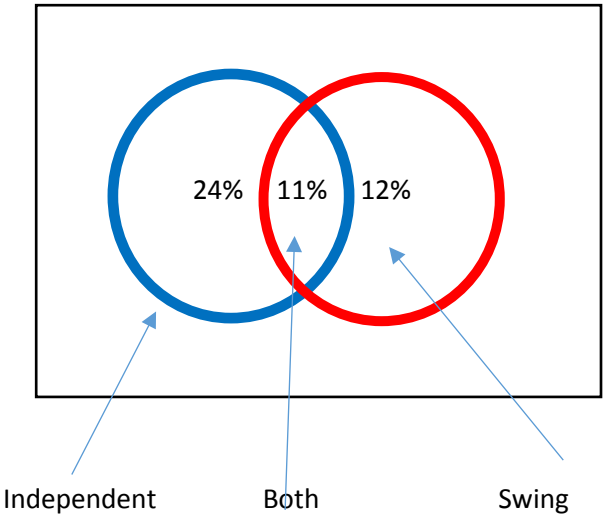
Assignment 1:

1)



a) Are being Independent and being a swing voter disjoint, i.e. mutually exclusive?
No, there are voters who are both Independent and swing. Therefore, they are not disjoint.

b) What percent of voters are Independent but not swing?
24% voters are independent but not swing (35-11=24%)

c) What percent of voters are Independent or swing voters?
Independent + Swing- (both independent & Swing)
35+23-11= 47%

d) What percent of voters are neither Independent nor swing voters?

100-(Independent or swing)

100-47= 53%

e) Is the event that someone is a swing voter independent of the event that someone is a political Independent?
P(Independent)* P (swing) = 0.35* 0.23 = 0.08
which does not equal P(Independent & swing) = 0.11, Therefore we can say that the events are **dependent**.

2)

Load the Felix Hernandez dataset in R.

```
Felix<-read.csv("FelixHernandez2015.csv")

head(Felix)
```

a) How many wins does Felix have this year?

Felix has 18 wins this year

Method 1:

```
sum(Felix[,"W"] == 1)
```

answer-> 18

Method 2:

```
table<-table(Felix[,"W"])

table
```

```
0  1
13 18
```

Frequency of 1 is 18

b) What is the mean, median, and mode number of strikeouts Felix threw over the 2015 season?

Mean:

```
> mean(Felix[,"SO"])
[1] 6.16129
```

Median:

```
> median(Felix[,"SO"])
[1] 6
```

Mode:

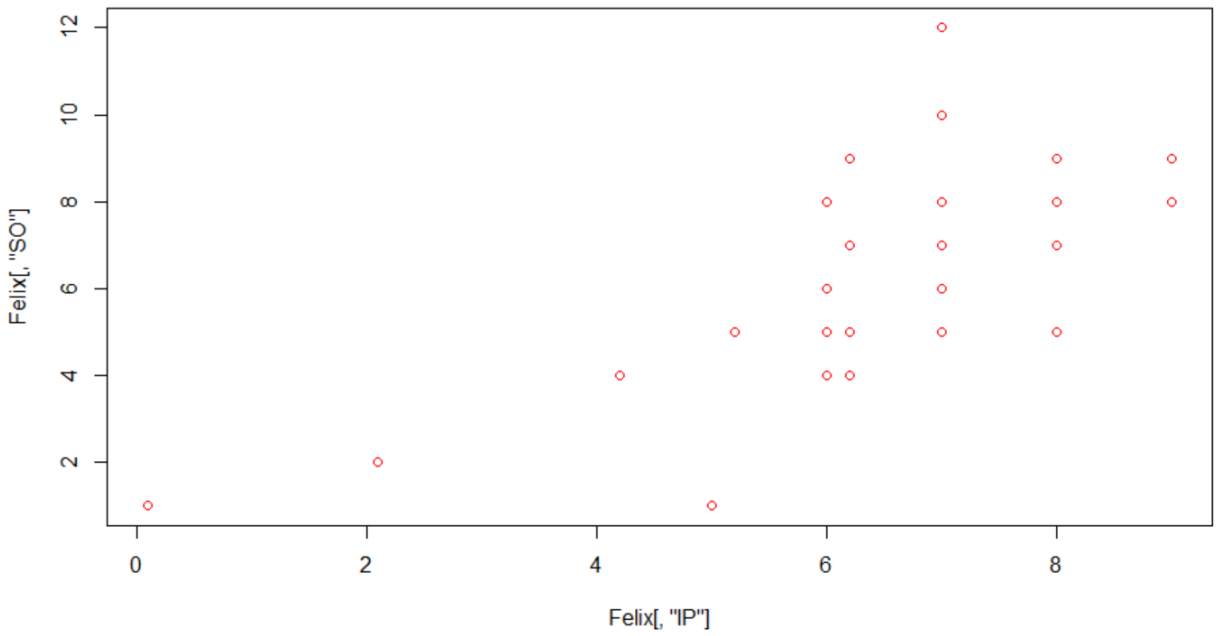
```
Mode <- function(x) {
```

```
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

> Mode(Felix[,"SO"])
[1] 5
```

c) Plot the relationship between innings pitched and strikeouts and between innings pitched and walks (base on balls). Describe the patterns you see (decreasing relationship? No relationship?).

Innings pitched and strikeouts:

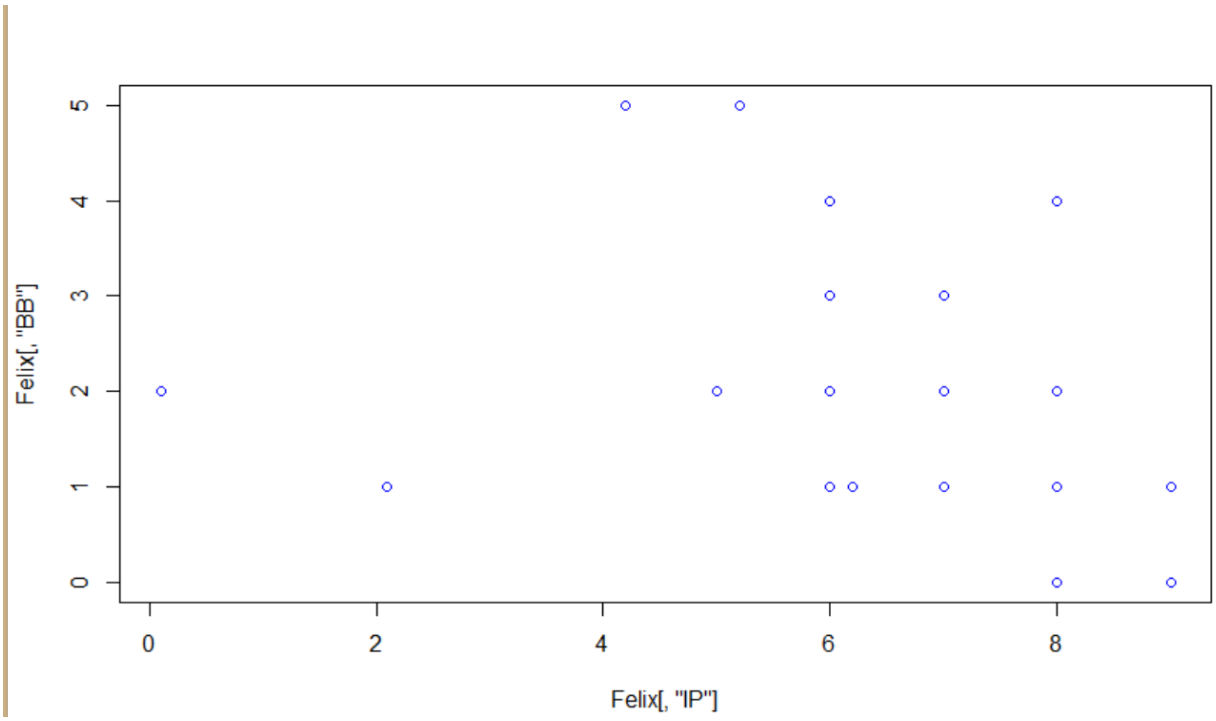
plot(Felix[,"IP"], Felix[,"SO"], col="red")



There is an increasing relationship- As the innings pitched increase, the strikeouts also increase.

Innings pitched and walks:

plot(Felix[,"IP"], Felix[,"BB"], col="blue")



There is a decreasing relationship. As the innings pitched increase, the base on balls decrease.

d) Calculate the correlation coefficient between innings pitched and strikeouts and between innings pitched and walks. Do these align with what you saw in the plots?

```
> cor(Felix[,"IP"], Felix[,"SO"])
[1] 0.6816081
```

Positive value indicates that it is an increasing relationship as observed from the plot.

```
> cor(Felix[,"IP"], Felix[,"BB"])
[1] -0.2638496
```

Negative value indicates that it is a decreasing relationship as observed from the plot.
Therefore it aligns with the plots.

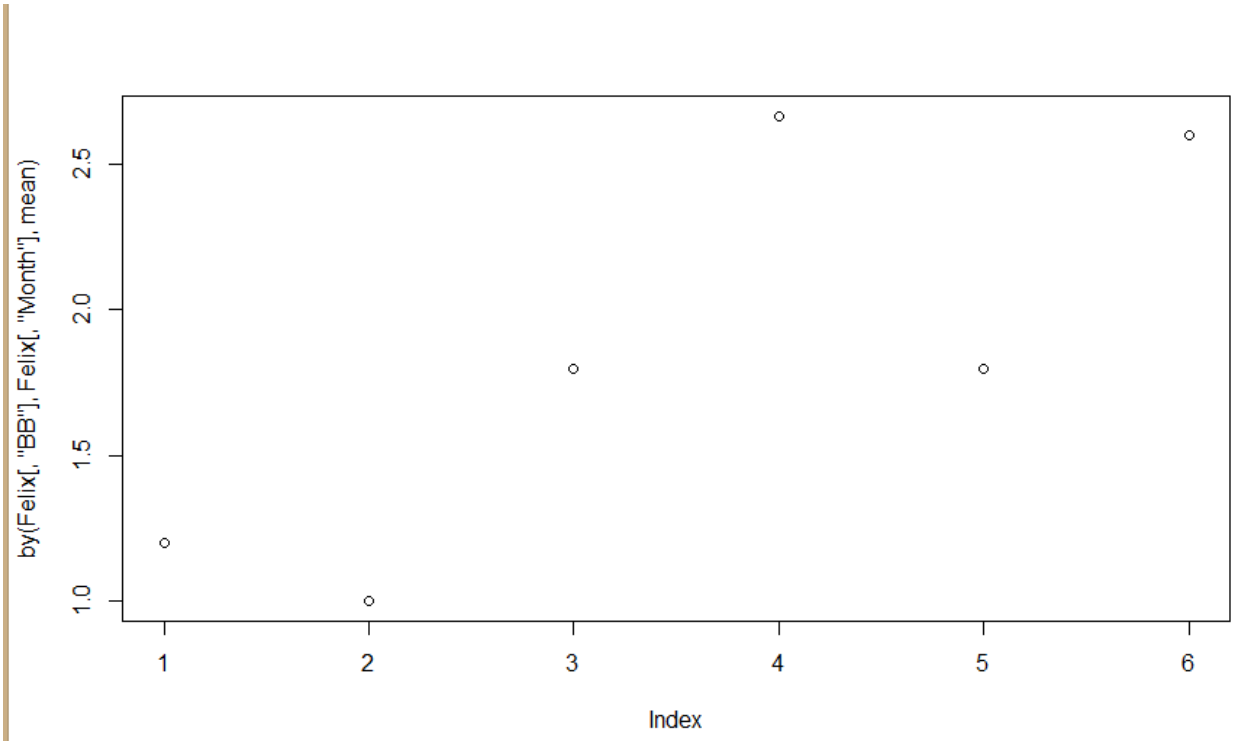
e) Calculate the mean and variance of walks by month (hint: use the by() function like in lab). Do you see changing mean walks over time? What about the variability over time? What might the pattern mean?

Mean of walks by month:

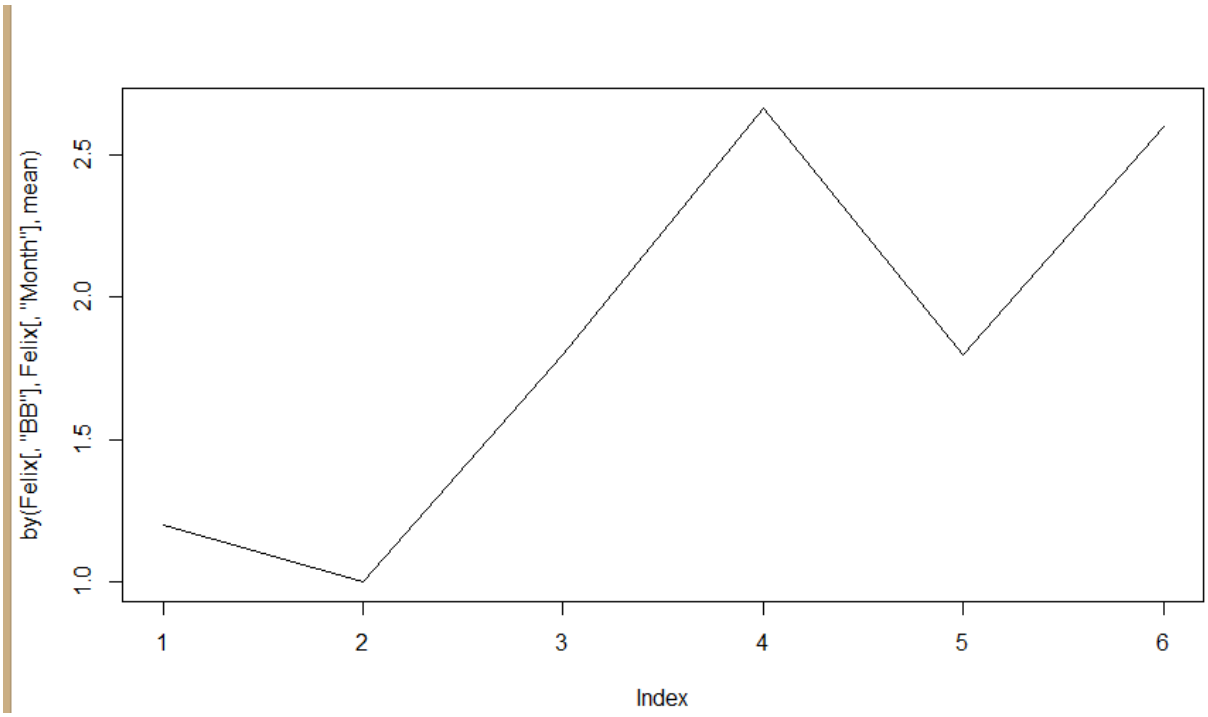
```
+ by(Felix[,"BB"], Felix[,"Month"], mean)

   Felix[, "Month"]: Apr
   [1] 1.2
```

```
------------------------------------------------------------------------
------------------------------------------------------
Felix[, "Month"]: Aug
[1] 1
------------------------------------------------------------------------
------------------------------------------------------
Felix[, "Month"]: Jul
[1] 1.8
------------------------------------------------------------------------
------------------------------------------------------
Felix[, "Month"]: Jun
[1] 2.666667
------------------------------------------------------------------------
------------------------------------------------------
Felix[, "Month"]: May
[1] 1.8
------------------------------------------------------------------------
------------------------------------------------------
Felix[, "Month"]: Sep
[1] 2.6
```

plot(by(Felix[,"BB"], Felix[,"Month"], mean))



plot(by(Felix[,"BB"], Felix[,"Month"], mean), type="l")

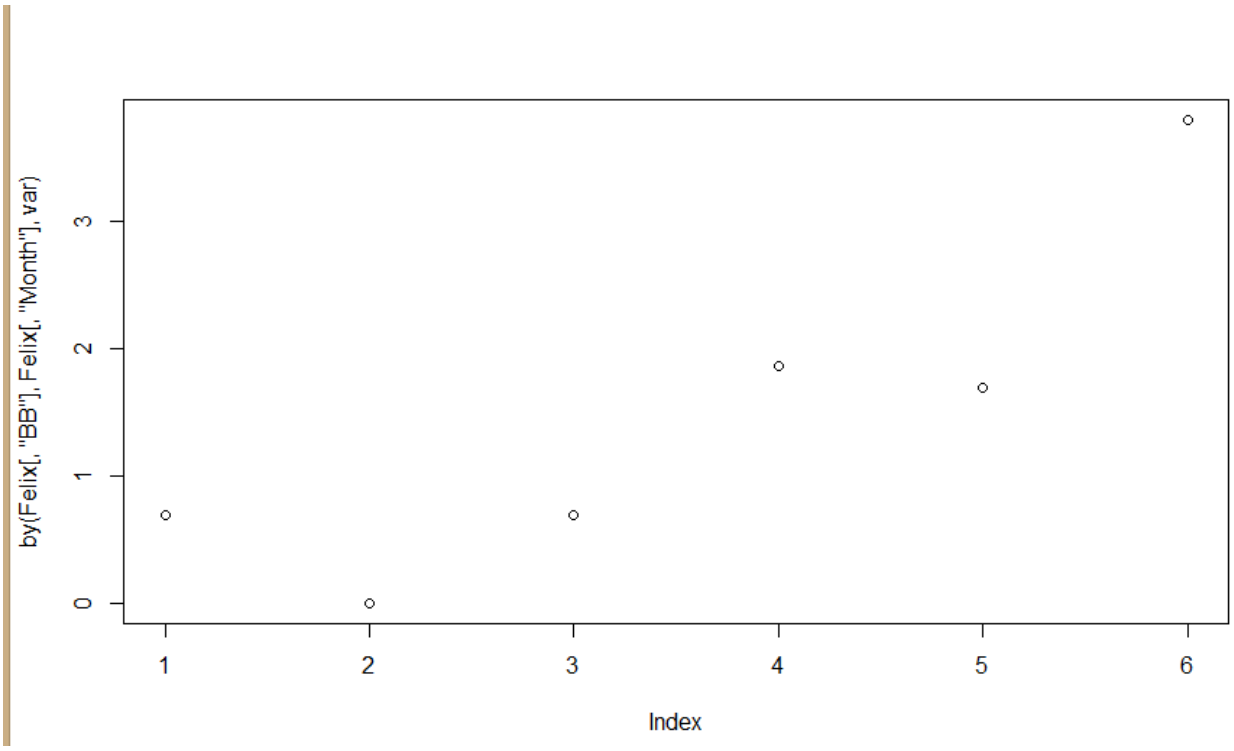


The mean is changing every month with lows and highs depending on average performance.

Variance of walks by month:

```
> by(Felix[,"BB"], Felix[,"Month"], var)
Felix[, "Month"]: Apr
[1] 0.7
------------------------------------------------------------------------------------------------
Felix[, "Month"]: Aug
[1] 0
------------------------------------------------------------------------------------------------
Felix[, "Month"]: Jul
[1] 0.7
------------------------------------------------------------------------------------------------
Felix[, "Month"]: Jun
[1] 1.866667
------------------------------------------------------------------------------------------------
Felix[, "Month"]: May
[1] 1.7
------------------------------------------------------------------------------------------------
Felix[, "Month"]: Sep
[1] 3.8


>
```

plot(by(Felix[,"BB"], Felix[,"Month"], var))



Since the mean is changing depending on the performance of the player, the variance which is the degree to which each point differs from mean, is also changing

f) Does Felix win more on the road or at home?

```
> table1<-table(Felix[,"away"],Felix[,"W"])
> table1

   0  1
0  6 11
1  7  7
```

Wins at home=11
Wins on the road=7
Therefore, **Felix wins more at home**.

g) Load the other data set containing similar records for Randy Johnson in 1995. Does Randy Johnson outperform Felix in terms of strikeouts across the 1995 season?

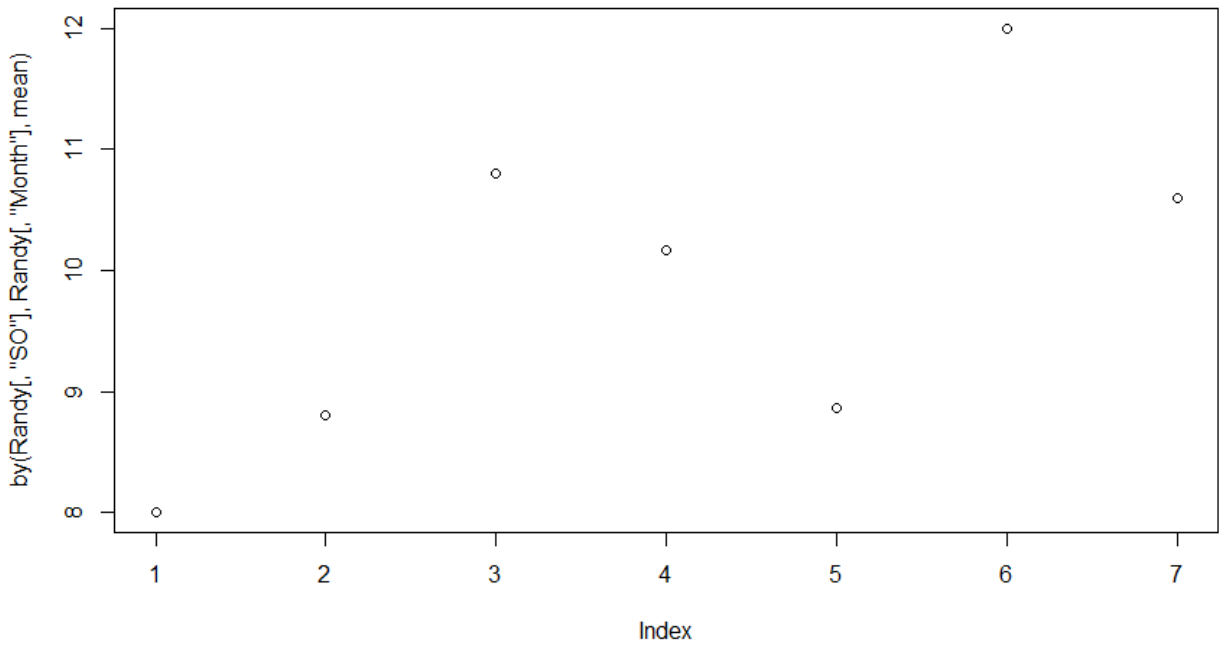
```
> Randy<-read.csv("RandyJohnson1995.csv")
> mean(Randy[,"SO"])
```

```
[1] 9.8
```

Mean by month:

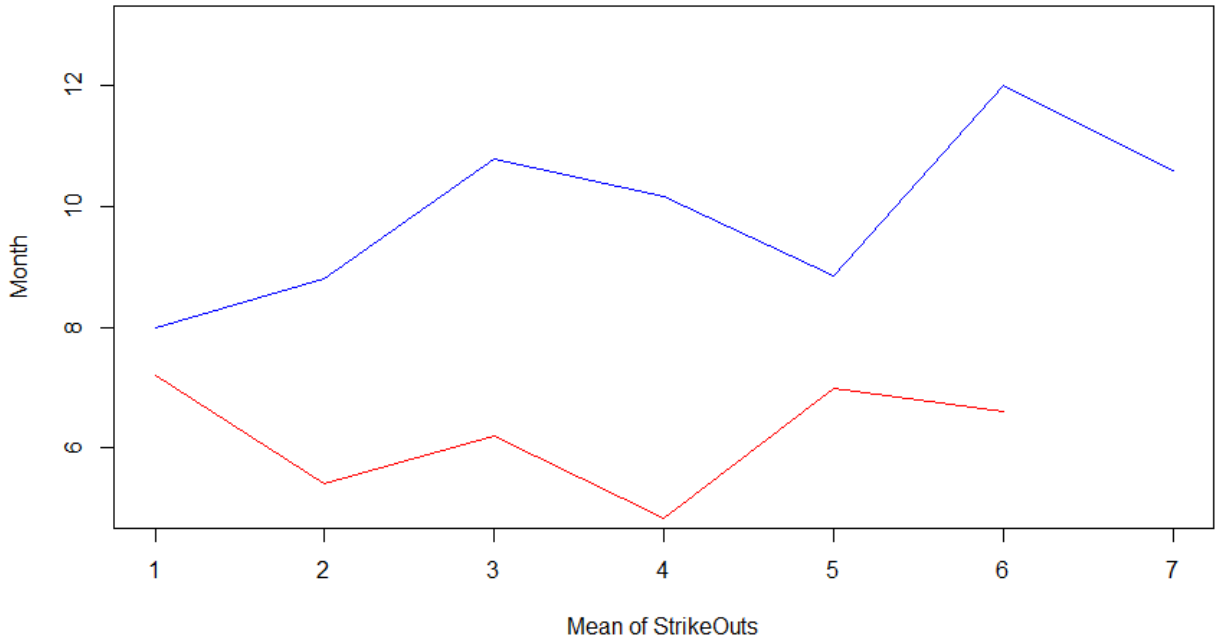
```
> by(Randy["SO"], Randy["Month"], mean)
Randy[, "Month"]: Apr
[1] 8
------------------------------------------------------------------------------
------------------------------------------------
Randy[, "Month"]: Aug
[1] 8.8
------------------------------------------------------------------------------
------------------------------------------------
Randy[, "Month"]: Jul
[1] 10.8
------------------------------------------------------------------------------
------------------------------------------------
Randy[, "Month"]: Jun
[1] 10.16667
------------------------------------------------------------------------------
------------------------------------------------
Randy[, "Month"]: May
[1] 8.857143
------------------------------------------------------------------------------
------------------------------------------------
Randy[, "Month"]: Oct
[1] 12
------------------------------------------------------------------------------
------------------------------------------------
Randy[, "Month"]: Sep
[1] 10.6
```

plot(by(Randy["SO"], Randy["Month"], mean))



Comparing both means:

```
by(Felix[,"SO"], Felix[,"Month"], mean)
plot(by(Felix[,"SO"], Felix[,"Month"], mean))

#change limits for Randy to match Felix
plot(by(Randy[,"SO"], Randy[,"Month"], mean), ylim=c(5,13),col='blue',type="l",xlab="Mean
of StrikeOuts", ylab="Month")

points(by(Felix[,"SO"], Felix[,"Month"], mean), col='red',type="l")
```



We can see Randy Johnson has outperformed Felix in terms of strikeouts across the 1995 season.

3) Sophia who took the Graduate Record Examination (GRE) scored 156 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal.

Verbal reasoning:
$\mu=151$
$\sigma=7$

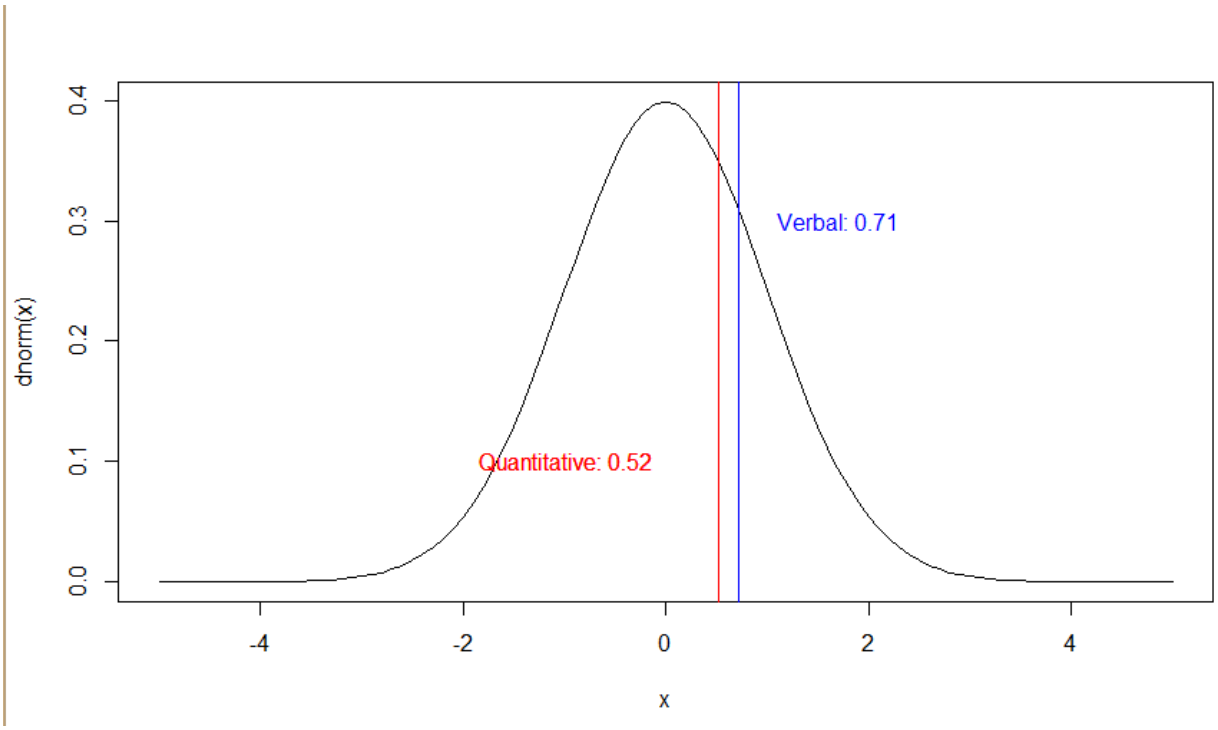
Quantitative reasoning:
μ=153
σ=7.67

a) What is Sophia's Z-score on the Verbal Reasoning section? On the Quantitative Reasoning section?

Verbal Reasoning:
$Z = (x - \mu)/\sigma = (156 - 151)/7$
$= (x - \mu)/\sigma = (156 - 151)/7 = 0.7142$
Quantitative Reasoning:
$Z = (x - \mu)/\sigma = (157 - 153)/7.67 = 0.5215$

b) Draw a standard normal distribution curve and mark these two Z-scores.

```
curve(dnorm, from = -5, to=5)

abline(v=0.7142, col="blue")

abline(v=0.5215, col="red")

text(0.7142857+1, 0.3, "Verbal: 0.71",col="blue")

text(0.5215124-1.5, 0.1, "Quantitative: 0.52", col="red")
```



c) Relative to others, which section did she do better on?
She did better on Verbal Reasoning, because that Z-score is higher.

d) Find her percentile scores for the two exams.

Verbal reasoning
```
+ pnorm(0.714)
[1] -0.7623864 =76%-- therefore it is 76th percentile
```

```
Quantative reasoning
> pnorm(0.5215)
[1] 0.6989907=70%-- 70th percentile
```

e) What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?

Verbal reasoning
```
> 1-pnorm(0.714)
[1] 0.2376136
```

About 24% did better on the Quantitative Reasoning section

Quantative reasoning

```
> 1-pnorm(0.5215)
[1] 0.3010093=30%
```

About 30% did better on the Quantitative Reasoning section.

f) Explain why simply comparing her raw scores from the two sections would lead to the incorrect conclusion that she did better on the Quantitative Reasoning section (2-3 sentences).

The raw scores are 156 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section- both values are very close
Also the mean and standard deviation values are close.
Currently the verbal score is lower than the raw quantitative score. It is possible that someone might have a higher raw score in one section than the other, but in fact do worse in that section. When we calculate the percentiles we get an understanding of her performance compared to others in class which is important to understand how she performed in the section.
The percentiles being 76 for verbal and 70 for quantitative, we can compare better.