BIOSCI738 Assignment 1

Haileigh Patterson

2022-03-23

Question 1 - In brief, discuss the benefits and potential drawbacks related to the use of the open source software R. [10 marks]

Software that is open source is able to be changed and manipulated by any user due to the accessibility of the source code.

- The benefits of open source software R:
- Designed specifically by and for statisticians.
- You aren't limited in what you can do with the program - you may change, edit and distribute it however you like, as long as you have the knowledge to do so.
- Collaboration is encouraged through the design of the software - open source is open for anyone to use and update the source code, and collaborate on ensuring the code is operating optimally. This also encourages a community to form around the software to help and encourage each other.
- Disadvantages of open source software R:
- As R was designed for statisticians, it was designed with flexibility in how you input phrases and functions - but this also means that code can be messy and a lack of discipline can lead to hard-to-follow code.
- There is little incentive for packages to work perfectly or to be continually updated or optimised; anything you install is likely made by someone out of the goodness of their heart and it is your responsibility to ensure it works and to troubleshoot any issues.
- If the person who is maintaining a package that you use is no longer able to maintain it, it could become obsolete or cease to be compatible with new versions of RStudio etc. This is also mitigated by the fact that it is open source and anyone can pick up this piece of code and attempt to update and redistribute.

Question 2 - Te Tiriti o Waitangi/Treaty of Waitangi obliges the Government to actively protect taonga, consult with Māori in respect of taonga, give effect to the principle of partnership and recognize Māori rangatiratanga over taonga. The Te Mana o te Raraunga Model was developed to align Māori concepts with data rights and interests, and guide agencies in the appropriate use of Māori data. In brief, discuss the relevance of and researchers' obligation to Māori data sovereignty and its importance when dealing with data in Aotearoa. [10 marks]

Aotearoa has a unique history of colonisation that requires a unique response to the issues that now arise from it. One of the issues is upholding the values that were outlined in te tiriti o waitangi; tino rangatiratanga

(self-governance) and the kaitiakitanga of taonga. I think all people who reside in Aotearoa have an obligation to uphold the values in te tiriti o waitangi as occupiers of a colonised whenua.

Data sovereignty can concern both of these principles in that data has a direct impact on informing decision-making in a range of areas. As an example, reports published on Māori health data may influence initiatives to improve Māori health outcomes. The way in which that data is collected and the results portrayed may influence the public knowledge or opinion on Māori so there are additional considerations that need to be taken into account. If the data or its use affects Māori interests or taonga, then the data should be collected and utilised in a way that is informed and consensual. For example, any samples and data collected or used from kauri trees in the Waitakere ranges is done with the permission of Te Kawerau ā Maki which is the iwi of that rohe.

I think that science that is done without the above considerations for the indigenous people to whom the data pertains is a form of secondary colonisation and contributes to oppression. It means that the right to control the discourse or outcomes of the data is taken away.

Question 3 - Read the paper (Peterson et al. 2020) and consider Figure 1 (see below also). Reproduce (as close as you can) this figure using your ggplot2 skills. Do you think this figure could be improved? Alongside your reproduced figure plot your improved figure. You should include your code for both plots in your answer (in printed code chunks). Your code must be reproducible by your peers. You may only assume that your peers have loaded the data as above. Note: don't forget to include package calls etc.

```
library(tidyverse)
library(dplyr)
library(scales)
library(readr)
```

```
## Rows: 24 Columns: 42
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Treat!
## dbl (41): Block!, Calluna vulgaris08, Exotic dicots08, Exotic monocots08, Na...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#url <- "https://raw.githubusercontent.com/STATS-UOA/databunker/master/data/dicots_proportions.csv" #ob
#data <- read_csv(url) #places data into variable "data"
```

Code chunk for Data Manipulation

```
#extract the calluna data from the dataset
callunavulgaris_data <- data %>% select(starts_with("Calluna"), `Treat!`)

#save dataframe in project folder
save(callunavulgaris_data, file = "callunavulgaris_data.RData")

# pivot long rather than wide
calluna_long <- pivot_longer(callunavulgaris_data,
                             cols = starts_with("Calluna"),
                                      names_to = "Year", values_to = "Spread")
```

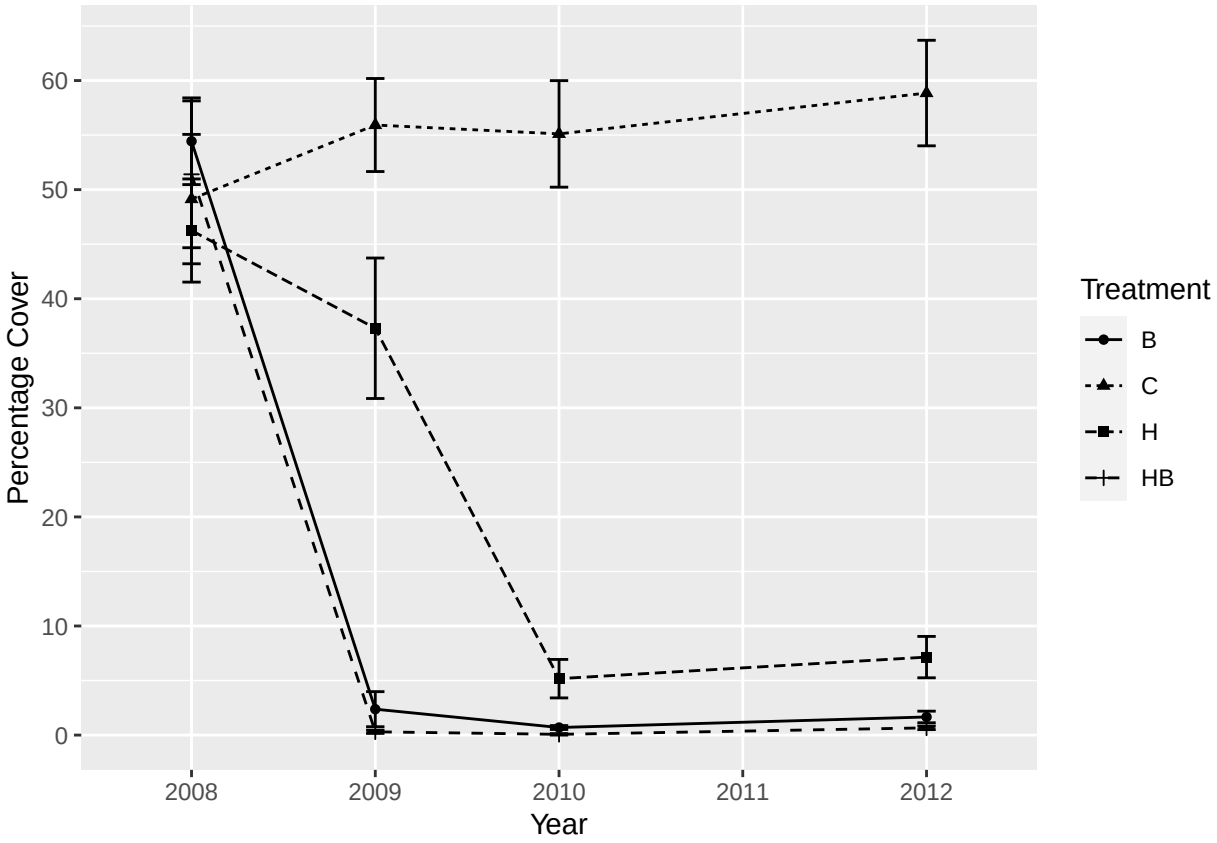
```r
## rename columns to match the dataset
names(calluna_long)[1] <- 'Treatment'

#find the index (place in the column) where the "Year" column ends with "08"
index_08 <- which(endsWith(calluna_long$Year,"08"))
print(index_08)

#for every index in the "Year" column. Change it to "2008"
for (index in index_08){
  calluna_long$Year[index] <- "2008"
}

#find the index (place in the column) where the "Year" column ends with "09"
index_09 <- which(endsWith(calluna_long$Year,"09"))
print(index_09)

#for every index in the "Year" column. Change it to "2009"
for (index in index_09){
  calluna_long$Year[index] <- "2009"
}

#find the index (place in the column) where the "Year" column ends with "10"
index_10 <- which(endsWith(calluna_long$Year,"10"))
print(index_10)

#for every index in the "Year" column. Change it to "2010"
for (index in index_10){
  calluna_long$Year[index] <- "2010"
}

#find the index (place in the column) where the "Year" column ends with "12"
index_12 <- which(endsWith(calluna_long$Year,"12"))
print(index_12)

#for every index in the "Year" column. Change it to "2012"
for (index in index_12){
  calluna_long$Year[index] <- "2012"
}

# wrangle the spread into percentage format
calluna_long <- calluna_long %>%
  select(Treatment, Year, Spread)%>%
  mutate(
    Spread = Spread * 100)

# find the mean and sd/se for the treatment type and year
calluna_mean <- calluna_long %>%
  group_by(Treatment, Year) %>%
  summarise(., mean_year_treatment = mean(Spread), sd = sd(Spread), n = n(), se = sd / sqrt(n))
```

```
## `summarise()` has grouped output by 'Treatment'. You can override using the
## `.groups` argument.
```

```
calluna_mean$mean_year_treatment = as.numeric(calluna_mean$mean_year_treatment)
```

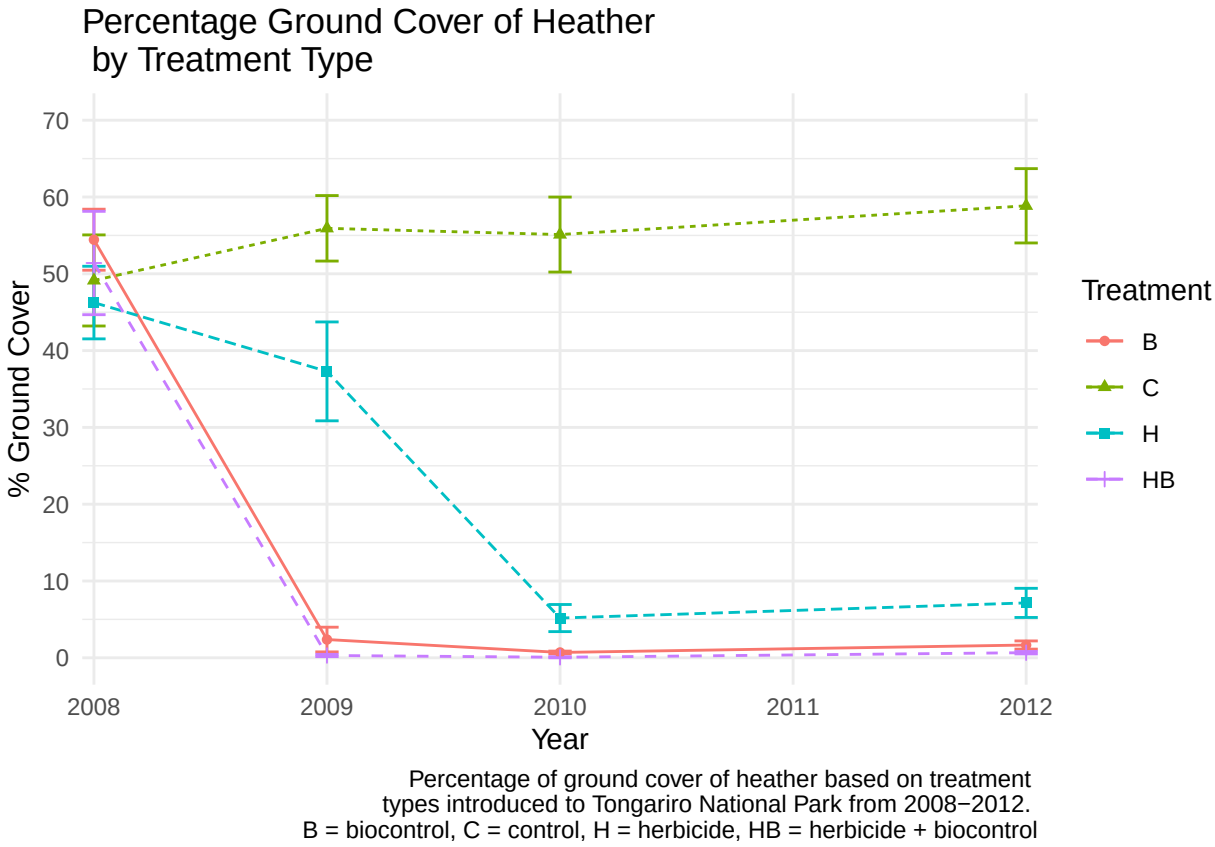
Code chunk for Reproduced Figure 1

```
## ggplot line graph
ggplot2::ggplot(data = calluna_mean, aes(x = Year, y = mean_year_treatment, group = Treatment)) +
  geom_errorbar(aes(ymin=mean_year_treatment-se, ymax=mean_year_treatment+se ), width=.1) +
  geom_line(aes(linetype = Treatment)) +
  geom_point(aes(shape = Treatment)) + xlab("Year") + ylab("Percentage Cover") +
  scale_x_discrete(limit = c("2008", "2009", "2010", "2011", "2012")) +
  scale_y_continuous(
    n.breaks = 7,
    limits = NULL,
    )
```



Code chunk for Improved Figure 1

```
## ggplot line graph IMPROVED GRAPH
ggplot2::ggplot(data = calluna_mean, aes(x = Year, y = mean_year_treatment, group = Treatment)) +
  geom_errorbar(aes(ymin=mean_year_treatment-se, ymax=mean_year_treatment+se, colour = Treatment ), wid
  geom_line(aes(linetype = Treatment, colour = Treatment)) +
  theme_minimal() +
  ggtitle("Percentage Ground Cover of Heather \n by Treatment Type")+
  labs(caption = "Percentage of ground cover of heather based on treatment \n types introduced to Tonga
  geom_point(aes(shape = Treatment, colour = Treatment)) + xlab("Year") + ylab("% Ground Cover") +
  scale_x_discrete(limit = c("2008", "2009", "2010", "2011", "2012"), expand = c(0,0)) +
  scale_y_continuous(
    n.breaks = 7,
    limits = c(0, 70),
  )
```



Question 4 - Letting μt:y be the mean percentage cover for treatment level t in year y test the following hypothesis. Use a randomisation test in each case with 49 and then again with 499 resamples. [15 marks]

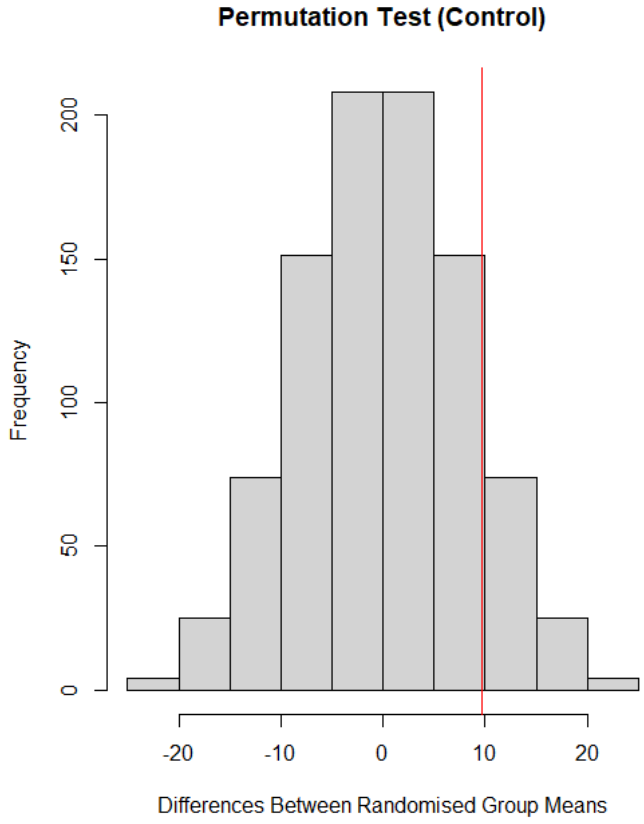
H0 : μC:2012 = μC:2008 vs. H1 : μC:2012 6= μC:2008

H0 : μB:2012 = μB:2008 vs. H1 : μB:2012 6= μB:2008

Treatment	Observed Test Statistic (difference in means)	p-value
Control	9.719599	0.2402597
Biocontrol	-52.77663	0.002164502

The above p-values show what you would expect from the results - that the control does not show a low p-value, i.e. the result we received was statistically likely to occur. Therefore, we accept the null hypothesis that there is no difference in means between the controls.

Conversely, the p-value for the biocontrol subset supports the hypothesis that there is a difference between the means between 2008 and 2012, showing the biocontrol treatment had a significant effect on the percentage of ground cover of heather. This is because the p-value is low, insinuating that the observed statistic is very unlikely to occur if the null hypothesis is true.



The red line in the histograms indicates the observed test statistics.

From the histograms I think it backs up the conclusions drawn from the p-values. The control histogram clearly shows the observed test statistic within the normal range of the expected observations, ie an expected result within the null hypothesis.

The histogram for the biocontrol subset, however, shows the observed test statistic outside of the normal range of expected results showing that something has acted to disprove the null hypothesis (ie the biocontrol treatment).