Chapter 5

Nick Lauerman

Started 5/19/2020

Contents

1 Data Setup **1**

2 Correlation Analysis **2**

3 A Word About Data Frames **2**

4 Testing Correlation with Randomization **3**

5 Exercises **5**

5.1 1 5

5.2 2 6

1 Data Setup

From Appendex C

```
text.v <- scan(file = "./SupportingMaterials/data/plainText/melville.txt",
               what = "character",
               sep = "\n")
start.v <- which(text.v == "CHAPTER 1. Loomings.")
end.v <- which(text.v == "orphan.")
novel.lines.v <- text.v[start.v:end.v]
novel.lines.v <- unlist(novel.lines.v)

chap.positions.v <- grep("^CHAPTER \\d", novel.lines.v)
last.position.v <- length(novel.lines.v)
chap.positions.v <- c(chap.positions.v,
                      last.position.v)
chapter.freqs.l <- list()
chapter.raws.l <- list()
for (i in 1:length(chap.positions.v)) {
  if(i != length(chap.positions.v)){
    chapter.title <- novel.lines.v[chap.positions.v[i]]
    start <- chap.positions.v[i] + 1
    end <- chap.positions.v[i + 1] - 1
    chapter.lines.v <- novel.lines.v[start:end]
    chapter.words.v <- tolower(paste(chapter.lines.v,
                                     collapse = " "))
    chapter.words.l <- strsplit(chapter.words.v,
                                "\\W")
    chapter.words.v <- unlist(chapter.words.l)
    chapter.words.v <- chapter.words.v[which(chapter.words.v != "")]
    chapter.freq.t <- table(chapter.words.v)
    chapter.raws.l[[chapter.title]] <- chapter.freq.t
    chapter.freqs.t.rel <- 100 * (chapter.freq.t/sum(chapter.freq.t))
    chapter.freqs.l[[chapter.title]] <- chapter.freqs.t.rel
  }
}
whale.l <- lapply(chapter.freqs.l, '[', 'whale')
whales.m <- do.call(rbind, whale.l)
ahab.l <- lapply(chapter.freqs.l, '[', 'ahab')
ahabs.m <- do.call(rbind, ahab.l)
whales.v <- as.vector(whales.m[,1])
ahabs.v <- as.vector(ahabs.m[,1])
whales.ahabs.m <- cbind(whales.v, ahabs.v)
colnames(whales.ahabs.m) <- c("whale",
                              "ahab")
```

2 Correlation Analysis

```
whales.ahabs.m[which(is.na(whales.ahabs.m))] <- 0
cor(whales.ahabs.m)
```

```
##            whale       ahab
## whale  1.0000000 -0.2411072
## ahab  -0.2411072  1.0000000
```

```
mycor <- cor(whales.ahabs.m[,"whale"],
             whales.ahabs.m[,"ahab"])
mycor
```

```
## [1] -0.2411072
```

3 A Word About Data Frames

```
x <- matrix(1, 3, 3)
class(x[1,2])
```

```
## [1] "numeric"
```

```
x[1,2] <- "Sam I am"
x
```

```
##      [,1] [,2]       [,3]
## [1,] "1"  "Sam I am" "1"
## [2,] "1"  "1"        "1"
## [3,] "1"  "1"        "1"
```

```
class(x[1,2])
```

```
## [1] "character"
```

```
class(x[1,3])
```

```
## [1] "character"
```

```r
x <- matrix(1, 3, 3)
x.df <- as.data.frame(x)
x.df
```

```
##   V1 V2 V3
## 1  1  1  1
## 2  1  1  1
## 3  1  1  1
```

```r
x.df[1,2] <- "Sam I am"
class(x.df[1,2])
```

```
## [1] "character"
```

```r
class(x.df[1,3])
```

```
## [1] "numeric"
```

```r
x.df
```

```
##   V1       V2 V3
## 1  1 Sam I am  1
## 2  1        1  1
## 3  1        1  1
```

4 Testing Correlation with Randomization

```r
cor.data.df <- as.data.frame(whales.ahabs.m)
cor(cor.data.df)
```

```
##            whale       ahab
## whale  1.0000000 -0.2411072
## ahab  -0.2411072  1.0000000
```

```r
sample(cor.data.df$whale)
```

```
##   [1] 0.60716454 1.00767754 1.15546218 1.26506024 0.39920160 0.00000000
##   [7] 0.38722168 0.89485459 0.15313936 0.10000000 0.08207934 0.69124424
##  [13] 0.24375381 0.34364261 0.58167717 0.69620253 0.00000000 2.07452939
##  [19] 1.24777184 0.18761726 0.15829046 0.76628352 1.26582278 0.17341040
##  [25] 0.77565632 0.80200501 0.11926058 1.76565008 0.21901007 0.15723270
##  [31] 0.83682008 0.24711697 0.06882312 0.00000000 0.88832487 0.00000000
##  [37] 0.23790642 0.41841004 0.16722408 0.00000000 2.06782465 0.07047216
##  [43] 0.10638298 0.50125313 0.15485869 0.29296875 0.64400716 0.21097046
##  [49] 0.55865922 0.82987552 0.00000000 1.04895105 0.24067389 0.66760365
##  [55] 0.00000000 1.29198966 2.02788340 0.00000000 0.96566524 0.00000000
##  [61] 0.00000000 0.00000000 0.28391557 0.13368984 0.00000000 0.00000000
##  [67] 0.17942584 0.54305663 0.44247788 0.16037063 0.22271715 0.87131367
##  [73] 0.67127746 0.32017076 0.00000000 0.00000000 0.89726335 0.39761431
##  [79] 0.46838407 0.00000000 0.78616352 1.07469103 0.07949126 0.00000000
##  [85] 0.83275503 1.35440181 0.00000000 0.06079027 0.13114754 0.27548209
##  [91] 0.08748906 1.02739726 0.61099796 1.51515152 0.81168831 0.00000000
##  [97] 0.41841004 0.00000000 0.35971223 0.29411765 0.82840237 0.71283096
## [103] 1.82481752 0.64878893 0.00000000 1.70807453 0.04448399 0.14035088
## [109] 0.10857763 0.00000000 0.11286682 0.96562379 0.00000000 0.00000000
## [115] 0.00000000 1.03578154 0.76965366 0.61892131 1.05485232 0.62176166
## [121] 0.06079027 0.87623220 0.00000000 0.11580776 0.19762846 0.94339623
## [127] 0.69930070 0.41793313 0.98159509 0.85653105 1.09151973 0.00000000
## [133] 0.16260163 0.56191467 0.30193237
```

```
cor(sample(cor.data.df$whale),
    cor.data.df$ahab)
```

```
## [1] -0.00117504
```

```
mycors.v <- NULL
for (i in 1:10000) {
  mycors.v <- c(mycors.v,
                cor(sample(cor.data.df$whale),
    cor.data.df$ahab))
}
min(mycors.v)
```

```
## [1] -0.2841935
```

```
max(mycors.v)
```

```
## [1] 0.3883957
```

```
range(mycors.v)
```

```
## [1] -0.2841935  0.3883957
```

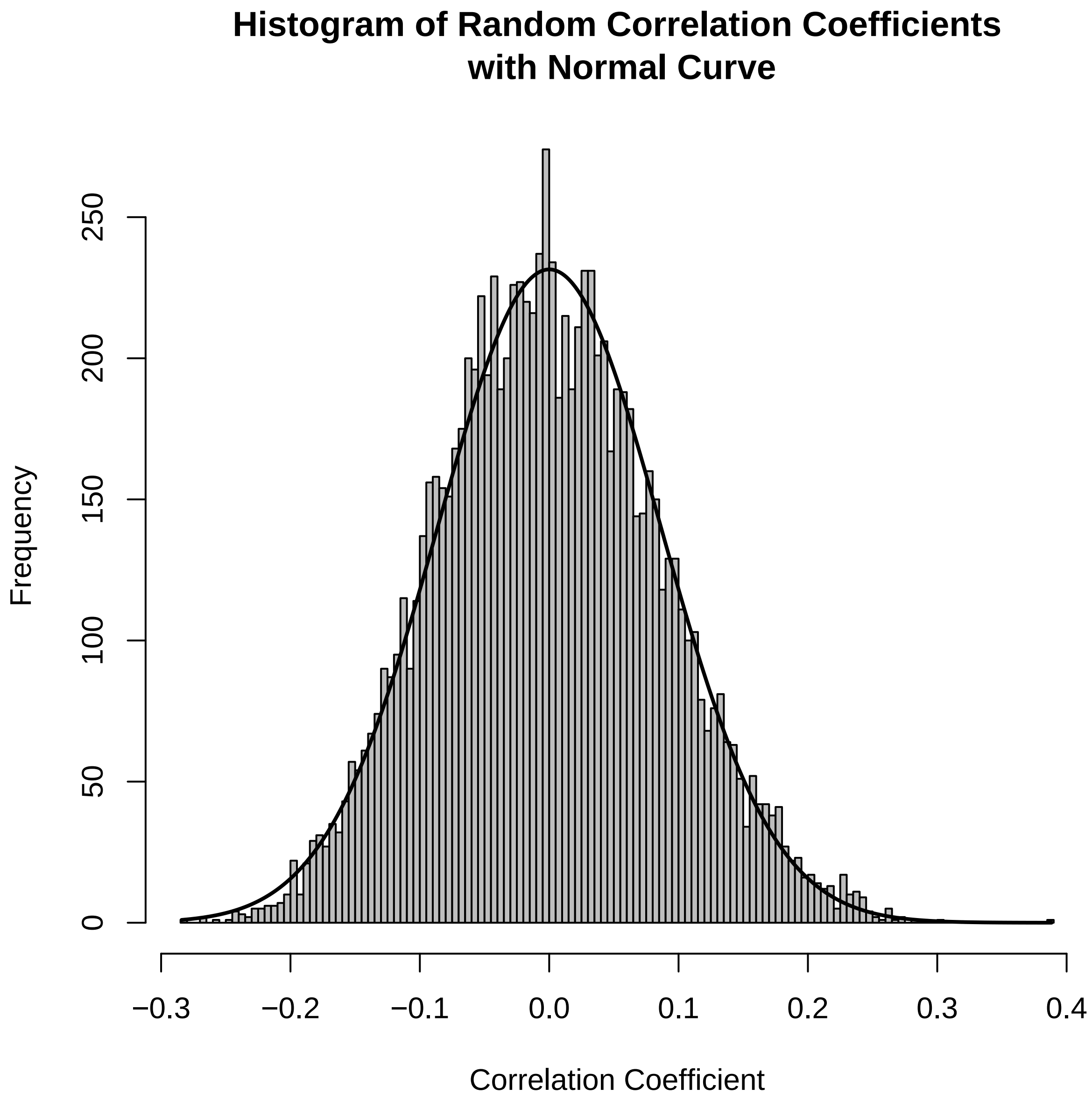
```
mean(mycors.v)
```

```
## [1] 7.275493e-05
```

```
sd(mycors.v)
```

```
## [1] 0.08617295
```

```
h <- hist(mycors.v,
          breaks = 100,
          col = "grey",
          xlab = "Correlation Coefficient",
          main = "Histogram of Random Correlation Coefficients\n with Normal Curve",
          plot = TRUE)
xfit <- seq(min(mycors.v), max(mycors.v), length = 1000)
yfit <- dnorm(xfit,
              mean = mean(mycors.v),
              sd = sd(mycors.v))
yfit <- yfit * diff(h$mids[1:2]) * length(mycors.v)
lines(xfit, yfit, col = "black", lwd = 2 )
```



5 Exercises

5.1 1

```
i.l <- lapply(chapter.freqs.l, '[', 'i')
i.m <- do.call(rbind, i.l)
i.v <- as.vector(i.m[,1])
i.v[which(is.na(i.v))] <- 0

cor.data.df$i <- i.v

my.l <- lapply(chapter.freqs.l, '[', 'my')
my.m <- do.call(rbind, my.l)
```

```r
my.v <- as.vector(my.m[,1])
my.v[which(is.na(my.v))] <- 0

cor.data.df$my <- my.v

cor(cor.data.df)
```

```
##            whale       ahab          i         my
## whale  1.0000000 -0.2411072 -0.2823192 -0.2567552
## ahab  -0.2411072  1.0000000  0.0709321  0.1047598
## i     -0.2823192  0.0709321  1.0000000  0.7739595
## my    -0.2567552  0.1047598  0.7739595  1.0000000
```

5.2 2

```r
my.i.m <- cbind(my.v, i.v)
my.i.cor.data.df <- as.data.frame(my.i.m)

cor(my.i.cor.data.df$i,
    my.i.cor.data.df$my)
```

```
## [1] 0.7739595
```

```r
i.my.cor.v <- NULL
for (i in 1:10000) {
  i.my.cor.v <- c(i.my.cor.v,
                  cor(sample(my.i.cor.data.df$i),
                      my.i.cor.data.df$my))
}
min(i.my.cor.v)
```

```
## [1] -0.2749866
```

```r
max(i.my.cor.v)
```

```
## [1] 0.3585906
```

```r
range(i.my.cor.v)
```

```
## [1] -0.2749866  0.3585906
```

```r
mean(i.my.cor.v)
```

```
## [1] -0.0003223358
```

```r
sd(i.my.cor.v)
```

```
## [1] 0.08644189
```