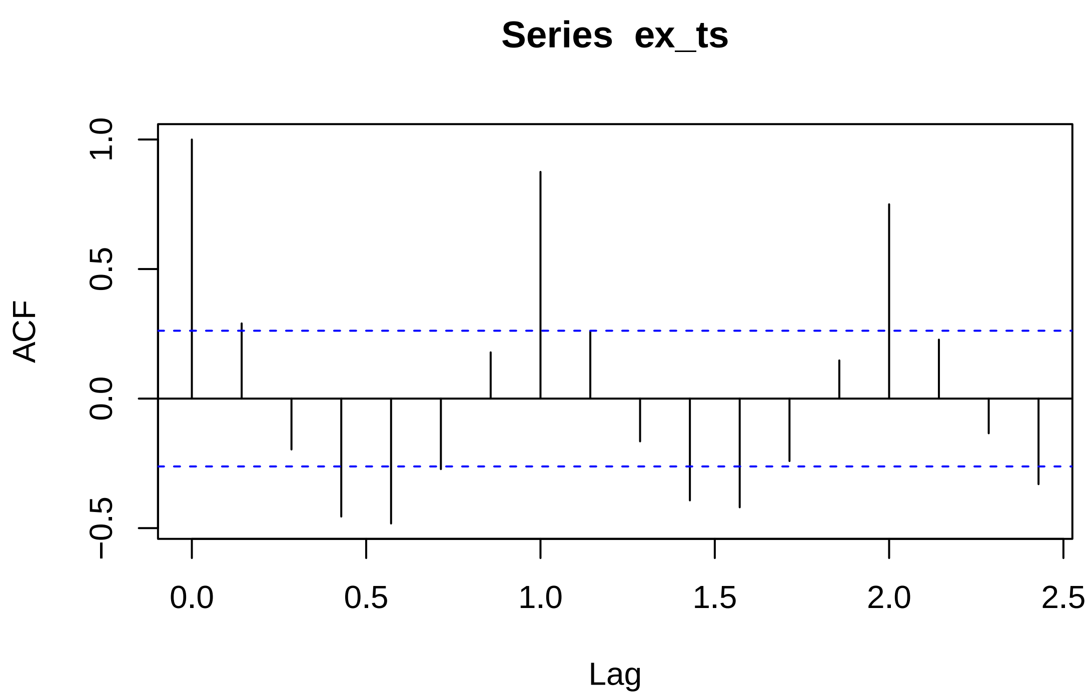
Name: Christian
Due: 2024/09/30

Homework 3 - Question 2 explanation

Be sure to submit **both** the .pdf and .qmd file to Canvas by Monday, September 30th at 11:59 pm.

2. [1 pt] The code below creates a time series out of the same seven values on repeat, and generates a correlogram for that series. Describe what is happening in the correlogram at time lags that are a multiple of seven. Why should that make sense?

```r
ex_ts <- ts(
  rep(-3:3, 8),
  start = c(1,1),
  end = c(8, 7),
  freq = 7
)
acf(ex_ts)
```



Observations that are seven time points apart are *identical*, so the correlation is extremely close to 1 (but not equal? See why below). This is one way to diagnose unaccounted for seasonal variability! You would expect to see large spikes in the ACF plot corresponding to the seasonal frequency.

So then why isn't the spike at lag 7 and lag 14 exactly equal to 1? Or, if that doesn't surprise you, why *is* it surprising that the autocorrelation at lag 7 and lag 14 are not equal to 1? First, let us look at the raw time series, $\{x_t\}$, and lagged series at 7 and 14, $\{x_{t+7}\}$ and $\{x_{t+14}\}$. Note that they are exactly equivalent.

```
knitr::kable(
  tibble(
    xt = c(ex_ts),
    xtp7 = c(rep(NA, 7), ex_ts[8:length(ex_ts)]),
    xtp14 = c(rep(NA, 14), ex_ts[15:length(ex_ts)])
  )
)
```

xt	xtp7	xtp14
-3	NA	NA
-2	NA	NA
-1	NA	NA
0	NA	NA
1	NA	NA
2	NA	NA
3	NA	NA
-3	-3	NA
-2	-2	NA
-1	-1	NA
0	0	NA
1	1	NA
2	2	NA
3	3	NA
-3	-3	-3
-2	-2	-2
-1	-1	-1
0	0	0
1	1	1
2	2	2
3	3	3
-3	-3	-3
-2	-2	-2
-1	-1	-1

xt	xtp7	xtp14
0	0	0
1	1	1
2	2	2
3	3	3
-3	-3	-3
-2	-2	-2
-1	-1	-1
0	0	0
1	1	1
2	2	2
3	3	3
-3	-3	-3
-2	-2	-2
-1	-1	-1
0	0	0
1	1	1
2	2	2
3	3	3
-3	-3	-3
-2	-2	-2
-1	-1	-1
0	0	0
1	1	1
2	2	2
3	3	3
-3	-3	-3
-2	-2	-2
-1	-1	-1
0	0	0
1	1	1
2	2	2
3	3	3

> Our first thought might be to calculate the correlation between each of these series using the `cor` function.

```r
# x_t and x_{t+7}
cor(ex_ts[1:(length(ex_ts)-7)], ex_ts[8:length(ex_ts)])
```

```
[1] 1
```

```
# x_t and x_{t+14}
cor(ex_ts[1:(length(ex_ts)-14)], ex_ts[15:length(ex_ts)])
```

```
[1] 1
```

So what gives? Why does the ACF plot not have these spikes at 1 for lags 7 and 14? Recall that *correlation* between variables x and y is calculated as $r = \frac{1}{(n-1)s_x s_y} \sum (x_i - \bar{x})(y_i - \bar{y})$ where s_x and s_y are the standard deviations, and the $n-1$ is computed based on the length of each series. This is demonstrated for $\{x_t\}$ and $\{x_{t+7}\}$ below.

```
# lets make some vectors to simplify our life
xt <- ex_ts[1:(length(ex_ts)-7)]
xtp7 <- ex_ts[8:length(ex_ts)]

# this is the "n"
length(xt);length(xtp7)
```

```
[1] 49
```

```
[1] 49
```

```
# lets calculate it
## standard deviations first
s1 <- sd(xt)
s2 <- sd(xtp7)

# calculate NORMAL CORRELATION
1/((length(xt)-1)*s1*s2) * sum((xt - mean(xt)) * (xtp7 - mean(xtp7)))
```

```
[1] 1
```

Indeed, the correlation between $\{x_t\}$ and $\{x_{t+7}\}$ *is* 1. But what about the *autocorrelation*? Recall that the autocorrelation is calculated as $r_k = \frac{1}{nc_0} \sum (x_t - \bar{x})(x_{t+k} - \bar{x})$, where $c_0 = \frac{1}{n} \sum (x_t - \bar{x})^2$ is the auto covariance at lag 0.

```
# lets make some vectors to simplify our life
xt <- ex_ts[1:(length(ex_ts)-7)]
xtp7 <- ex_ts[8:length(ex_ts)]

# for autocorrelation, we use the length of the total series as n
# and the variance of the total series for s^2
n <- length(ex_ts)
s2 <- var(ex_ts)
```

```r
xbar <- mean(ex_ts)

# calculate NORMAL CORRELATION
c0 <- (1/n) * sum((ex_ts - xbar) * (ex_ts - xbar))
1/(n*c0) * sum((xt - xbar) * (xtp7 - xbar))
```

```
[1] 0.875
```

So what is the takeaway? Even though $\{x_t\}$ and $\{x_{t+7}\}$ are perfectly *correlated*, their *autocorrelation* is less than 1, because we calculate autocorrelation based on summaries of the entire time series $(n, c_0, \bar{x})$. This results in an autocorrelation of less than 1 when we compare a series to a k-lagged version of itself, which is k time points shorter. We can verify these calculations by looking at the output of `acf`.

```r
# lets make some vectors to simplify our life
tmp <- acf(ex_ts, plot = F)
knitr::kable(
  tibble(
    lag = 0:(length(c(tmp$acf))-1),
    acf = c(tmp$acf)
  )
)
```

lag	acf
0	1.0000000
1	0.2901786
2	-0.1964286
3	-0.4553571
4	-0.4821429
5	-0.2723214
6	0.1785714
7	0.8750000
8	0.2589286
9	-0.1651786
10	-0.3928571
11	-0.4196429
12	-0.2410714
13	0.1473214
14	0.7500000
15	0.2276786
16	-0.1339286
17	-0.3303571