Assignment-1-Solution

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?

No they are not disjoint, as it is indicated that 11% of the registered voters have identified as both.

b) What percent of voters are Independent but not swing

24% of voters are Independent but not swing.

c) What percent of voters are Independent or swing voters?

As the Independent and swing variable are disjoint,

P(I or S) = P(I) + P(S) - P(I and S)

where P(I or S) = Probability of voters that are Independent or swing voters P(I) = Probability of voters that are Independent P(S) = Probability of voters that are swing P(I and S) = Probability of voters that are both

Therefore,

P(I or S) = 0.35 + 0.23 - 0.11 P(I or S) = 0.47

Percent of voters that are Independent or swing voters = 47%

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

P(not I or not S) = 1 - P(I or S) = 1 - 0.47 = 0.53

Percent of voters that are neither Independent nor swing voters = 53%

e) Is the event that someone is a swing voter independent of the event that someone is a political Independent?

To verify if two variables are independent, consider :

P(I) * P(S) = 0.35 * 0.23 = 0.0805

P(I and S) = 0.11

As, P(I) * P(S) != P(I and S)
The two events are dependent.

2) Felix Hernandez dataset

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

a) How many wins does Felix have this year?

```
table(FH$W)

##
##  0  1
## 13 18

nrow(FH[FH[,"W"] == 1,])

## [1] 18
```

Felix has 18 Wins this year.

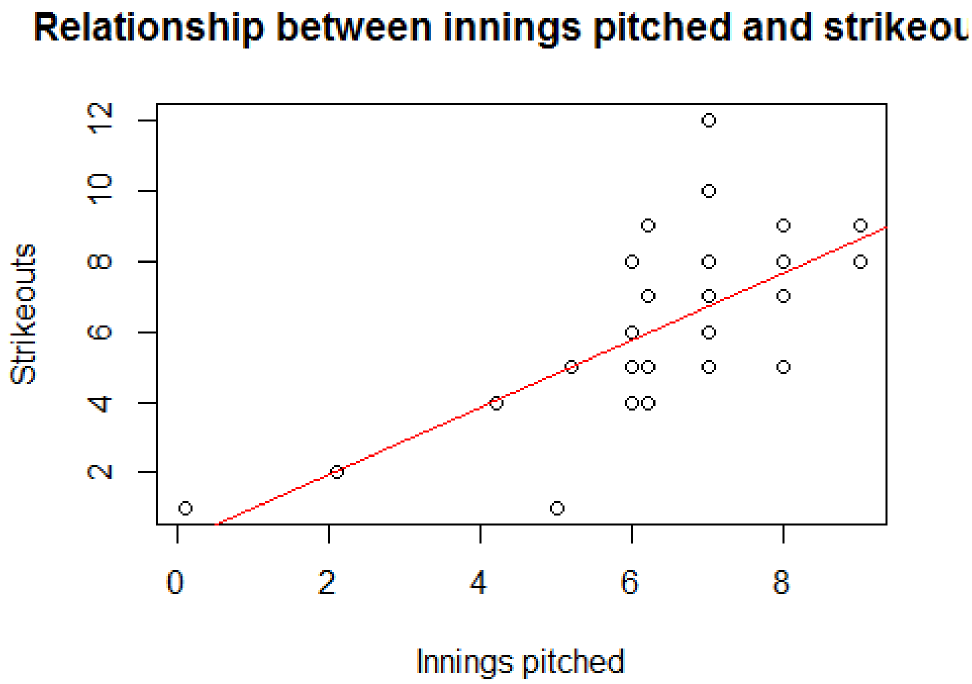
b) What is the mean, median, and mode number of strikeouts Felix threw over the 2015 season? Use this function to calculate the mode:
Mode <- function(x) {
ux <- unique(x)
ux[which.max(tabulate(match(x, ux)))]
}

```
summary(FH$SO)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.000   6.000   6.161   8.000  12.000

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

Mode(FH$SO)

## [1] 5
```

Mean = 6.161
Median = 6
Mode = 5

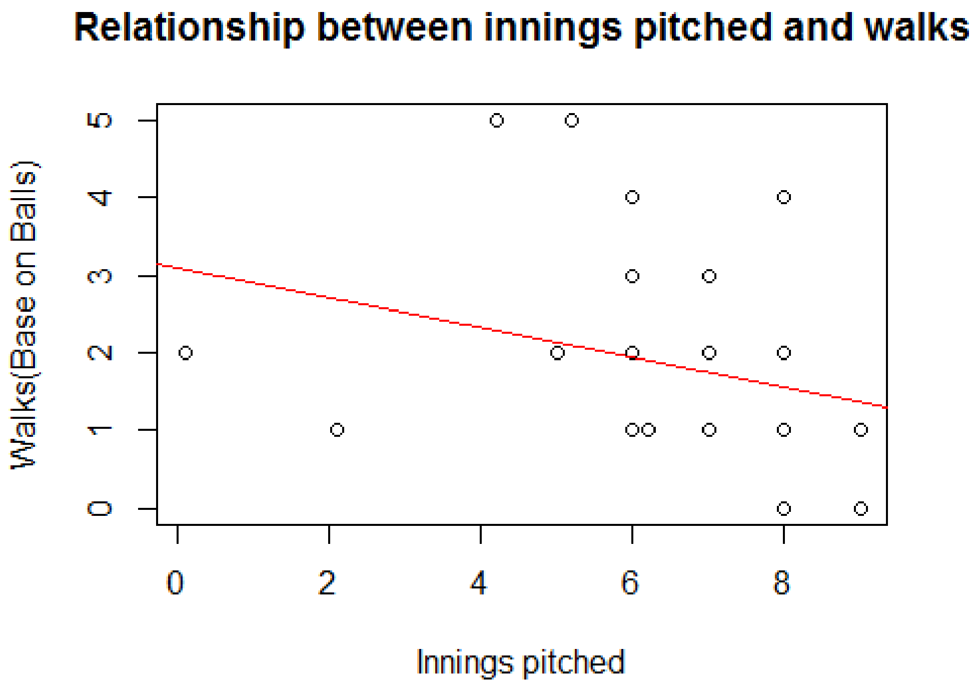
c) Plot the relationship between innings pitched and strikeouts and between innings pitched and walks (base on balls). Describe the patterns you see (decreasing relationship? No relationship?).

```
plot(FH$IP, FH$SO,
     xlab = "Innings pitched", ylab = "Strikeouts", main = "Relationship
between innings pitched and strikeouts")
mod1 <- lm(FH$SO ~ FH$IP)
abline(mod1, lwd = 1, col = 2)
```



Innings pitched and strikeouts have a positive (increasing) relationship.

```
plot(FH$IP, FH$BB,
     xlab = "Innings pitched", ylab = "Walks(Base on Balls)", main =
"Relationship between innings pitched and walks")

mod2 <- lm(FH$BB ~ FH$IP)
abline(mod2, lwd = 1, col = 2)
```



Innings pitched and walks have a negative (decreasing) relationship.

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?

```
round(cor(FH$IP, FH$SO, use="everything", method = c("pearson")),2)

## [1] 0.68
```

Based on the rule of thumb for interpreting the size of a Correlation Coefficient, 0.68 indicates a moderate linear correlation.

```
round(cor(FH$IP, FH$BB, use="everything", method = c("pearson")),2)

## [1] -0.26
```

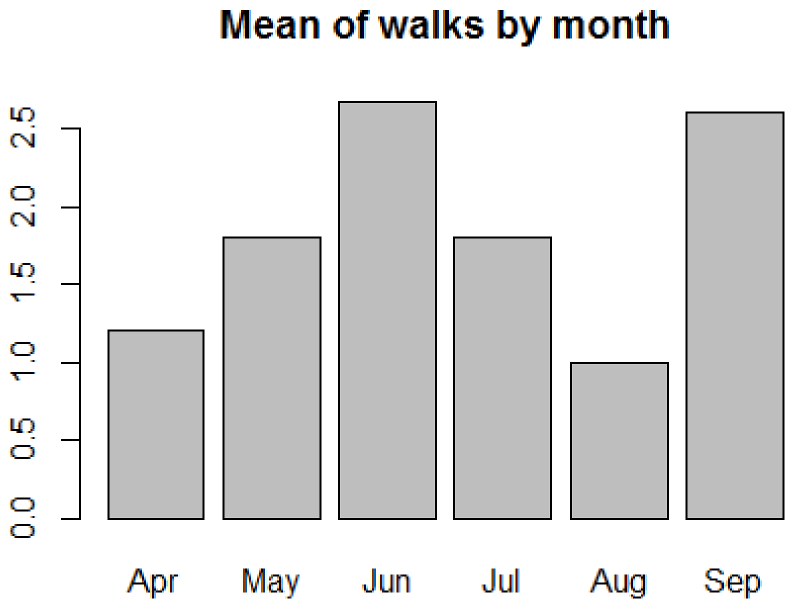
Based on the rule of thumb for interpreting the size of a Correlation Coefficient, -0.26 indicates little (weak) if any negative correlation.

The correlation coeefficient aligns with the visual interpretation of the plots. However, it helped determine the strength of the correlation.

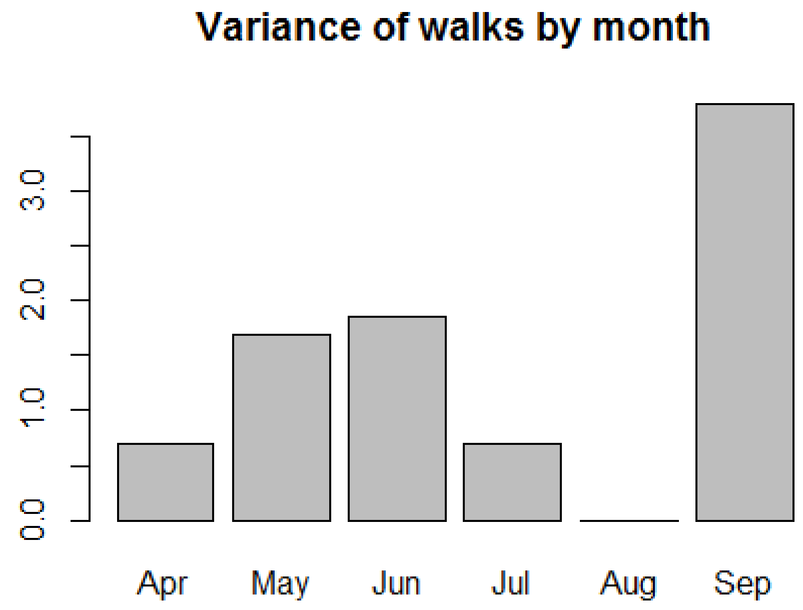
e) Calculate the mean and variance of walks by month (hint: use the by() function like in lab). Do you see changing mean walks over time? What about the variability over time? What might the pattern mean?

```
month <- ordered(FH$Month, levels = c("Apr", "May", "Jun", "Jul",
"Aug","Sep"))
```

```r
a <- by(FH[,"BB"], month, mean)
barplot(a, main = "Mean of walks by month")
```



```r
b <- by(FH[,"BB"], month, var)
barplot(b, main = "Variance of walks by month")
```



The mean walks change over time. There is a upward trend from the months Apr to Jun, followed by a downward slope from Jun to Aug with a peak at the end of baseball season in Sept. The variance value also seems to vary over the Baseball season in the same pattern as the mean. There is a high variance observed in the month of Sept. The peak in Sept is possibly as its the end of the baseball season.

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

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

FH %>%
group_by(away) %>%
summarise(mean(W))

## # A tibble: 2 × 2
##    away `mean(W)`
##   <int>     <dbl>
## 1      0 0.6470588
## 2      1 0.5000000
```

Felix wins more at home. The number of wins at home are 11/17 = 64.7% The number of wins away are 7/14 = 50%

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?

```
RJ <- read.csv("RandyJohnson1995.csv")

sum(FH$SO)

## [1] 191

sum(RJ$SO)

## [1] 294
```

Yes, Randy Johnson outperforms Felix in terms of strikeouts. Randy has 103 more strikeouts than Felix in the 1995 season.

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.

Given data: V = 160 Q = 157

Population Mean V = 151 Population s.d V = 7

Population Mean Q = 153 Population s.d Q = 7.67

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

```
pop_mean_v <- 151
pop_sd_v <- 7
z_v <- (160 - pop_mean_v)/pop_sd_v

pop_mean_q <- 153
pop_sd_q <- 7.67
z_q <- (157 - pop_mean_q)/pop_sd_q
```

Z-score on the Verbal Reasoning

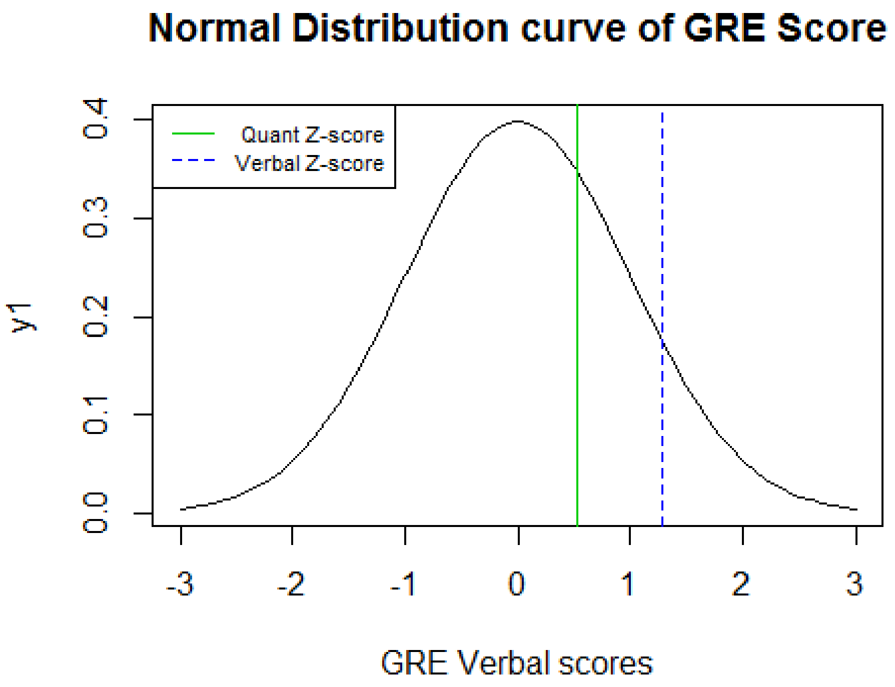
```
z_v

## [1] 1.285714
```

Z-score on the Quantitative Reasoning

```
z_q

## [1] 0.5215124
```

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

```
x1   <- seq(-3.0,3.0, by=0.1)
y1   <- dnorm(x1,mean=0, sd=1)
plot(x1,y1,
     type="l",
     main = "Normal Distribution curve of GRE Score",
     xlab = "GRE Verbal scores")
legend("topleft", cex = 0.70,legend=c(" Quant Z-score", "Verbal Z-
score"),lty=c(1,2), lwd = c(1,1), col= c(3,4))
abline(v=z_q, lty = 1, lwd=1, col = 3)
abline(v=z_v, lty = 2, lwd=1, col = 4)
```



c) Relative to others, which section did she do better on?

She did better in the Verbal section which is indicated through a higher z-score of 1.286 as compared to the zscore of her quantitative score.

d) Find her percentile scores for the two exams.

Sophia's Verbal Percentile score is

```
pnorm(z_v) * 100

## [1] 90.07286
```

Sophia's Quant Percentile score is

```
pnorm(z_q) * 100

## [1] 69.89951
```

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

Percent of test takers better that did better than Sophia on Verbal Reasoning

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

## [1] 9.92714
```

Percent of test takers better that did better than Sophia on Quantitative Reasoning

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

## [1] 30.10049
```

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 very close in value and not conclusive that she did better in one of the sections. It clearly does not indicate her performance in comparison to the other test takers. However, calculating percentiles helps us understand that she performed better than 90 percent of other test takers in Verbal and 70 percent better than the others in quantitative which is a considerable difference.