**Pivot Tables with dplyr**

*## attach libraries*

library(tidyverse)

library(readxl)

library(here)

library(skimr) *# install.packages('skimr')*

library(kableExtra) *# install.packages('kableExtra')*

*## read in data*

lobsters <- read\_xlsx(here("data/lobsters.xlsx"), skip=4) – to skip the first 4 line in the excel file

To look at summary statistics we’ve used summary, which is good for numeric columns, but it doesn’t give a lot of useful information for non-numeric data. So it means it wouldn’t tell us how many unique sites there are in this dataset. To have a look there I like using the skimr package:

*# explore data*

skimr::skim(lobsters)

This skimr:: notation is a reminder to me that skim is from the skimr package. It is a nice convention: it’s a reminder to others (especially you!).

skim lets us look more at each variable. Here we can look at our character variables and see that there are 5 unique sites (in the n\_unique output). Also, I particularly like looking at missing data. There are 6 missing values in the size\_mm variable.

## **group\_by() %>% summarize()**

In R, we can create the functionality of pivot tables with the same logic: we will tell R to group by something and then summarize by something. Visually, it looks like this:

data %>%

group\_by() %>%

summarize()

### **group\_by one variable**

Let’s use group\_by() %>% summarize() with our lobsters data, just like we did in Excel. We will first group\_by year and then summarize by count, using the function n() (in the dplyr package). n() counts the number of times an observation shows up, and since this is uncounted data, this will count each row.

We can say this out loud while we write it: “take the lobsters data and then group\_by year and then summarize by count in a new column we’ll call count\_by\_year.”

lobsters %>%

group\_by(year) %>%

summarize(count\_by\_year = n())

### group\_by multiple variables

Great. Now let’s summarize by both year and site like we did in the pivot table. We are able to group\_by more than one variable. Let’s do this together:

lobsters %>%

group\_by(site, year) %>%

summarize(count\_by\_siteyear = n())

We put the site first because that is what we want as an end product. But we could easily have put year first. We saw visually what would happen when we did this in the Pivot Table.

### summarize multiple variables

We can summarize multiple variables at a time.

So far we’ve summarized the count of lobster observations. Let’s also calculate the mean and standard deviation. First let’s use the mean() function to calculate the mean. We do this within the same summarize() function, but we can add a new line to make it easier to read. Notice how when you put your curser within the parenthesis and hit return, the indentation will automatically align.

lobsters %>%

group\_by(site, year) %>%

summarize(count\_by\_siteyear = n(),

mean\_size\_mm = mean(size\_mm, na.rm=TRUE),

sd\_size\_mm = sd(size\_mm, na.rm=TRUE))

### Table formatting with kable()

There are several options for formatting tables in RMarkdown; we’ll show one here from the kableExtra package and learn more about it tomorrow.

It works nicely with the pipe operator, so we can build do this from our new object:

*## make a table with our new variable*

siteyear\_summary %>%

kable()

Write this **in Markdown** but replace the # with a backtick (`): “There are #r nrow(lobsters)# total lobsters included in this report.” Let’s knit to see what happens.

I hope you can start to imagine the possibilities. If you wanted to write which year had the most observations, or which site had a decreasing trend, you would be able to.

### Activity

1. Build from our analysis and calculate the median lobster size for each site year. Your calculation will use the size\_mm variable and function to calculate the median (Hint: ?median)
2. create and ggsave() a plot.

Then, save, commit, and push your .Rmd, .html, and .png.

Solution (no peeking):

siteyear\_summary <- lobsters %>%

group\_by(site, year) %>%

summarize(count\_by\_siteyear = n(),

mean\_size\_mm = mean(size\_mm, na.rm = TRUE),

sd\_size\_mm = sd(size\_mm, na.rm = TRUE),

median\_size\_mm = median(size\_mm, na.rm = TRUE))

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

*## a ggplot option:*

ggplot(data = siteyear\_summary, aes(x = year, y = median\_size\_mm, color = site)) +

geom\_line()