W271-2 – Spring 2016 – **HW 5**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

March 9, 2016

Contents

Exercises

Question 1

1. **Install the library "astsa" using the function: `install.packages("astsa")`**

```
# Check if already installed; if not, install it
if (!"astsa" %in% installed.packages()[, "Package"]) install.packages("astsa")
```

2. **Load the library: `library(astsa)`**

```
# Load the library: library(astsa)
library(astsa)
# Last two commands can be substituted by simply...
if (!require(astsa)) install.packages("astsa")
```

3. **Use the function `str()` to see the information of a particular data series, such as `str(EQ5)` for the Seismic Trace of Earthquake number 5 series**

```
str(EQ5)
```

```
##  Time-Series [1:2048] from 1 to 2048: 0.01749 0.01139 0.01512 0.01477 0.00651 ...
```

```
str(flu)
```

```
##  Time-Series [1:132] from 1968 to 1979: 0.811 0.446 0.342 0.277 0.248 ...
```

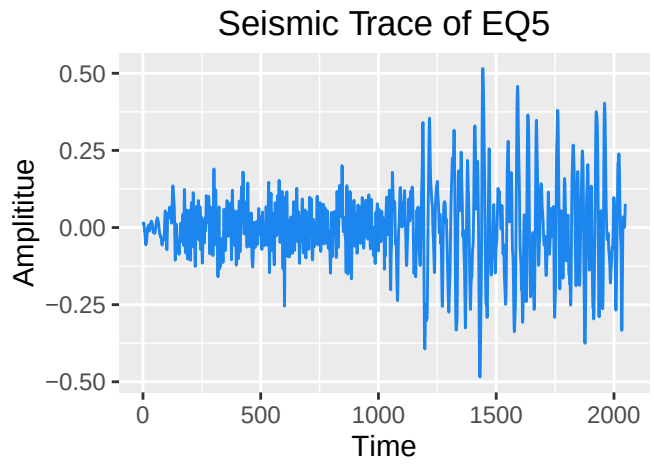
```
str(gas)
```

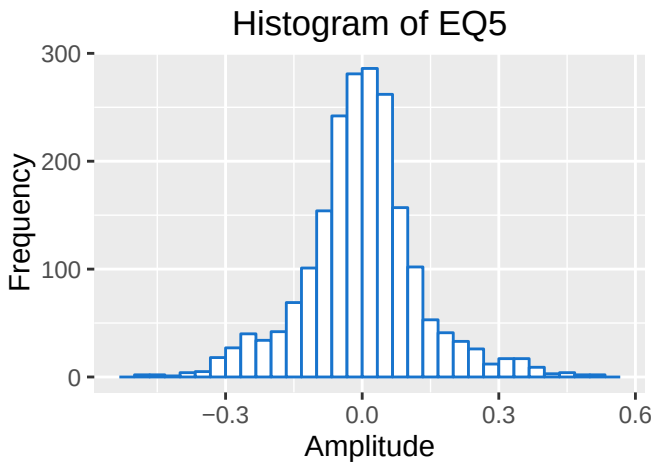
```
##  Time-Series [1:545] from 2000 to 2010: 70.6 71 68.5 65.1 67.9 ...
```

According to that package documentation, `EQ5` corresponds to the *Seismic trace of an earthquake [two phases or arrivals along the surface, the primary wave (t = 1, . . . , 1024) and the shear wave (t = 1025, . . . , 2048)] recorded at a seismic station.*

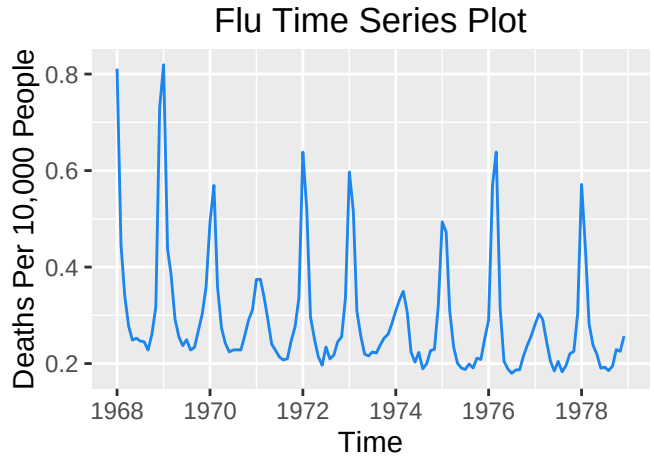
4. **Plot the time series plots and histograms of the following 3 series. Feel free to use the codes provided in the R scripts. Make sure that each of your graph has a title, the axis ticks are clear, the axes are well-labelled, and use color intelligently.**

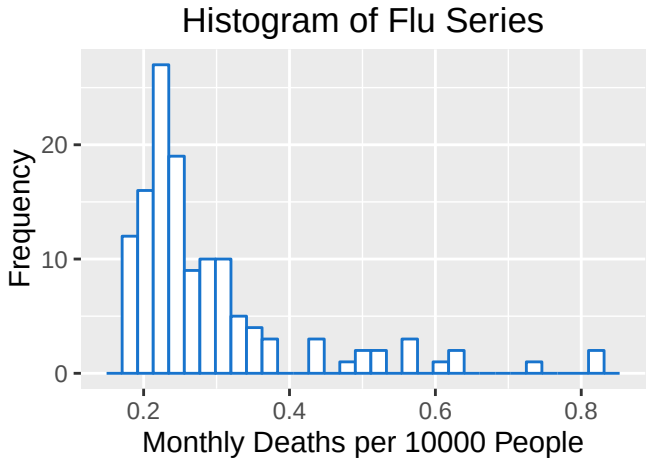
```
# Time series and histogram Plots for the EQ5 seismic trace
autoplot(EQ5, main='Seismic Trace of EQ5', ts.colour= 'dodgerblue2',
         xlab='Time', ylab='Amplititue')
qplot(EQ5, geom="histogram",  main='Histogram of EQ5', xlab='Amplitude',
      ylab='Frequency', colour = I('dodgerblue3'), fill = I("white") )
```



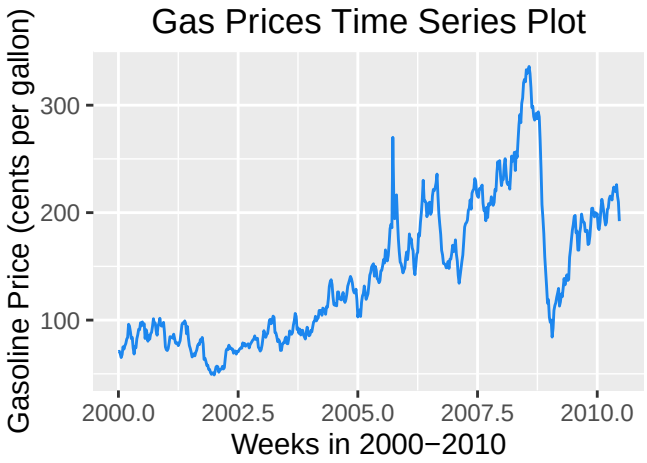


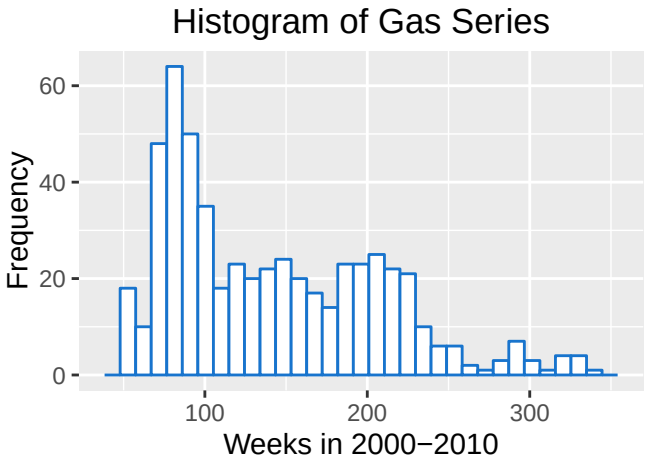

```
# Time series and histogram Plots for the flu data series
autoplot(flu, main='Flu Time Series Plot', ts.colour = 'dodgerblue2',
         xlab='Time', ylab='Deaths Per 10,000 People' )
qplot(flu, geom="histogram",  main='Histogram of Flu Series', xlab='Monthly Deaths per 10000 People',
      ylab='Frequency', colour = I('dodgerblue3'), fill = I("white") )
```






```
# Time series and histogram plots for the gas price series
autoplot(gas, main='Gas Prices Time Series Plot',  ts.colour = 'dodgerblue2', xlab='Weeks in 2000-2010'
         ylab='Gasoline Price (cents per gallon)')
qplot(gas, geom="histogram",  main='Histogram of Gas Series', xlab='Weeks in 2000-2010',
      ylab='Frequency', colour = I('dodgerblue3'), fill = I("white") )
```





5. **Write a few sentences to describe each of the series.**

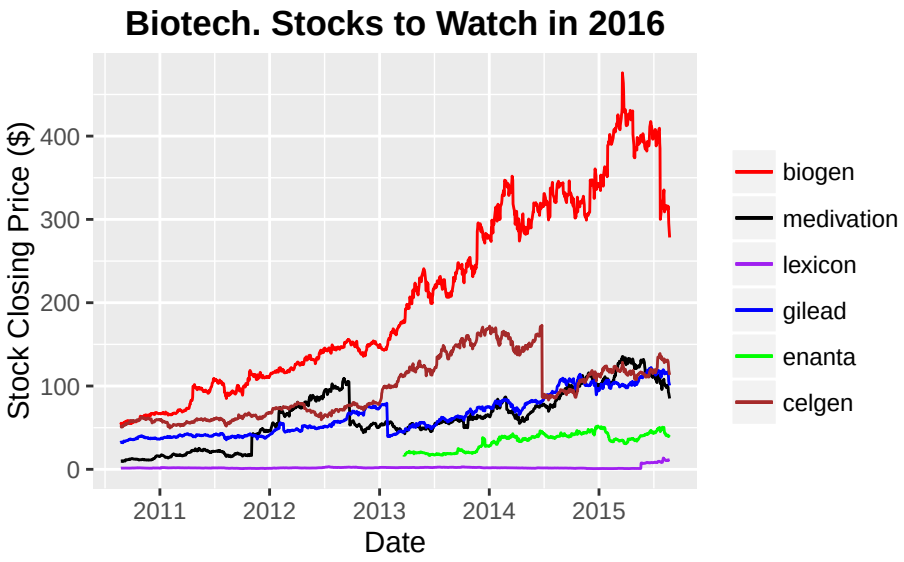
- **EQ5:** The EQ5 data encodes the seismic trace of an earthquake saved as the ts data class. It include two arrive phases: primary wave or P phase (t=1, . . . ,1024) and the shear wave or S phase (t= 1025, . . . , 2048)]. In this series, the amplitude of the P phase is much smaller than the amplitude of the S phase. The amplitude mean of P phase appears to be around 0. S phase is quite volatility. In addition, as shown in the above time series plot, it seems that the S phase of EQ5 series is periodic as well because it contains several similar segments with the equal length. The whole distribution of EQ5 data follows a normal distribution.
- **flu:** The flu data records monthly deaths per 10,000 people due to pneumonia and influenza in the United States for a period of 132 months(1968-1978). This series also exhibits an annually seasonal pattern that the death rate tends to be the highest in January or spring, ten decreases gradually till October or Noverber, and finall start to level up in the last three months. it distribution shows postive skew with longer right tail.
- **gas:** The gas data collects the weekly price of gasoline in cents per gallon at New York Harbor from 2000 to mid-2010. According to the data, the gas price first slight leveled off in the first two years and then started to gradually level up from 2002 to 2008. In 2009 the price dropped significantly around the 4th quarter of 2009 and then gradualy recovered..

Question 2

Describe 3 examples you have used in your work or encounter in real life. Ideally, you can even load at least one of these time series, plot it, and the write a few statements to describe its characteristics.

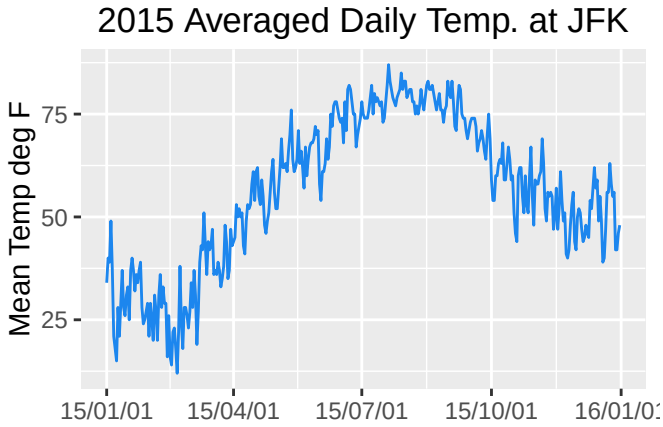
- **the Example of Biotech Stocks:** In the first example, we pull out the stock data (from August 2010 to August 2015) of some biotech companie from yahoo financial web links. Biogen's closing stock price is used as the example here. Starting from 2010, it started to gradually go up because of the promising clinical results of aducanumab for Alzheimer treatment in March, 2013. However, the stock began to drop in 2015 due to slowing sales of its multiple sclerosis drug, Tecfidera.

```
# define the variable to get access to the yahoo finacial stock data
biogen_stock_url <- "http://real-chart.finance.yahoo.com/table.csv?s=BIIB
&a=07&b=24&c=2010&d=07&e=24&f=2015&g=d&ignore=.csv"
mdvn_stock_url <- "http://real-chart.finance.yahoo.com/table.csv?s=MDVN
&a=07&b=24&c=2010&d=07&e=24&f=2015&g=d&ignore=.csv"
lexicon_stock_url <- "http://real-chart.finance.yahoo.com/table.csv?s=LXRX
&a=07&b=24&c=2010&d=07&e=24&f=2015&g=d&ignore=.csv"
gilead_stock_url <- "http://real-chart.finance.yahoo.com/table.csv?s=GILD
&a=07&b=24&c=2010&d=07&e=24&f=2015&g=d&ignore=.csv"
enanta_stock_url <- "http://real-chart.finance.yahoo.com/table.csv?s=ENTA
&a=07&b=24&c=2010&d=07&e=24&f=2015&g=d&ignore=.csv"
celgen_stock_url <- "http://real-chart.finance.yahoo.com/table.csv?s=CELG
&a=07&b=24&c=2010&d=07&e=24&f=2015&g=d&ignore=.csv"

# define function to read financial data through url
yahoo.read <- function(url){
  dat <- read.table(url,header=TRUE,sep=",")
  df <- dat[,c(1,5)]
  df$Date <- as.Date(as.character(df$Date))
  return(df)}

# grap the stock data from 2010 to 2016 for those companies
biogen  <- yahoo.read(biogen_stock_url)
medivation <- yahoo.read(mdvn_stock_url)
lexicon <- yahoo.read(lexicon_stock_url)
gilead <- yahoo.read(gilead_stock_url)
enanta <- yahoo.read(enanta_stock_url)
celgen <- yahoo.read(celgen_stock_url)

# time series plot for those stocks
ggplot(biogen,aes(Date,Close)) +
  geom_line(aes(color="biogen")) +
  geom_line(data=medivation,aes(color="medivation")) +
  geom_line(data=lexicon,aes(color="lexicon")) +
  geom_line(data=gilead,aes(color="gilead")) +
  geom_line(data=enanta,aes(color="enanta")) +
  geom_line(data=celgen,aes(color="celgen")) +
  labs(color="Legend") +
  scale_colour_manual("", breaks = c("biogen", "medivation","lexicon","gilead","enanta","celgen"),
                      values = c("red", "brown", "green", "blue","purple","black")) +
  ggtitle("Biotech. Stocks to Watch in 2016") +
  theme(plot.title = element_text(lineheight=.7, face="bold")) +
  labs(y = "Stock Closing Price ($)")
```



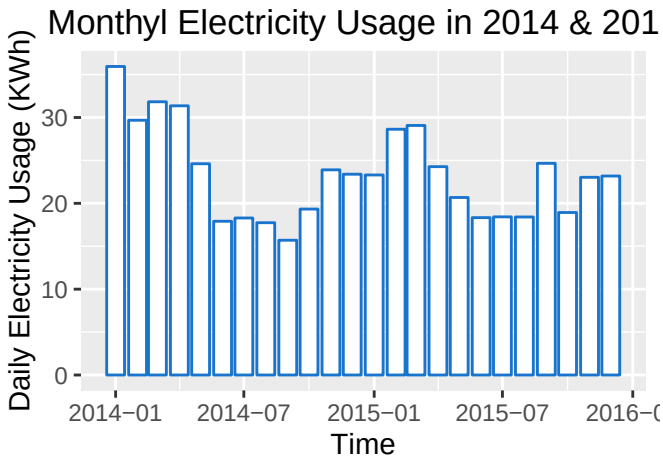
- **the Example of Daily Averaged Tempreture at JFK airport:** The 2nd data is the daily averaged tempretaure recorded at JFK airport weather station. The data is directly pulled out through the WeatherData package. Similar to stock price, it is also non-regular time series data. But the tempreture at a centain time point would be highly relevant to the tempreture of previous times.

```
# get access to the weather data through weatherdata package (need the scales package as well)
W_KJFK_2015 <- getWeatherForYear("KJFK",2015)
W_KJFK_2015$Date <- as.Date(W_KJFK_2015$Date,format="%y-%m-%d")

ggplot(W_KJFK_2015, aes(Date, Mean_TemperatureF)) + geom_line(colour = 'dodgerblue2') +
  scale_x_date(labels=date_format("%y/%m/%d")) + xlab("") + ylab("Mean Temp deg F") +
  ggtitle("2015 Averaged Daily Temp. at JFK")
```



- **the Example Monthly Averaged Electricity Usage Example:** The 3rd example shows the monthly averaged electricity usage of my house between 2014 and 2015. This periodic series shows that summers, in general, tended to have lower monthly averaged electricity usage (<20KWh) than the other seasons, e.g. winter.

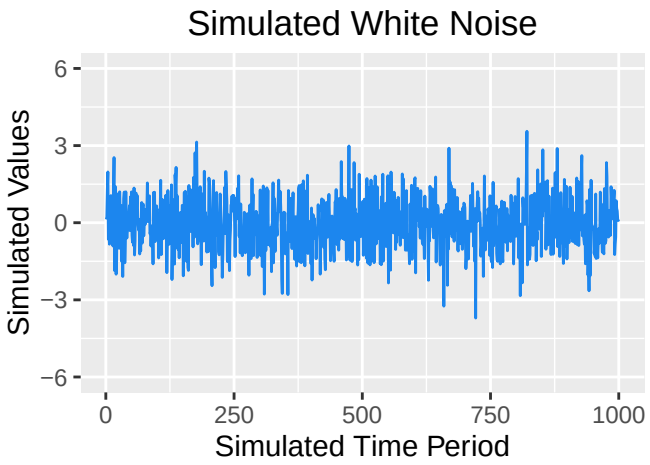
```
# load monthyl averaged electricity usage for months in 2014 and 2015
elec_usage = c(35.94, 29.68, 31.83, 31.36, 24.61, 17.91, 18.29, 17.74, 15.70,
               +19.33, 23.90, 23.39, 23.30, 28.63, 29.07, 24.27, 20.68, 18.33,
               +18.42, 18.41, 24.66, 18.93, 23.03, 23.18)
elec_usage_ts <- ts(elec_usage, start=c(2014, 1), end=c(2015, 12), frequency=12)
autoplot(elec_usage_ts, main='Monthyl Electricity Usage in 2014 & 2015',geom = "bar",
         xlab='Time', ylab='Daily Electricity Usage (KWh)',colour = I('dodgerblue3'),
         fill = I("white"))
```

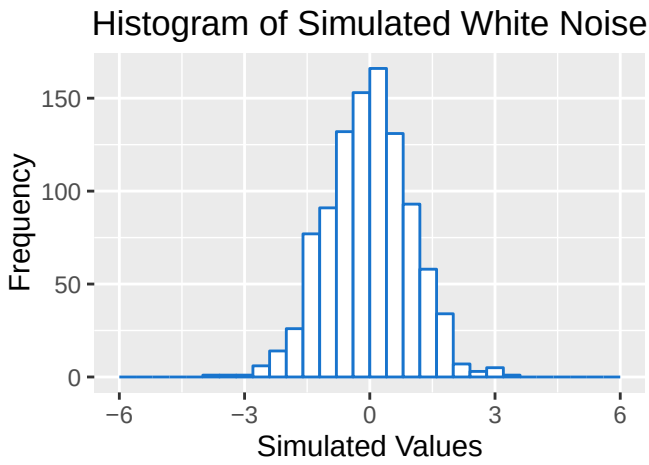


Question 3

Simulate a white noise series with 1000 random draws and plot (1) a time series plot and (2) a histogram. The usual reqirements on graphics (described) in Question 1) applied.

```
rand_draw <- rnorm(1000) # 1000 random draw
rand_draw_ts <- ts(rand_draw)
autoplot(rand_draw_ts , xlab = "Simulated Time Period", ylab = "Simulated Values",
         main="Simulated White Noise", ts.colour = 'dodgerblue2', ylim=c(-6,6) )
qplot(rand_draw_ts, geom="histogram",  main='Histogram of Simulated White Noise',
      ylab='Frequency', xlab='Simulated Values', colour = I('dodgerblue3'),
      fill = I("white"), xlim=c(-6,6) )
```





Question 4

Simulate (with 1000 random draws) the following two zero-mean autoregressive model with order 1 (i.e. AR(1)) models:

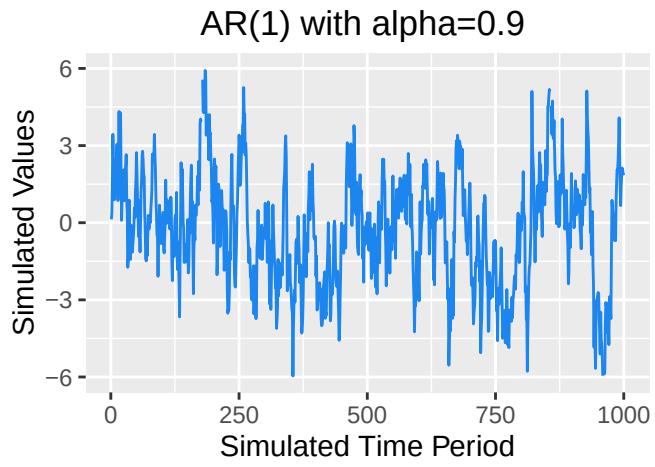
$$\mathbf{y_t = 0.9y_{t1} + w}$$

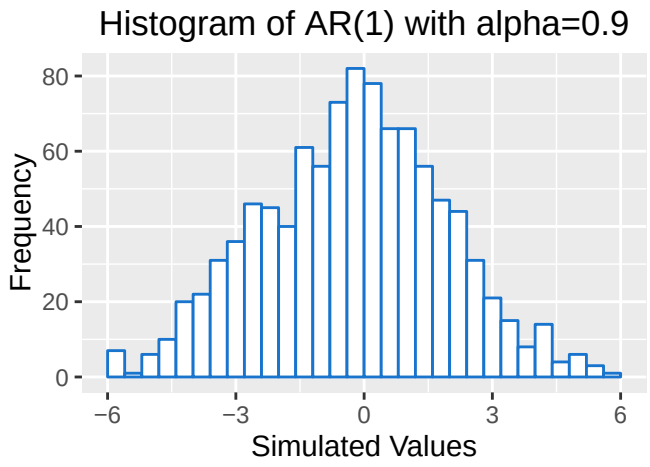
$$\mathbf{y_t = 0.2y_{t1} + w}$$

Plot a time plot for each of the simulated series. Graph a histogram for each of the simulated series. Write a few statements to compare the two series.

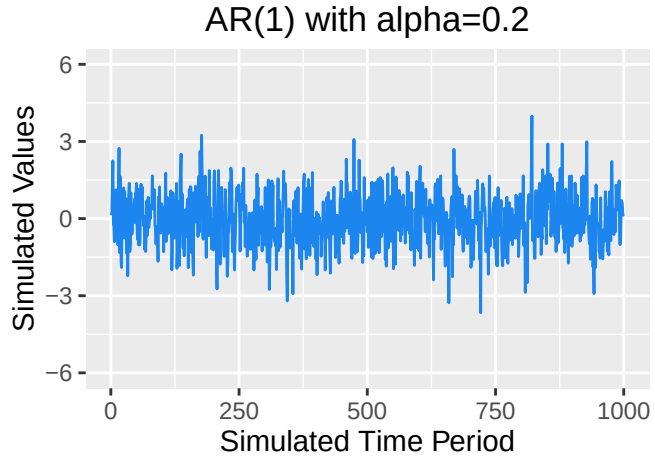
```
# Generate two simulated time series
z1 <- rand_draw
for (t in 2:length(rand_draw)){
  z1[t] <- 0.9 * z1[t-1] + z1[t] # use the same random normal sequence generated above
}
z1_ts <- ts(z1)
z2 <- rand_draw
for (t in 2:length(rand_draw)){
  z2[t] <- 0.2 * z2[t-1] + z2[t] # use the same random normal sequence generated above
}
z2_ts <- ts(z2)
```

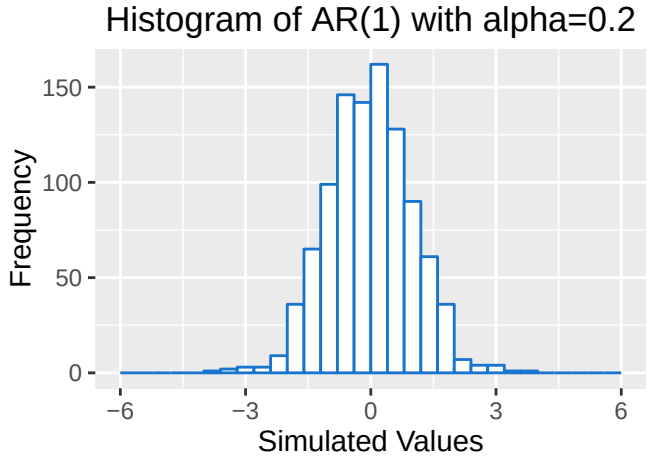
```
# the time series plot and histogram of 1st series
#par(mfrow=c(1,2))
autoplot(z1_ts , xlab = "Simulated Time Period", ylab = "Simulated Values",
         main="AR(1) with alpha=0.9 ", ts.colour = 'dodgerblue2', ylim=c(-6,6) )
qplot(z1, geom="histogram",  main='Histogram of AR(1) with alpha=0.9',
      ylab='Frequency', xlab='Simulated Values', colour = I('dodgerblue3'),
      fill = I("white"), xlim=c(-6,6))
```






```
# the time series plot and histogram of 2nd series
#par(mfrow=c(1,2))
autoplot(z2_ts , xlab = "Simulated Time Period", ylab = "Simulated Values",
         main="AR(1) with alpha=0.2", ts.colour = 'dodgerblue2', ylim=c(-6,6))
qplot(z2_ts, geom="histogram",  main='Histogram of AR(1) with alpha=0.2',
      ylab='Frequency', xlab='Simulated Values', colour = I('dodgerblue3'),
      fill = I("white"),xlim=c(-6,6) )
```





- **The Comparision between Two Simulated synthetic Series:** The first synthetic series (AR(1) with $alpha = 0.9$) is more volatile than the second series (AR(1) with $alpha = 0.2$). Due to the larger model parameter, the value at certain time t in the first series tend to be more statistically dependent on (or corrlated with) values at preceding times. In addition, although both series follow the normal distrubution, the first sereis have a larger standand deviation.

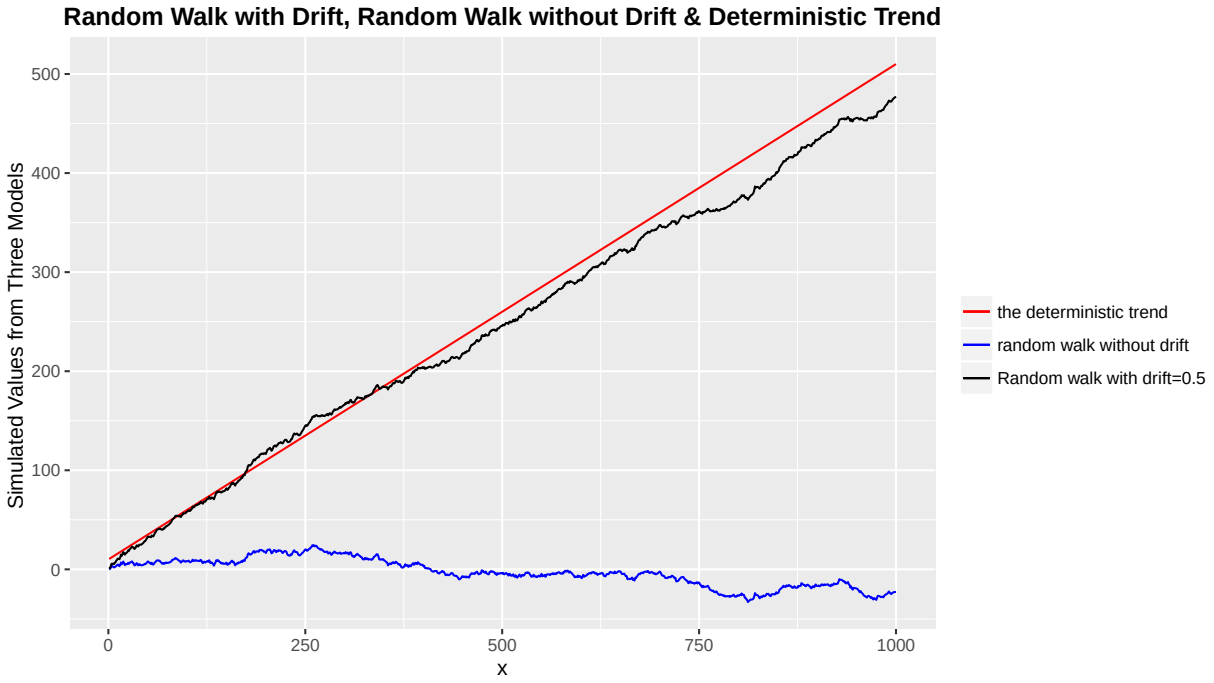
Question 5

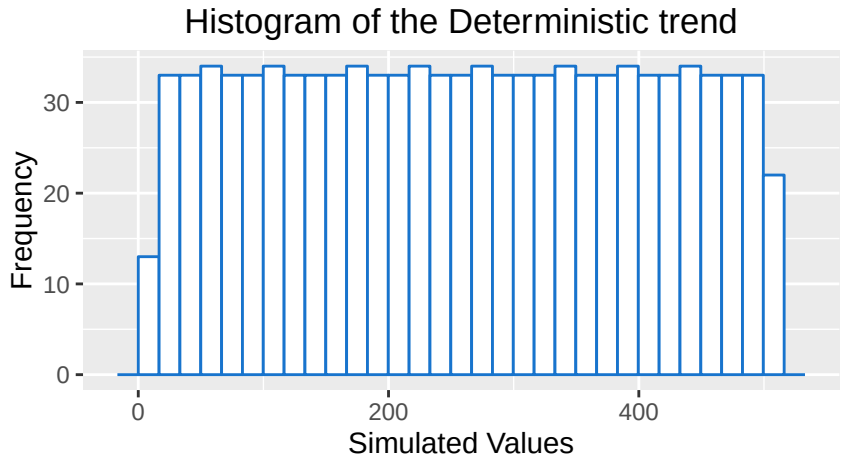
Simulate (with 1000 random draws) the following 3 models:

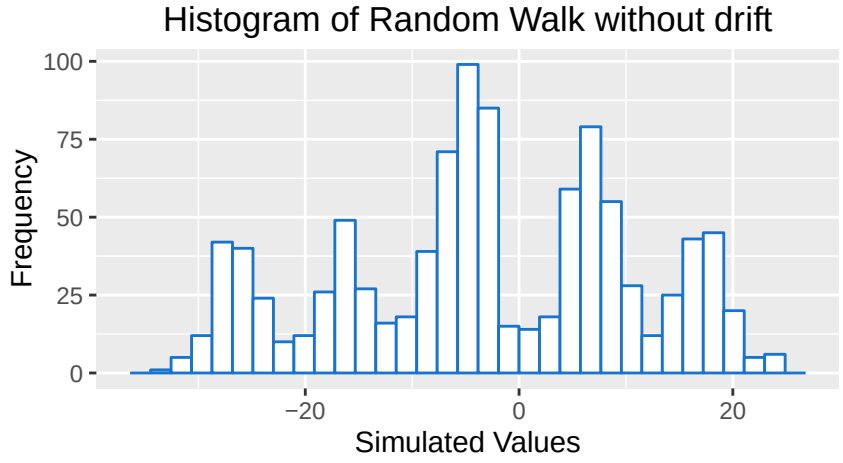
1. **A deterministic linear (time) trend of the form:** $y_t = 10 + 0.5t$
2. **Random walk without drift**
3. **Random walk with drift = 0.5**

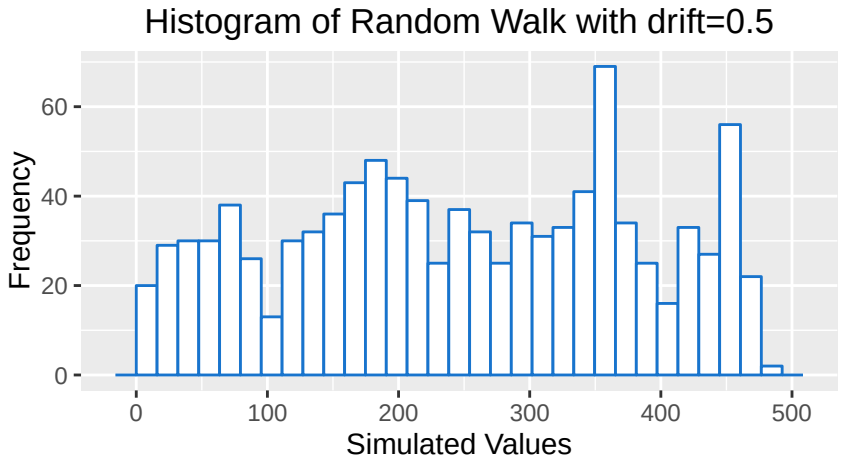
Plot a time plot for each of the simulated series. Graph a histogram for each of the simulated series. Write a few statements to compare the two series.

```
# Simulate (with 1000 random draws) the following 3 models:
# 1. A deterministic linear (time) trend of the form: yt = 10 + 0.5t
# 2. Random walk without drift
# 3. Random walk with drift = 0.5
# Plot a time plot for each of the simulated series.
# Graph a histogram for each of the simulated series.

# Generate a deterministic linear series
Linear_trend <- seq(1, 1000) * 0.5 + 10
Linear_trend <-ts(Linear_trend)
mod1 <- data.frame( x=as.integer(time(Linear_trend)), y=as.matrix(Linear_trend) )

# Generate a random walk without drift
randwalk=cumsum(rand_draw)
randwalk=ts(randwalk)
mod2 <- data.frame( x=as.integer(time(randwalk)), y=as.matrix(randwalk) )

# Generate a Random walk with drift = 0.5
randwalk_d = 0.5 + rand_draw;
randwalk_d = cumsum(randwalk_d)
mod3 <- data.frame( x=as.yearqtr(time(randwalk_d)), y=as.matrix(randwalk_d) )

ggplot(mod1,aes(x,y)) +
  geom_line(aes(color="the deterministic trend")) +
  geom_line(data=mod2,aes(color="random walk without drift")) +
  geom_line(data=mod3,aes(color="Random walk with drift=0.5")) +
  labs(color="Legend") +
  scale_colour_manual("", breaks = c("the deterministic trend",
                      "random walk without drift", "Random walk with drift=0.5"),
                      values = c("black", "blue", "red")) +
  ggtitle("Random Walk with Drift, Random Walk without Drift & Deterministic Trend") +
  theme(plot.title = element_text(lineheight=.7, face="bold")) +
  labs(y = "Simulated Values from Three Models ")
```


Random Walk with Drift, Random Walk without Drift & Deterministic Trend
Simulated Values from Three Models
500
400
300
200
100
0
0
250
500
750
1000
x
the deterministic trend
random walk without drift
Random walk with drift=0.5


Histogram of the Deterministic trend
Frequency
30
20
10
0
0
200
400
Simulated Values


Histogram of Random Walk without drift
Frequency
100
75
50
25
0
−20
0
20
Simulated Values


Histogram of Random Walk with drift=0.5
Frequency
60
40
20
0
0
100
200
300
400
500
Simulated Values
