

Week 9 Assignment

David Russo

3/12/2017

10.7 #7

```
# load the data
data("USArrests")

# scale the data so that each variable has a mean of 0 and a variance of 1
usa_scaled <- as.data.frame(scale(USArrests))
# verify that the scaled data's variables all have mean 0 and variance 1
apply(usa_scaled, 2, function(x) round(c(mean(x), var(x)), 2))

##      Murder Assault UrbanPop Rape
## [1,]      0      0      0      0
## [2,]      1      1      1      1

# create data frame of combinations of rows of usa_scaled data set
# there are choose(50, 2) = 1225 ways to choose 2 observations out of 50
combs <- t(combn(nrow(usa_scaled), 2))
head(combs)

##      [,1] [,2]
## [1,]    1    2
## [2,]    1    3
## [3,]    1    4
## [4,]    1    5
## [5,]    1    6
## [6,]    1    7

###----- Determine Correlations -----

# cycle through combs to get each correlation between rows
one_minus_cors <- rep(0, nrow(combs))

# create function for getting correlations between each observation
one_minus_cor_usa <- function(x){
  res <- cor(as.numeric(usa_scaled[combs[x, 1], ]),
             as.numeric(usa_scaled[combs[x, 2], ]))
  res <- 1 - res
  res
}

# determine 1 - r_ij for each pair of observations
one_minus_cors <- sapply(1:nrow(combs), function(x) one_minus_cor_usa(x))

###----- Determine Squared Euclidean Distances -----

# cycle through combs to get distance between rows
distances <- dist(usa_scaled)^2

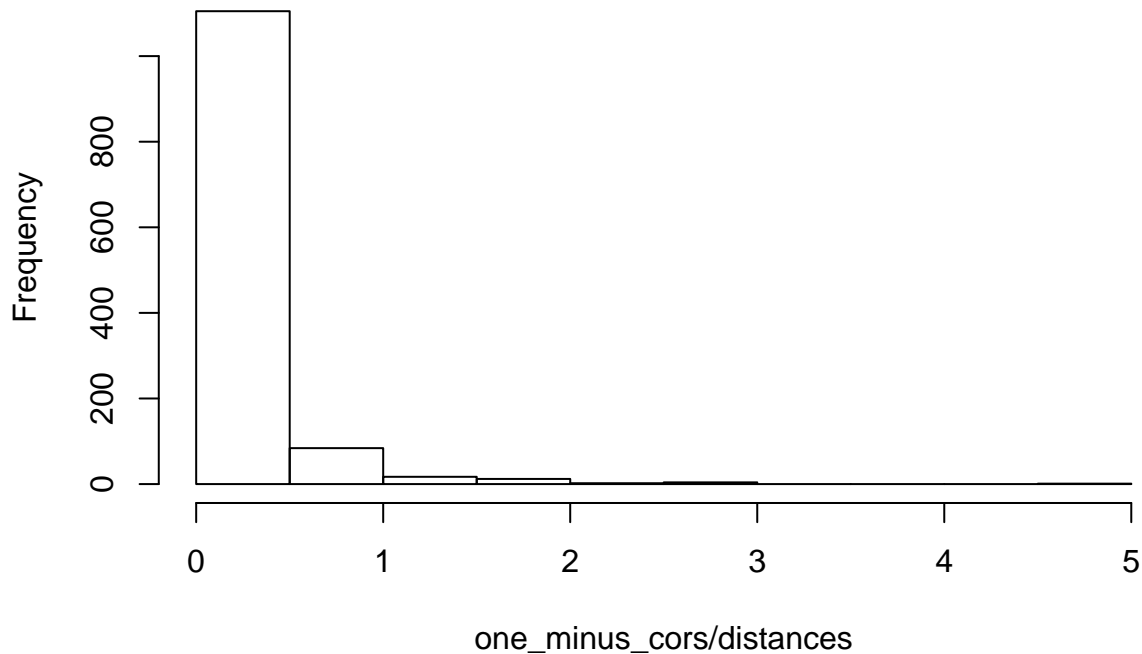
###----- Determine Proportionality -----
```

```
summary(one_minus_cors/distances)
```

```
##      Min.   1st Qu.   Median     Mean 3rd Qu.     Max.
## 0.000086 0.069140 0.133900 0.234200 0.262600 4.888000
```

```
hist(one_minus_cors/distances)
```

Histogram of one_minus_cors/distances



As can be seen from the summary above, there is no evidence of proportionality. If there were proportionality, there would only be one value. I believe the wording of this question is either wrong or simply hard to understand. The question refers to distances and correlations between the i^{th} and j^{th} observations, which would lead me to believe it is referring to the rows of the data set (i.e., 50 rows; one per state). If the authors meant to refer to the distances and correlations between i^{th} and j^{th} variables (i.e., the 4 variables of Murder, Assault, UrbanPop, and Rape), then there is constant proportionality. This is shown below.

```
# squared distance matrix
```

```
distance_matrix <- dist(t(usa_scaled))^2
distance_matrix
```

```
##           Murder  Assault UrbanPop
## Assault  19.41642
## UrbanPop  91.18188 72.63057
## Rape      42.76927 32.80636 57.68856
```

```
# 1 - correlation matrix
```

```
one_minus_cor_mat <- as.dist(1-cor(usa_scaled))
one_minus_cor_mat
```

```
##           Murder  Assault UrbanPop
## Assault  0.1981267
## UrbanPop 0.9304274 0.7411283
## Rape     0.4364212 0.3347588 0.5886588
```

```
# demonstration of proportionality  
distance_matrix/one_minus_cor_mat
```

```
##           Murder Assault UrbanPop  
## Assault           98  
## UrbanPop        98      98  
## Rape            98      98      98
```