Problem Set 2

Olivia Bogiages

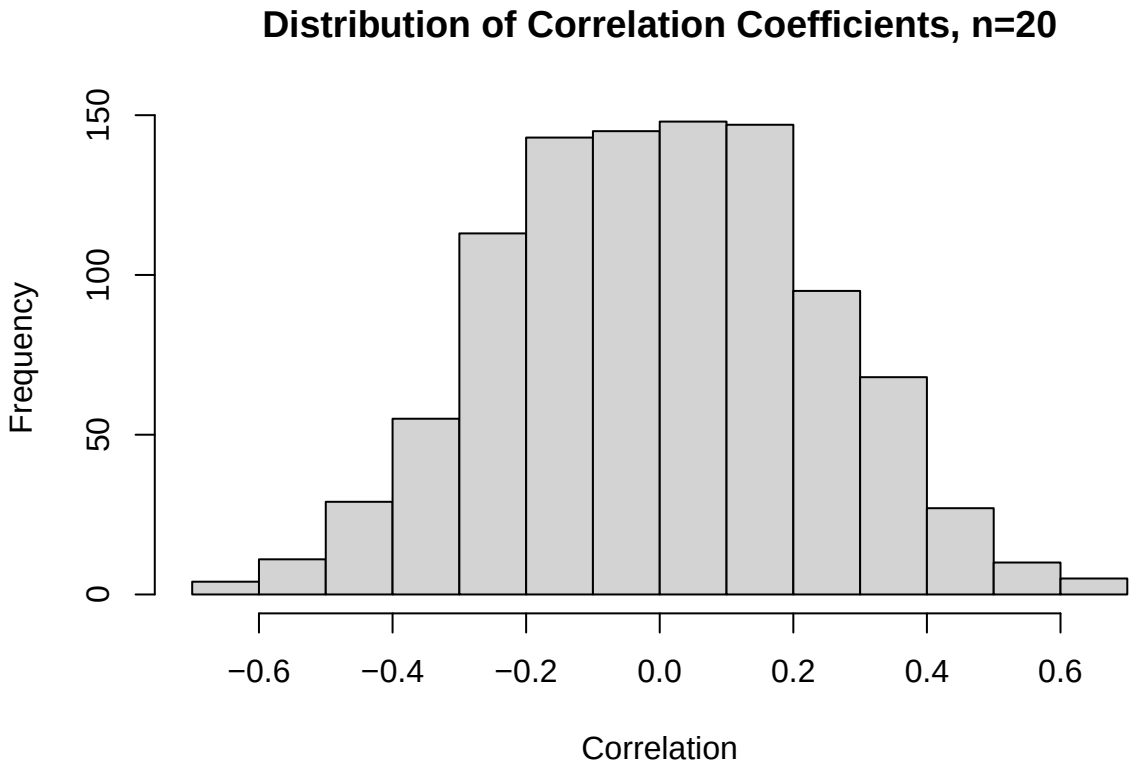
2025-10-21

```r
##1. Use rnorm() to create 2 random variables with 20 observations each
x<-rnorm(20)
y<-rnorm(20)
##Calculate the correlation between the two variables
cor(x,y)
```

```
## [1] 0.3416194
```

```r
#Repeat process many times
set.seed(123)
correlations<-replicate(1000,{
  x<-rnorm(20)
  y<-rnorm(20)
  cor(x,y)})

#Plot distribution of correlation coefficients and report standard deviation
hist(correlations, main="Distribution of Correlation Coefficients, n=20",
     xlab="Correlation")
```



```r
sd(correlations)
```

```
## [1] 0.2378933
```

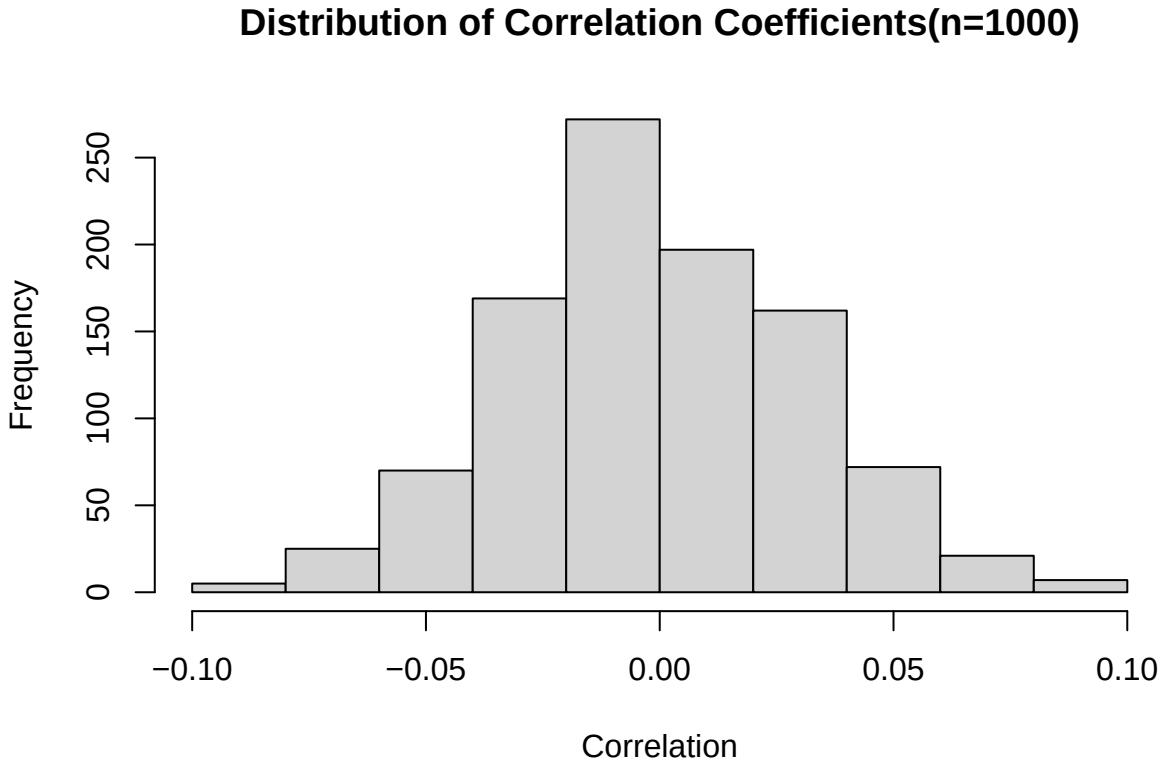
```r
#What do we expect the correlation between the two variables to be?
#We would expect the correlation between these two random variables,
#on average, to be 0.
#They are independent of each other and have no causal relationship and
#so as one increases or
#decreases we would not expect this to impact the other variable.

#What does this distribution tell us about sample
#estimates of population parameters? This distribution tells us that
#sample estimates of the population vary.
#The distribution shows, when pulled from n=20, that there are observations of
#correlation coefficients between
#-.6 and .6 depending on the iteration. A single sample estimate is, therefore,
#insufficient for accurately capturing the
#population parameter especially when the sample is small.


#2 Repeat the previous step with n=1000
x<-rnorm(1000)
y<-rnorm(1000)
##Calculate the correlation between the two variables
cor(x,y)
```

```
## [1] 0.002733113
```

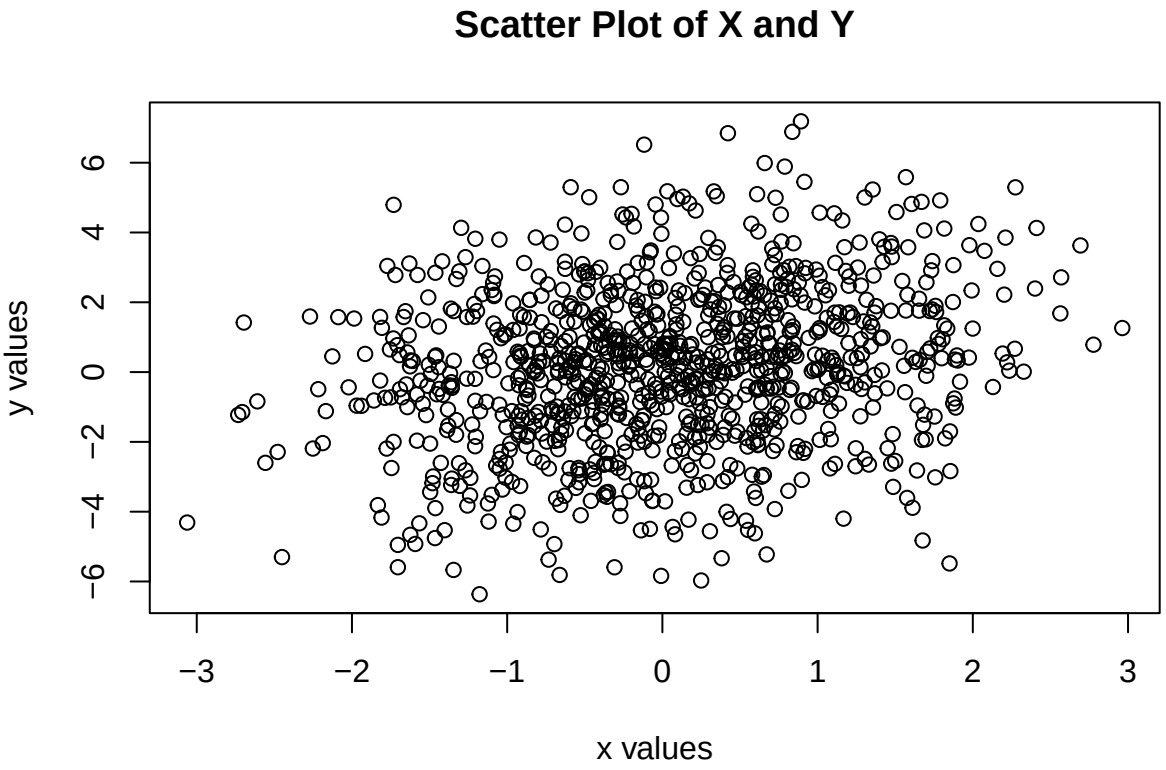
```r
#Repeat process many times
set.seed(123)
correlations<-replicate(1000,{
  x<-rnorm(1000)
  y_<-rnorm(1000)
  cor(x,y)})
#plot distribution
hist(correlations, main="Distribution of Correlation Coefficients(n=1000)",
     xlab="Correlation")
```




```
#Substantive interpretation of how results differ?
#The distribution for the larger sample (n=1000) is more evenly
#distributed than the smaller sample (n=20) which displays more variance.
#The distribution of the smaller sample size of n=20 shows the data as having
#a larger standard deviation than that of the n=1000 sample. This demonstrates
#that smaller samples sizes will have more variance because they are not as
#representative of the true population.
#Each observation in the smaller sample is a greater average
#distance from the mean than the observations in a larger sample size,
##which are a smaller average distance from the mean.

#3 create 3 random variables, z causes x and y
z<-rnorm(1000)
x<-.3*z+rnorm(1000)
y<-2*z+rnorm(1000)

###Plot x and y on a scatter plot
plot (x,y, main="Scatter Plot of X and Y", xlab="x values", ylab="y values" )
```




```
#Report correlation
cor(x,y)
```

```
## [1] 0.227019
```

```
#What does this tell us about reporting correlation?
#This demonstrates that unrelated variables can still be shown as
#having a correlation. Therefore, we should be careful about associating a
#correlation with a causal relationship as both x and y in this case are not
#causally related, and yet, they have a positive correlation. We should also
#consider possible variables that may confound results. In this case, where z
#has a causal relationship with both x and y but x and y have no causal
#relationship, a non-zero positive correlation coefficient exists
#between x and y. Overall, we should not equate correlation as causation and
#we should consider confounding variables as factors in
#interpretations of correlations.
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

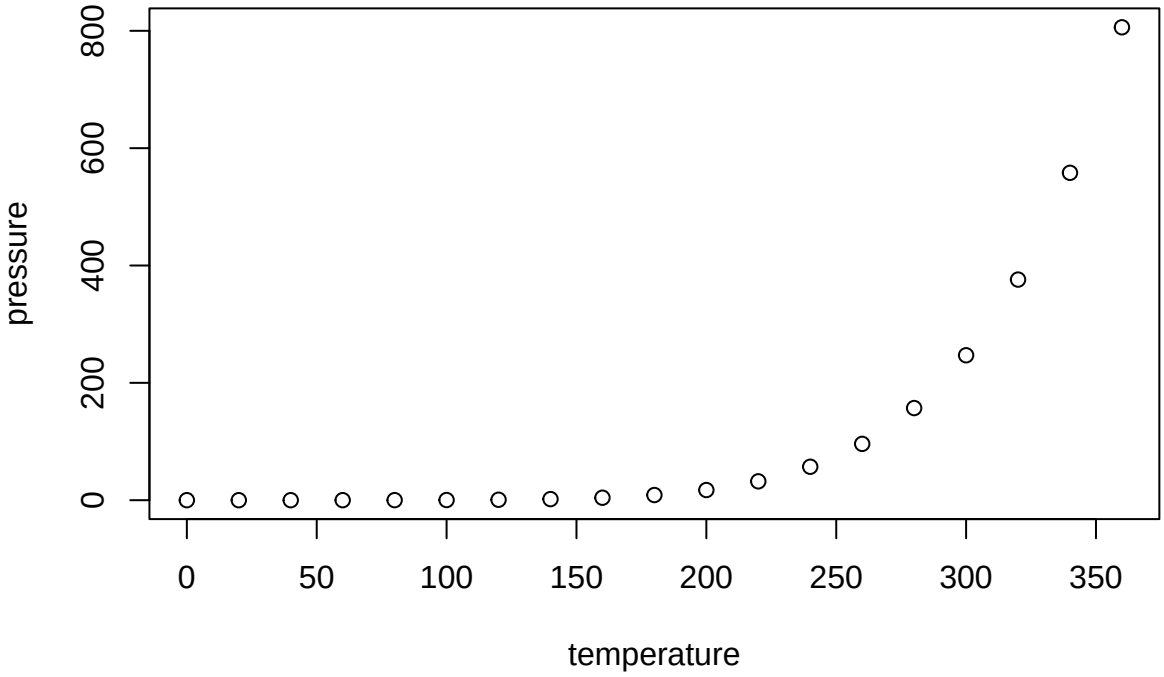
```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
```

```
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.