Species accumulation curves and diversity indices

Clarke van Steenderen

2025-02

Using the *vegan* package to compute species accumulation curves

A species accumulation curve (SAC), also sometimes called a species richness curve, is a means of estimating species richness in a particular area as sampling effort increased. The x-axis shows the cumulative number of surveys/collections, and the y-axis shows the cumulative number of species found. A curve that reaches an asymptote suggests that further sampling is unlikely to yield further species, while a curve that is steadily increasing suggests that further sampling effort is required.

The first section of this tutorial has been adapted from Guy Sutton's blogpost.

We will be using a data set containing the insect community associated with the *Lycium ferocissimum* shrub (African boxthorn) native to South Africa. Here, we want to find out whether the sampling effort so far has likely found all the potential insect biocontrol agents on this shrub.

Let's load up the data into R!

```
# install the required packages, if not available already
if (!require("pacman")) install.packages("pacman")
```

```
## Warning: package 'pacman' was built under R version 4.3.3
```

```
pacman::p_load(tidyverse,
               tidyr,
               janitor,
               vegan,
               readr,
               magrittr)

# Read in the data file
sp_comm <- readr::read_csv("data/species_abundance_matrix.csv") %>%
# Clean column names
  janitor::clean_names() %>%
  dplyr::mutate(season = dplyr::if_else(season == 1, "Summer", "Winter"))

# Check the data contents
dplyr::glimpse(sp_comm)
```

```
## Rows: 56
## Columns: 56
## $ provinces                                        <chr> "Eastern Cape", "Eastern C~
## $ climatic_zones                                   <chr> "Cfb", "Cfb", "Cfa", "Cfa"~
## $ site                                             <chr> "EC1", "EC1", "EC7", "EC7"~
## $ season                                      <chr> "Summer", "Winter", "Summe~
## $ haplotype                                   <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5,~
## $ cleta_eckloni                               <dbl> 23, 8, 11, 0, 0, 0, 1, 2, ~
## $ pseudambonea_capeni_schuhistes_lekkersingia <dbl> 4, 28, 0, 0, 0, 0, 3, 20, ~
## $ acanthocoris_spinosus                       <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ antestiopsis_thunbergii                     <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ cassida_distinguenda                        <dbl> 0, 3, 0, 0, 0, 0, 0, 1, 0,~
## $ epilachna_sp_1                              <dbl> 0, 4, 0, 0, 0, 0, 0, 0, 0,~
## $ cleta_sp_1                                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ cleta_sp_2                                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ exochomus_flavipes                          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ scymnus_sp                                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ cheilomenes_lunata                          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ cheilomenes_sulphurea                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ cf_nephus_sp                                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1,~
## $ chnootriba_sp                               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ oenopia_cinctella                           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ hippodamia_variegate                        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ cassida_melanophthalma                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ cassida_reticulipennis                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ macetes_sp                                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1,~
## $ cryptocephalus_nr_liturellus                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ epitrix_sp                                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ chrysomelidae_sp                            <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0,~
## $ sulcobruchus_longipennis                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ monolepta_bioculata                         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ eurytomidae_sp                              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ pachycnema_sp                               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ scarabaeidae_sp                             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ neoplatygaster_serieturberculata            <dbl> 0, 8, 0, 0, 0, 0, 0, 0, 0,~
## $ sciobius_sp                                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ lixini_sp                                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ beaufortiana_cornuta                        <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0,~
## $ pentatomidae_sp                             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ apalochrus_sp                               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ hylomela_sexpunctata                        <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0,~
## $ anthripidae_sp                              <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0,~
## $ cenaeus_carnifex                            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ thrips_simplex                              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ ceratitis_sp                                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ brachymeria_sp                              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ syrphidae_sp                                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ cicadidae_sp                                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ apis_mellifera                              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ pteromalidae_sp                             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ chrysopidae_sp                              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ pamphagidae_sp                              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ amphipsocidae_sp                            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ diptera_sp                                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ hymenoptera_sp                              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ decapotoma_lunata                           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ pyrrhocordae_sp_1                           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ mylabris_oculata                            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```
# You can also view the first six rows
head(sp_comm)
```

```
## # A tibble: 6 x 56
##   provinces    climatic_zones site  season haplotype cleta_eckloni
##   <chr>        <chr>          <chr> <chr>      <dbl>         <dbl>
## 1 Eastern Cape Cfb            EC1   Summer         5            23
## 2 Eastern Cape Cfb            EC1   Winter         5             8
## 3 Eastern Cape Cfa            EC7   Summer         5            11
## 4 Eastern Cape Cfa            EC7   Winter         5             0
## 5 Eastern Cape Bsk            EC8   Summer         5             0
## 6 Eastern Cape Bsk            EC8   Winter         5             0
## # i 50 more variables: pseudambonea_capeni_schuhistes_lekkersingia <dbl>,
## #   acanthocoris_spinosus <dbl>, antestiopsis_thunbergii <dbl>,
## #   cassida_distinguenda <dbl>, epilachna_sp_1 <dbl>, cleta_sp_1 <dbl>,
## #   cleta_sp_2 <dbl>, exochomus_flavipes <dbl>, scymnus_sp <dbl>,
## #   cheilomenes_lunata <dbl>, cheilomenes_sulphurea <dbl>, cf_nephus_sp <dbl>,
## #   chnootriba_sp <dbl>, oenopia_cinctella <dbl>, hippodamia_variegate <dbl>,
## #   cassida_melanophthalma <dbl>, cassida_reticulipennis <dbl>, ...
```

Note the use of the pipe operator above (%>%). This comes from the **magrittr** package, and it allows for a series of functions to be applied to an object. Much like pushing something through a pipe, where it gets modified along its path. For example, if we create a variable called "var", and we want to first assign it a value of 1, then add 5 to it, and then square root it, we could write this as:

```
var = 1 %>%
+ 5 %>%
sqrt()

var
```

```
## [1] 2.44949
```

We can now run a species accumulation curve (SAC) analysis. We first need to remove the first five columns of the dataset, so that we are left with only the species abundance values. Once this has been done, we can use the **poolaccum()** function from the **vegan** package:

```
sac_raw <- sp_comm %>%
  # Remove site description variables
  dplyr::select(-c(provinces, climatic_zones, site, season, haplotype)) %>%
  # Compute SAC
  vegan::poolaccum()
```

In this output:

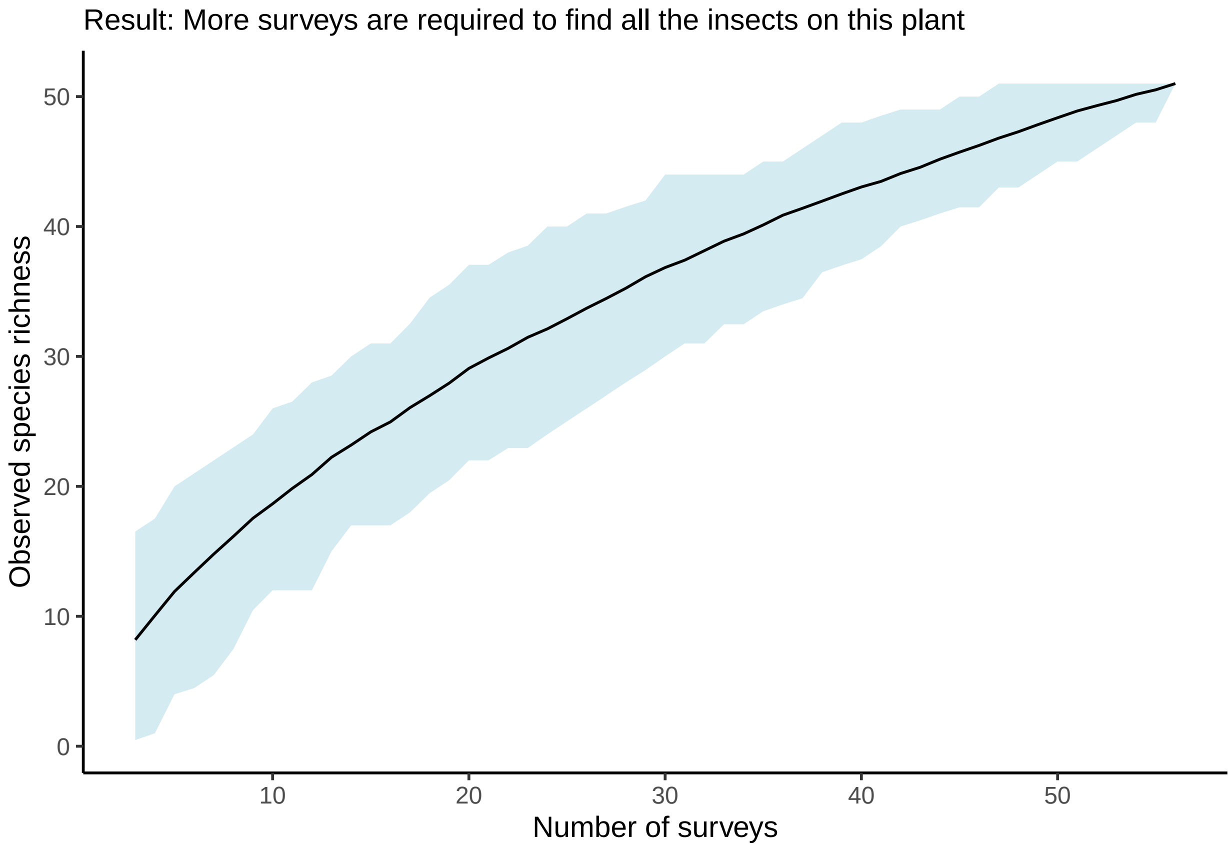
- N = number of surveys (sampling effort)
- S = observed species richness
- lower2.5 = lower 95% confidence interval of S
- upper97.5 - upper 95% confidence interval of S

```
# Extract observed richness (S) estimate
obs <- data.frame(summary(sac_raw)$S, check.names = FALSE)
colnames(obs) <- c("N", "S", "lower2.5", "higher97.5", "std")
head(obs)
```

```
##   N     S lower2.5 higher97.5      std
## 1 3  8.19    0.475     16.525 3.922726
## 2 4 10.06    1.000     17.525 4.352684
## 3 5 11.91    4.000     20.000 4.401779
## 4 6 13.36    4.475     21.000 4.427919
## 5 7 14.79    5.475     22.000 4.325856
## 6 8 16.15    7.475     23.000 4.472983
```

Now we can plot sampling effort (N) against observed species richness (S) using **ggplot**. The geom_ribbon() line adds the confidence intervals as a shaded band to the trend line.

```
ggplot(data = obs, aes(x = N, y = S)) +
  # Add confidence intervals
  geom_ribbon(aes(ymin = lower2.5, ymax = higher97.5), alpha = 0.5, fill = "lightblue") +
  # Add observed richness line
  geom_line() +
  labs(x = "Number of surveys",
       y = "Observed species richness",
       subtitle = "Result: More surveys are required to find all the insects on this plant") +
  theme_classic()
```



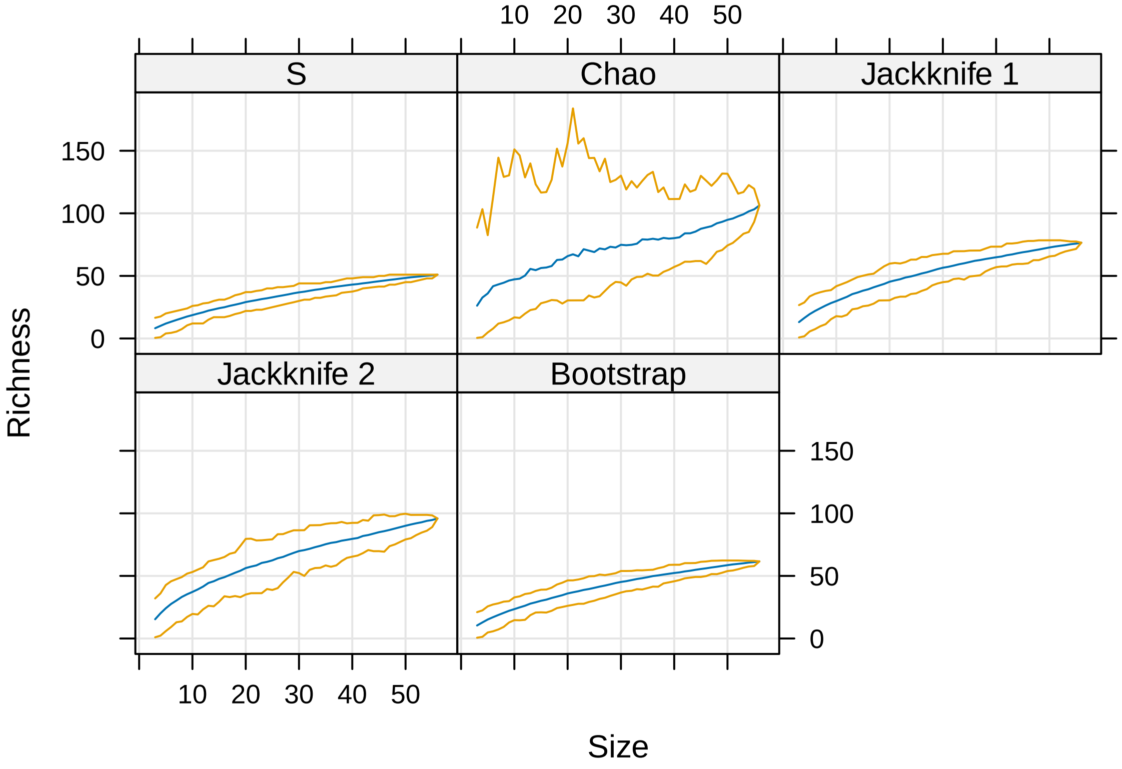
This SAC shows that additional surveys are likely to yield a greater species diversity, since the curve has not reached an asymptote yet.

Extrapolating -> how do our observed species richness values compare to extrapolations?

```
class(sac_raw)
```

```
## [1] "poolaccum"
```

```
plot(sac_raw)
```



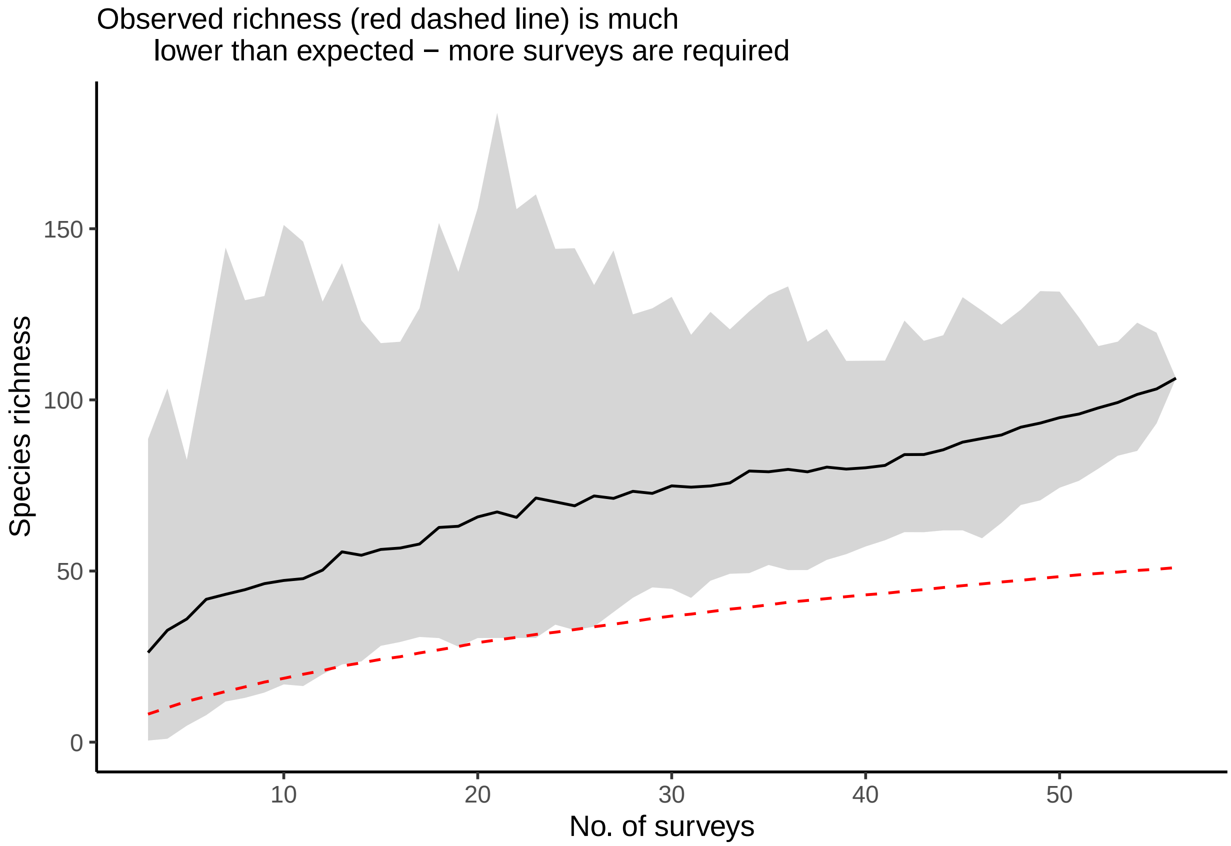
```
# Extract chao -> one of the measures of extrapolated species richness
chao <- data.frame(summary(sac_raw)$chao, check.names = FALSE)
colnames(chao) <- c("N", "S", "lower2.5", "higher97.5", "std")
head(chao)
```

```
##   N        S  lower2.5 higher97.5      std
## 1 3 26.18194  0.475000   88.55833 21.86667
## 2 4 32.68594  1.000000  103.32812 25.15152
## 3 5 35.98275  4.800000   82.52500 23.47098
## 4 6 41.73898  7.884375  112.67083 29.32223
## 5 7 43.18731 11.869643  144.48929 31.84297
## 6 8 44.54512 12.943750  129.12031 27.50060
```

```
# Extract S -> observed spp richness
obs <- data.frame(summary(sac_raw)$S, check.names = FALSE)
colnames(obs) <- c("N", "S", "lower2.5", "higher97.5", "std")
head(obs)
```

```
##   N     S lower2.5 higher97.5      std
## 1 3  8.19    0.475     16.525 3.922726
## 2 4 10.06    1.000     17.525 4.352684
## 3 5 11.91    4.000     20.000 4.401779
## 4 6 13.36    4.475     21.000 4.427919
## 5 7 14.79    5.475     22.000 4.325856
## 6 8 16.15    7.475     23.000 4.472983
```

```
chao %>%
  ggplot(data = ., aes(x = N, y = S)) +
  # Add confidence intervals
  geom_ribbon(aes(ymin = lower2.5, ymax = higher97.5), alpha = 0.2) +
  geom_line() +
  # Add S richness -> our observed spp richness
  geom_line(data= obs, aes(x = N, S), linetype = "dashed", col = "red") +
  labs(x = "No. of surveys", y = "Species richness",
       subtitle = "Observed richness (red dashed line) is much
       lower than expected - more surveys are required") +
  theme_classic()
```



Species diversity

What about species diversity across provinces, climatic zone, and season? Let's calculate species numbers:

```
# subset the data, so that it contains only abundance values, not grouping variables
sp_num_input = sp_comm %>%
  # Remove site description variables
  dplyr::select(-c(provinces, climatic_zones, site, season, haplotype))

# get species numbers by:

# per row/sampling event
overall_sp_num = vegan::specnumber(sp_num_input)

# climatic zones
clim_zones = vegan::specnumber(sp_num_input, group = sp_comm$climatic_zones)

# provinces
provs = vegan::specnumber(sp_num_input, group = sp_comm$provinces)

# seasons
seasons = vegan::specnumber(sp_num_input, group = sp_comm$season)
```

We can now create plots for each of these variables, starting with province.

```
# create a new dataframe containing the species numbers with additional information
overall_sp_num_df = overall_sp_num %>%
                    as.data.frame() %>%
                    dplyr::mutate(province = sp_comm$provinces,
                                  clim = sp_comm$climatic_zones,
                                  season = sp_comm$season)

colnames(overall_sp_num_df) = c("sp_number", "province", "clim", "season")

# create factors
overall_sp_num_df$province = as.factor(overall_sp_num_df$province)
overall_sp_num_df$clim = as.factor(overall_sp_num_df$clim)
overall_sp_num_df$season = as.factor(overall_sp_num_df$season)

head(overall_sp_num_df)
```

```
##   sp_number     province clim season
## 1         4 Eastern Cape  Cfb Summer
## 2         5 Eastern Cape  Cfb Winter
## 3         1 Eastern Cape  Cfa Summer
## 4         0 Eastern Cape  Cfa Winter
## 5         3 Eastern Cape  Bsk Summer
## 6         0 Eastern Cape  Bsk Winter
```

```
# Let's quickly run some analysis of variance tests to have a quick look at whether
# there are differences across provinces, climates, and seasons

# is there a difference in species numbers across provinces?
summary(aov(sp_number ~ province, data = overall_sp_num_df) )
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## province     1    0.7   0.656   0.096  0.757
## Residuals   54  367.5   6.805
```

```r
# between climatic zones?
summary(aov(sp_number ~ clim, data = overall_sp_num_df) )
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## clim         5   42.4   8.485   1.303  0.278
## Residuals   50  325.7   6.514
```

```r
# seasons?
summary(aov(sp_number ~ season, data = overall_sp_num_df) )
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## season       1    4.0   4.018   0.596  0.444
## Residuals   54  364.1   6.743
```

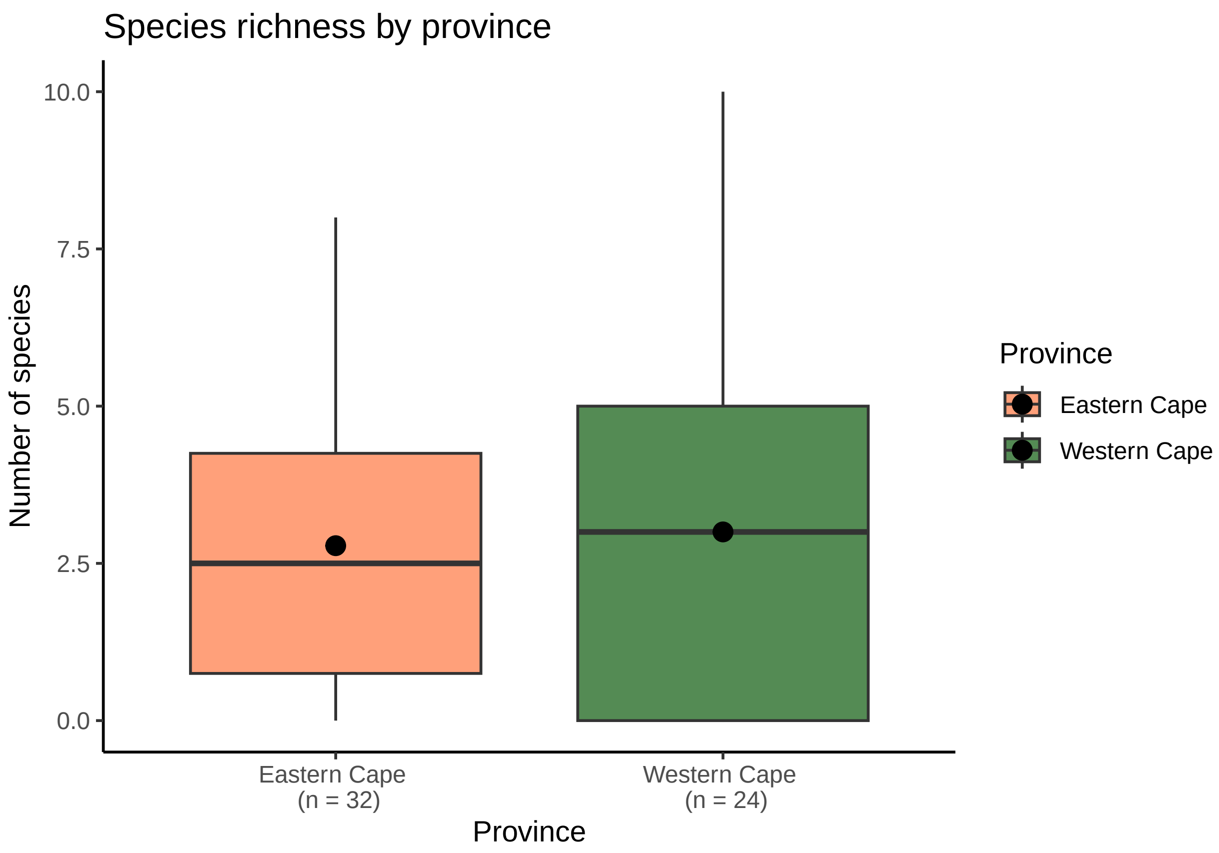
```r
# Let's get a quick stats summary across provinces:
Rmisc::summarySE(data = overall_sp_num_df, measurevar = "sp_number",
                                 groupvars = "province")
```

```
##       province  N sp_number       sd        se        ci
## 1 Eastern Cape 32   2.78125 2.324267 0.4108763 0.8379877
## 2 Western Cape 24   3.00000 2.948839 0.6019293 1.2451856
```

```r
# Choose some colours for each province. Remember that R colours groups alphabetically, so the Eastern
pal.prov <- c("lightsalmon1", "palegreen4")

plot_by_province <- ggplot2::ggplot(overall_sp_num_df, aes(x = province,
                                                           y = sp_number,
                                                           fill = province)) +
  geom_boxplot() +
  scale_fill_manual(values = pal.prov) +
  scale_x_discrete(labels = c("Eastern Cape \n (n = 32)", "Western Cape \n (n = 24)")) +
  # add black circles for means
  stat_summary(fun = mean, geom = "point", color = "black", size = 3) +
  labs(x = "Province",
       y = "Number of species",
       title = "Species richness by province",
       # change the legend title
       fill = "Province") +
  theme_classic()

plot_by_province
```



Let’s do the same for climatic zones:

```
# Let's get a quick stats summary across climates:
Rmisc::summarySE(data = overall_sp_num_df, measurevar = "sp_number",
                                  groupvars = "clim")
```

```
##   clim  N sp_number       sd        se        ci
## 1  Bsh 12  1.916667 2.193309 0.6331539  1.393562
## 2  Bsk  2  1.500000 2.121320 1.5000000 19.059307
## 3  Cfa 14  2.642857 2.817723 0.7530680  1.626905
## 4  Cfb 10  3.100000 1.595131 0.5044249  1.141088
## 5  Csa  6  2.500000 2.810694 1.1474610  2.949642
## 6  Csb 12  4.333333 3.055050 0.8819171  1.941086
```

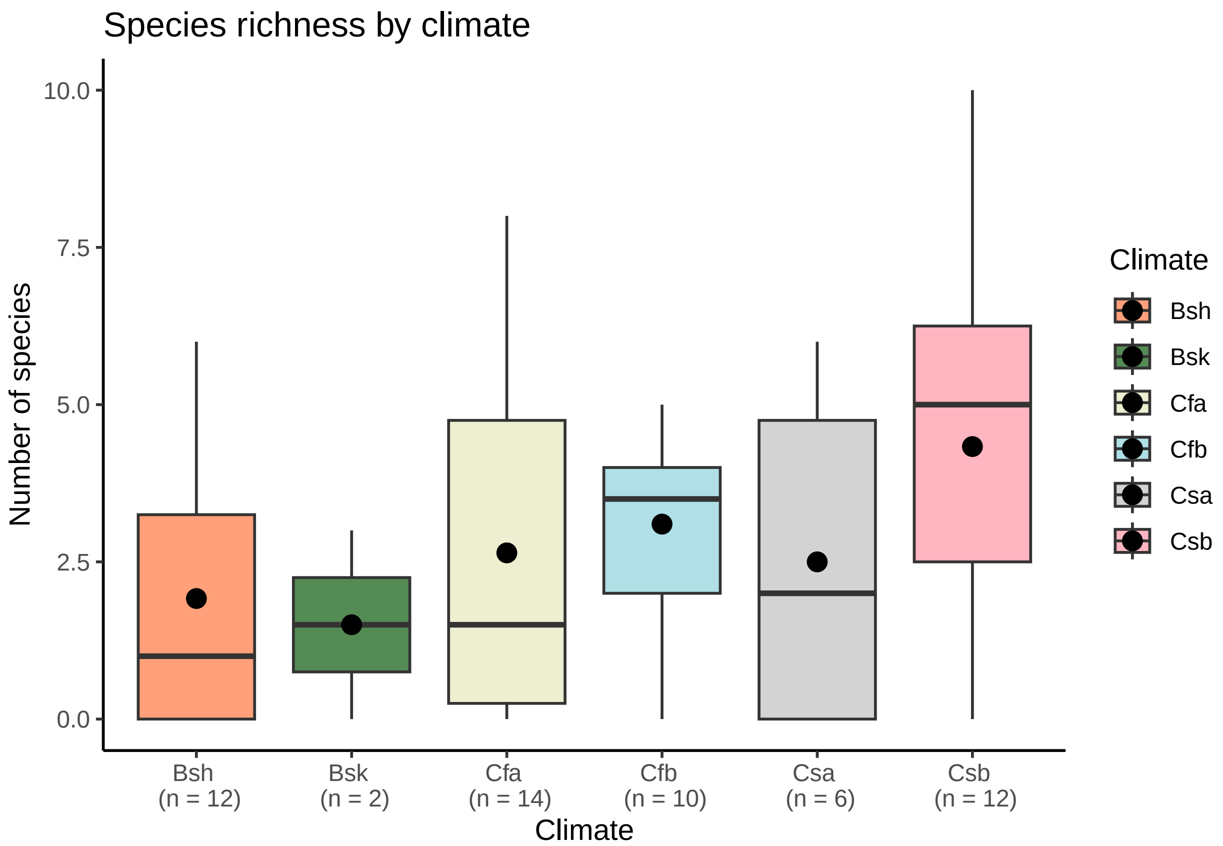
```
# check how many groups there are in clim
levels(overall_sp_num_df$clim)
```

```
## [1] "Bsh" "Bsk" "Cfa" "Cfb" "Csa" "Csb"
```

```
# Choose some colours for each climate type
pal.clim <- c("lightsalmon1", "palegreen4", "lightyellow2",
         "powderblue", "lightgrey", "lightpink")

plot_by_clim <- ggplot2::ggplot(overall_sp_num_df, aes(x = clim,
                                                       y = sp_number,
                                                           fill = clim)) +
  geom_boxplot() +
  scale_fill_manual(values = pal.clim) +
  scale_x_discrete(labels = c("Bsh \n (n = 12)", "Bsk \n (n = 2)",
                              "Cfa \n (n = 14)", "Cfb \n (n = 10)",
                              "Csa \n (n = 6)", "Csb \n (n = 12)")) +
  # add black circles for means
  stat_summary(fun = mean, geom = "point", color = "black", size = 3) +
  labs(x = "Climate",
       y = "Number of species",
       title = "Species richness by climate",
       # change the legend title
       fill = "Climate") +
  theme_classic()

plot_by_clim
```



Try to plot the same as above, but for season.

Diversity Indices

Let's calculate Shannon diversity index, using the original species count data:

```
shan.div = vegan::diversity(sp_num_input, index = "shannon")
head(shan.div)
```

```
## [1] 0.6893115 1.2766530 0.0000000 0.0000000 1.0986123 0.0000000
```

```
# create a new dataframe containing the diversity indices with additional information
shan.div.df = shan.div %>%
              as.data.frame() %>%
              dplyr::mutate(province = sp_comm$provinces,
                            clim = sp_comm$climatic_zones,
                            season = sp_comm$season)

colnames(shan.div.df) = c("shannon", "province", "clim", "season")

# create factors
shan.div.df$province = as.factor(shan.div.df$province)
shan.div.df$clim = as.factor(shan.div.df$clim)
shan.div.df$season = as.factor(shan.div.df$season)

# is there a difference in Shannon diversity across provinces?
summary(aov(shannon ~ province, data = shan.div.df) )
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## province     1   0.01  0.0101   0.024  0.878
## Residuals   54  23.16  0.4289
```

```
# between climatic zones?
summary(aov(shannon ~ clim, data = shan.div.df) )
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## clim         5  1.993  0.3987   0.941  0.463
## Residuals   50 21.176  0.4235
```

```
# seasons?
summary(aov(shannon ~ season, data = shan.div.df) )
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## season       1  0.166  0.1664   0.391  0.535
## Residuals   54 23.003  0.4260
```

Can you create some box plots for the Shannon diversity indices across the different groups?