**Introduction**

library(FSAdata) # for data

library(ggplot2)

I am **finally** learning ggplot2 for elegant graphics. One of the first plots that I wanted to make was a length frequency histogram. As it turns out, there are a few “tricks” to make the histogram appear as I expect most fisheries folks would want it to appear – primarily, left-inclusive (i.e., 100 would be in the 100-110 bin and not the 90-100 bin). Below are length frequency histograms that I like.

The data I use are lengths of Lake Erie Walleye (*Sander vitreus*) captured during October-November, 2003-2014. These data are available in my FSAdata package .My primary interest is in the tl (total length in mm), sex, and loc variables and I will focus on 2010 (as an example).

data(WalleyeErie2)

WE <- dplyr::filter(WalleyeErie2,year==2010)

**Basic Length Frequency**

Making the histogram begins by identifying the data.frame to use in data= and the tl variable to use for the x-axis as an aes()thetic in ggplot(). The histogram is then constructed with geom\_hist(), which I customize as follows:

* Set the width of the length bins with binwidth=.
* By default the bins are centered on breaks created from binwidth=. The bins can be changed to begin on these breaks by using boundary=. The value that boundary=, which is set to the beginning of a first break, regardless of whether that break is in the data or not. I use boundary=0 so that bins will start on breaks that make sense relative to binwidth= (e.g., 0, 25, 50, 75, etc.).
* Bins are left-exclusive and right-inclusive by default, but including closed="left" will make the bins the desired left-inclusive and right-exclusive.
* The fill color of the bins is set with fill= (I prefer a slight gray).
* The outline color of the bins is set with color= (defaults to the same as fill=; I prefer a dark boundary to make the bins obvious).

The scale\_y\_continuous() and scale\_x\_continuous() are primarily used to provide labels (i.e., names) for the y- and x-axes, respectively. By default, the bins of the histogram will “hover” slightly above the x-axis, which I find annoying. The expand= in scale\_y\_continuous() is used to expand the lower limit of the y-axis by a multiple of 0 (thus, not expand the lower-limit) and expand the upper limit of the y-axis by a multiple of 0.05 (thus, the upper-limit will by 5% higher than the tallest bin so that the top frame of the plot will not touch the tallest bin). Finally, theme\_bw() gives a classic “black-and-white” feel to the plot (rather than the default plot with a gray background).

lenfreq1 <- ggplot(data=WE,aes(x=tl)) +

geom\_histogram(binwidth=25,boundary=0,closed="left",

fill="gray80",color="black") +

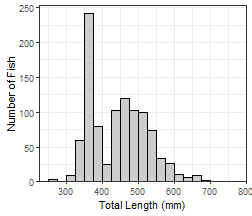
scale\_y\_continuous(name="Number of Fish",expand=expand\_scale(mult=c(0,0.05))) +

scale\_x\_continuous(name="Total Length (mm)") +

theme\_bw()

Note that the resultant plot was assigned to an object. Thus, the object name must be run to see the plot.

lenfreq1



This base object/plot can also be modified by adding (using +) to it as demonstrated later.

**Bins Stacked by Another Variable**

It may be useful to see the distribution of categories of fish (e.g., sex) within the length frequency bins. To do this, move the fill= in geom\_histogram() to an aes()thetic in geom\_histogram() and set it equal to the variable that will identify the separation within each bin (e.g., sex). The bins will be stacked by this variable if position="stacked" in geom\_histogram() (this is the default and would not need to be explicitly set below). The fill colors for each group can be set in a number of ways, but they are set manually below with scale\_fill\_manual().

lenfreq2 <- ggplot(data=WE,aes(x=tl)) +

geom\_histogram(aes(fill=sex),binwidth=25,boundary=0,closed="left",

color="black",position="stack") +

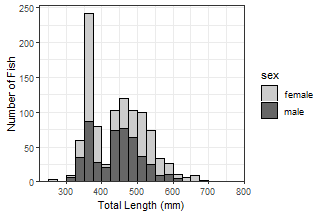
scale\_fill\_manual(values=c("gray80","gray40")) +

scale\_y\_continuous(name="Number of Fish",expand=expand\_scale(mult=c(0,0.05))) +

scale\_x\_continuous(name="Total Length (mm)") +

theme\_bw()

lenfreq2

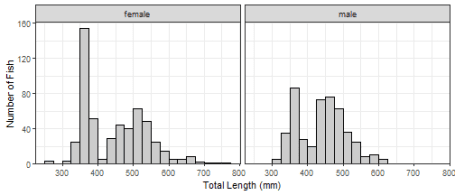


Stacked histograms are difficult to interpret in my opinion. In a future post, I will show how to use empirical density functions to examine distributions among categories. For the time being, see below.

**Separated by Other Variable(s)**

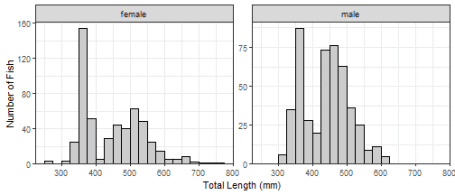
A strength of ggplot2 is that it can easily make the same plot for several different levels of another variable; e.g., separate length frequency histograms by sex. The plot can be separated into different “facets” with facet\_wrap()m which takes the variable to separate by within vars() as the first argument.

lenfreq1 + facet\_wrap(vars(sex))



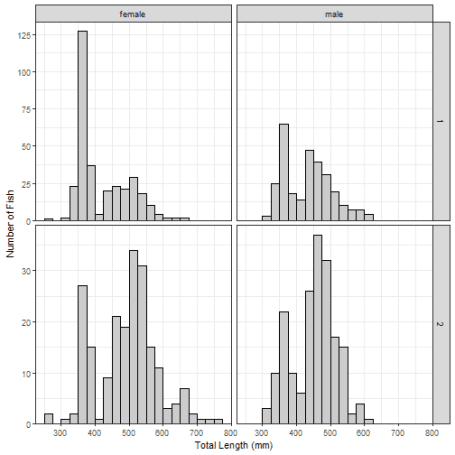
If the faceted groups have very different sample sizes then it may be useful to use a potentially different y-axis scale for each facet by including scales="free\_y" in facet\_wrap(). Similarly, a potentially different scale can be used for each x-axis with scales="free\_x" or for both axes with scales="free".

lenfreq1 + facet\_wrap(vars(sex),scales="free\_y")



Plots may be faceted over multiple variables with facet\_grid(), where the variables that identify the rows and variables for a grid of facets are included (within vars()) in rows= and cols=, respectively. Both scales can not be “free” with facet\_grid() and the scale is only “free” within a row or column.

lenfreq1 + facet\_grid(rows=vars(loc),cols=vars(sex),scales="free\_y")



**Final Thoughts**

This post is likely not news to those of you that are familiar with ggplot2. However, I am going to try to post some examples here as I learn ggplot2 in hopes that hit will help others. This is the first of what I hope will be more frequent posts.