Exploring Similarities and Differences in Ecosubsections

```r
# Import data
spatial <- read_csv("../data/plot_level/plt_spatial.csv")
response <- read_csv("../data/plot_level/plot_response.csv")

# Join data
## Keep only observations in both `spatial` and `response`
dat <- inner_join(spatial, response,
                  by = c("PLT_CN" = "PLT_CN",
                         "INVYR" = "INVYR"))

# Create columns for province, sections, and subsections
dat <- dat %>%
  mutate(
    subsection = ECOSUBCD.x,
    section = str_remove_all(ECOSUBCD.x, "[:lower:]"),
    province = str_sub(section, end = -2)
  )

# Select small subset of columns to work with for this EDA
dat_small <- dat %>%
  select(PLT_CN, INVYR, PLOT.x, LON_PUBLIC.x, LAT_PUBLIC.x, LON_PUBLIC.y, LAT_PUBLIC.y,
         ELEV_PUBLIC.x, ELEV_PUBLIC.y, forgrp, forprob, nlcd11, demLF, evtLF, forbio,
         BALIVE_TPA, CNTLIVE_TPA, BIOLIVE_TPA, VOLNLIVE_TPA, subsection, section, province)

# Remove redundent columns, rename columns for ease of use
dat_small <- dat_small %>%
  select(-LON_PUBLIC.y, -LAT_PUBLIC.y, -ELEV_PUBLIC.y) %>%
  rename(PLOT = PLOT.x,
         LON_PUBLIC = LON_PUBLIC.x,
         LAT_PUBLIC = LAT_PUBLIC.x,
         ELEV_PUBLIC = ELEV_PUBLIC.x)
```

```r
n_subsections <- dat_small %>%
  group_by(section, subsection) %>%
  summarize(n()) %>%
  group_by(section) %>%
  summarize(number_of_subsections = n())
```

```
## `summarise()` regrouping output by 'section' (override with `.groups` argument)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
head(n_subsections)
```

```
## # A tibble: 6 x 2
##   section number_of_subsections
##   <chr>                   <int>
## 1 313A                       19
## 2 313B                        7
## 3 313C                        4
## 4 313D                        5
## 5 315A                        3
## 6 315B                        4
## Does this make sense?
sum(n_subsections$number_of_subsections)
```

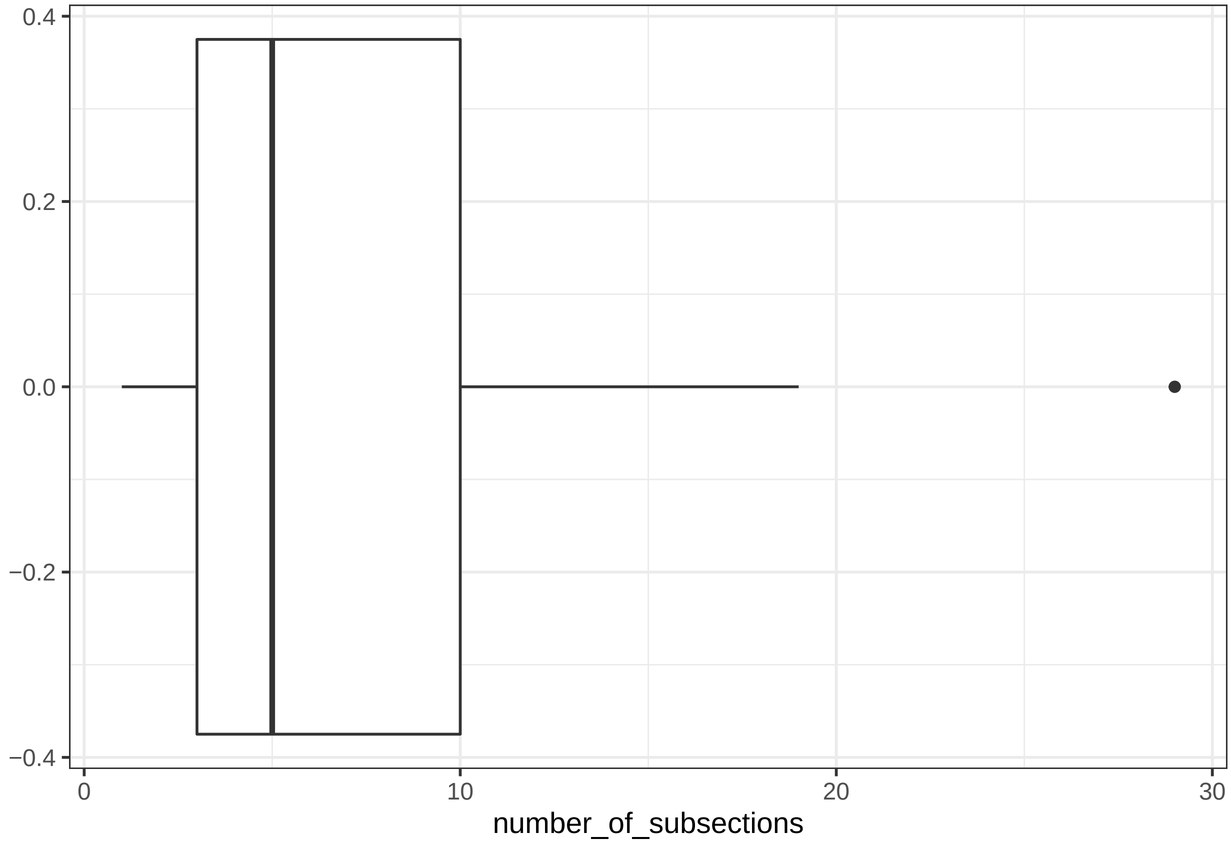
```
## [1] 480
```

```
length(unique(dat_small$subsection))
```

```
## [1] 480
```

```
## Yes!

## Let's look at this distribution:
ggplot(n_subsections, aes(x = number_of_subsections)) +
  geom_boxplot() +
  theme_bw()
```



```
mean(n_subsections$number_of_subsections)
```

```
## [1] 7.058824
```

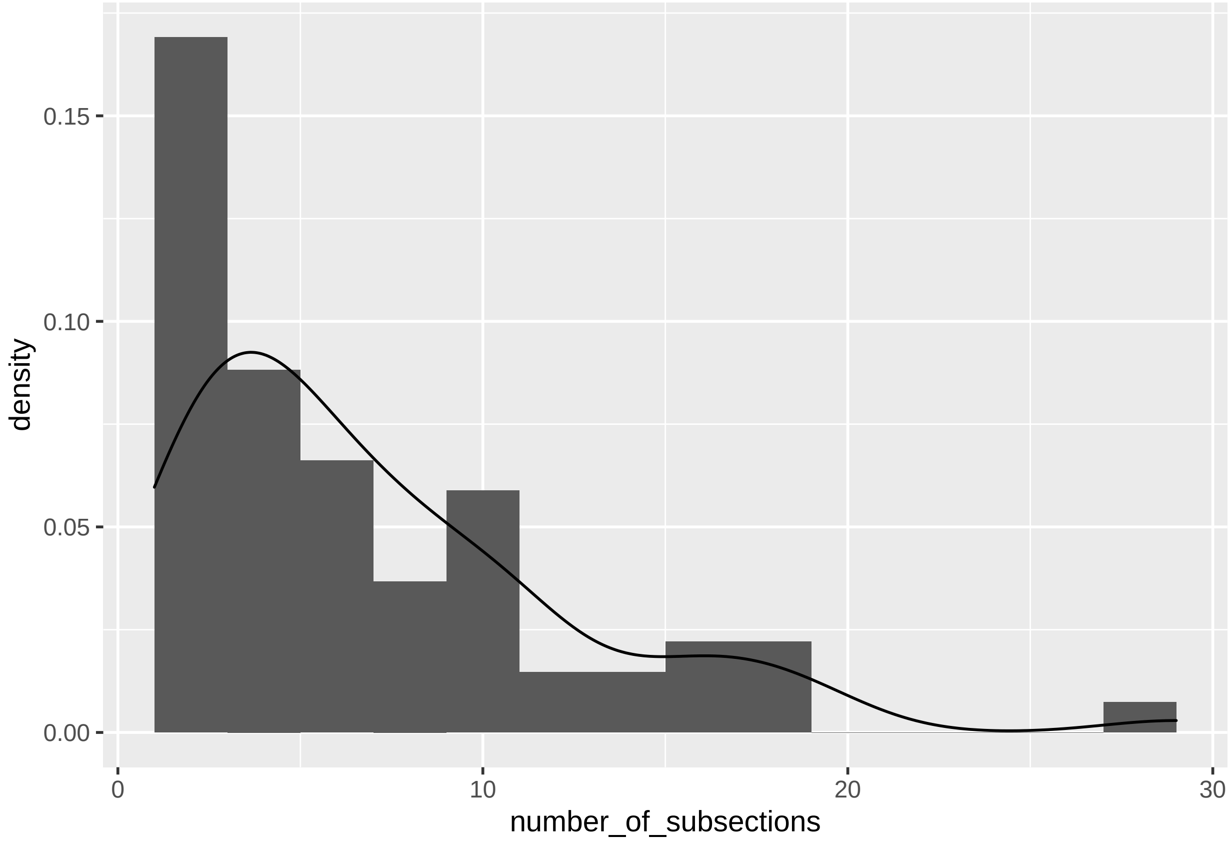
```
sd(n_subsections$number_of_subsections)
```

```
## [1] 5.5068
```

```
median(n_subsections$number_of_subsections)
```

```
## [1] 5
```

```
ggplot(n_subsections, aes(x = number_of_subsections)) +
  geom_histogram(bins = 15, aes(y = ..density..)) +
  geom_density()
```



```
## What about sections in provinces?
n_sections <- dat_small %>%
  group_by(province, section) %>%
  summarize(n()) %>%
  group_by(province) %>%
  summarize(number_of_sections = n())
```

```
## `summarise()` regrouping output by 'province' (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(n_sections)
```

```
## # A tibble: 6 x 2
##   province number_of_sections
##   <chr>                 <int>
## 1 313                       4
## 2 315                       3
## 3 321                       1
## 4 322                       3
## 5 331                      14
## 6 341                       7
```

```
mean(n_sections$number_of_sections)
```

```
## [1] 4.857143
```

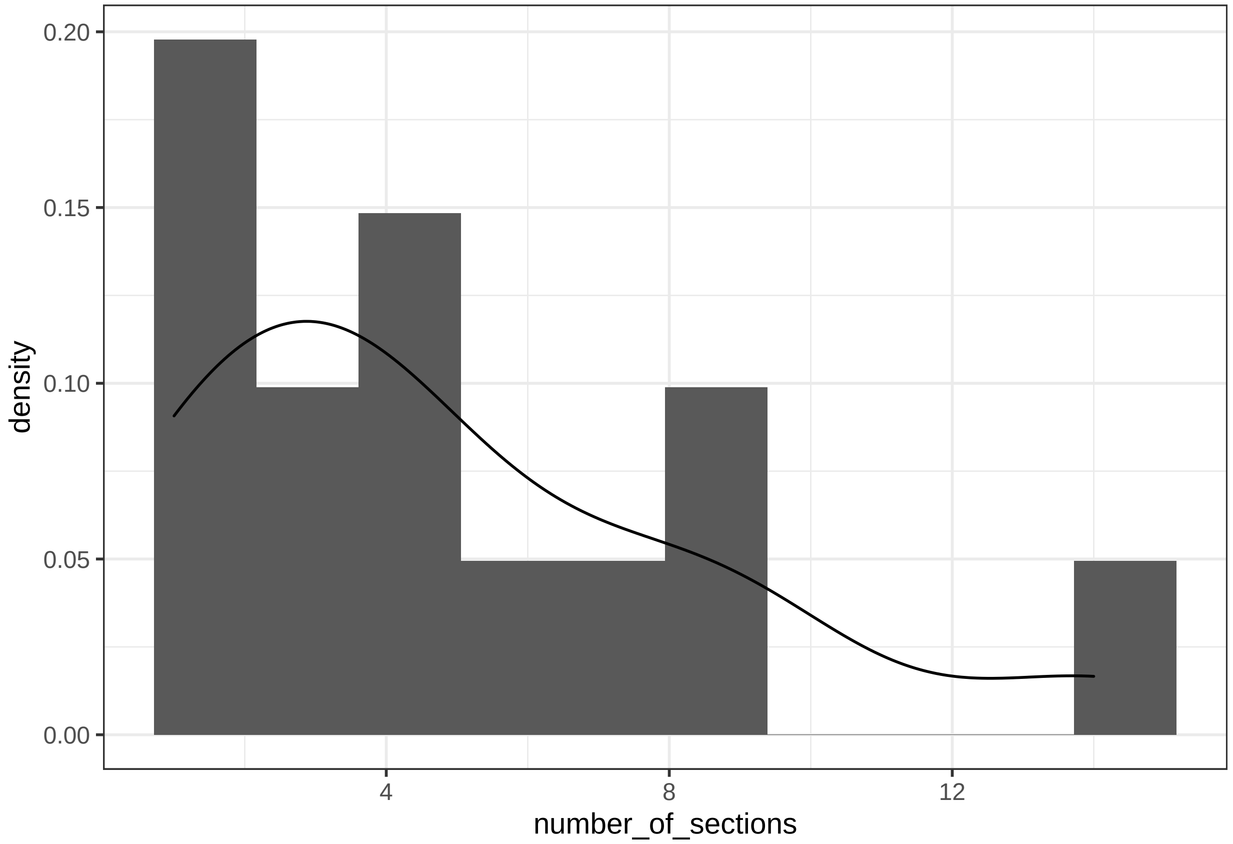
```
median(n_sections$number_of_sections)
```

```
## [1] 4
```

```
sd(n_sections$number_of_sections)
```

```
## [1] 3.779645
```

```
ggplot(n_sections, aes(x = number_of_sections)) +
  geom_histogram(bins = 10, aes(y = ..density..)) +
  geom_density() +
  theme_bw()
```



From this, we see that the average number of subsections in a section is about 7, with a median of 5 giving us a right-skewed distribution. We have a similar right-skewed distribution with numbers of sections in provinces, with a smaller mean and median. This makes me think we may be able to take info from subsection up to province, however we must first look at how similar or different attributes are in each subsection/section/province.

Aside: There is one outlying section which I will investigate now:

```
# M332A: Idaho Batholith, "The batholith section is a large, contiguous uplifted area of granitic pluto
library(concaveman)
library(sf)
```

```
## Linking to GEOS 3.7.2, GDAL 2.4.2, PROJ 5.2.0
```

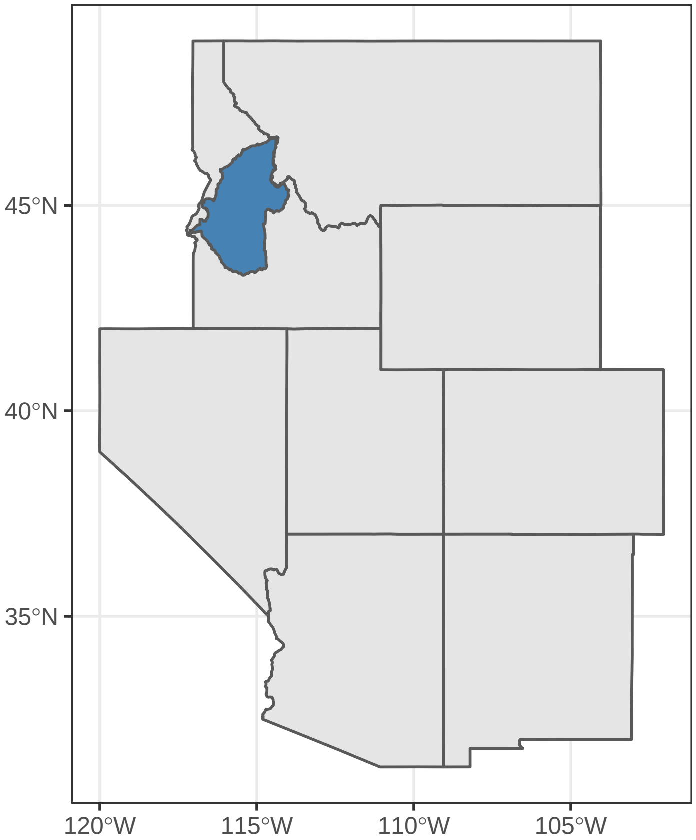
```
library(USAboundaries)
`%ni%` <- Negate(`%in%`)
```

```r
interior_west <- c("AZ", "CO", "ID", "MT", "NV", "NM", "UT", "WY")

states <- data.frame(state.abb) %>%
  filter(state.abb %ni% interior_west &
           state.abb %ni% c("AK", "HI")) %>% pull()


m332a_poly <- dat_small %>%
  filter(section == "M332A") %>%
  st_as_sf(coords = c("LON_PUBLIC", "LAT_PUBLIC"), crs = 4326) %>%
  concaveman()

ggplot() +
  geom_sf(data = us_boundaries(type = "state",
                               states = interior_west)) +
  geom_sf(data = m332a_poly,
          fill = "steelblue") +
  theme_bw()
```



```r
# Okay, so it is a big section, but is it way bigger than others?
# interior_west_sf <- dat_small %>%
#   st_as_sf(coords = c("LON_PUBLIC", "LAT_PUBLIC"),
#            crs = 4326) %>%
#   concaveman()
#
# total_area <- st_area(interior_west_sf)
# m332a_area <- st_area(m332a_poly)
#
# m332a_area / total_area # This is the proportion of total area m332a takes up
#
#
# 1 / length(unique(dat_small$section)) # This is the proportion of total area an "average" section wou

# M332A takes up more area than average but not *way* more. This means that it likely has some small su

### this doesn't seem right based on the picture. revisit this with fresh eyes tomorrow.
```

This section has many subsections (29), but based on the map we can tell that it is a very large section.

Quantifying Homogeneity in Ecosubsections

```
means <- dat_small %>%
  group_by(subsection, section) %>%
  summarize(avg_balive = mean(BALIVE_TPA),
            avg_cntlive = mean(CNTLIVE_TPA),
            avg_biolive = mean(BIOLIVE_TPA),
            avg_volnlive = mean(VOLNLIVE_TPA),
            avg_forbio = mean(forbio))
```

```
## `summarise()` regrouping output by 'subsection' (override with `.groups` argument)
```

```
# Is the variation of means within sections less variable than the variation of means overall?

overall_balive <- sd(means$avg_balive, na.rm = TRUE)
overall_cntlive <- sd(means$avg_cntlive, na.rm = TRUE)
overall_biolive <- sd(means$avg_biolive, na.rm = TRUE)
overall_volnlive <- sd(means$avg_volnlive, na.rm = TRUE)
overall_forbio <- sd(means$avg_forbio, na.rm = TRUE)

meta_sd <- means %>%
  group_by(section) %>%
  summarize(
    sd_mean_balive = sd(avg_balive, na.rm = TRUE),
    sd_mean_cntlive = sd(avg_cntlive, na.rm = TRUE),
    sd_mean_biolive = sd(avg_biolive, na.rm = TRUE),
    sd_mean_volnlive = sd(avg_volnlive, na.rm = TRUE),
    sd_mean_forbio = sd(avg_forbio, na.rm = TRUE)
  ) %>%
  summarize(
    mean(sd_mean_balive, na.rm = TRUE),
    mean(sd_mean_cntlive, na.rm = TRUE),
    mean(sd_mean_biolive, na.rm = TRUE),
    mean(sd_mean_volnlive, na.rm = TRUE),
    mean(sd_mean_forbio, na.rm = TRUE)
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
comparison <- tibble(
  overall_sd_of_mean = c(overall_balive, overall_cntlive, overall_biolive, overall_volnlive, overall_fo
  sd_of_mean_by_section = c(pull(meta_sd[,1]), pull(meta_sd[,2]), pull(meta_sd[,3]),
                            pull(meta_sd[,4]), pull(meta_sd[,5])),
  variable_name = c("balive", "cntlive", "biolive", "volnlive", "forbio")
)
```

Holy hell. Okay, so we first took the mean of variables by subsection. That is the `means` dataframe. Then we took the standard deviation of the means of the subsections. So, this is looking at how different the means are from each other throughout each subsection.

Now, this is where it gets crazy. We group by section, and then take the standard deviation of the means of each section. Finally we take the mean of that standard deviation to give us something like the "average standard deviation of the means within a section." My god. There has to be a better way to do this. Or a less complicated way to explain what has been done. Or a simpler statistic.