Assignment_1

Abhinav Garg

1/15/2017

1

Pew Research Survey

1) A 2012 Pew Research survey asked 2,373 randomly sampled registered voters their political affiliation (Republican, Democrat, or Independent) and whether or not they identify as swing voters.35% of respondents identified as Independent, 23% identified as swing voters, and 11% identified as both.

a) Are being Independent and being a swing voter disjoint, i.e. mutually exclusive? Ans : For two events A and B to be disjoint, P(AB) = P(A)*P(B) In this scenario, Let I = being Independent S = being Swing Voter P(IS) = 0.11, P(I) = .35, P(S) = 0.23 P(IS) is not P(I)*P(S) Therefore these two events are not disjoint

b) Percent of voters that are Independent but not swing, is percent of Independent minus the percent of Independent and swing voters 35 % - 11 % = 24 % = 0.24

c) Percent of swing voters or Independent voters P(I U S) = P(I) + P(S) - P(IS) = 35% + 23% - 11% = 47% = .47

d) Percent of neither Independent nor swing voters = P(I U S)’ = 1 - P(I U S) = 1 - 0.47 = .53 = 53 %

e) The two events are not independent of each other since there is an intersection between the two events of 11 %

2

Loading Data

```
setwd("/Users/abhi/Documents/UW/Courses/Winter_Quarter_17/INFX_573/Assignments/
Assignment_1")
d1 <- read.csv("FelixHernandez2015.csv")
```

Number of Wins

```
sum(d1$W)
```

```
## [1] 18
```

Mean, Median and Mode

```
#Mean
mean(d1$SO)
```

```
## [1] 6.16129
```

```
#Median
median(d1$SO)
```

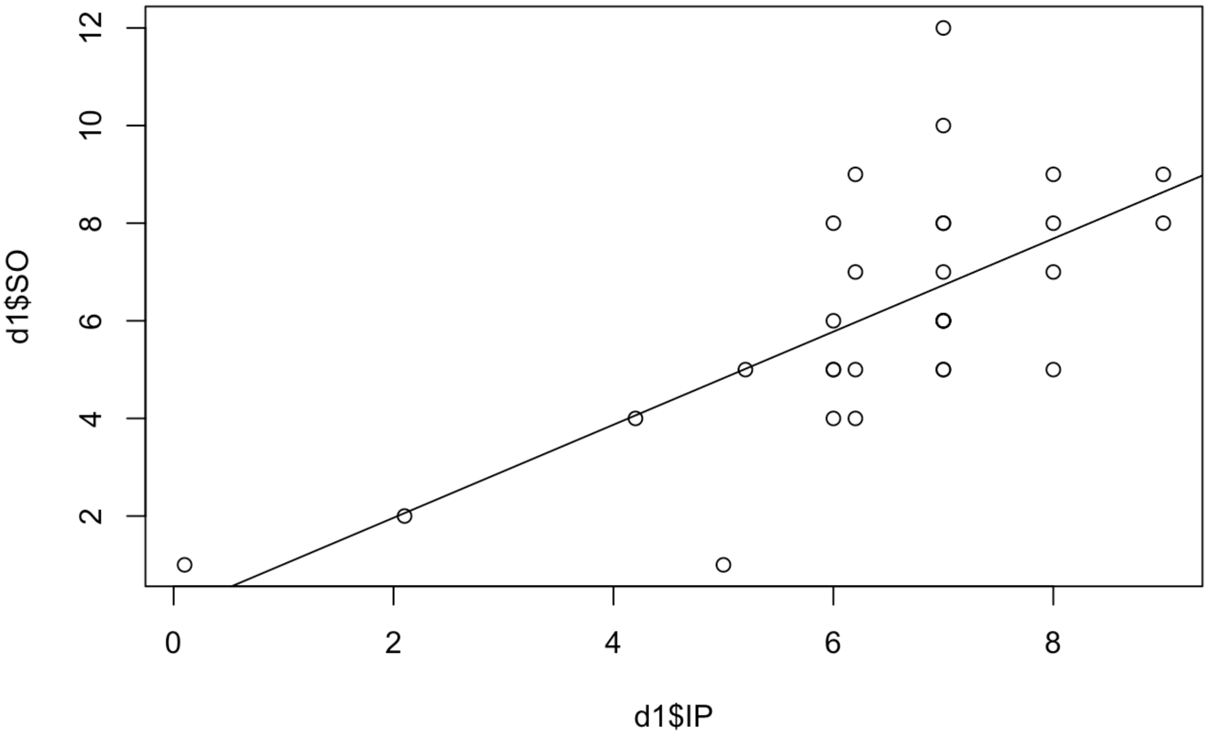
```
## [1] 6
```

```
Mode <- function(x) {
    ux <- unique(x)
    ux[which.max(tabulate(match(x, ux)))]
}
Mode(d1$SO)
```

```
## [1] 5
```

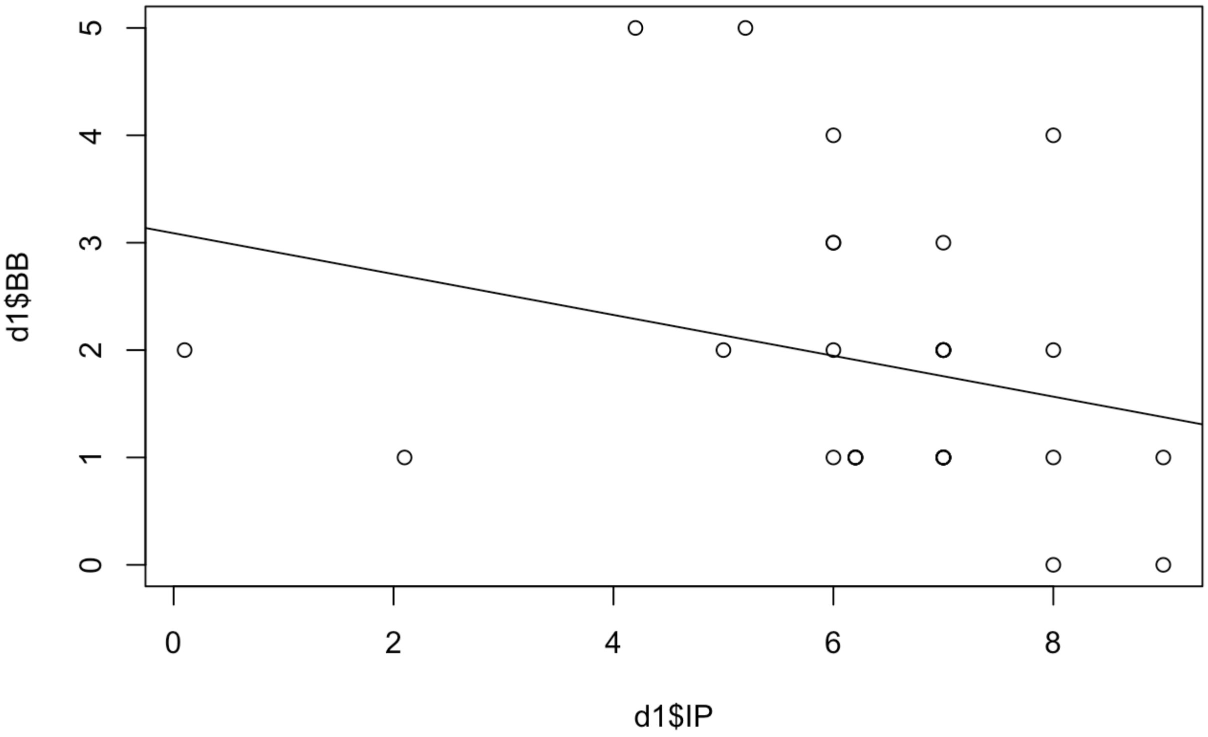
Plotting Relationships

```
plot(d1$IP,d1$SO)
abline(lm(d1$SO ~ d1$IP))
```



As we can see there is an increase in Strikeouts as the Number of Innings pitched increases

```
plot(d1$IP,d1$BB)
abline(lm(d1$BB ~ d1$IP))
```



As we can see there is a decrease in Base on balls (walks) as the Number of Innings pitched increases

Correlation Coefficient

```
cor(d1$IP,d1$SO)
```

```
## [1] 0.6816081
```

```
cor(d1$IP,d1$BB)
```

```
## [1] -0.2638496
```

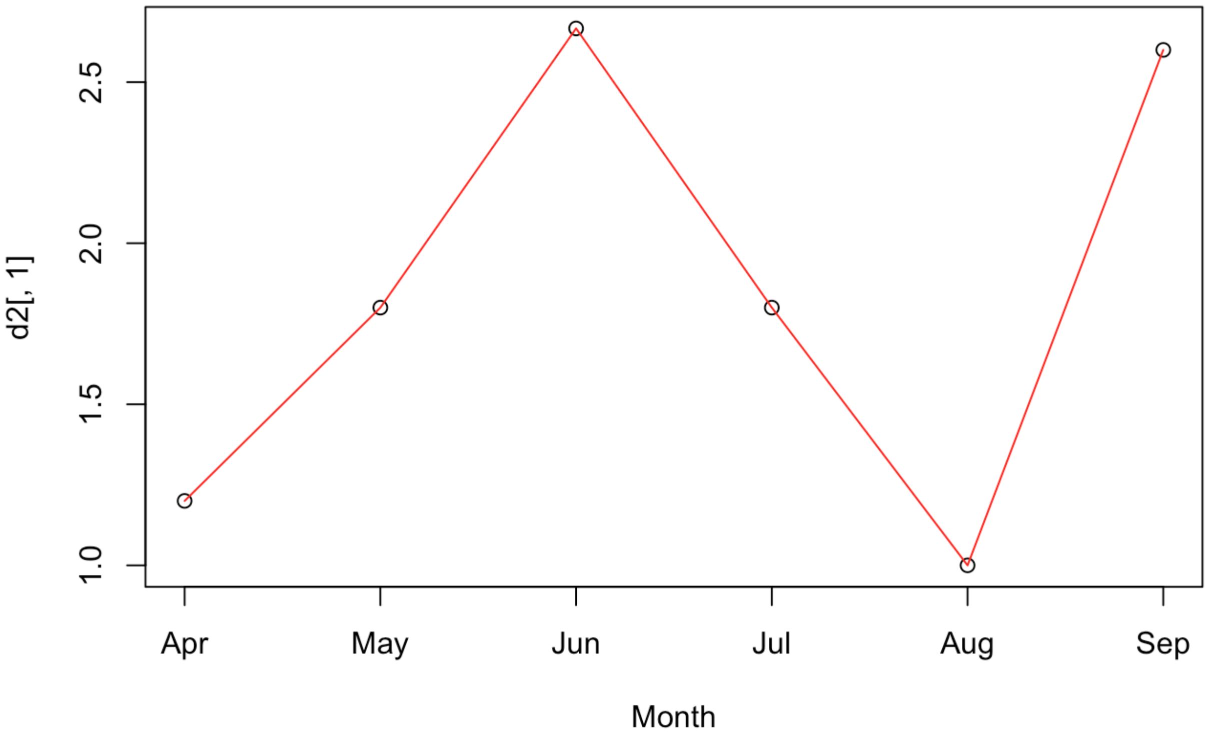
As we can see both correlation coefficients align with what we saw in the plots

Mean and Variance by Month

```
d2 = as.matrix(by(d1$BB, d1$Month,mean))
mo2Num <- function(x) match(tolower(x), tolower(month.abb))
d2 = cbind(d2,mo2Num(levels(factor(d1$Month))))
d2 = d2[order(d2[,2]),]
d2
```

```
##          [,1] [,2]
## Apr 1.200000    4
## May 1.800000    5
## Jun 2.666667    6
## Jul 1.800000    7
## Aug 1.000000    8
## Sep 2.600000    9
```

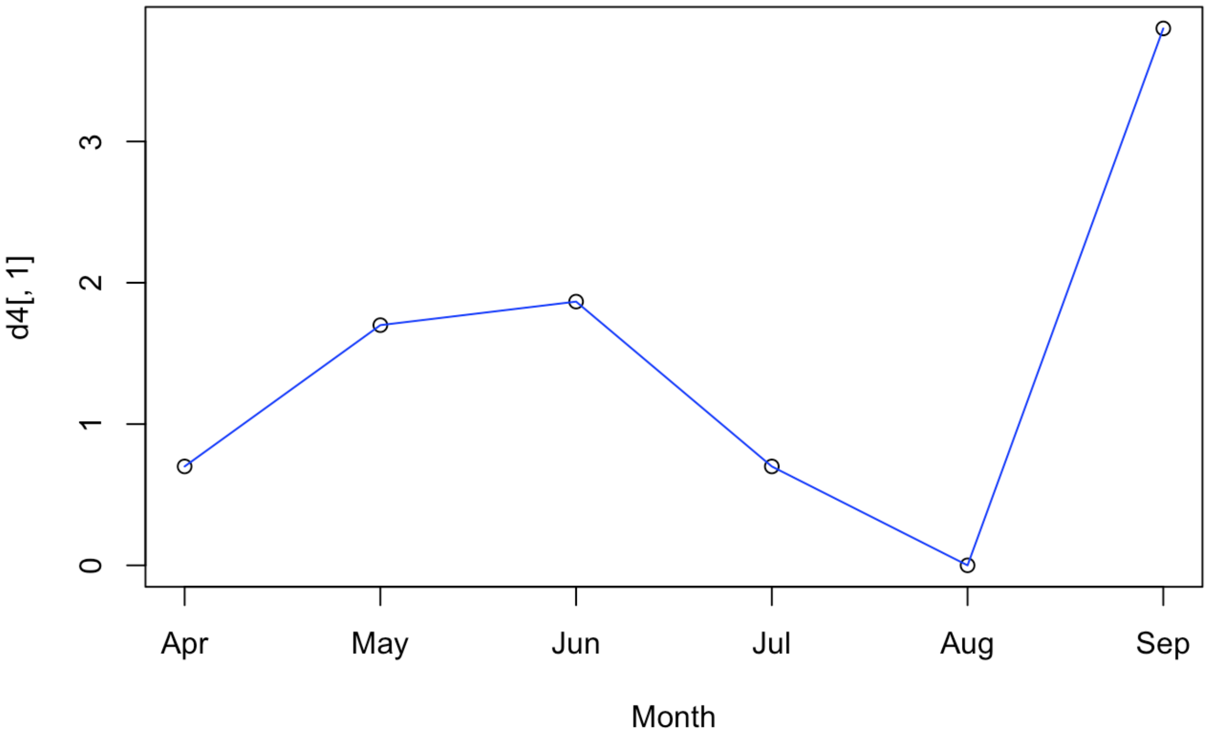
```
plot(d2[,1],xlab = "Month", xaxt="n")
axis(1,at = 1:6, labels=c("Apr","May","Jun","Jul","Aug","Sep"))
lines(d2[,1], col="red")
```



```
d4 = as.matrix(by(d1$BB, d1$Month,var))
mo2Num <- function(x) match(tolower(x), tolower(month.abb))
d4 = cbind(d4,mo2Num(levels(factor(d1$Month))))
d4 = d4[order(d4[,2]),]
d4
```

```
##         [,1] [,2]
## Apr 0.700000    4
## May 1.700000    5
## Jun 1.866667    6
## Jul 0.700000    7
## Aug 0.000000    8
## Sep 3.800000    9
```

```
plot(d4[,1],xlab = "Month", xaxt="n")
axis(1,at = 1:6, labels=c("Apr","May","Jun","Jul","Aug","Sep"))
lines(d4[,1], col="Blue")
```



The mean and variance both vary over time. The pattern that we see is a fluctuation in mean performance over the season. However Felix's performance varies a lot from the mean in the month of

September. Which shows that towards the end of the season his performance fluctuated quite a lot. Also as the mean number of walks increases, there is a decline in performance. So in the plot with mean of walks per month, we see that his performance declines in the month of June and September.

Wins by Home or Away

```
by(d1$W,d1$away,sum)
```

```
## d1$away: 0
## [1] 11
## --------------------------------------------------------
## d1$away: 1
## [1] 7
```

From this we see that Felix wins more at Home : 11 than Away: 7

Loading from other data set

```
d2 <- read.csv("RandyJohnson1995.csv")
Felix <- sum(d1$SO)
Randy <- sum(d2$SO)
df <- cbind(Felix,Randy)
df
```

```
##      Felix Randy
## [1,]   191   294
```

As we can see Randy outperforms Felix in terms of Strike Outs

3

Sophia who took the Graduate Record Examination (GRE) scored 160 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.

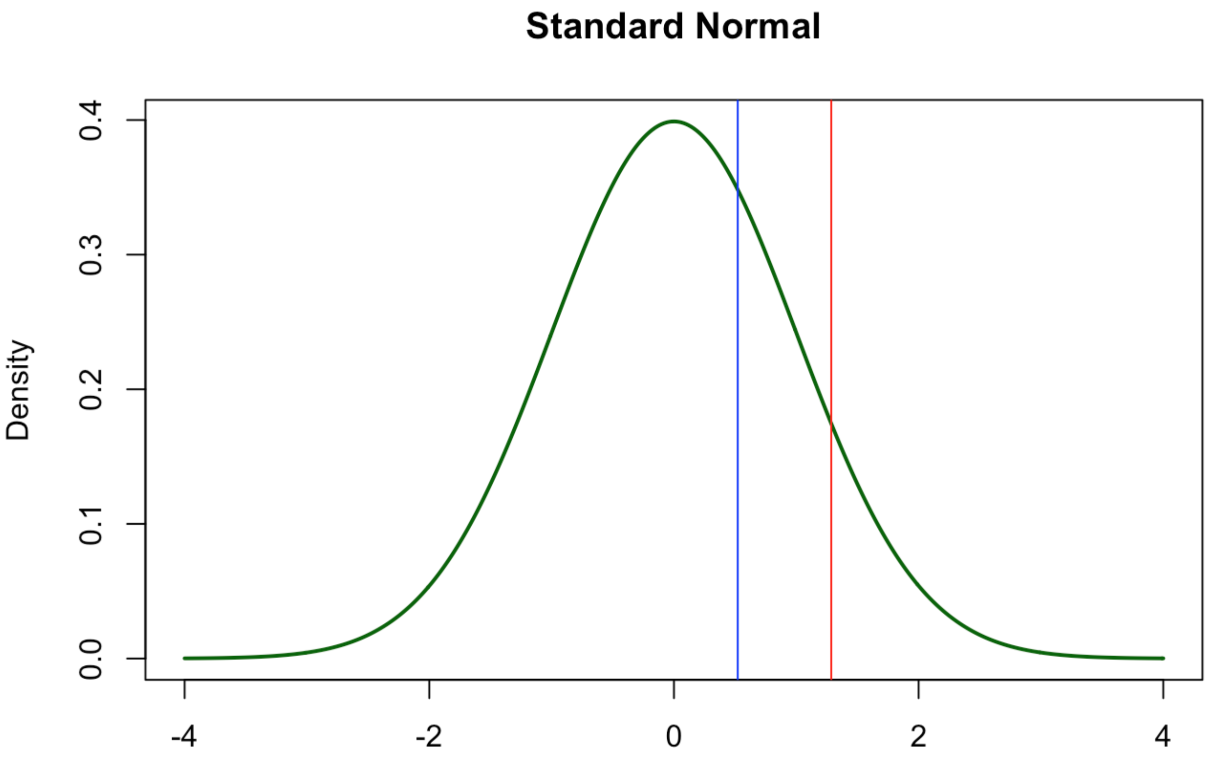
a) Sophia’s Z-Score

```
On the Verbal Reasoning Section is :
160 - 151 / 7
= 1.2857

On the Quantitaive Reasoning Section is :
157 - 153 / 7.67
= 0.5215
```

b)

```
xseq <- seq(-4,4,.01)
densities  <- dnorm(xseq, 0, 1)
plot(xseq, densities, col = "darkgreen", xlab = "", ylab = "Density", type = "l
", lwd = 2, main = "Standard Normal")
legend(2000,9.5,c("Verbal,Quantitative"),lty=c(1,1),lwd=c(2.5,2.5), col=c("red"
,"blue"))
abline(v = 1.2857,col = "red")
abline(v = 0.5215,col = "blue")
```



c) Relative to others she did well on the Verbal Section d)

Percentile for Verbal Section

```
pnorm(1.2857)*100
```

```
## [1] 90.07261
```

Percentile for Quantitative Section

```
pnorm(0.5215)*100
```

```
## [1] 69.89907
```

e)

Percent of people who did better on Verbal Section

```
(1 - pnorm(1.2857))*100
```

```
## [1] 9.927389
```

Percent of people who did better on Quantitative Section

```
(1 - pnorm(0.5215))*100
```

```
## [1] 30.10093
```

f)

Simply comparing the Raw scores would lead to an incorrect solution that she did better on the Quantitative Section because the mean for the Quantitative Section is higher than the mean for the Verbal section however we also have to consider the standard deviations for the two which helps us get a more better statistic in distinguishing her performance relevant to the other students