Personal log of R commands

(your name here)

TASK: Your own personal log of commands

(*Delete this opening from your final document.*) Generate for yourself examples of commands you've seen, ones which perform various useful tasks. The first few pages provide examples of examples done in R. On the final page are 10 exercises to which you should respond by generating examples with adequate explanations as appropriate—as if **the audience is yourself** several weeks or months from now. This is a quiz, but you can find examples adequate to your needs in the daily class notes.

The product of this work is the Quarto (.qdf) document itself, along with the .pdf obtained from knitting it. Both should be handed in, but you should also keep them for later reference, perhaps also expanding it as you learn more commands in the upcoming weeks. Upon rendering the .qmd, both files will be in your file space on the server. You'll need to download them to a local computer, and then upload them to MOM as your homework submission.

Basic operations on data frames

If you have the direct url to a data set in .csv format, it can be imported using the url in `read.csv()`:

```
mlbSal = read.csv("https://www.lock5stat.com/datasets3e/BaseballSalaries2019.csv")
```

It is possible to read into R any .csv file using this command, not merely those delivered via a web address. If, for instance, you create a .csv file using a spreadsheet program like Microsoft Excel, you can read the data into R so long as the file can be accessed in your workspace.

For data frames, either those imported as above, or ones obtained in other ways (through the loading of a package, for instance), you can learn its basic structure through commands like `nrow()` (tells how many rows/cases), `names()` (tells the names of columns/variables), or `head()` (to view the first few cases).

```
nrow(mlbSal)
```

```
[1] 877
```

```
names(mlbSal)
```

```
[1] "Name"   "Salary" "Team"   "POS"
```

```
head(mlbSal, n=3)    # specify the number of cases with 'n=_'
```

```
               Name Salary Team POS
1      Max Scherzer 42.143 WSH   SP
2 Stephen Strasburg 36.429 WSH   SP
3        Mike Trout 34.083 LAA   CF
```

Building a data frame within R: `data.frame()`

Some problems are stated with their own small data sets, not easily found as .csv files. The data in Exercise 1.20, p. 15 of the Lock5 text is such a problem. There are two variables measured on 8 cases. If there is reason to make the data available in R, it might be easiest to build it with lines like these:

```
ex1.20data = data.frame(
  sex = c("Male", "Male", "Male", "Male", "Male", rep("Female", 3)),
  time = c(40, 87, 78, 106, 67, 70, 153, 81)
)
ex1.20data
```

```
     sex time
1   Male   40
2   Male   87
3   Male   78
4   Male  106
5   Male   67
6 Female   70
7 Female  153
8 Female   81
```

You can access the time values using the `dFrameName$colName` notation:

```
ex1.20data$time
```

```
[1]  40  87  78 106  67  70 153  81
```

Selecting cases from data frame: `subset()`

You can select out cases in a data frame based on specified criteria. It's not the same as choosing only to see the values in a certain column. Rather, we want to see full case information, but perhaps only ones applying to "Female" rowers, or only those who made the Atlantic transit in between 70 and 100 days.

```
subset(ex1.20data, sex=="Female")   # Note the double-equal signs
```

```
     sex time
6 Female   70
7 Female  153
8 Female   81
```

```
subset(ex1.20data, time <= 100 & time >= 70)
```

```
     sex time
2   Male   87
3   Male   78
6 Female   70
8 Female   81
```

Compute a mean/average: `mean()`

Working with the iris data set, I may want to know the mean petal length (variable is named `Petal.Length`). To obtain one average for all plants in the data frame:

```
mean(~ Petal.Length, data=iris)
```

```
[1] 3.758
```

If what I really want is average petal lengths broken down by species, that becomes a bivariate issue; I am, perhaps, interested in whether there is an association between `Species` and `Petal.Length`. A natural first step might be to look at the sort of differences which exist in sample means:

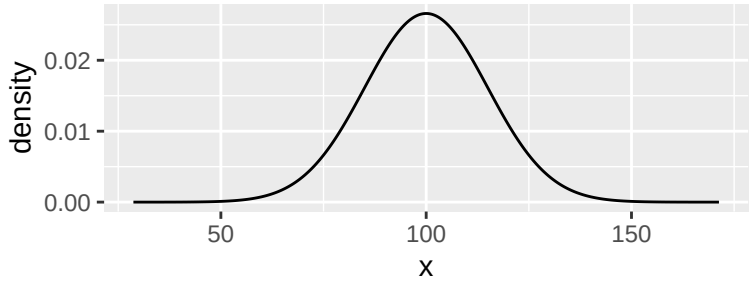
```
mean(~ Petal.Length | Species, data=iris)
```

```
    setosa versicolor  virginica
     1.462      4.260      5.552
```

View the distribution from a named family: `gf_dist()`

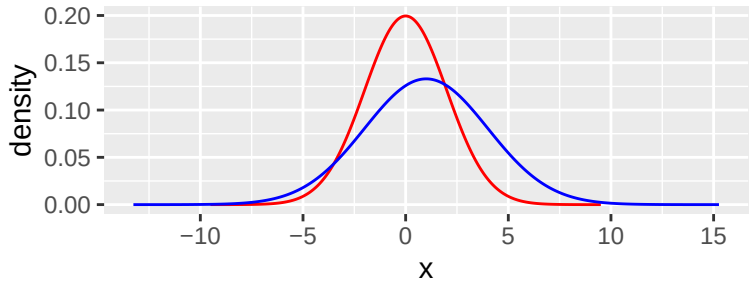
We have encountered two types of random variables considered important enough to specify them by names: **binomial** and **normal**. There are many different instances within these families. The Norm(100,15) distribution is the one typically used as a model for how different values of IQ appear in a population. It can be displayed using

```
gf_dist("norm", mean=100, sd=15)
```



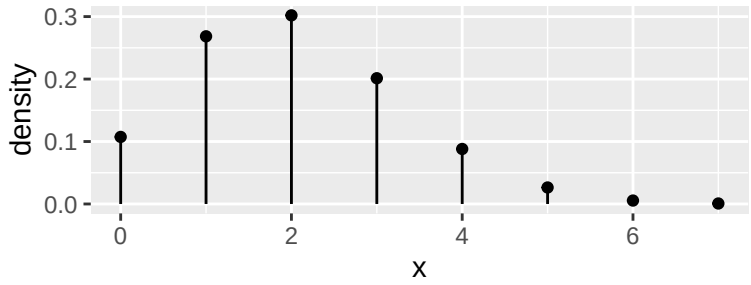
Here, using a pipe, I overlay a second normal distribution onto a first one:

```
gf_dist("norm", mean=0, sd=2, color="red") |>
 gf_dist("norm", mean=1, sd=3, color="blue")
```



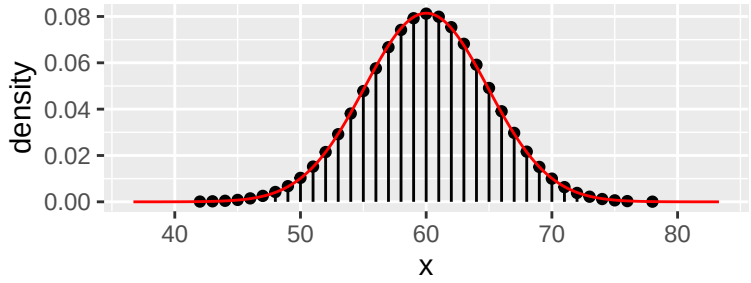
To display a binomial distribution, Binom(10, 0.2) for instance:

```
gf_dist("binom", size=10, prob=0.2)
```



When $X \sim \text{Binom}(n, p)$, we know (from Chapter 11) that the mean and standard deviation of X are $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$. The specific r.v. $X \sim \text{Binom}(100,\ 0.6)$ will have $\mu_X = (100)(0.6) = 60$ and $\sigma_X = \sqrt{100(0.6)(0.4)} \doteq 4.90$. It is interesting to compare the distributions Binom(100, 0.6) and Norm(60, 4.9):

```
gf_dist("binom", size=100, prob=0.6) |>
  gf_dist("norm", mean=60, sd=4.9, color="red")
```



You take it from here. In this order, provide instructions to yourself for doing these things in R:

1. Building a frequency table (univariate data) from raw data, as well as how to display it as a bar graph.
2. Building a two-way table from raw data, much as Table 2.5 was built using the data in **StudentSurvey**, a data frame from the Lock5withR package.
3. How to draw both an SRS and an iid sample from a collection of values.
4. How to produce quantiles-to-order; that is, if it is desired to learn the 5th, 23rd, and 81st percentile in a sample of values, how to get these efficiently.
5. How to generate a scatterplot for bivariate quantitative data.
6. How to produce side-by-side boxplots, perhaps in the case of `Sepal.Length`, giving one boxplot per `Species` (iris data set).
7. How to produce a histogram with bins that are of a specified width.
8. R commands that can be used to produce values such as those found in the standard normal table at this link https://math.arizona.edu/~jwatkins/normal-table.pdf. For instance, the linked table indicates the cumulative probability up to $Z = -0.92$ is 0.1788. On the other hand, if you wish to know the value of Z at which the cumulative probability is 0.6368, the table shows this occurs at $Z = 0.35$.
9. How to use `dbinom()` and `pbinom()`, and how to understand their results.
10. How to calculate the correlation coefficient for bivariate data.