Data Visualization, Part 2

February 5, 2018

Warm Up

I have loaded a data set called `senate_113` with information about the senators in the 113th US Senate (this is the senate that came into session in January 2013)

Here's a first look at the data set:

```
dim(senate_113)
```

```
## [1] 100   7
```

```
head(senate_113)
```

```
## # A tibble: 6 x 7
##   firstname middlename  lastname   birthday state party   age
##       <chr>      <chr>     <chr>     <date> <chr> <chr> <dbl>
## 1    Dianne       <NA> Feinstein 1933-06-22    CA     D  79.5
## 2   Charles         E.  Grassley 1933-09-17    IA     R  79.3
## 3     Orrin         G.     Hatch 1934-03-22    UT     R  78.8
## 4   Richard         C.    Shelby 1934-05-06    AL     R  78.7
## 5      Carl       <NA>     Levin 1934-06-28    MI     D  78.5
## 6     James         M.    Inhofe 1934-11-17    OK     R  78.1
```

1. How many observational units are in the data set? What is each observational unit?

2. What are the variables in the data set? What type of variable is each?

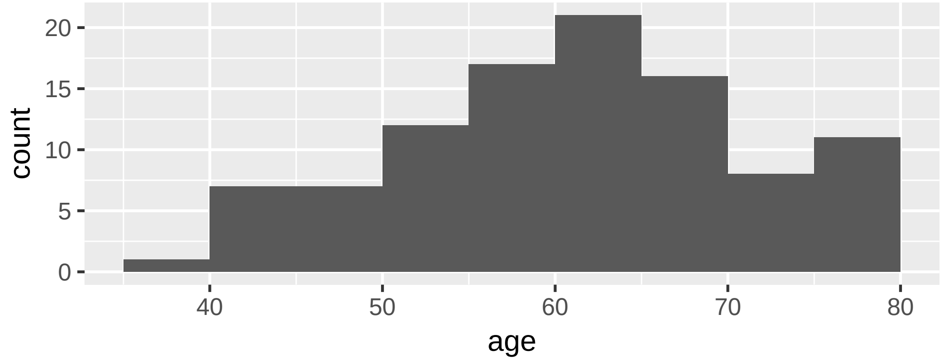
Histograms

Histograms are a common type of plot for displaying a quantitative variable. The horizontal axis is divided into bins of equal width, and the height of each bar represents the number of units in that bin. There are two common types of histograms, where the vertical axis means different things:

1. **count**: The height of each bar is the number of observational units in that bin.
2. **density**: The area of each bar is the **proportion** of observational units in that bin. (The height is whatever it needs to be to make the area work out correctly).

3. Here is a histogram where the vertical axis is a count. Based on this plot, how many senators were aged between 40 and 50?

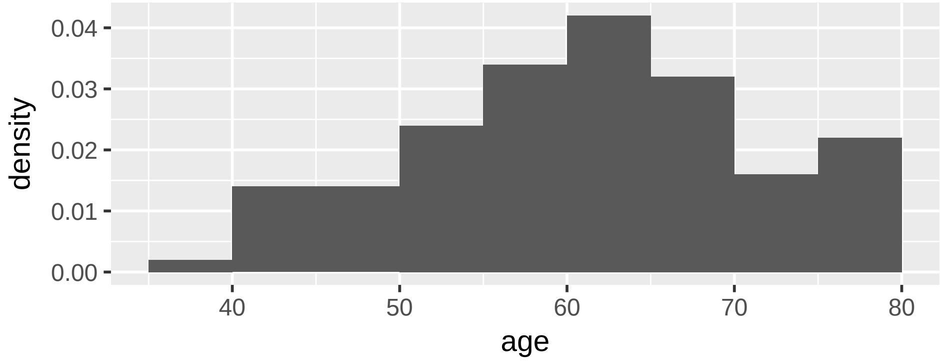
```r
ggplot(data = senate_113, mapping = aes(x = age)) +
  geom_histogram(binwidth = 5, boundary = 40)
```



4. Here is a histogram where the vertical axis is density. Based on this plot, approximately what proportion of senators were aged between 55 and 60?

Note that in this plot, each bin has width 5. Remember that the formula for area is width × height.

```r
ggplot(data = senate_113, mapping = aes(x = age, y = ..density..)) +
  geom_histogram(binwidth = 5, boundary = 40)
```

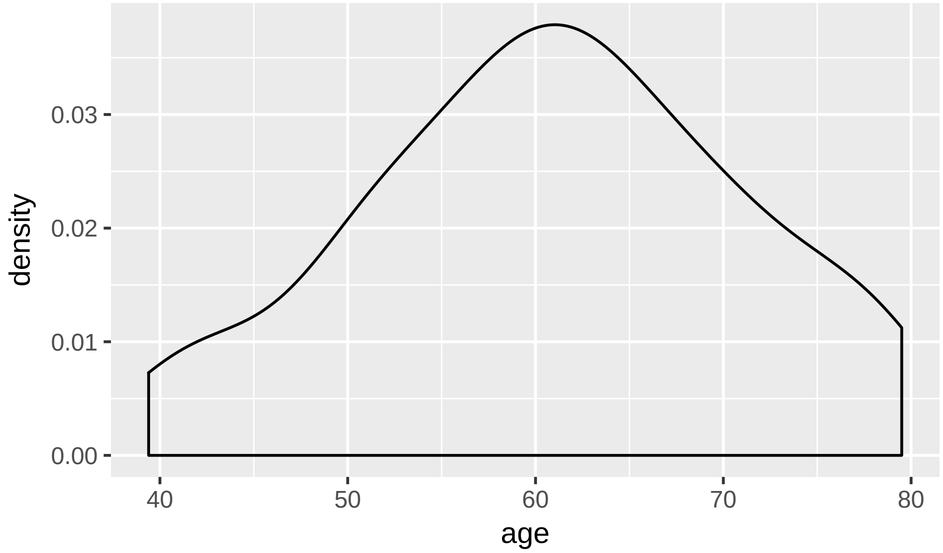


Note that I will never again ask you to actually calculate a proportion like this, **but** it is important to know how you would, conceptually.

Density Plots

A density plot is basically a smoothed density histogram.

```
ggplot(data = senate_113, mapping = aes(x = age)) +
  geom_density()
```



5. What would the area under the curve between 40 and 50 tell you?

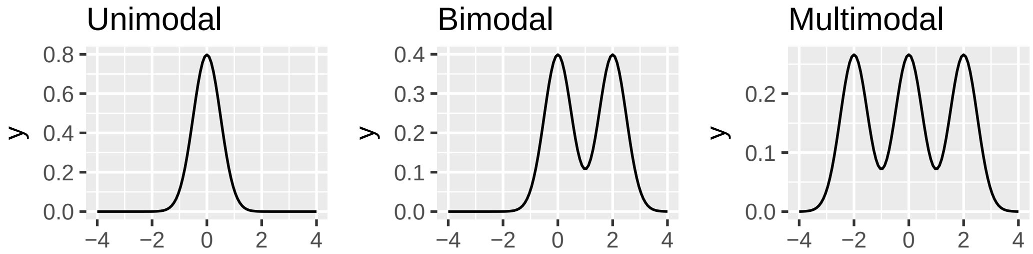
As with the density histogram, we won't actually calculate the area under a density curve to answer questions like this – but, we **will** need to know what the area corresponds to.

Describing the Shape

When I look at a histogram or density plot, I'm evaluating three characteristics of the plot:

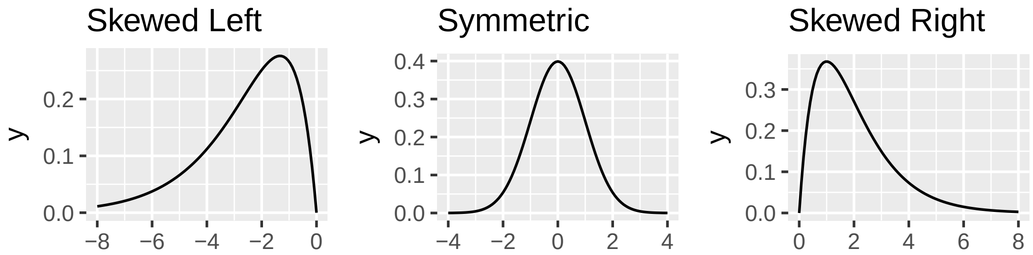
A. Unimodal or Multimodal?

A **mode** is a local peak in the distribution.



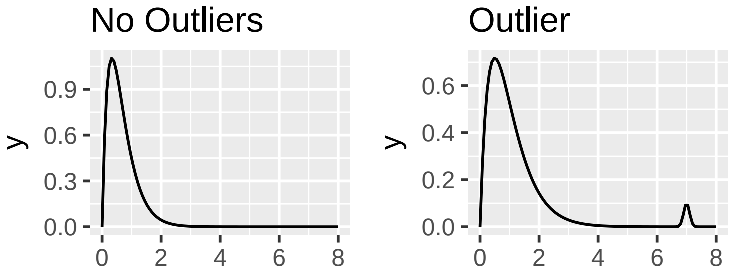
B. Symmetric or skewed?

If a distribution is skewed, it's skewed **towards the tail**.



C. Are there any gaps or outliers?

An **outlier** is a data value that "doesn't fit" with the rest of the data.

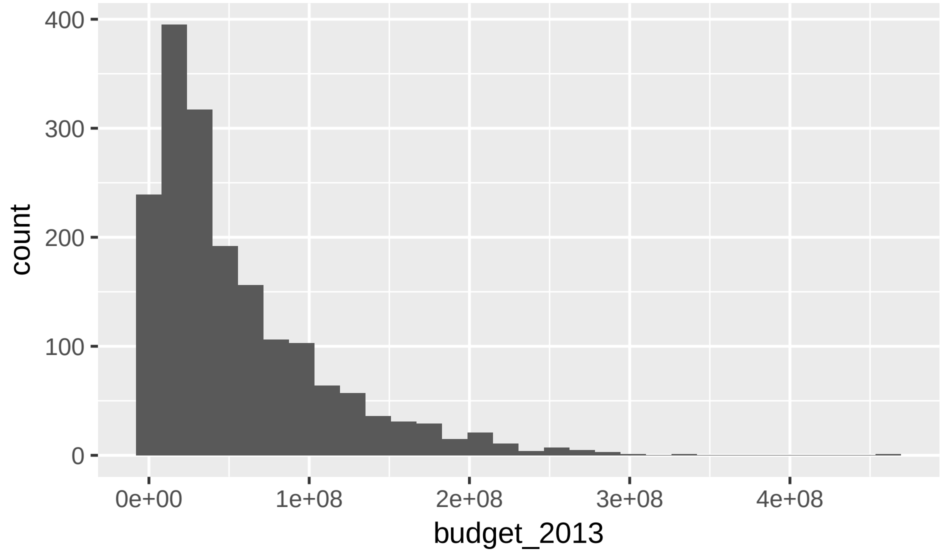


6. Describe each of the following distributions; for each, discuss whether it is unimodal or multimodal, symmetric or skewed, and whether there are any outliers.

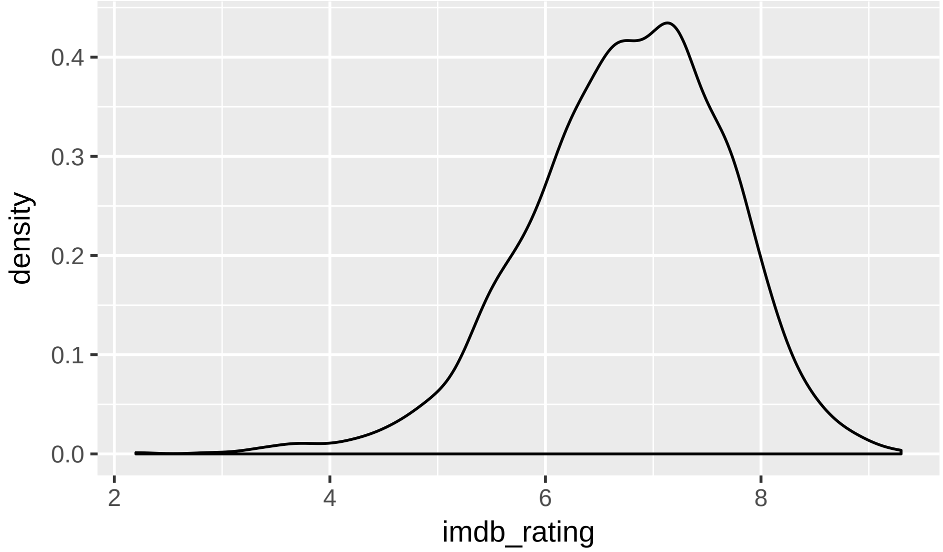
I've taken these examples from the Bechdel Test data. In the first example, look carefully at the histogram!

```
ggplot(data = movies, mapping = aes(x = budget_2013)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



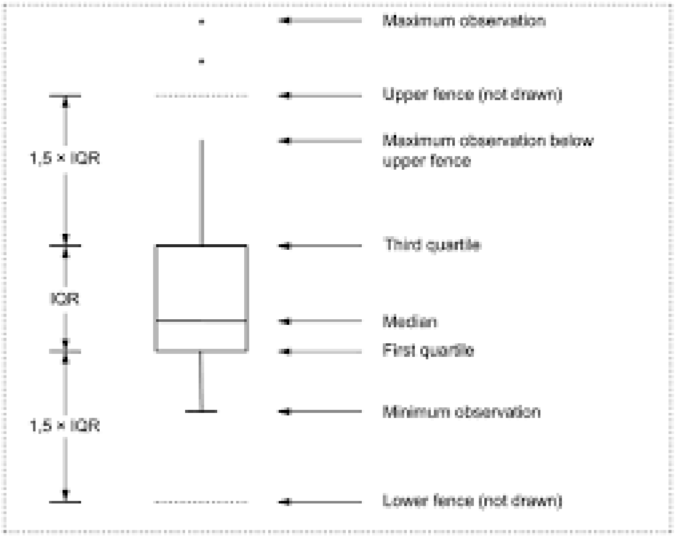
```
ggplot(data = movies, mapping = aes(x = imdb_rating)) +
  geom_density()
```



Box Plots

A box plot is a graphical representation of the **5 number summary** of a data set. The five numbers in the five number summary are:

1. The maximum
2. The 75th percentile (the number such that 75% of the data are less than that value, and 25% are greater than that value)
3. The median (the number such that half of the data are less than that value and half are greater than that value)
4. The 25th percentile (the number such that 25% of the data are less than that value, and 75% are greater than that value)
5. The minimum



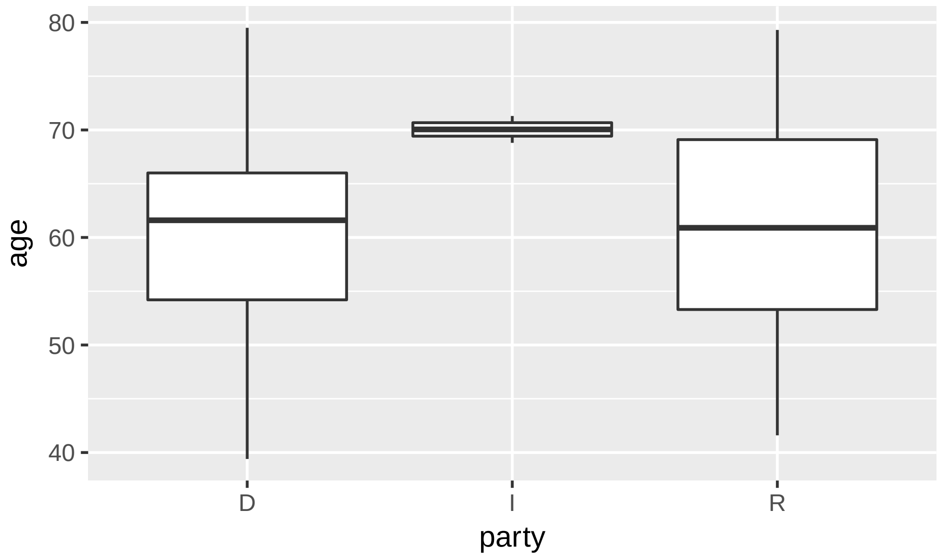
Some additional terminology:

- The 25th percentile is also called the **first quartile**
- The 75th percentile is also called the **third quartile**
- The difference between the 75th percentile and the 25th percentile is called the Interquartile range (IQR). It gives the width of an interval containing the middle half of the data.

In practice, we basically only use box plots to compare the distribution of values across multiple groups. There are better plots to use for examining the distribution for only one or two groups.

Let's go back to looking at the ages of members of the US Senate.

```
ggplot(data = senate_113, mapping = aes(y = age, x = party)) +
  geom_boxplot()
```



Here are some questions you should be able to answer based on the box plots above:

7. Which party had the highest median age?

8. The youngest member of the senate belonged to which party?

9. 75% of Republican senators were younger than what age?

10. How wide of an interval would you need to cover the ages of the middle half of Democratic senators?