## 6-14

Data: 2.13, 2.96, 3.02, 1.82, 1.15, 1.37, 2.04, 2.47, 2.60. (n=9)

### a

Calculate sample mean and standard deviation by hand

$$\bar{x} = \frac{2.13+2.96+3.02+1.82+1.15+1.37+2.04+2.47+2.6}{9} = \frac{5.09+4.84+1.15+3.41+5.07}{9} = \frac{8.5+4.84+1.15+5.07}{9} = \frac{13.57+5.99}{9} = \frac{19.56}{9} = 2\frac{1.56}{9} = 2.1733$$

$$s^2 = \frac{(2.13-2.1733)^2+(2.96-2.1733)^2+(3.02-2.1733)^2+(1.82-2.1733)^2+(1.15-2.1733^2)+(1.37-2.1733)^2+(2.04-2.1733)^2+(2.47-2.173}{8}$$

$$\frac{(-0.043333)^2+(0.786667)^2+(0.846667)^2+(-0.353333)^2+(-1.023333)^2+(-0.803333)^2+(-0.133333)^2+(0.296667)^2+(0.426667)^2}{8} =$$

$$\frac{0.001877778+0.61884444+0.71684444+0.124844442+1.0472111+0.64534444+0.01777777+0.08801111+0.18204444}{8} =$$

$$\frac{0.00187489+0.61889689+0.7169+0.12482+1.04714+0.64529+0.0177689+0.08803+0.18207}{8} = 0.43035$$

$$s^2 = 0.43035$$

$$s = \sqrt{0.43035} = 0.65601$$

### b

Calculate sample meadian by hand

2.13, 2.96, 3.02, 1.82, 1.15, 1.37, 2.04, 2.47, 2.60 -> 1.15, 1.37, 1.82, 2.04, 2.13, 2.47, 2.6, 2.96, 3.02 -> 1.37, 1.82, 2.04, 2.13, 2.47, 2.6, 2.96 -> 1.82, 2.04, 2.13, 2.47, 2.6, -> 2.04, 2.13, 2.47, -> 2.13

### c

Repeat above using R

```r
data <- c(2.13, 2.96, 3.02, 1.82, 1.15, 1.37, 2.04, 2.47, 2.60)

print(paste("Sample Mean:", mean(data)))
```

```
[1] "Sample Mean: 2.17333333333333"
```

```r
print(paste("Sample Standard Deviation:", sd(data)))
```

```
[1] "Sample Standard Deviation: 0.656010670644922"
```

```r
print(paste("Sample Median:", median(data)))
```

```
[1] "Sample Median: 2.13"
```

## 6-44

### a

Comment on the shape of the distribution

> There are a lot of data points in the 400s, and the data is much more sparse at higher numbers. This means the data is skewed left (smaller)

### b

Comment on the outliers of the data (DO NOT USE 1.5 IQR Rule)

> The outliers of this distribution would include 3469 and 3227. They are much higher than the median which is probably around 1000 or 1200.

### c

Which do you think has a higher value, sample mean or median? (EXPLAIN)

> The mean is probably greater than the median because the mean is more affected by large outliers whereas the median is not.

### d

Do you think the sample standard deviation is big or small? (EXPLAIN)

> The sample standard deviation is probably big because there is a very large range in the data. If the data was clustered around 1000, then the std would be much lower.

### e

Find the 3rd quartile and 80th percentile by hand

$n = 27$

Sorted: 450 450 452 453 457 473 507 1066 1085 1111 1145 1215 1254 1256 1364 1396 1575 1617 1733 1911 2588 2635 2725 2753 3186 3227 3469

Q3: $28 * 0.75 = 21$

21st number is 1911 so that is Q3

80th: $28 * 0.8 = 22.4$

22nd and 23rd number are 2588 and 2635 so 80th percentile is 2611.5

### f

Repeat part e using R

```
data <- c(450, 450, 473, 507, 457, 452, 453, 1215, 1256, 1145, 1085, 1066, 1111, 1364, 1254,
1575, 1617, 1733, 2753, 3186, 3227, 3469, 1911, 2588, 2635, 2725)

quantile(data, probs=c(0.75,0.8))
```

```
   75%    80%
2249.5 2625.6
```

```
print(length(data))
```

```
[1] 27
```

```
print(sort(data))
```

```
 [1]  450  450  452  453  457  473  507 1066 1085 1111 1145 1215 1254 1256 1364
[16] 1396 1575 1617 1733 1911 2588 2635 2725 2753 3186 3227 3469
```

## 6-42

### a

Use R to find the 5 number summary

```
str = "680 669 719 699 670 710 722 663 658 634 720 690 677 669 700 718 690 681 702 696 692 69
data = c(as.numeric(strsplit(str, " ")[[1]]))
summary(data)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  634.0   667.8   683.0   686.8   703.2   763.0
```

```
print(sort(data))
```

```
  [1] 634 637 642 644 648 649 649 652 652 653 655 656 656 658 659 659 660 660
 [19] 660 660 660 660 661 662 663 663 664 665 665 667 668 668 668 669 669 670
 [37] 670 672 672 674 675 675 675 676 677 678 679 679 680 680 680 680 681 681
 [55] 681 681 682 683 683 683 683 683 684 685 688 690 690 690 690 691 691 692
 [73] 693 694 695 695 695 695 696 697 697 698 698 699 700 701 701 702 702 703
 [91] 704 704 704 705 705 706 710 710 715 717 718 718 719 720 720 720 720 721
[109] 722 722 723 724 724 724 727 735 739 746 748 763
```

### b

Identify any outliers by hand using the 1.5 IQR Rule

Q1 = 667.8, Q2 = 703.2 $\rightarrow$ IQR = 35.4

1.5 IQR = 53.1

Therefore the range is between 614.7 and 756.3

$MIN = 634, MAX = 748$

Outliers are all values outside of this range, which include: 763.

### c

Construct a box plot by hand based on your results in pars a and b.

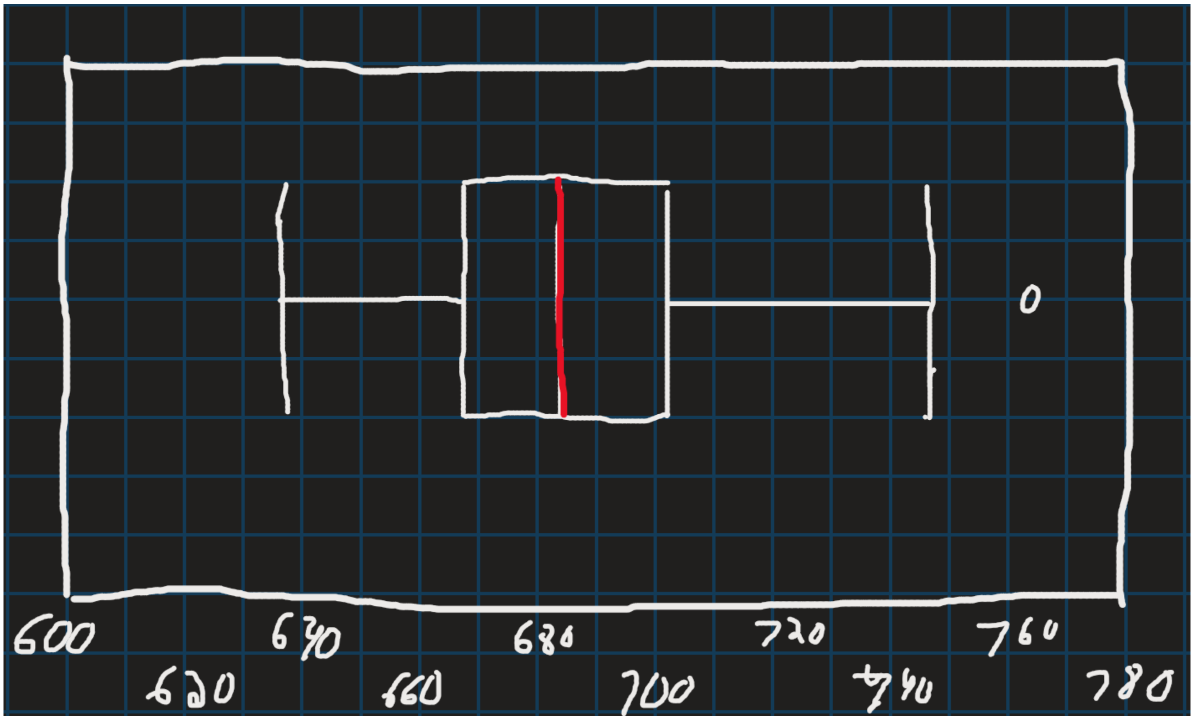



Figure 1: Box Plot

### d

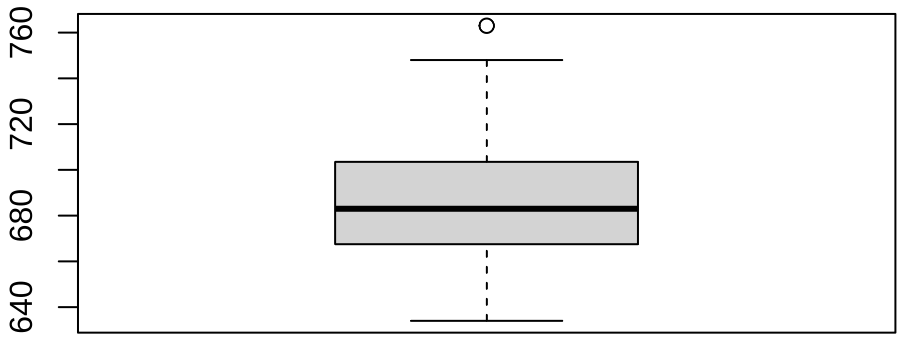
Describe the shape of the data distribution based on the boxplot that you created in part c

> The left whisker is much smaller than the right whisker and the median is slightly to the left of the box (Q1, Q3) so the data is skewed right.

### e

Repeat part c using R

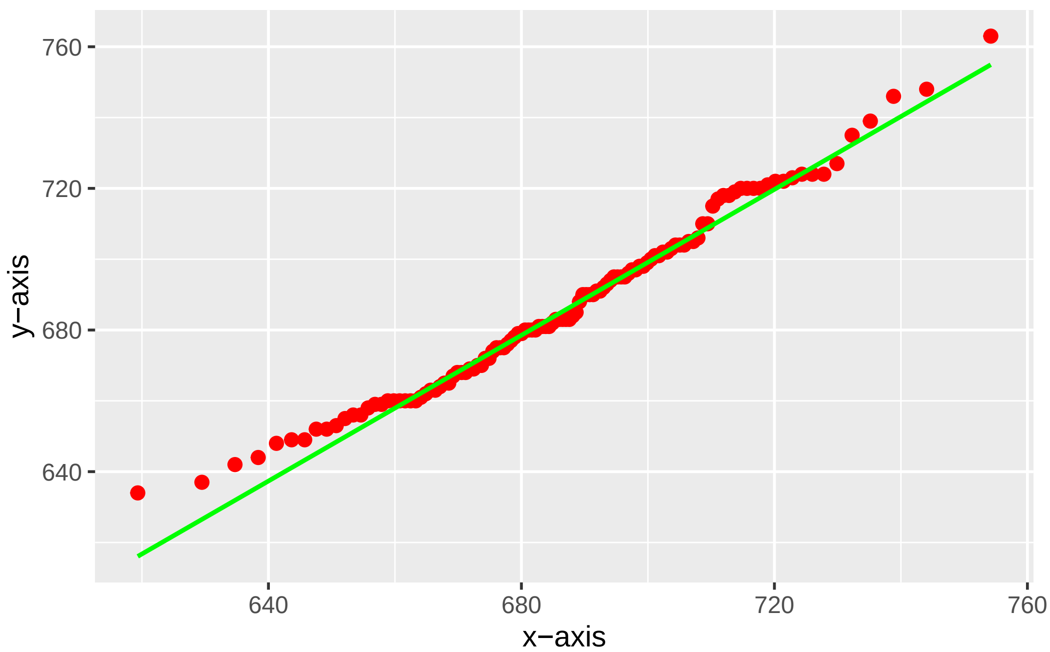
```
boxplot(data)
```




## f

Construct a normal probability plot for the data using R

```
ggplot(mapping = aes(sample = data)) + stat_qq_point(size = 2,color = "red") + stat_qq_line(c
```




### g

Is it reasonable to assume the data is normally distributed? Why or why not?

> The data is not normally distributed and is instead skewed right. This is because the box plot indicates a right skew and the normal probability plot also indicates a slight right skew.