James Scruggs

M01255052

Today we will examine measures of center and measures of variation.

Mean and Median

1. Create three vectors, *var1*: 1 2 3 4 5 6 7 8 9 10

var2: 1 2 3 4 5 6 7 8 9 50

var3: 1 2 3 4 5 6 7 8 9 100

In R:

var1<-1:10

var2<-c(1:9,50)

var3<-c(1:9,100)

Calculate the mean and median of each of these vectors *summary(var1); summary(var2); summary(var3)*. Put your results below:

First set mean: 5.50	First set median: 5.50
Second set mean: 9.50	Second set median: 5.50
Third set mean: 14.50	Third set median: 5.50

2. What is the difference between the mean and the median? Why do the means change for these data sets, while the medians stay the same?

The **mean** is the total sum divided by number of elements.

The **median** is the middle element (if data set is odd) or the mean of the middle two elements (if the data set is even).

The means change because the sum of all the elements change by replacing 10 with a larger value. The median stays the same because the middle two values are still the same.

We will use the notation $\overline{X}$ for the mean of a sample.

Range and Standard Deviation

3. Now look at the range and the standard deviation for each of the three data sets.

In R

sd(var1); range(var1)

sd(var2); range(var2)

sd(var3); range(var3)

Put the results below:

First set range: 1 10	First set standard deviation: 3.02765
Second set range: 1 50	Second set standard deviation: 14.4626
Third set range: 1 100	Third set standard deviation: 30.15239

The range and the standard deviation both are measures of how spread out a data set is. The range command in R gives the largest and smallest number. The statistic, “range,” takes the largest value and subtracts the smallest value, so that there is only one number. The standard deviation can be thought of as an average of the distances between the data points and the mean.

4. In most applications of statistics, the standard deviation is used as the measure of variation in a data set instead of the range. To see why, enter the following data sets in the *var4* and *var5*:

var4: 10 20 30 40 50 60 70 80 90 100 var5: 10 52 53 54 55 55 56 57 58 100

In R

```
var4<-seq(from=10,to=100,by=10)
```

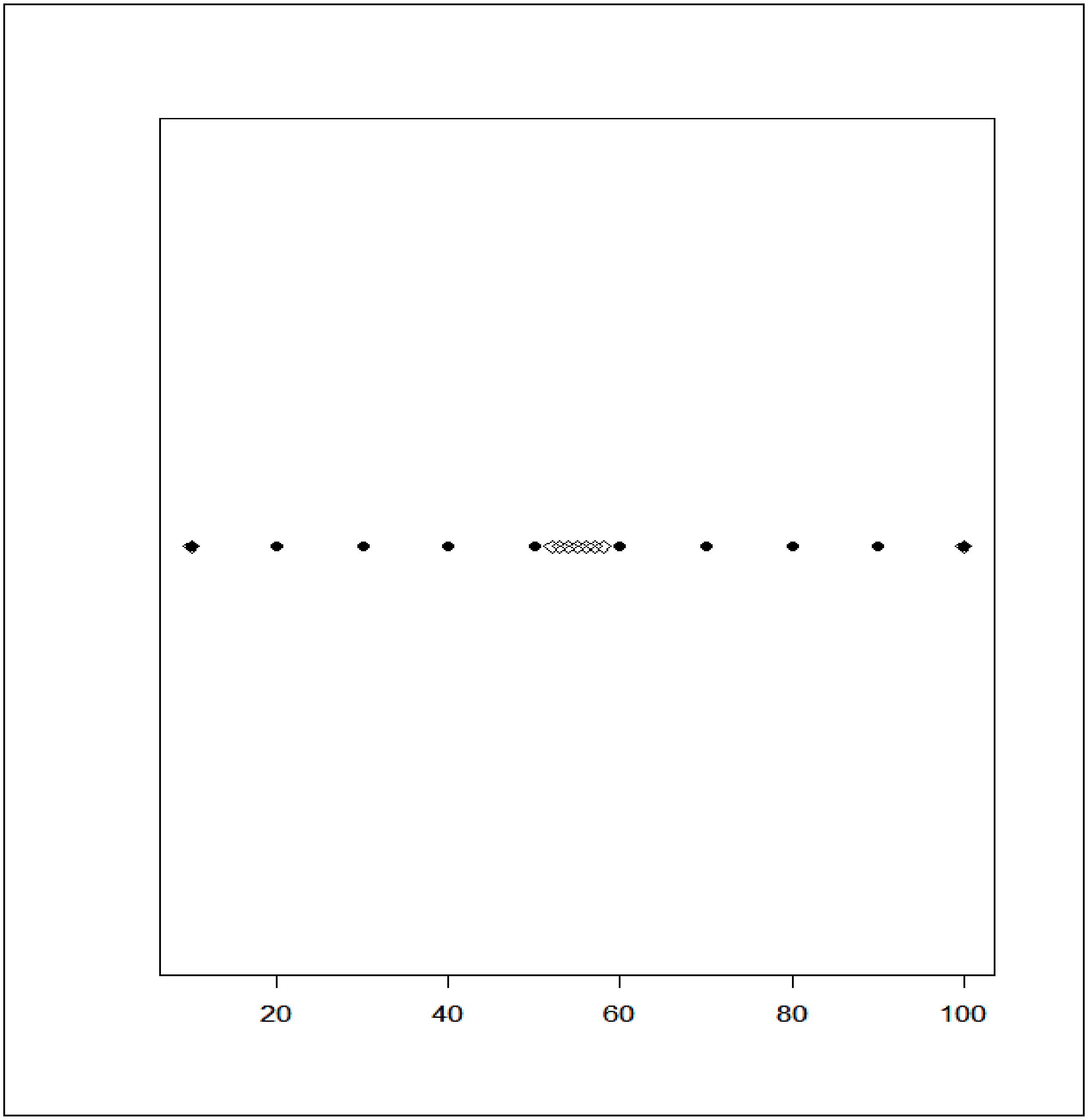
```
var5<-c(10,52:58,100)
```

Create dot plots for the two data sets

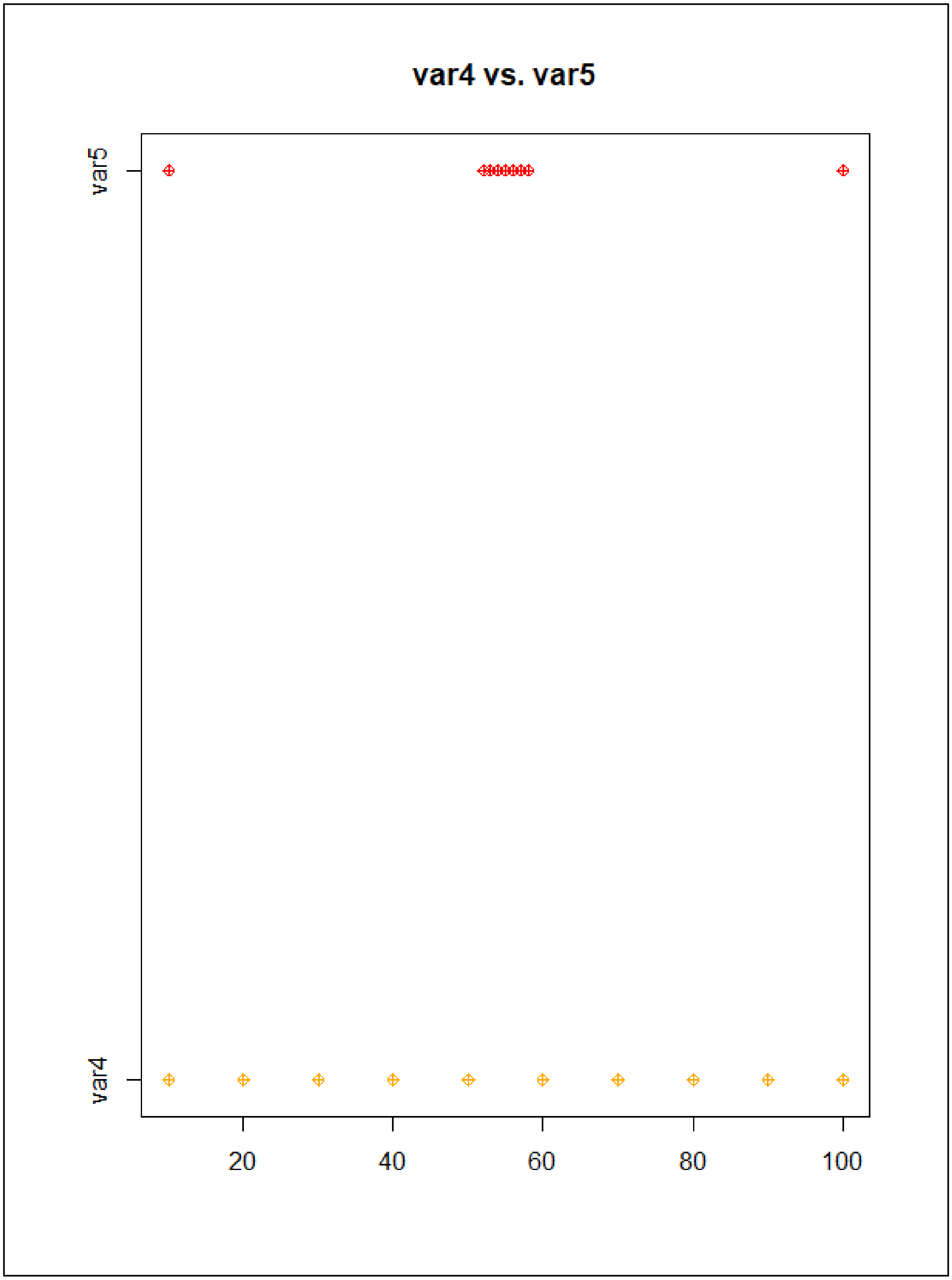
```
stripchart(var4,method="stack",pch=19);
```

```
par(new=TRUE) # optional, can put the two graphs on the same axis
```

```
stripchart(var5,method="stack",pch=23)
```


20
40
60
80
100

Use list(var4, var5) and plot stripchart()



and calculate the range and standard deviations for the two sets (*max(var4)-min(var4); sd(var4); max(var5)-min(var5);sd(var5)*). Paste these below.

var4 range: 90 var4 standard deviation: 30.2765

var5 range: 90 var5 standard deviation: 22.57764

5. Based on the graphs, which set has more variation (ie: is furthest from the middle on average)? Which measure (range or standard deviation) better reflects the variation?

The var4 data set has more variation with respect to the mean. Standard deviation better reflects the variation.

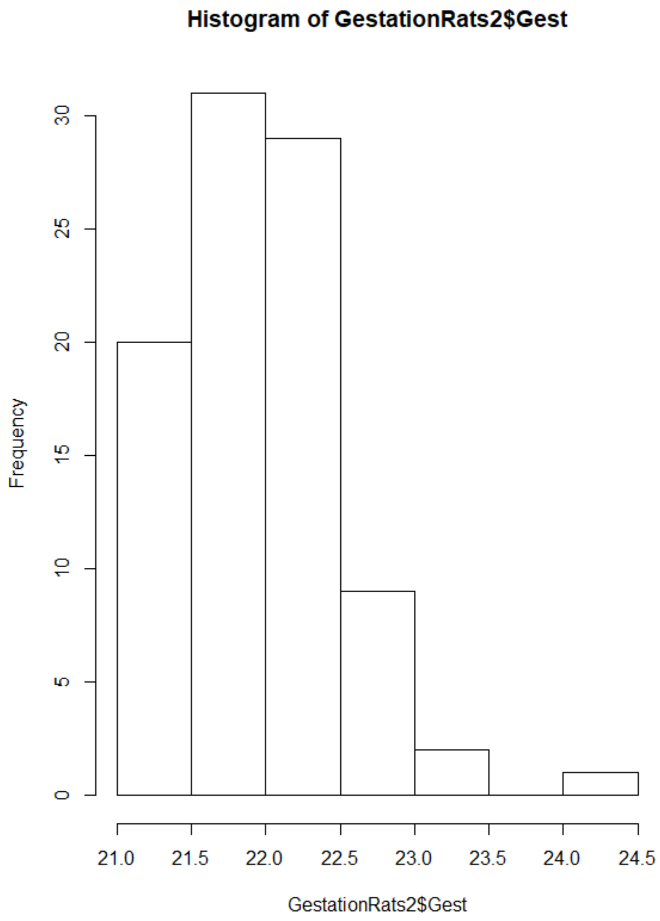
Using Mean and Standard Deviation Together: The Empirical Rule

The mean and the standard deviation together describe the center and the variation of a data set. In fact, in many data sets, these two numbers used together are sufficient to describe the data set. The reason for this is the empirical rule:

> If the data set comes from a population that is mound-shaped, then:
>
> - Approximately 68% of the data are within one standard deviation of the mean.
> - Approximately 95% of the data are within two standard deviations of the mean.
> - Approximately 99.7% of the data are within three standard deviations of the mean.

We need to understand the different parts of this rule. Load the data set *gestation rats 2.csv (GestationRats2<-read.csv(file.choose(),header=TRUE)).* This file contains the gestation period (time spent pregnant, to the nearest half day) and the total weight of male offspring (in grams) for 92 mother rats.

6. Create a histogram of the gestation period (*hist(GestationRats2$Gest))* and calculate the mean and standard deviation of the gestation period (*mean(GestationRats2$Gest); sd(GestationRats2$Gest)*).



Notice that the histogram is larger in the middle and smaller on the ends. This is an example of the mound shape referenced in the rule. A perfect mound shape would look like a bell (you've heard of the "bell curve"), but it is rare to get perfection in a real data set.

Mean: 22.20109	Standard Deviation: 0.5648095

7. The rule talks about numbers that are "within one standard deviation of the mean". To find this set of numbers, calculate the two numbers below.

Mean – standard deviation = $(\overline{X} - s) = 21.63628$

Mean + standard deviation $= (\overline{X} + s) = 22.7659$

The set of numbers between the two values you just calculated are within one standard deviation of the mean.

8. Count how many data values are within one standard deviation of the mean.

In R:

```
MeanGest<-mean(GestationRats2$Gest)

SdGest<-sd(GestationRats2$Gest)

sum(GestationRats2$Gest<MeanGest+SdGest & GestationRats2$Gest>MeanGest-SdGest)

# in R, the logical TRUE is counted as 1, the logical FALSE is counted as 0. So, the sum is the
number of TRUES.
```

Enter the value below:

60

9. Find the values that are two standard deviations from the mean, and count how many data points are within that interval:

Mean – 2 * standard deviation ($\overline{X} - 2s$) = 21.07147

Mean + 2 * standard deviation ($\overline{X} + 2s$) = 23.33071

Number of values in this interval = 88

10. Find the values that are two standard deviations from the mean, and count how many data points are within that interval:

Mean – 3 * standard deviation ($\overline{X} - 3s$) = 20.50666

Mean + 3 * standard deviation ($\overline{X} + 3s$) = 23.8952

Number of values in this interval = 91

11. Recall that there are 92 values in this data set, and calculate the percentage of the data set that are within one, two, and three standard deviations of the mean. Observe the similarity between these results and the results predicted by the Empirical Rule.

Percentage within one standard deviation: 65%

Percentage within two standard deviations: 95%

Percentage within three standard deviations: 99%

12. The Empirical Rule does have a requirement. It only applies when data sets are mound-shaped. Do you think the rule will apply to the weight of the litters? Support your answer.

Yes, the histogram of the weight appears to have a mound shaped histogram (maybe even more so than the Gest histogram).

13. Load the data set *presidentdays.csv*. Do you think the Empirical Rule will apply to this data set? Support your answer. Find out what percentage of the data set is within one, two, and three standard deviations of the mean. Note that even though the rule doesn't apply officially, we still end up with a decent approximation.

I do not think it will. The Empirical Rule is only applied to normal distribution, and from looking at the histogram of the data, it does not appear to have a normal distribution. I was shocked to find out that the standard deviations were 58%, 95%, and 100% respectively. The only part that did not fit the Empirical Rule standard was the percentage of data 1 standard deviation from the mean. This is probably because a good portion of data is still outside of one standard deviation.