**INITIAL OBSERVATIONS YEAR**

## ###PACKAGES USED##

install.packages("ggplot2")

install.packages("esquisse")

install.packages("shiny")

install.packages ("dplyr")

library("esquisse")

library("ggplot2")

library("shiny")

**library("dplyr")**

# ⚠ TASKS TO DO FROM NOW ON

## STEP #1 GRPH

Create a plot for each number/case combinations.

First, create a table of frequencies based on a +2 occurrence subset.

### #these create the data frames to draw the dot plot

genplu.gph1 <- YrGenPlu.wide %>% filter(tot.occ >= 2) %>%

group\_by(tot.occ, tot.form) %>%

summarise(n = n())

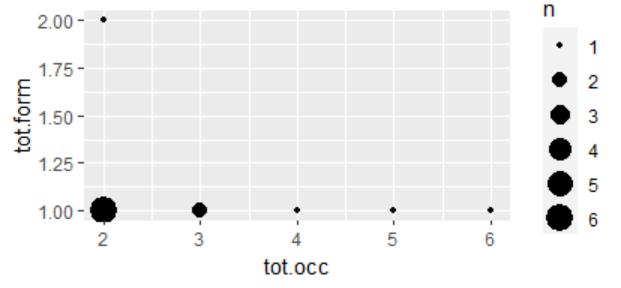
ggplot(data = genplu.gph1, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

Afterwards, create a graph of scattered plots where size is adjusted for frequency. We want to answer two questions:  
 a) How many people have used more than 2 occurrences for each parameter?

b) Which percentage of the +2 occurrences (i.e. use) use +1 form?

# NOM SG

Only one person seems to have used more than one form, so this dataset is not interesting. 

## Code

# Filter the data and assign it to a new dataframe

nomsg.gph1 <- YrNomSin.wide %>%

filter(tot.occ >= 2)

# Group by variables and calculate the count

nomsg.gph2 <- nomsg.gph1 %>%

group\_by(tot.occ, tot.form) %>%

tally(name = "n")

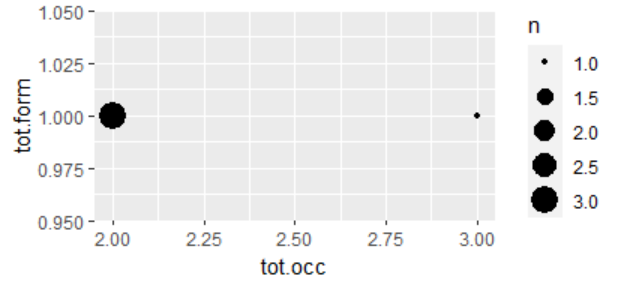
# Plot the data

ggplot(data = nomsg.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# NOM PL

This form is not interesting at all. Not many have used it more than twice and in any case, everybody has been consistent with their chosen form.



## Code

YRnompl.gph1 <- YrNomPlu.wide %>%

filter(tot.occ >= 2)

YRnompl.gph2 <- YRnompl.gph1 %>%

group\_by(tot.occ, tot.form) %>%

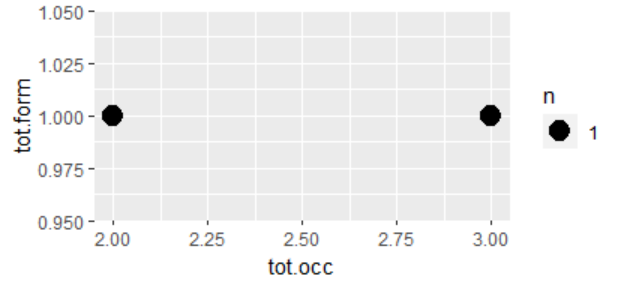
tally(name = "n")

ggplot(data = YRnompl.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# ACC.SG

This form is not very interesting. Nobody who has used more than two ACC.SG-s has used a different form.



## Code

YRaccsg.gph1 <- YrAccSin.wide %>%

filter(tot.occ >= 2)

YRaccsg.gph2 <- YRaccsg.gph1 %>%

group\_by(tot.occ, tot.form) %>%

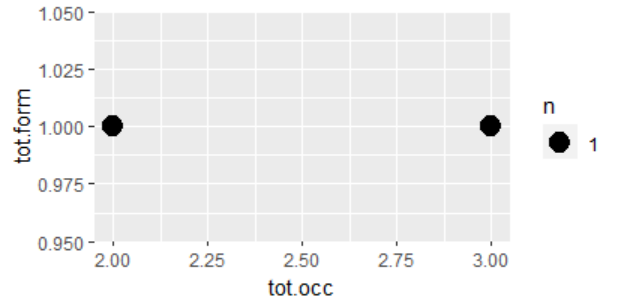
tally(name = "n")

ggplot(data = YRaccsg.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# ACC.PL

The data for this parameter are also uninteresting. People have used different forms, but speakers adhere to one single form. Therefore, we shouldn’t study it further.



## Code

YRAccPlu.gph1 <- YrAccPlu.wide %>%

filter(tot.occ >= 2)

YRAccPlu.gph2 <- YRAccPlu.gph1 %>%

group\_by(tot.occ, tot.form) %>%

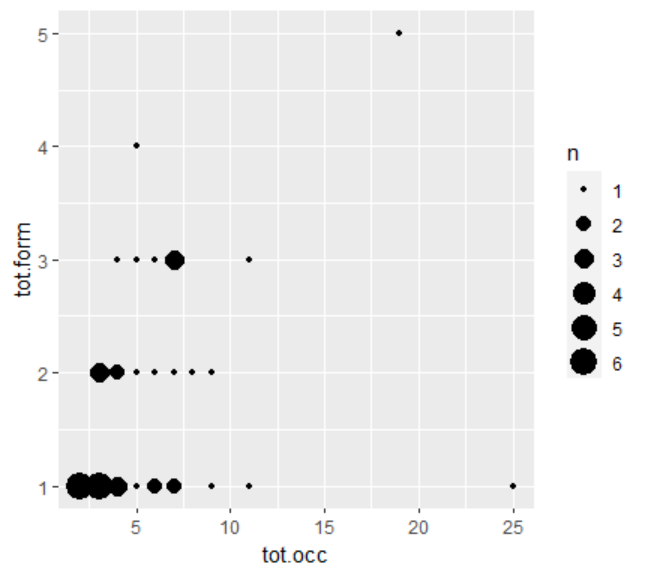
tally(name = "n")

ggplot(data = YRAccPlu.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# GENITIVE PL

Step #1 Create data frames



## Code

genplu.gph1 <- YrGenPlu.wide %>% filter(tot.occ >= 2) %>%

group\_by(tot.occ, tot.form) %>%

summarise(n = n())

ggplot(data = genplu.gph1, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

## Percentage

In order to ask about the percentage of individuals using the GEN PL more than once, and who also use more than one form, this is the code I have used. The answer is 45.24 %

### Code

# Filter for individuals with 2 or more occurrences

genplu.gph1 <- YrGenPlu.wide %>%

filter(tot.occ >= 2)

# Calculate the number of individuals with each combination of occurrences and forms

genplu.gph2 <- genplu.gph1 %>%

group\_by(tot.occ, tot.form) %>%

summarise(n = n())

# Calculate the percentage of individuals with 2 or more forms out of all individuals with 2 or more occurrences

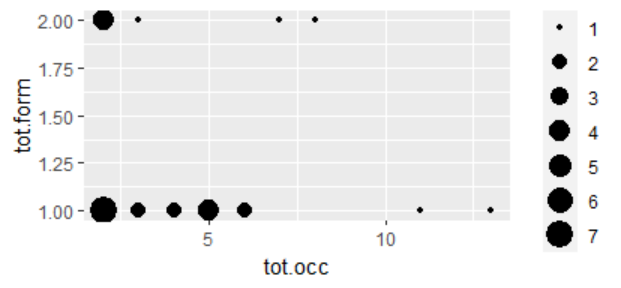
perc <- sum(genplu.gph2[genplu.gph2$tot.form >= 2, "n"]) / sum(genplu.gph2$n) \* 100

# Print the percentage

cat("Percentage of individuals with 2 or more forms out of all individuals with 2 or more occurrences:", round(perc, 2), "%\n")

# GEN.SG

In this dataset, people have been a bit less coherent than with others, although, not as interesting. In any case, after running the plot, I want to calculate the percentage of people using more than one form for the gen.sg.



## Code

YRgensg.gph1 <- YrGenSin.wide %>%

filter(tot.occ >= 2)

YRgensg.gph2 <- YRgensg.gph1 %>%

group\_by(tot.occ, tot.form) %>%

tally(name = "n")

ggplot(data = YRgensg.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

## Percentage

The percentage of people using more than 2 forms, who have used more than 2 gen.sg, is **26.92 %**

### Code

# Calculate the percentage of individuals with 2 or more forms out of all individuals with 2 or more occurrences

perc <- sum(YRgensg.gph2[YRgensg.gph2$tot.form >= 2, "n"]) / sum(YRgensg.gph2$n) \* 100

# Print the percentage

cat("Percentage of individuals with 2 or more forms out of all individuals with 2 or more occurrences:", round(perc, 2), "%\n")

# INS.PL

I have tried running a plot with the subset but it comes as empty. I have checked the data (especially, the long table) and there are very few instances of ins.pl, therefore, I assume nobody has used it more than once.

## Code

YRInsPl.gph1 <- YrINSPlu.wide %>%

filter(tot.occ >= 2)

YRInsPl.gph2 <- YRInsPl.gph1 %>%

group\_by(tot.occ, tot.form) %>%

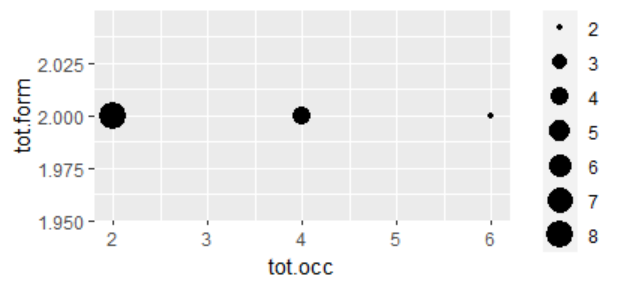
tally(name = "n")

ggplot(data = YRInsPl.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# ADNM

There are not many people using it, but all of them seem to have used two different forms, at least.



## Code

YRadnm.gph1 <- YrADMN.wide %>%

filter(tot.occ >= 2)

YRadnm.gph2 <- YRadnm.gph1 %>%

group\_by(tot.occ, tot.form) %>%

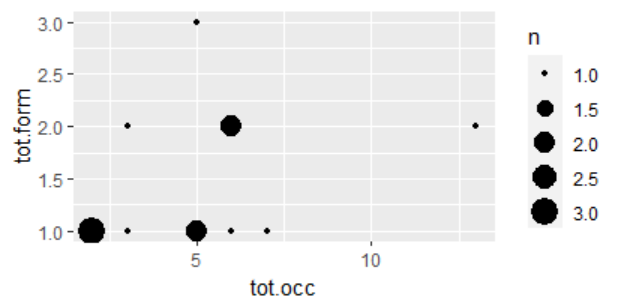
tally(name = "n")

ggplot(data = YRadnm.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# LOC.SG

This parameter has up to 8 different forms, so we would expect diverse and “beautiful” results, which are confirmed by the graphs. Still many people are coherent with the form they have always used, but there is more room for variation. However, only one person seems to have used 3 different forms, the rest have only used 2.



## Code

YRLocSin.gph1 <- YrLocSin.wide %>%

filter(tot.occ >= 2)

YRLocSin.gph2 <- YRLocSin.gph1 %>%

group\_by(tot.occ, tot.form) %>%

tally(name = "n")

ggplot(data = YRLocSin.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

## Percentage

Now, we want to know the percentage of people who have used more than one form (out of those producing at least 2 loc.sg), which, according to R, is 38.46 %

### Code

# Calculate the percentage of individuals with 2 or more forms out of all individuals with 2 or more occurrences

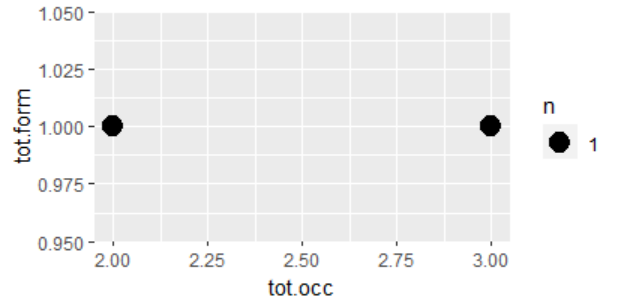
perc <- sum(YRLocSin.gph2[YRLocSin.gph2$tot.form >= 2, "n"]) / sum(YRLocSin.gph2$n) \* 100

# Print the percentage

cat("Percentage of individuals with 2 or more forms out of all individuals with 2 or more occurrences:", round(perc, 2), "%\n")

# LOC.PL

This parameter is also not interesting. People have been coherent with the forms they have used.



## Code

YRLocPlu.gph1 <- YrLocPlu.wide %>%

filter(tot.occ >= 2)

YRLocPlu.gph2 <- YRLocPlu.gph1 %>%

group\_by(tot.occ, tot.form) %>%

tally(name = "n")

ggplot(data = YRLocPlu.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# STEP #2

We will be using multilevel logistic regressions (instead of Poisson) to see which form is used in each dependent variable.

For that, we will use the “long” datasets, but given that the presence seems a bit problematic, we will first create a subset which excludes all the false values for presence. ⚠ IMPORTANT, save the subset as a new data frame!

Then, we will go to the “Esquisse” package and using “esquisser” we open the modified data frames and use the following values for the arguments:

x= form; facet= “Village”

We will replace the facet argument depending on the dependent we want to study.