Day 2 Worksheet

Numerical and graphical summary of birth weight data

Today, we will use data from the CDC on births within the US from 2007 to 2015. These data were downloaded from wonder.cdc.gov, where you can go to see these and other similar data. The data are aggregated so that individual babies cannot be identified (for privacy reasons). The data are aggregated by the following factors – State, Year, Gender and Age of Mother. For each combination of those factors, we know how many babies were born, what their mean birth weight and mean gestational age were. Each row of the dataset represents one such combination of factors and the associated summaries of births. There are 29659 rows in this data set, trying to manually go through every row to create summary statistics would be impossible. So we are going to leverage R to summarize and graph these data.

Part 1: Load and Subset the data

Download CDC natality data from my website Day 2 (<http://www.raphswall.com/big-data-in-biology.html>) and put it into your working directory. Like yesterday’s abiotic data, the file you need is a .csv file. Use the ***read.csv ()*** function to read the data. Use the <- to assign your csv to a variable (any variable with more than one value is called a vector in R), you can name this variable whatever is easiest for you to remember. Use the ***head ()*** and ***str ()*** function to begin to acquaint yourself with the data frame. Use the following questions to practice methods to summarize a data set. Below are some commonly used R functions, but I also suggest you have your script from yesterday open for reference, for any relevant action you can copy from your old script to make a new script. Here is a brief summary of some of the R functions you might want to use;

* To understand the structure/column titles of a dataset: ***str ()***, ***head ()***
* Target a specific aspect of the dataset*:* ***subset ()***
* Statistically summarize a column*:* ***mean (), sd (), var (), min (), max (), summary ()***
* Manipulate the dataset: ***table (), aggregate ()***

1. How many boys were born to 26 year old mothers in Arkansas in 2009?
2. Of all female babies born to 30 year old mothers in 2010, which states had average birth weights higher than 3350?
3. In Oregon what was the average (mean) birth rate of females and what was it for males?
4. In Oregon, what was the minimum, mean, and maximum birth weight?
5. In Oregon how many males and females were born? For this question you can use the table () function. Typically, we use the table function to give us an idea of how often something happens, in this case it can count the number of male and female births if you tell R to focus on the gender column.
6. In Texas what was the mean gestation length?
7. In Texas what was the minimum, mean, and maximum birth weight?
8. In Texas how many males and females were born?
9. In Texas what was the mean birth weight for mothers that were 25 years old.
10. Use the table function to calculate how many male and female babies were born in Vermont in 2010?

Part 2: Making Graphs

Making graphs is a key method that scientists use to visually summarize data. Graphing is a critical skill that integrates data summary with visually communicating data in a clear and simple format. Making graphs is a little complex in base R. However, there are multiple packages available for making really nice graphs. One of the best and easiest to use is called ggplot2. We are going to use this package for graphing in this class (most versions of R already have ggplot2 loaded, but if yours does not, you can use the following script to get the ggplot package:

install.packages (“ggplot2”)

library (ggplot2)

The syntax of ggplot2 takes a little getting used to, but the basic idea is that you make a plot in two steps – first defining the data that will be used in the graph and how they will be related to the position, color and shape of any visual elements, and then by defining the visual elements that will be used to represent those data (these are called *geometries*). For example:  
myGraph = ggplot(data = d, aes(x = Species, y = Height))

This says that I’ll be making a graph using data stored in an object called d. The aes() function does what is called “aesthetic mapping”. That means that we can tell it which columns of the data should be used to define which aspects of the graph. Here, I am telling it that the column “Species” should be arrayed along the x axis and “Height” along the y. You can also map variables to colors, symbols and point sizes.

That doesn’t actually make a graph, though. To make a graph, you need to add on geometries – how these data will be visualized. Here are a few different geometries that you could use to make your graphs in R:

myGraph + geom\_point()

myGraph + geom\_jitter()

myGraph + geom\_col()

myGraph + geom\_boxplot()

myGraph + geom\_violin()

*Functions you might need:* ***ggplot(), geom\_points(), geom\_smooth(), geom\_histogram (), geom\_density (), geom\_boxplot (), geom\_bar ()***

*\*there is a really nice cheat sheet for different functions in ggplot2 at* [*https://statsandr.com/blog/files/ggplot2-cheatsheet.pdf*](https://statsandr.com/blog/files/ggplot2-cheatsheet.pdf)

1. Make a histogram of gestation lengths for the entire data set (the geom to use is ***geom\_histogram ()***). A histogram presents a graph of the frequency of the data. The tallest bar indicates the bin of data points that occurs the most often. A histogram is a great way to visualize how often something happens.
2. Make another histogram of baby weight. Do you see how this graph shows you which weight is most common within the entire dataset.
3. We can also make a different type of graph to help us visualize the data. A box plot shows us the 1st and 3rd quartiles and the tips of the lines shows us the outliers. The width of the box effectively shows us the variance of the data set. Make a box plot ***geom\_boxplot ()*** of the entire dataset for gestation length. Compare this to your histogram. Which data visualization do you like better?

Part 3: Calculate summary statistics for a subset

For this part, we will focus just on California. Make a new data set by subsetting from the whole dataset just those rows containing California data. How many data points are in your new data frame?

1. How many male and female babies were born in California in 2015?
2. What was the smallest, average, and largest birth weight in California in 2011?
3. Use ***aggregate*** ***()*** to see how birth weight in California changes as a function of the mother’s age. Plot the relationship between mother’s age and birth weight. When we do one variable by another we use the ~ symbol. So in this case we might write Weight ~ Age. Aggregate also needs to know the dataset to look in data=ca (my dataset for California is called ca). Aggregate also wants to know what function to use. In our case we want to use the ***mean ()***, but you could use any of our statistical methods including; ***sd (), max (), min ()***.
4. For the Californian data of baby weight by mother’s age create a bar plot.
5. For the Californian data of baby weight by mother’s age create a line graph.
6. For the Californian data of baby weight by mother’s age create a scatter plot with points.
7. Which graph type do you like the most?
8. Do this for another state of your choice. I choose to do Hawaii.
9. Add some labels to the graph. It is always important to tell you reader what they are looking at. + ***labs(x= "Age", y= "Weight", title= "HI Weight by Age")***. This will add Age as the label for your x axis, and Weight as your label for the y-axis. It will also add a title to the top of your graph.
10. I want my title centered, so you can do a horizontal adjustment. Try adding this element to your graph to center the title + ***theme(plot.title = element\_text(hjust = 0.5))***
11. ggplot2 uses a concept of themes to define the non-data plotting elements. These include:

theme\_gray(): the default theme  
 theme\_bw(): a classic theme with a white background  
 theme\_minimal(): a minimalistic theme  
 theme\_classic(): close to the style of basic R plotting

1. Try these themes out by adding +theme\_NAME() to your plotting line for the last graph. Which theme you like best? I like the classic theme, but how your graph looks is up to you!

Bonus Activities: More Plotting using GGplot!

1. Try making a plot that compares the birth weight of California to that of Hawaii (or the state you choose). To do this you will need to build a data frame with both the data for California and Hawaii put together. I would recommend adding a column with the title of state to each individual data frame (define state with an abbreviation for each dataset. Then bind the two data frames together using the function ***rbind ()***. As part of your aes (), you are going to want to add a feature to separate the two states with different color lines. You can do this by telling ggplot that color is determined by state, color=State. Do you see how you could use this script to add as many lines/states as you want to this graph!

Now try to visualize the whole dataset.

1. Make a boxplot showing the distribution of birthweights across all states (so you should end up with 51 boxes (including Washington DC)). Notice that the axis labels are probably overlapping. Rotate them to vertical by adding this to the plotting code:

+ theme(axis.text.x=element\_text(angle = 90,vjust = 0))

When making a nice graph, you should write it as a pdf file. This will create the highest resolution version possible. The format of this typically looks like:

pdf("name of your graph.pdf")

ggplot( … whatever …) +

whatever geoms and stuff you want  
dev.off()

The pdf() function creates and opens a file, and as long as that file is open any plots you make will go to that and not into your RStudio plot window. This will happen until you close that file with dev.off(). At this point, you can open your file and see how it turned out. You might want to resize the window. By default, it will be 7 x 7 inches. If, for example, you wanted it to be 5 x 10, you could use:

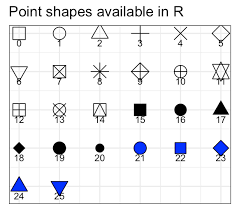
pdf("name of your graph.pdf",height = 5, width = 10)

1. Now, remake the above graph where each state’s box is split by the baby’s gender (do this by specifying color = Gender). Change the colors that have been assigned to the boxplots by using:

+scale\_color\_manual(values=c("a color", "another color"))

***colors()*** will provide you a list of all of the color names you can choose from.

1. Using just the California data, plot the relationship between a mother’s age (x), gestation length (y) and birth weight (color). Change the plotting symbol and/or size by adding a shape and/or size argument to geom\_point(), like geom\_point(shape = 2, size = 3). The shape codes are below:



When color is being used to indicate a continuous (rather than categorical) variable, we need to specify a continuous color gradient, e.g.:

+ scale\_color\_gradient(low="blue", high="red")

Adjust the symbol shape, size and colors in the plot.

1. Using the entire data set, compute the mean birthweight of babies aggregated by mother’s Age, baby’s Gender and the birth Year. Then, make a scatterplot showing the relationship between Age and Weight, using some combination of symbol color, shape and size to indicate Year and Gender. If using color, adjust the color palettes and choose a theme. Manipulate all of the variables we have been working on to make the graph look great! Save this as a pdf so you can show people the amazing graph you just made!