Math 321 – Spring 2019 – R Activity 2

This is an exercise in understanding the "shape of data". You do not need to turn anything in for this assignment. It is just meant to give you additional practice with R.

First enter this data, get its summary statistics, and plot a histogram:

```
> x = c(0,0,0,1,1,1,1)
  summary(x)
  hist(x)
```

R hint: To enter multiple lines at once, hit ***Shift-Enter*** *to go to the next line without executing the one above. Then after you have entered all three lines of R commands, you can hit* ***Enter*** *to evaluate them all.*

Is this data skewed? Is it symmetric?

By the comparison of the mean and median, you might think it is left-skewed (negatively-skewed). However, it is really just two bins of nearly equal height. Any dataset of zeros and ones with only an additional one will look similar.

Try this one. It will look symmetric, but the mean is still less than the median!

```
> x=c(rep(-1,20),rep(0,30), rep(1,31),rep(2,20))
  summary(x)
  hist(x,breaks=12)
```

Don't worry about understanding the commands, but `rep(a,n)` makes a list that repeats the value a n times: $\{a, a, \ldots, a\}$.

Now try this. Replace the last 1 in the first dataset by a 4:

```
> x = c(0,0,0,1,1,1,4)
  summary(x)
  hist(x,breaks=14)
```

Now, you should see that the mean and median are identical, but the histogram clearly does not look symmetric!

What's going on?

Perfect symmetry is really very rare in data, and the mean and median being identical does not truly determine symmetry. Also, skewness has several different definitions. Comparing the mean and median is only one way to think about skewness. For our purposes, you can call data skewed by comparing the mean and median, or by visually seeing a tail that stretches out to one side on a histogram.

Now we will use some advanced R commands to generate "simulated data". Don't worry about understanding these commands. The purpose here is to just explore some different datasets to get a feel for how symmetry and skewness plays out.

To see a histogram from a large dataset that is approximately symmetric type:

```
> x=rnorm(5000)
  summary(x)
  hist(x,breaks=24)
```

The median and the mean will not be identical, but they will be close, and the left and right sides of the histogram should be close to mirror images, but not perfectly so.

Here is an example that looks very right-skewed, but the mean and median are still fairly close:

```
> x=rchisq(5000,10)
  summary(x)
  hist(x,breaks=32)
```

Here is an example that looks extremely right-skewed, and the mean and median are much further away from each other:

```
> x=rexp(5000,0.1)
  summary(x)
  hist(x,breaks=32)
```