R Lab #5b - Mean vs Median

Contents

Comparing mean versus median

For many variables, frequency distributions for the population has one mode. The mean and median are common statistics used to identify the "middle" of a distribution, and hence the position of the mode. When the spread of the data is relatively symmetrical around the mode, the mean and median values will be similar to each other. However, when the distribution is skewed (either positive or negative) then the mean will be "pulled" in the direction of the skew whereas the median will be closer to the mode. In this lab we'll demonstrate this by analyzing salary data.

Massachusetts state employee salary data

The salaries of Massachusetts state employees is public record. I downloaded the data for 2018, and filtered it to retain only full time employees of the UMass system. Use read.csv() to import the data and summary() to get an idea of the different variables included.

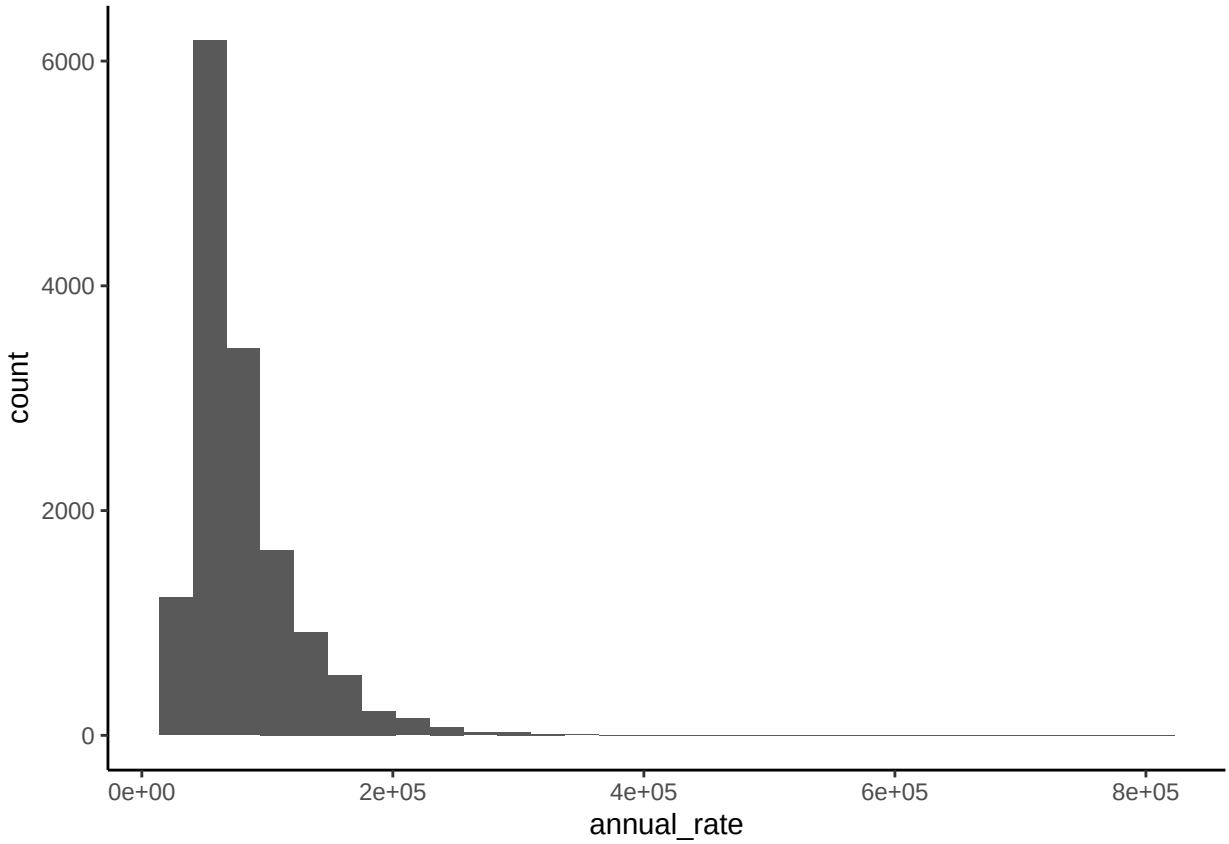
```
umassData <- read.csv("Massachusetts_Payrollv2_2018_UMS.csv",
    stringsAsFactors = T)
summary(umassData)
##       Year        name_last          name_first
##  Min.   :2018   Smith  :   56   Michael    :  245
##  1st Qu.:2018   Wang   :   49   David      :  189
##  Median :2018   Chen   :   42   John       :  184
##  Mean   :2018   Nguyen :   39   James      :  143
##  3rd Qu.:2018   Li     :   38   Robert     :  139
##  Max.   :2018   Miller :   38   Christopher:  125
##                 (Other):14224   (Other)    :13461
##                  department_location            position_title
##  U Mass Amherst             :5353     Maintainer I       :  517
##  U Mass Boston              :1808     Associate Professor:  419
##  U Mass Central Office      : 256     Clerk IV           :  416
##  U Mass Dartmouth           :1216     Assistant Professor:  413
##  U Mass Lowell              :1898     Professor          :  412
##  U Mass Other Administrative:  39     Professor U of M   :  371
##  U Mass Worcester           :3916     (Other)            :11938
```

```
##   annual_rate
##  Min.   : 17576
##  1st Qu.: 50538
##  Median : 67000
##  Mean   : 79753
##  3rd Qu.: 94637
##  Max.   :800000
##
```

Plotting salaries

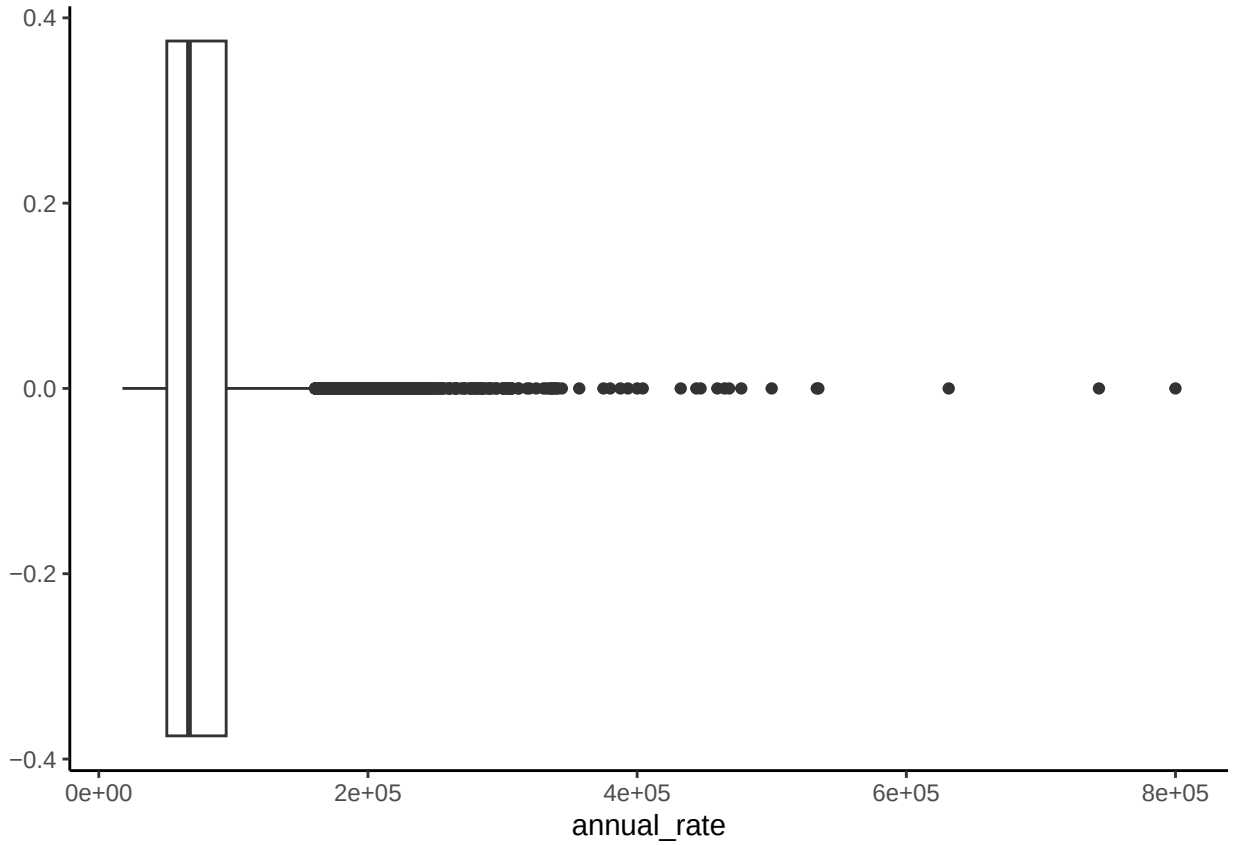
Let's now use ggplot2 to make a frequency distribution plot of the salary data (annual_rate). You will first need to import ggplot2 with library() if you haven't done it yet in your current R session.

```
library(ggplot2)
ggplot(umassData, aes(x = annual_rate)) + geom_histogram() +
    theme_classic()
```



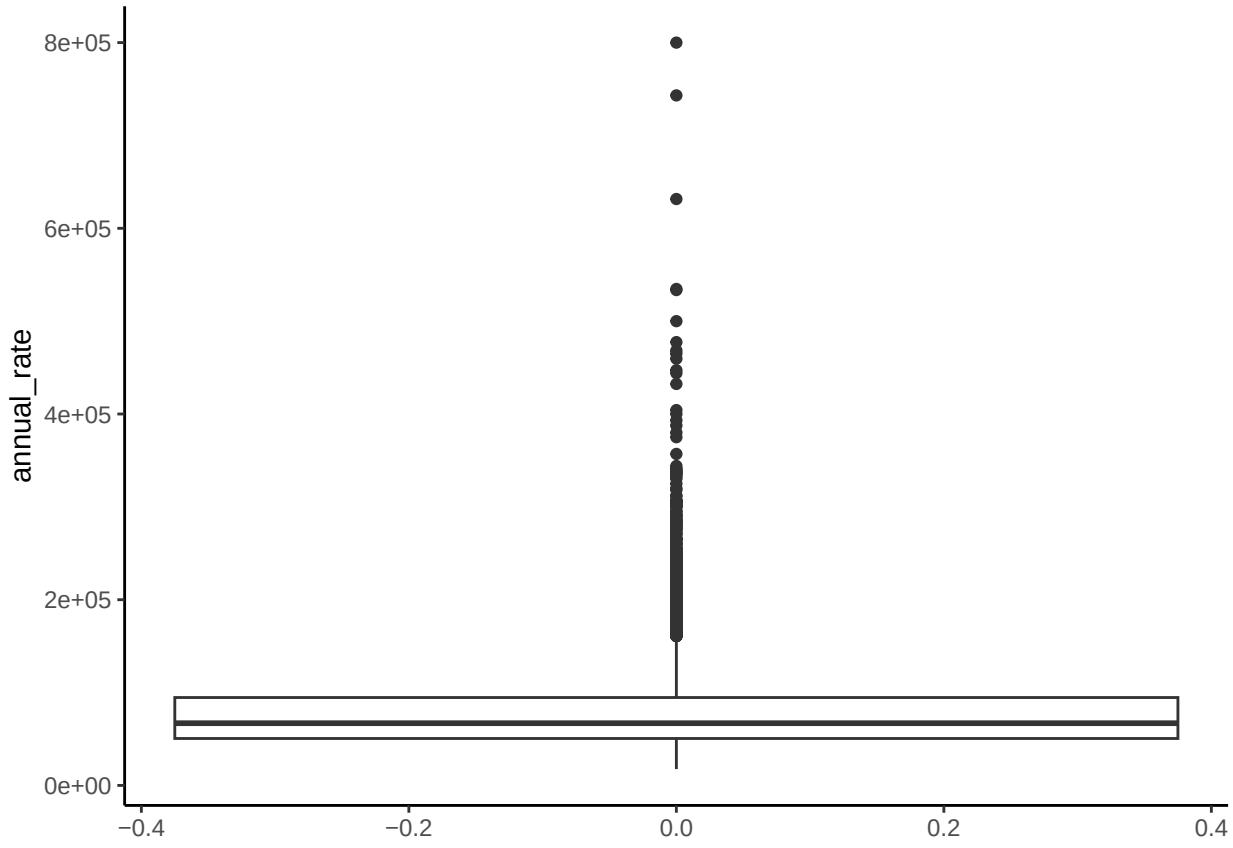
Notice that there is a lot of white space to the right. This is because there is a small number of high salaries that you can barely see, but they introduce positive (right) skew. Let's make a boxplot as another way to visualize the salary data.

```
ggplot(umassData, aes(x = annual_rate)) + geom_boxplot() + theme_classic()
```



Each dot to the right is a sampling unit that is greater than Q3+(1.5*IQR), and you can better see the right skew. If you prefer a vertical plot then just change the annual_rate from x to y:

```
ggplot(umassData, aes(y = annual_rate)) + geom_boxplot() + theme_classic()
```



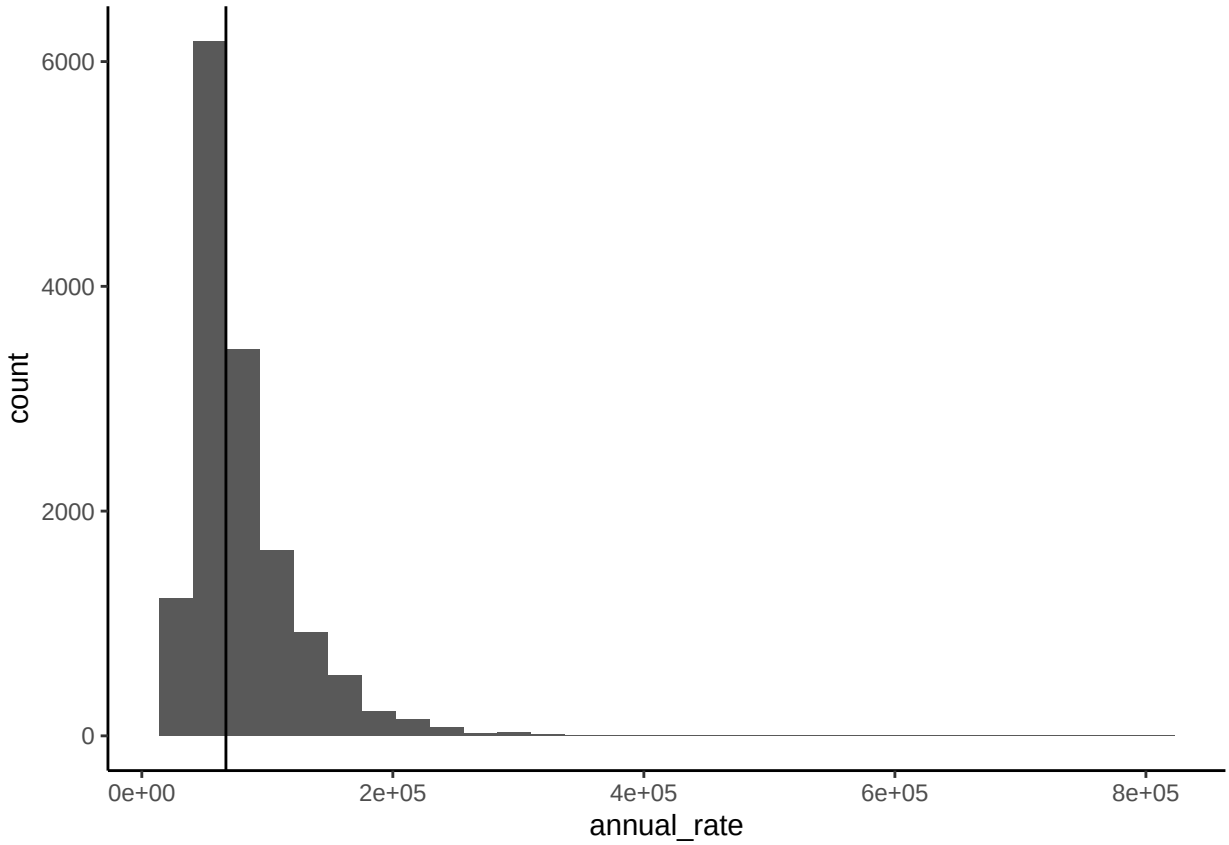
Comparing mean and median

So here we are dealing with a distribution that has considerable positive/right skew. These (relatively) small number of high values are going to affect the mean more than the median. These high values "pull" the mean in the direction. Let's calculate these statistics with mean() and median().

```
mean(umassData$annual_rate)
## [1] 79752.93
median(umassData$annual_rate)
## [1] 67000.18
```

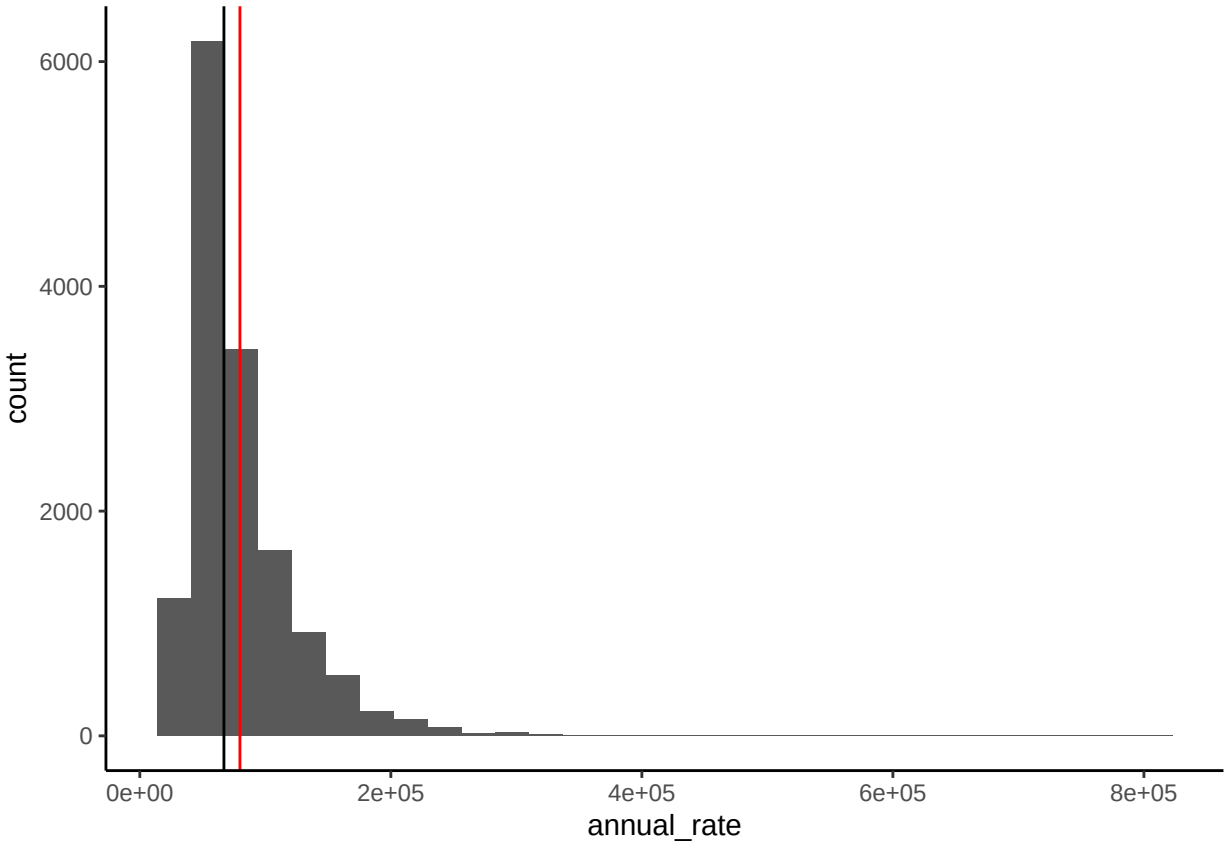
Notice that the mean is over 12700 higher than the median, but the median is closer to the mode (given the positive skew). We can visualize this by adding vertical lines showing the placement of the mean and median on the histogram. A vertical line can be added to a ggplot with the function geom_vline(), which just needs the x-intercept. First let's add the median...

```
ggplot(umassData, aes(x = annual_rate)) + geom_histogram() +
    theme_classic() + geom_vline(xintercept = median(umassData$annual_rate))
```



Now let's add a second line for the mean and change the color of this line to red...

```
ggplot(umassData, aes(x = annual_rate)) + geom_histogram() +
    theme_classic() + geom_vline(xintercept = median(umassData$annual_rate)) +
    geom_vline(xintercept = mean(umassData$annual_rate), col = "red")
```



Now you can visualize the difference. If there was negative/left skew than the mean would be lower than the median, and if there is no skew then the mean and median would be about the same.

R commands summary

- **ggplot histogram**
  - geom_histogram()
- **ggplot vertical line**
  - geom_vline()