Worksheet 04 - Intro to R programming - NCBS MSc WL (Answers)

Akshay Surendra, Anand Osuri

13 October 2020

The problem set is built around a dataset of tree species occurrences within three habitats. While the dataset itself is not real, it is inspired by real species and places. Import the file `prob_set_data.csv` and store it within an object named `treedat`. Each row corresponds to individual trees of different Species recorded within three Habitats – `Type_1`, `Type_2` and `Type_3`

```
rm(list = ls()) # clears your R Environment = equal to clicking the broom symbol
library(tidyverse)
setwd("D:/2020_IntroToR_NCBS/IntroR_2020_NCBS_content/Worksheet_04/")
treedat <- read_csv(file = "Send_To_Students/prob_set_data.csv")
```

Question 1 We need to add a column describing the year during which each habitat type was sampled. `Type_2` was sampled in 2017, and `Type_1` and `Type_3` during 2018. Can you add this information to the tibble using `mutate()` in conjunction with a conditional statement? Call the new column `Year_conditional`

```
# method 1, using case_when
treedat <-
  treedat %>%
  mutate(Year_conditional = case_when(Habitat == "Type_1" ~ 2018,
                                      Habitat == "Type_2" ~ 2017,
                                      Habitat == "Type_3" ~ 2018))

# method 2, using ifelse()
# treedat <-
#   treedat %>%
#   mutate(Year_conditional = ifelse(test = Habitat == "Type_2",
#                                    yes =  2017,
#                                    no = 2018))
```

Question 2 Can you also achieve the above task using a join function? In this case, call the new column for year `Year_join`

[hint: first, create a new tibble `Year_tib <- tibble(Habitat = c("Type_1", "Type_2", "Type_3"), Year_join = c(2018, 2017, 2018)`]

```
Year_tib <- tibble(Habitat = c("Type_1","Type_2","Type_3"),
                   Year_join = c(2018,2017,2018))

treedat <- left_join(x = treedat,
```

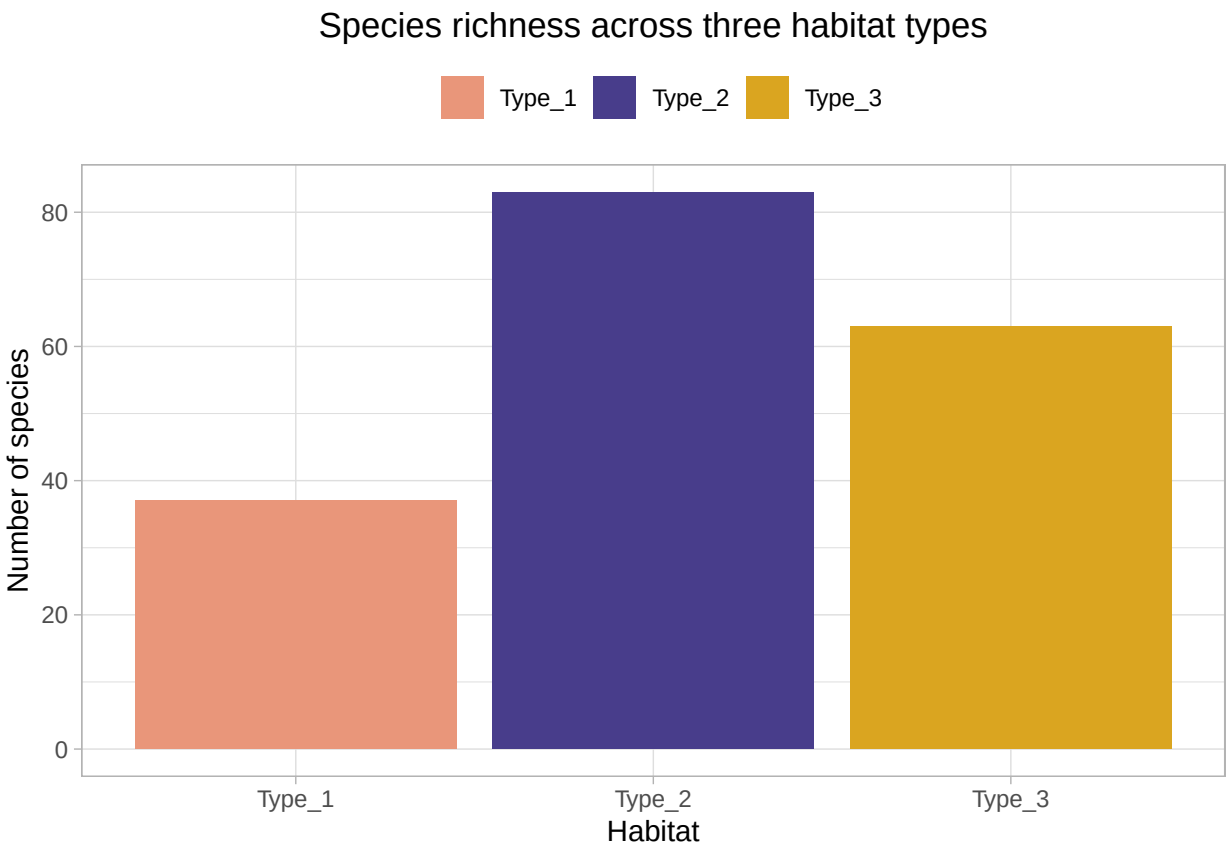
```
                      y = Year_tib,
                      by = c("Year_conditional" = "Year_join",
                             "Habitat" = "Habitat"))
# join more than one column this way (otherwise you get two copies)

# you can also pipe data frames into left_join:
# treedat <- treedat %>% left_join(y = Year_tib,
#                                  by = c("Year_conditional" = "Year_join"))
```

Question 3 Using what you have learned previously in `dplyr`, generate a table reporting the number of trees recorded within each habitat type (i.e., the numbers of rows of data within each habitat type) [hint: `group_by()`, `summarize()`]

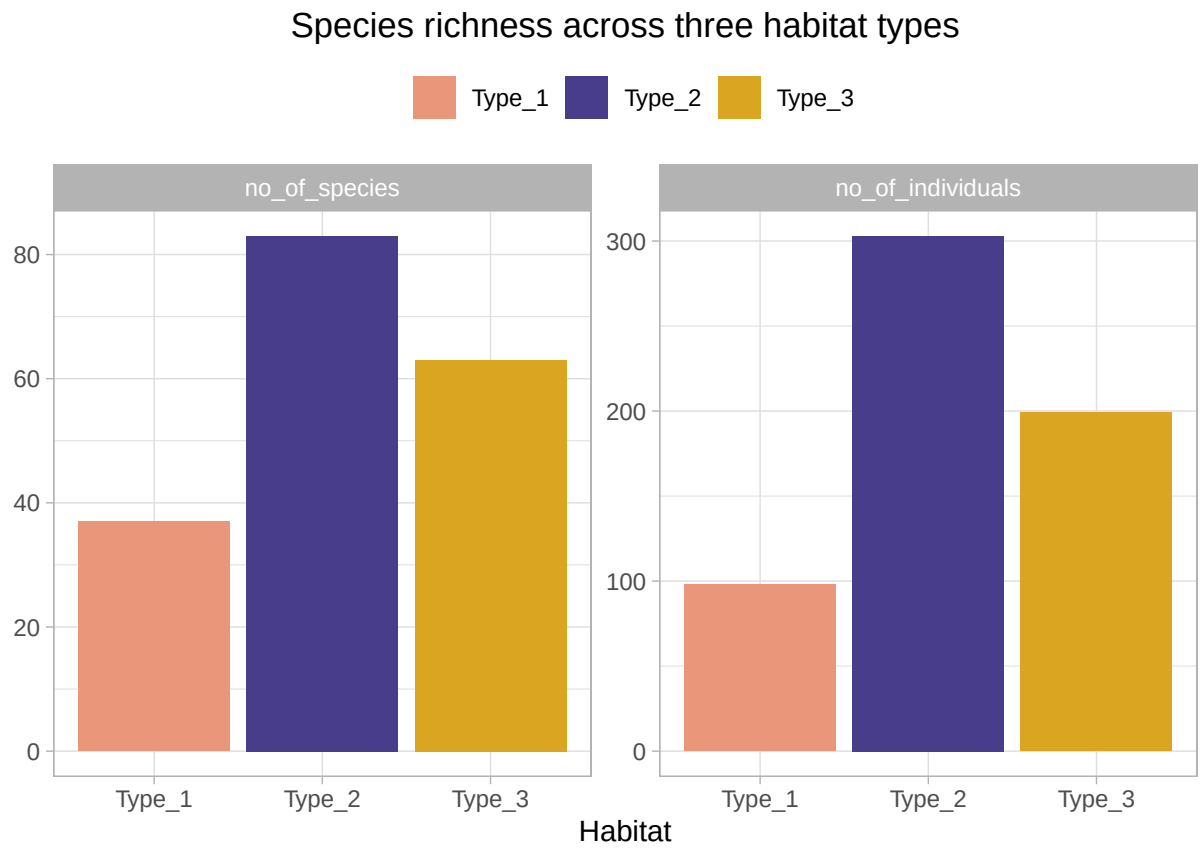
```
table_q3 <-
  treedat %>%
  group_by(Habitat) %>%
  summarise(no_of_species = n_distinct(Species))

# replacing n_distinct(Species) with n() will give you no of individual trees
```

Question 4 Again, using dplyr functions, generate a table reporting the number of species recorded in each habitat type. Also generate a bar graph showing the species richness of the three habitats. [hint: `n_distinct()`]

```
table_q3 <-
  treedat %>%
  group_by(Habitat) %>%
  summarise(no_of_species = n_distinct(Species))

plot1 <-
  ggplot(data = table_q3) +
  geom_bar(mapping = aes(x=Habitat,
                         y = no_of_species),stat = "identity")
# note the use of stat = "identity"

plot1_beautified <-
ggplot(data = table_q3) +
  geom_bar(mapping = aes(x=Habitat,
                         y = no_of_species,
                         fill = Habitat),
           stat = "identity") +
  ylab("Number of species") +
  ggtitle("Species richness across three habitat types") +
  theme_light() + # cleaner look, try other theme_...() functions
  theme(legend.position = "top", # moves legend from right side to the top
        plot.title = element_text(hjust = 0.5)) + # centre-align plot
  scale_fill_manual(name="",
                    values = c("darksalmon","darkslateblue","goldenrod"))

plot1_beautified
```



Q: Can you comment on which of the three habitats harbours the most species? A: Based on raw counts, you may notice that Habitat Type_2 has the most species, followed by Type_3 and then by Type_1. You may also notice a similar ranking in terms of overall counts of trees (Type_2 > Type_3 > Type_1).



In making comparisons between habitats, sometimes we may be more interested in knowing the numbers of species for a given number of individuals in each habitat, than in the overall counts of species per habitat. This approach is called **individual-based rarefaction**

Question 5 Use `dplyr` tools to create a subset of the dataset comprising only data from habitat Type_1. Store this in an object habitat_1. From within habitat_1, select a random sample (without replacement) of 50 individuals, and count how many species this sample contains. Repeat the selection step five times, and as a comment in your code report the number of species present in each selection. Do you get the same number of species each time?

```
habitat_1 <- treedat %>% filter(Habitat=="Type_1")

tmp <- sample(x = 1:nrow(habitat_1),size = 50,replace = FALSE)
tmpdat <- habitat_1[tmp,]
tmp_nsp <- tmpdat %>% pull(Species) %>% unique() %>% length()
# same as doing: length(unique(tmpdat$Species))

tmp1 <- habitat_1[sample(x = 1:nrow(habitat_1),size = 50,replace = FALSE),] %>%
  distinct(Species) %>% nrow()
tmp2 <- habitat_1[sample(x = 1:nrow(habitat_1),size = 50,replace = FALSE),] %>%
  distinct(Species) %>% nrow()
tmp3 <- habitat_1[sample(x = 1:nrow(habitat_1),size = 50,replace = FALSE),] %>%
  distinct(Species) %>% nrow()
tmp4 <- habitat_1[sample(x = 1:nrow(habitat_1),size = 50,replace = FALSE),] %>%
  distinct(Species) %>% nrow()
tmp5 <- habitat_1[sample(x = 1:nrow(habitat_1),size = 50,replace = FALSE),] %>%
```

```
  distinct(Species) %>% nrow()

print(x = paste("Sample1:",tmp1,"species",sep = " ")) # 28
```

```
## [1] "Sample1: 26 species"
```

```
print(x = paste("Sample2:",tmp2,"species",sep = " ")) # 26
```

```
## [1] "Sample2: 27 species"
```

```
print(x = paste("Sample3:",tmp3,"species",sep = " ")) # 25
```

```
## [1] "Sample3: 23 species"
```

```
print(x = paste("Sample4:",tmp4,"species",sep = " ")) # 29
```

```
## [1] "Sample4: 27 species"
```

```
print(x = paste("Sample5:",tmp5,"species",sep = " ")) # 26
```

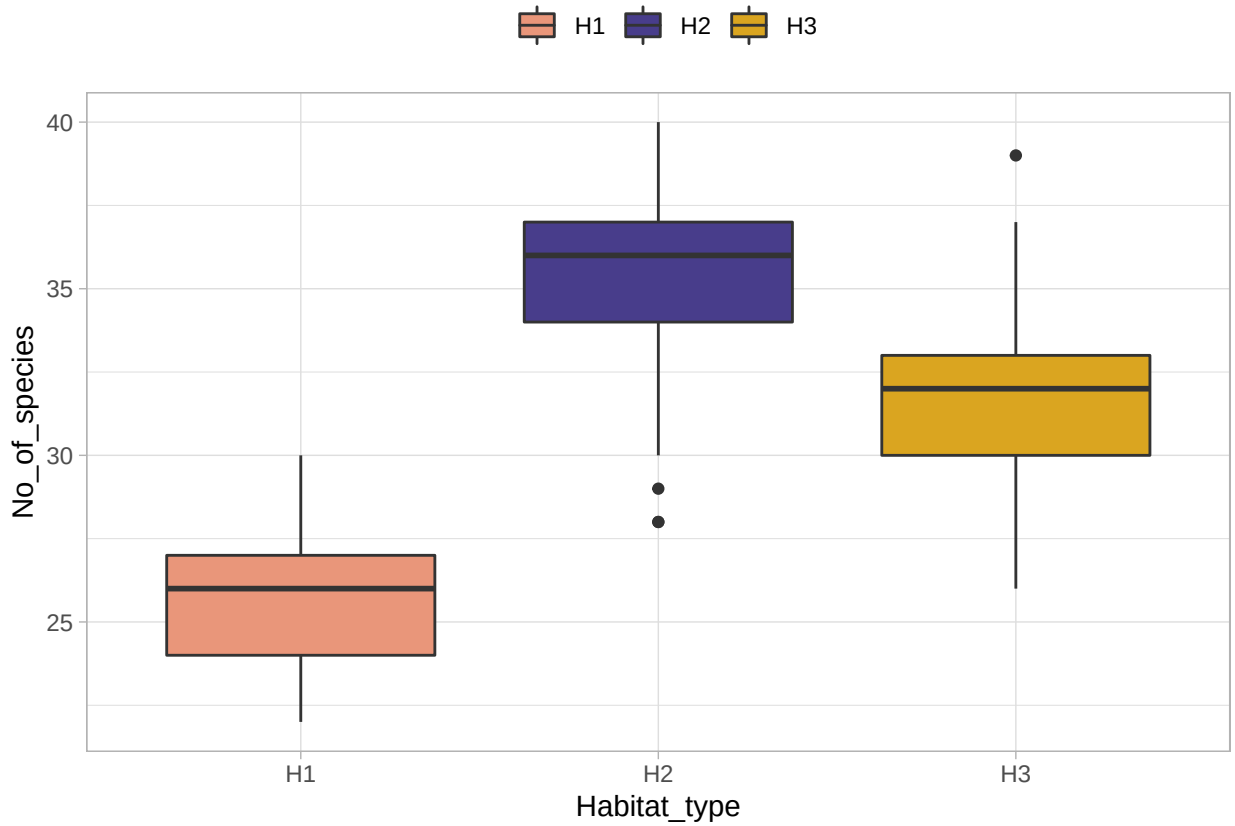
```
## [1] "Sample5: 25 species"
```

```
# We don't get the same number of species each time (we haven't set seed)
```

Question 6 As above, create data subsets corresponding to habitat `Type_2` and `Type_3`, and name them `habitat_2` and `habitat_3`, respectively. Use a `for()` loop to perform 100 iterations of individual-based rarefactions for each habitat. In each iteration, draw a sample of 50 individuals and count the numbers of species contained within the sample for each habitat. Plot the comparison of rarefied species richness across the three habitats using a box plot.

```
habitat_1 <- treedat %>% filter(Habitat=="Type_1")
habitat_2 <- treedat %>% filter(Habitat=="Type_2")
habitat_3 <- treedat %>% filter(Habitat=="Type_3")

iter <- 100
nsize <- 50
nspecies_h1 <- rep(NA,iter)
nspecies_h2 <- rep(NA,iter)
nspecies_h3 <- rep(NA,iter)

for(i in 1:iter)
{

  nspecies_h1[i] <-
    habitat_1[sample(x = 1:nrow(habitat_1),size = nsize,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()
```

```
  nspecies_h2[i] <-
    habitat_2[sample(x = 1:nrow(habitat_2),size = nsize,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  nspecies_h3[i] <-
    habitat_3[sample(x = 1:nrow(habitat_3),size = nsize,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  # print(x = paste("Currently running iteration number:",i))
}

raredat <-
  tibble(H1 = nspecies_h1,
         H2 = nspecies_h2,
         H3 = nspecies_h3) %>%
    pivot_longer(cols = c("H1","H2","H3"),
                 names_to = "Habitat_type",
                 values_to = "No_of_species")

ggplot(data = raredat) +
  geom_boxplot(mapping = aes(y = No_of_species,
                             x = Habitat_type,
                             fill = Habitat_type)) +
  theme_light() +
  theme(legend.position = "top") +
  scale_fill_manual(name="",
                    values = c("darksalmon","darkslateblue","goldenrod"))
```




```
habitat_1 <- treedat %>% filter(Habitat=="Type_1")
habitat_2 <- treedat %>% filter(Habitat=="Type_2")
habitat_3 <- treedat %>% filter(Habitat=="Type_3")

iter <- 100
nspecies_h1_25ind <- rep(NA,iter)
nspecies_h2_25ind <- rep(NA,iter)
nspecies_h3_25ind <- rep(NA,iter)

nspecies_h1_50ind <- rep(NA,iter)
nspecies_h2_50ind <- rep(NA,iter)
nspecies_h3_50ind <- rep(NA,iter)

nspecies_h1_75ind <- rep(NA,iter)
nspecies_h2_75ind <- rep(NA,iter)
nspecies_h3_75ind <- rep(NA,iter)

for(i in 1:iter)
{

  # 25 individuals per habitat
  nspecies_h1_25ind[i] <-
    habitat_1[sample(x = 1:nrow(habitat_1),size = 25,replace = FALSE),] %>%
```

```
    distinct(Species) %>% nrow()

  nspecies_h2_25ind[i] <-
    habitat_2[sample(x = 1:nrow(habitat_2),size = 25,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  nspecies_h3_25ind[i] <-
    habitat_3[sample(x = 1:nrow(habitat_3),size = 25,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  # 50 individuals per habitat
  nspecies_h1_50ind[i] <-
    habitat_1[sample(x = 1:nrow(habitat_1),size = 50,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  nspecies_h2_50ind[i] <-
    habitat_2[sample(x = 1:nrow(habitat_2),size = 50,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  nspecies_h3_50ind[i] <-
    habitat_3[sample(x = 1:nrow(habitat_3),size = 50,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  # 75 individuals per habitat
  nspecies_h1_75ind[i] <-
    habitat_1[sample(x = 1:nrow(habitat_1),size = 75,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  nspecies_h2_75ind[i] <-
    habitat_2[sample(x = 1:nrow(habitat_2),size = 75,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  nspecies_h3_75ind[i] <-
    habitat_3[sample(x = 1:nrow(habitat_3),size = 75,replace = FALSE),] %>%
  distinct(Species) %>%
    nrow()

  # print(x = paste("Currently running iteration number:",i))
}

raredat <-
  tibble(H1_25 = nspecies_h1_25ind,H2_25 = nspecies_h2_25ind,H3_25 = nspecies_h3_25ind,
         H1_50 = nspecies_h1_50ind,H2_50 = nspecies_h2_50ind,H3_50 = nspecies_h3_50ind,
         H1_75 = nspecies_h1_75ind,H2_75 = nspecies_h2_75ind,H3_75 = nspecies_h3_75ind
         ) %>%
    pivot_longer(cols = c("H1_25","H2_25","H3_25",
```

```
                          "H1_50","H2_50","H3_50",
                          "H1_75","H2_75","H3_75"),
                 names_to = "Habitat_type",
                 values_to = "No_of_species") %>%
  separate(col = Habitat_type,into = c("Habitat","SampleSize"),sep = "_",
           remove = TRUE) %>%
  arrange(Habitat,SampleSize)
# separates the habitat_type columns into two using the underscore sign
# removes the old combined columns
```

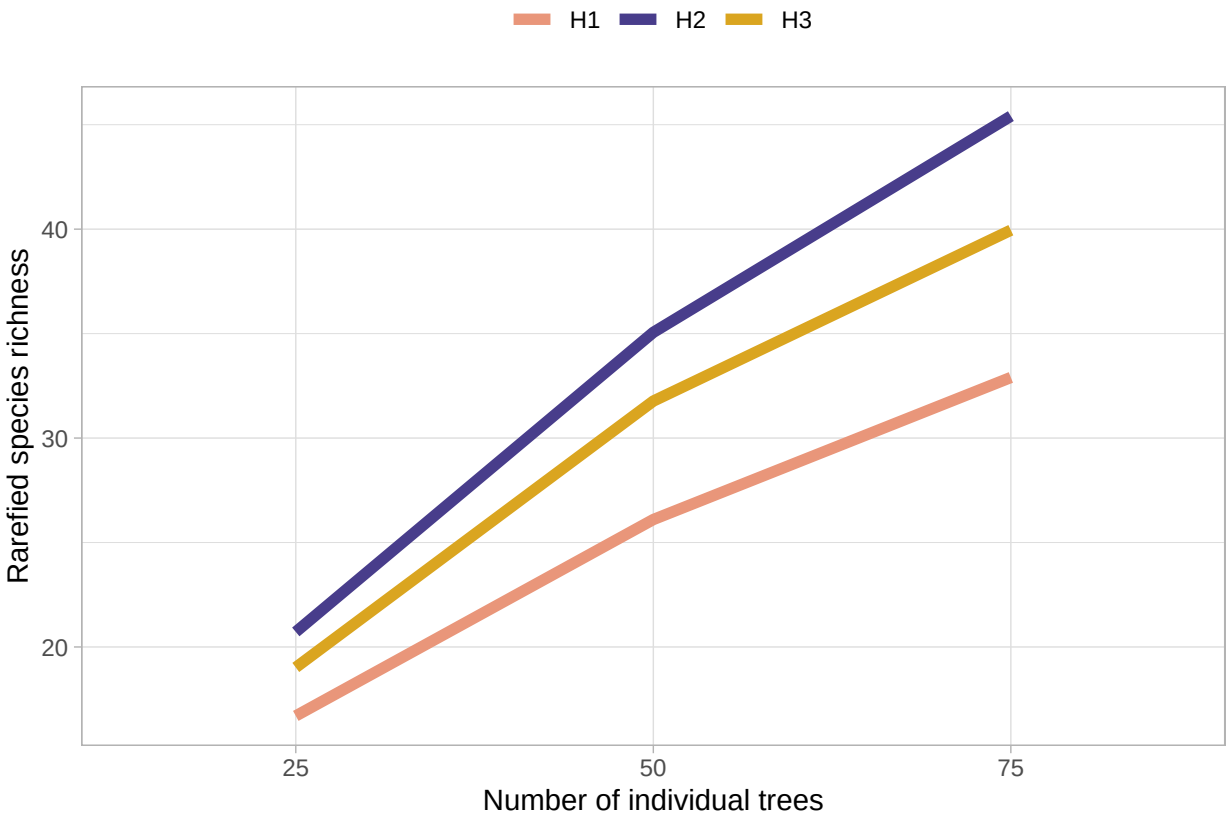
Bonus question Average number of species per habitat, per sample size:

```
raredat_tab <-
  raredat %>%
  group_by(Habitat,SampleSize) %>%
  summarise(mean_SR = mean(No_of_species),
            sd_SR = sd(No_of_species),
            niter = n()) %>%
  arrange(SampleSize)
raredat_tab
```

```
## # A tibble: 9 x 5
## # Groups:   Habitat [3]
##   Habitat SampleSize mean_SR sd_SR niter
##   <chr>   <chr>        <dbl> <dbl> <int>
## 1 H1      25            16.7  1.78   100
## 2 H2      25            20.7  1.56   100
## 3 H3      25            19.0  1.78   100
## 4 H1      50            26.1  2.04   100
## 5 H2      50            35.1  2.51   100
## 6 H3      50            31.8  2.39   100
## 7 H1      75            32.9  1.73   100
## 8 H2      75            45.4  2.89   100
## 9 H3      75            40.0  2.94   100
```

Line graph of species accumulation:

```
ggplot(data = raredat_tab) +
  geom_line(mapping = aes(y = mean_SR,
                          x = SampleSize,
                          group = Habitat,
                          colour = Habitat),size = 2) +
  theme_light() +
  theme(legend.position = "top") +
  scale_colour_manual(name="",
                      values = c("darksalmon","darkslateblue","goldenrod")) +
  ylab("Rarefied species richness") + xlab("Number of individual trees")
```



Note, At each point on the x-axis (25,50,75), the corresponding y-value i.e. species richness, is comparable - they've been calculated for the **same** number of individual trees