chi-square test of independence, though we may be able to use Fisher's exact test if the dataset is small enough.

We often are interested in the case where we want to see the relationship between **one quantitative variable and one qualitative variable.** In this case, we most commonly use some variation of a t-test if we have only have 2 groups we are considering, and some variation of an ANOVA test if we have more than 2. We will get into more detail about ANOVA tests in a future session.

7.2.2 Paired vs Unpaired

Paired and unpaired tests refer to whether or not there is a 1:1 correspondence between our different observations. Experiments which involve measuring the same set of biological samples, often as before and after some kind of treatment, are paired. In paired experiments we can look at each observation, see whether it individually changed between groups.

In unpaired tests we consider our samples to be independent across groups. This is the case if we have two different groups, such as a control group and a treatment group.

Performing a paired or unpaired test can be set as an argument in R's `t.test` function, but nonparametric tests have different names, the Mann-Whitney U test for unpaired samples and the Wilcoxon signed-rank test for paired samples in tests with 2 groups, and the Kruskal-Wallis test and Friedman test for more than two groups.

7.2.3 Parametric vs Non-Parametric

So far, we have only seen parametric tests. These are tests which are based on a statistical distribution, and thus depends on having defined parameters. These tests inherently assume that the collected data follows some distribution, typically a normal distribution as discussed above.

A nonparametric test makes many fewer assumptions about the distribution of our data. Instead of dealing with values directly, they typically perform their calculations on rank. This makes them especially good at dealing with extreme values and outliers. However, they are typically less powerful than parametric tests; they will be less likely to reject the null hypothesis (return a higher p-value) if the data did follow a normal distribution and you had performed a parametric test on it. Thus, they should only be used if necessary.

A typical rule of thumb is that around 30 samples is enough to not have to worry about the underlying distribution of your data. However, they are types of data, such as directly collecting ranking data or ratings, which should be analyzed with nonparametric methods.

7.2.4 One-tailed and Two-tailed tests

All tests have one-tailed and two-tailed versions. A two-tailed test considers a result significant if it is extreme in either direction; it can be higher or lower than what would be expected under the null hypothesis. A one-tailed test will only consider a single direction, either higher or lower. Usually, the p value for the two-tailed test is twice as large as that for the one-tailed test, which reflects the fact that an extreme value is less surprising since it could have occurred in either direction.

How do you choose whether to use a one-tailed versus a two-tailed test? The two-tailed test is always going to be more conservative, so it's always a good bet to use that one, unless you had a very strong prior reason for using a one-tailed test. This is set through the `alternative` argument in `t.test`.

7.2.5 Variance

Another underlying assumption of many statistical tests is that different groups have the same variance. The t-test will perform a slightly more conservative calculation if equal variance is not assumed (called Welch's t-test instead of Student's t-test). This can be set as the `var.equal` argument of `t.test`.

We often can assume equal variance, but as we will see in a later session, many modern sequencing technologies can produce data with patterns in its variance we will have to adjust for.

7.2.6 How Many Variables of Interest?

All of the above discussion is for experiments with where we are interested in looking at the relationship between two variables. These, slightly confusingly, are called 2 sample tests, and line up with the classical experimental paradigm of a single dependent and a single independent variable. However, there are other options.

- One Sample: Instead of wanting to compare how a categorical variable (like treatment) affects some outcome variable, we could imagine comparing against some known value. When we considered whether or not a coin was fair, we were not comparing two coins, but instead comparing the output of one coin against a known value.

- More than two samples: Modern observational studies often, by necessity, need to consider how many variables affect some outcome. These analyses are performed via regression models, multiple linear regression for a quantitative dependent variable and logistic regression for a qualitative dependent variable.

8 Problem Set 1

8.1 Problem 1

R can generate numbers from all known distributions. We now know how to generate random discrete data using the specialized R functions tailored for each type of distribution. We use the functions that start with an `r` as in `rXXXX`, where `XXXX` could be `pois`, `binom`, `multinom`. If we need a theoretical computation of a probability under one of these models, we use the functions `dXXXX`, such as `dbinom`, which computes the probabilities of events in the discrete binomial distribution, and `dnorm`, which computes the probability density function for the continuous normal distribution. When computing tail probabilities such as $P(X > a)$ it is convenient to use the cumulative distribution functions, which are called `pXXXX`. Find two other discrete distributions that could replace the `XXXX` above.

8.2 Problem 2

How would you calculate the *probability mass* at the value $X = 2$ for a binomial $B(10, 0.3)$ with `dbinom`? Use `dbinom` to compute the *cumulative* distribution at the value 2, corresponding to $P(X \leq 2)$, and check your answer with another R function. *Hint: You will probably want to use the* `sum` *function.*

8.3 Problem 3

In the epitope example (Section 5.1), use a simulation to find the probability of having a maximum of 9 or larger in 100 trials. How many simulations do you need if you would like to prove that "the probability is smaller than 0.000001"?

8.4 Problem 4

Find a paper in your research area which uses a hypothesis test. Cite the paper and note:

- The null hypothesis.

- The alternative hypothesis.
- Was the test two-tailed or one-tailed?
- What types of variables were compared?
- Was the test parametric or non-parametric?
- Can we safely assume equal variance?
- What was the sample size?

If the necessary details to determine any of the above are not in the paper, you can note that instead.

Given what you've written and the author's decisions, do you agree with the choice of hypothesis test and the conclusions drawn?
