

Time Series Midterm Review

David Josephs

2019-06-27

Contents

1	Expected Values, Variance, etc.	5
1.1	Notation and Some Definitions	5
1.2	Population Means and Variances	5
1.3	Estimating Mean and Variance for Multiple Realizations	5
1.4	Expected Value	6
1.5	Variance and Covariance	7
2	A Brief Discussion of Stationarity	9
2.1	Condition One: Constant Mean	9
2.2	Condition Two: Constant Variance	11
2.3	Condition 3: Constant autocorrelation	11
3	Autocorrelation	13
3.1	Independence	13
3.2	Serial Dependence / Autocorrelation	13
4	Autocorrelation Concepts and Notation	19
4.1	Theoretical concept of ρ	19
4.2	autocovariance	19
4.3	autocorrelation	19
4.4	Stationary Covariance	19
4.5	Stationary Time Series: Putting it Together	24
5	Practical Autocorrelation	25
5.1	Estimation!	25
6	The Frequency Domain	27
6.1	Review of sine and cosine	27
6.2	Frequency and Period	28
6.3	psueudo-periodicity	29
6.4	Aperiodicity	29
6.5	Frequency	29
6.6	Frequency Terminology	29
6.7	Composite sine function	29

Chapter 1

Expected Values, Variance, etc.

1.1 Notation and Some Definitions

In time series, we denote the response variable as X_t , and clearly the explanatory variable as t . To denote a specific point in time, let us say $t = 9$, we would say X_9 .

The collection of all X_n is a stochastic process known as a time series.

1.1.1 Realizations and Ensembles

A **realization** is a particular instance of a time series. It is one of the infinitely many possible time series that could have been observed. As a time series is a stochastic process, each one could be entirely different. The **realization** is the one that was *actually* observed.

Sometimes a realization is the only realization possible. We can theorize what an infinite population of those realizations, but it is often the case that there is only one.

An **ensemble** is the theoretical totality of all possible realizations. It is the *ensemble of realizations*. We can think of that as the *population* of realizations. When we are thinking of means and variances, we think of them in terms of the ensemble, even though we may only be able to look at a single realization.

1.2 Population Means and Variances

It is important to note that when dealing with an ensemble of realizations, each time point is allowed to have its own mean and variance.

1.2.1 The Mean of X_t

μ_t is the mean of all possible realizations of X_t *for a fixed t* .

1.2.2 Variance of X_t

σ_t^2 is the variance of all possible realizations of X_t *for a fixed t*

1.3 Estimating Mean and Variance for Multiple Realizations

Let us say we have n realizations of X_t at time t . The average can be then written as

$$\hat{\mu}_t = \frac{1}{n} \sum_{i=1}^n X_{t_n}$$

Variance is calculated the same way. We are making sample means and sample variances. If we assume the population is normally distributed, this allows us to take confidence intervals on the mean of our set of realizations.

However, We have a problem: What to do when we only have one realization?

1.4 Expected Value

In short, the expected value ($E[X]$) of a random variable (RV) denoted by X is the **mean** or more intuitively the *long run average* of the event that variable represents.

1.4.1 Discrete RVs

A discrete RV is an RV in which X cannot take on any value, it has specific values it can exist at and that is it.

The formula for EV of a discrete RV is as follows:

$$E[X] = \sum X P(X) = \mu$$

1.4.2 Continuous RVs

A continuous RV has instead of discrete values a *probability distribution/density function*, $f(x)$.

The formula for EV of a continuous RV is as follows:

$$E[X] = \int_a^b x f(x) dx = \mu$$

Note that is directly analogous to the discrete RV, and that a and b can span to $\pm\infty$

Important Note

$$E[X] = \mu$$

That is the expected value of a RV is equivalent to the population mean

1.4.3 Some Rules

Let a and b be constants

$$E[a] = a$$

$$E[aX] = aE[X] = a\mu$$

$$E[aX + b] = aE[X] + E[b] = aE[x] + b = a\mu + b$$

Knowing these rules, lets try out a challenge:

1.4.4 A challenge in Expected Values

Find $E[\bar{x}]$ where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

E.G find the expected value of an average

1.4.5 A solution

First, we can factor out the $1/n$, so we have

$$\frac{1}{n} E \left[\sum_{i=1}^n x_i \right]$$

then we can say that the sum is just $x_1 + x_2 + x_3 + \dots + x_n$. This would lead us to the conclusion that

$$E[\bar{x}] = \frac{1}{n} \sum_{i=1}^n E[x_i]$$

If we let each x_i be a constant then we have:

$$E[\bar{x}] = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

This makes sense as \bar{x} is the average value of x , and so the expected value of x for a very high n is the average value of x

1.5 Variance and Covariance

1.5.1 Variance

The variance is the *dispersion*. Variance is defined as

$$\text{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \sigma^2$$

$$\widehat{\text{Var}(X)} = \sum (x_i - \bar{x})^2$$

1.5.2 Covariance

Assume we have two RVs, X and Y :

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

$$\widehat{\text{Cov}(X, Y)} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

The variance is, in essence, the sum of cross products.

1.5.3 Some generalizations on covariance:

If y tends to increase with x , then $\text{Cov}(X, Y) > 0$

If Y tends to decrease with X , then $\text{Cov}(X, Y)$ is **Less than 0**

If it appears as a random cloud, then $\text{Cov}(X, Y)$ is approximately 0

1.5.4 A challenge

Let X be a RV. What is $\text{Cov}(X, X)$?

1.5.5 A solution

$$\text{Cov}(X, X) = E[(X - \mu_x)(X - \mu_x)]$$

We see that the arguments in the left and right parens are exactly the same. Thus, we can rewrite the equation as:

$$\text{Cov}(X, X) = E[(X - \mu_x)^2] = \text{Var}(X)$$

1.5.6 A note on covariance and correlation

Correlation is covariance but standardized.

$$\begin{aligned} \text{Corr}(X, Y) &= \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)} \\ &= \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \\ \widehat{\text{Corr}(X, Y)} &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{ns_x s_y} \end{aligned}$$

Chapter 2

A Brief Discussion of Stationarity

In order for a time series to be considered stationary, it must satisfy *three* conditions:

- Constant Mean with Time
- Constant Variance with Time
- Constant Autocorellation with Time

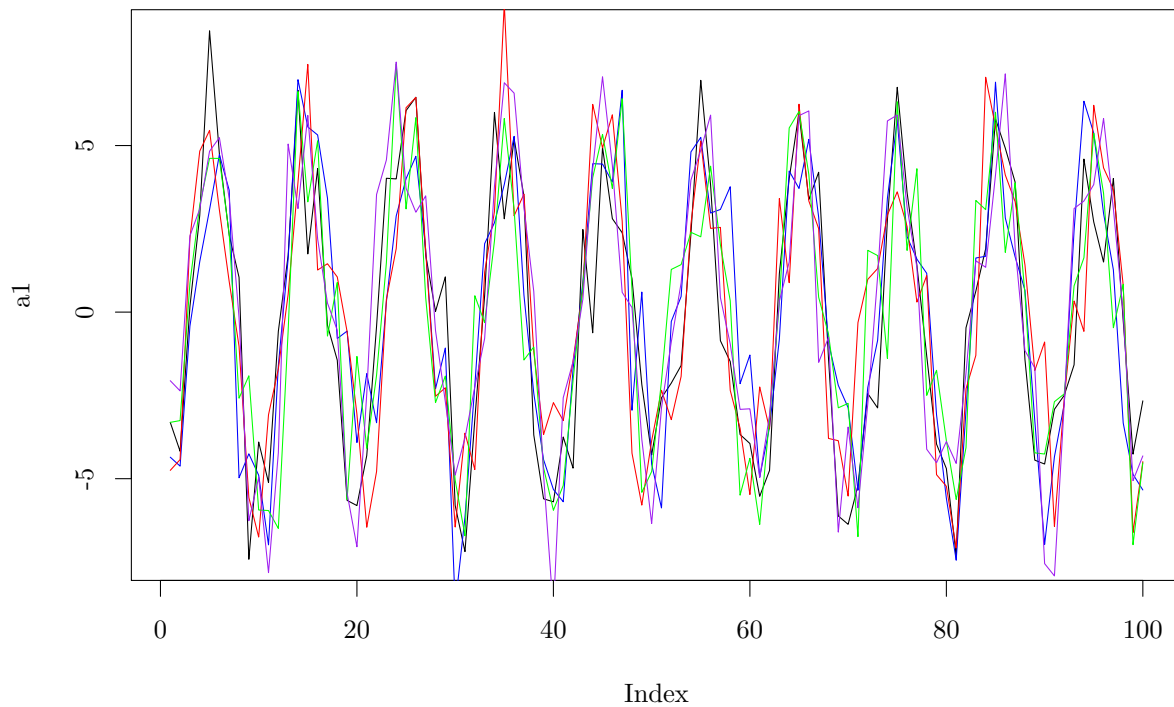
2.1 Condition One: Constant Mean

Mean is constant with time. That is $E[X_t] = \mu$. Note the lack of a little tiny t in mu! It is indepoent of tine! That is stationarity condition number 1!

Important result: If we assume constant mean, we can use *all* of the data to estimate the mean

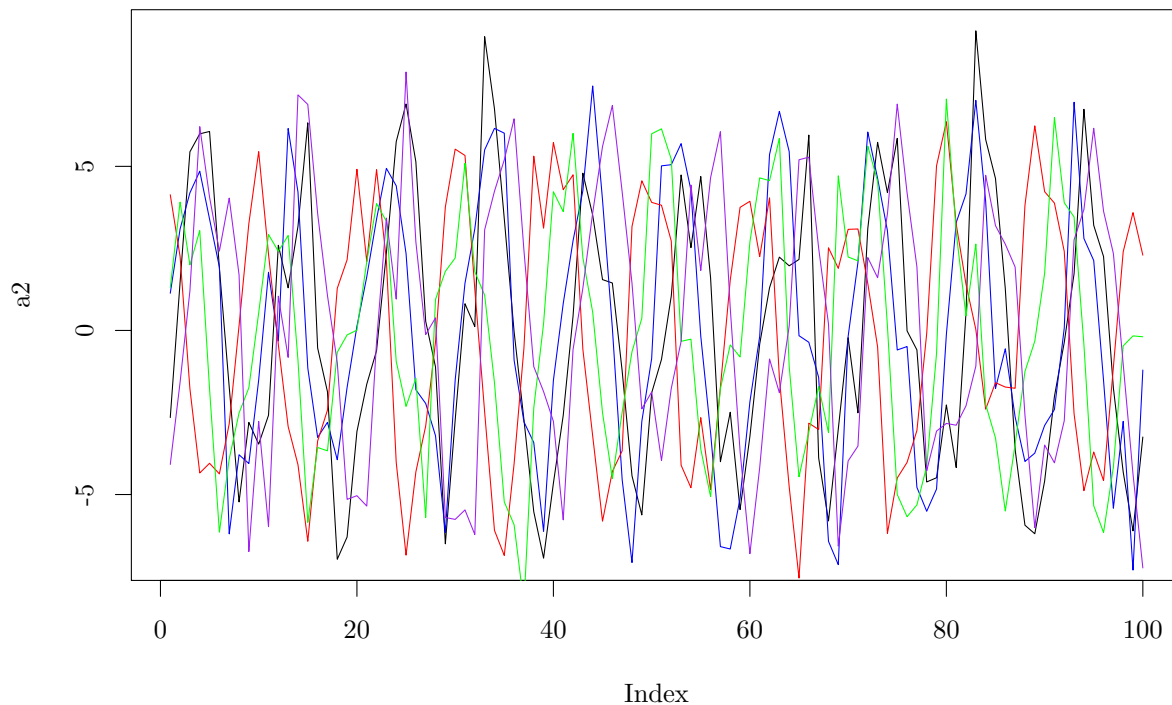
Here we can see after 5 realizations that the mean is clearly not constant with time

```
library(tswge)
a1 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(3,
0), vara = 3, plot = F)
b1 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(3,
0), vara = 3, plot = F)
c1 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(3,
0), vara = 3, plot = F)
d1 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(3,
0), vara = 3, plot = F)
e1 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(3,
0), vara = 3, plot = F)
```



Now let's look at 5 realizations which come from the same realization, which might not have a constant mean:

```
a2 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(runif(1,
0, 2 * pi), 0), vara = 3, plot = F)
b2 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(runif(1,
0, 2 * pi), 0), vara = 3, plot = F)
c2 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(runif(1,
0, 2 * pi), 0), vara = 3, plot = F)
d2 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(runif(1,
0, 2 * pi), 0), vara = 3, plot = F)
e2 = gen.sigplusnoise.wge(100, coef = c(5, 0), freq = c(0.1, 0), psi = c(runif(1,
0, 2 * pi), 0), vara = 3, plot = F)
```



We cannot say for sure whether the mean is constant or not in this case!

2.2 Condition Two: Constant Variance

Variance does not depend on time

Variance is *constant* and finite*

That is :

$$\text{Var}[X_t] = \sigma^2 \neq \infty$$

If we can make this assumption, then we can use all of the data to make the variance. This is typically the hardest one to tell, and it will take some practice to see.

2.3 Condition 3: Constant autocorrelation

Correlation of X_{t_1} and X_{t_2} only depends on $t_2 - t_1$

That is, the correlation between points depends only on how far apart they are in time, not where they are in time.

To describe this mathematically:

Let

$$h = t_2 - t_1$$

then

$$\text{Cor}(X_t, X_{t+h}) = \rho_h$$

ρ represents the population correlation coefficient.

Chapter 3

Autocorrelation

```
library(tswge)
library(ggplot2)
acdf <- function(vec) {
  vacf <- acf(vec, plot = F)
  with(vacf, data.frame(lag, acf))
}
ggacf <- function(vec) {
  ac <- acdf(vec)
  ggplot(data = ac, aes(x = lag, y = acf)) + geom_hline(aes(yintercept = 0)) +
    geom_segment(mapping = aes(xend = lag, yend = 0))
}

tplot <- function(vec) {
  df <- data.frame(X = vec, t = seq_along(vec))
  ggplot(data = df, aes(x = t, y = X)) + geom_point() + geom_line()
}
```

3.1 Independence

Two events are independent in a time series if the probability that an event at time t occurs in no way depends on the occurrence of any event in the past or affects any event in the future. Mathematically, this is written as:

$$\text{Independence : } P(x_{t+1}|X_t) = P(X_{t+1})$$

3.2 Serial Dependence / Autocorrelation

3.2.1 A definition

If two events are independent, their correlation is 0. That is if X_t and X_{t+k} are independent, $\rho_{x_t, x_{t+k}} = 0$

Corollary: If the correlation between two variables is not zero, then they are not independent

In other words if $\rho_{x_t, x_{t+k}} \neq 0$, they are *not* independent.

3.2.2 Autocorrelation Plots

3.2.3 Dependent(ish) data

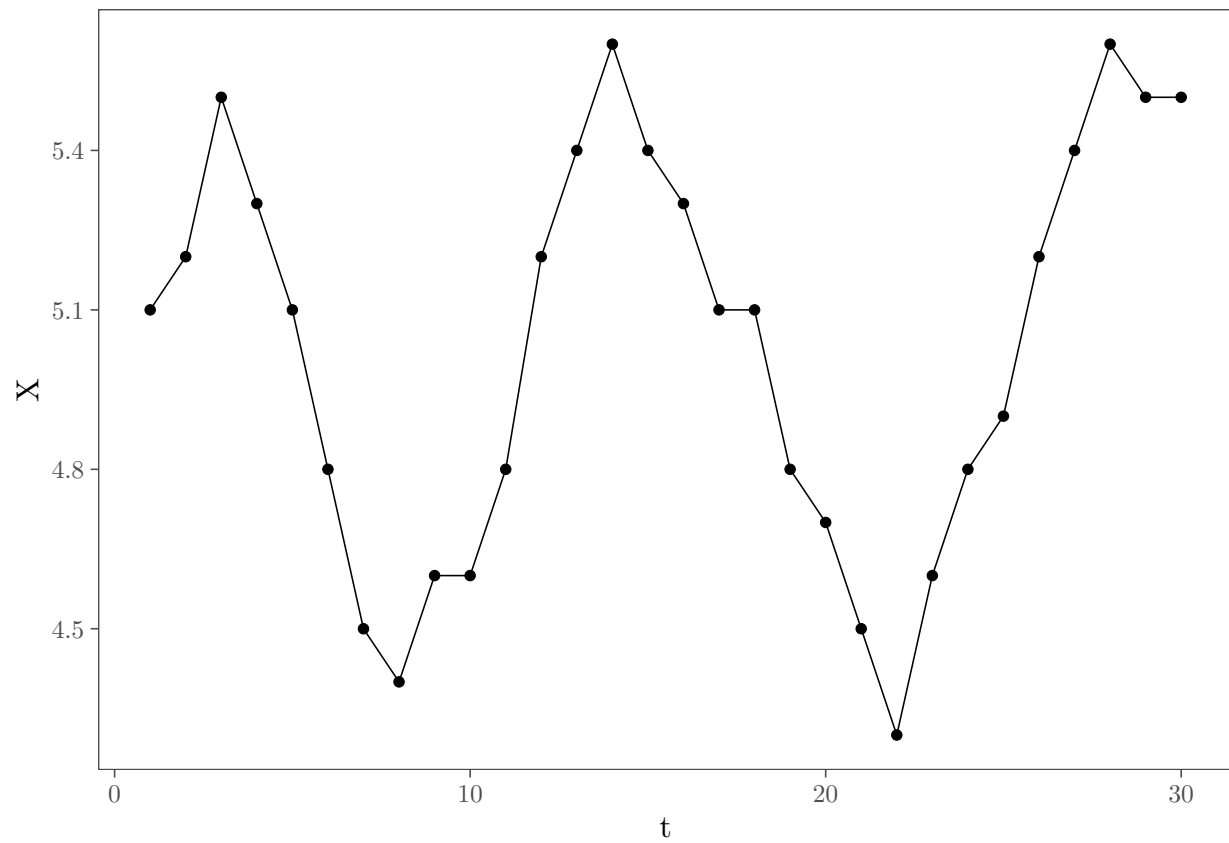
In time series we look at the autocorrelation between X_t and X_{t+1} etc (with t and $t+1$ it is lag 1 autocorrelation)

For example, visually, let us define a vector, $Y5$ and take its autocorrelation:

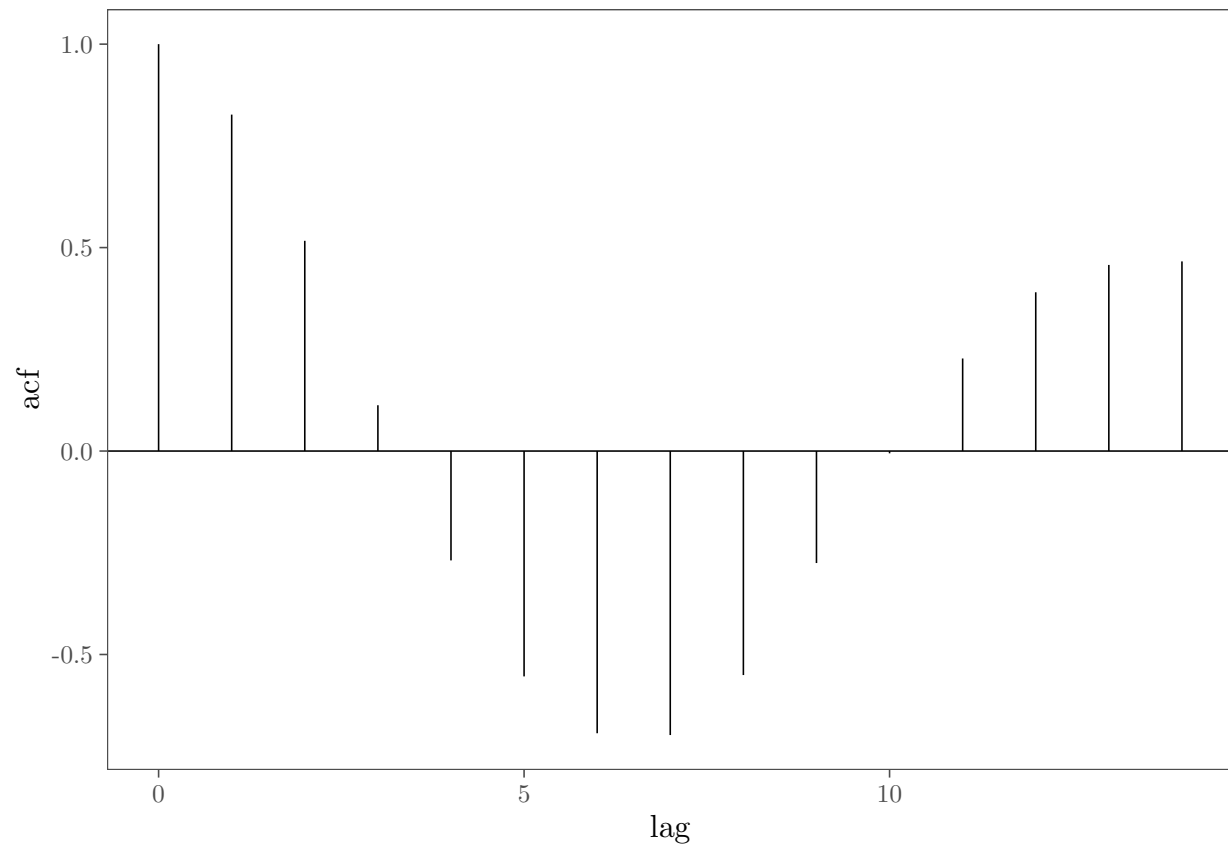
```
Y5 <- c(5.1, 5.2, 5.5, 5.3, 5.1, 4.8, 4.5, 4.4, 4.6, 4.6, 4.8, 5.2, 5.4, 5.6,
        5.4, 5.3, 5.1, 5.1, 4.8, 4.7, 4.5, 4.3, 4.6, 4.8, 4.9, 5.2, 5.4, 5.6, 5.5,
        5.5)
knitr::kable(acfdf(Y5))
```

lag	acf
0	1.0000000
1	0.8268482
2	0.5166514
3	0.1125421
4	-0.2687774
5	-0.5538901
6	-0.6935454
7	-0.6979912
8	-0.5505557
9	-0.2752524
10	-0.0056082
11	0.2274804
12	0.3902111
13	0.4573468
14	0.4660549

```
tplot(Y5) + ggthemes::theme_few()
```



```
ggacf(Y5) + ggthemes::theme_few()
```



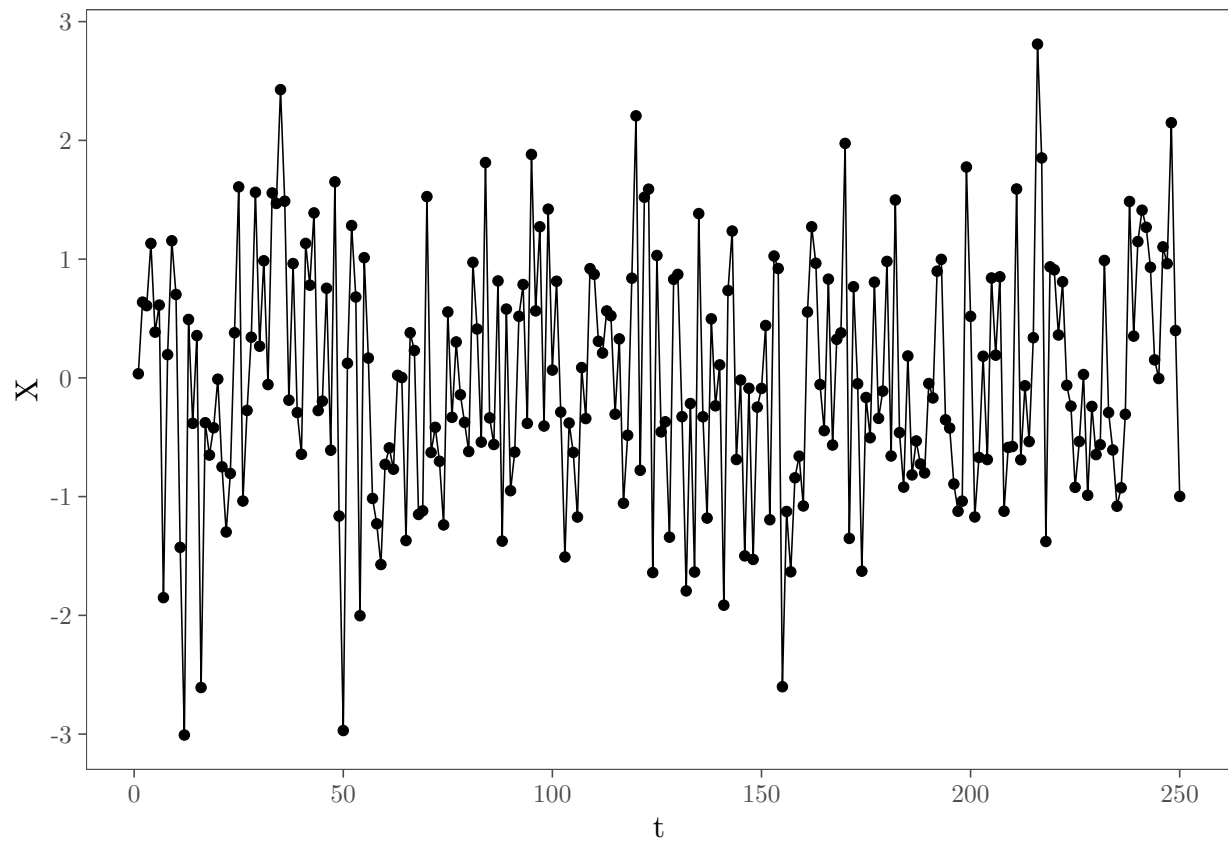
3.2.4 Independent(ish) data!

Now let us look at autocorrelation of *independent* data

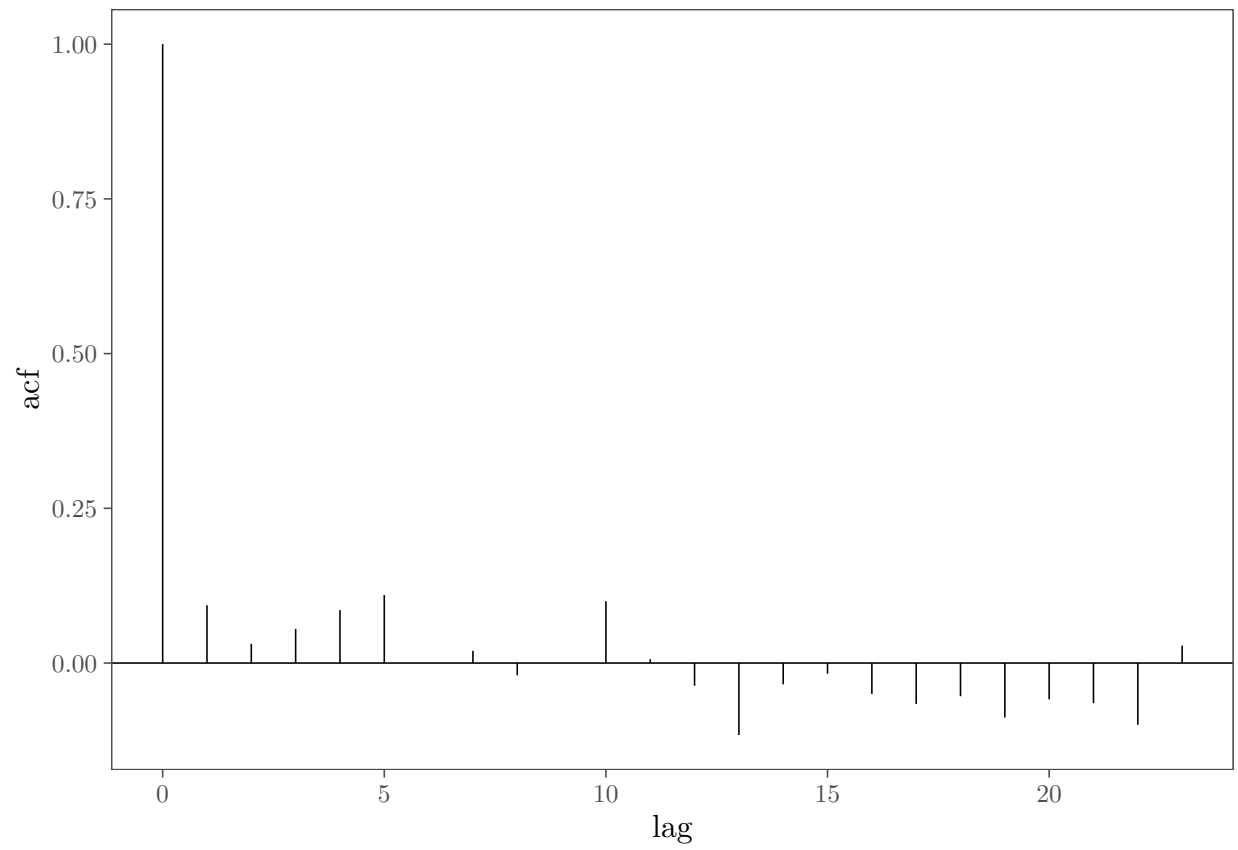
```
xs = gen.arma.wge(n = 250)
```

Let's check it out

```
tplot(xs) + ggthemes::theme_few()
```

```
ggacf(xs) + ggthemes::theme_few()
```



We see that the autocorrelation is more or less zero

Chapter 4

Autocorrelation Concepts and Notation

4.1 Theoretical concept of ρ

$$\rho = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{\text{covariance between } X \text{ and } Y}{\sigma_X \sigma_Y}$$

In a stationary time series, we have autocorrelation and autocovariance

4.2 autocovariance

$$\gamma_k = E[(X_t - \mu)(X_{t+k} - \mu)]$$

4.3 autocorrelation

$$\rho_k = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma_{X_t} \sigma_{X_{t+k}}} = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma_X^2} = \frac{\gamma_k}{\sigma_X^2}$$

This works because of constant variance and constant mean!

Note that in a stationary time series:

$$\sigma_X^2 = 1E[(X_t - \mu)^2] = E[(X_t - \mu)(X_t - \mu)] = \gamma_0$$

Therefore:

$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

4.4 Stationary Covariance

Correlation is not affected by where, only by how far apart, that is:

$$\rho_h = \text{Cor}(X_t, X_{t+h})$$

Let us try it out with this little snippet:

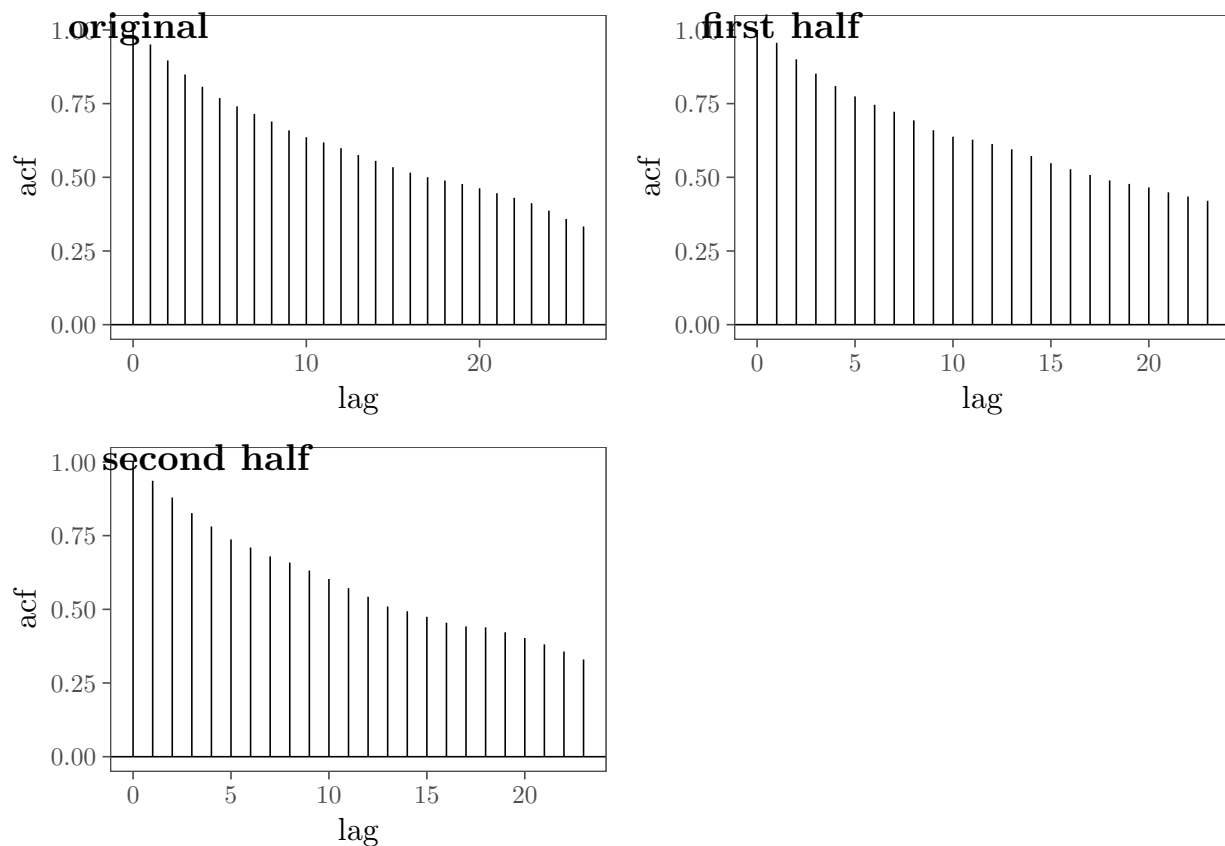
```

ggsplitacf <- function(vec) {
  h1 <- vec[1:(length(vec)/2)]
  h2 <- vec[(length(vec)/2):length(vec)]
  first <- ggacf(vec) + ggthemes::theme_few()
  second <- ggacf(h1) + ggthemes::theme_few()
  third <- ggacf(h2) + ggthemes::theme_few()
  cowplot::plot_grid(first, second, third, labels = c("original", "first half",
    "second half"), nrow = 2, align = "v")
}

```

```
Realize = gen.arma.wge(500, 0.95, 0, plot = T, sn = 784)
```

```
ggsplitacf(Realize)
```

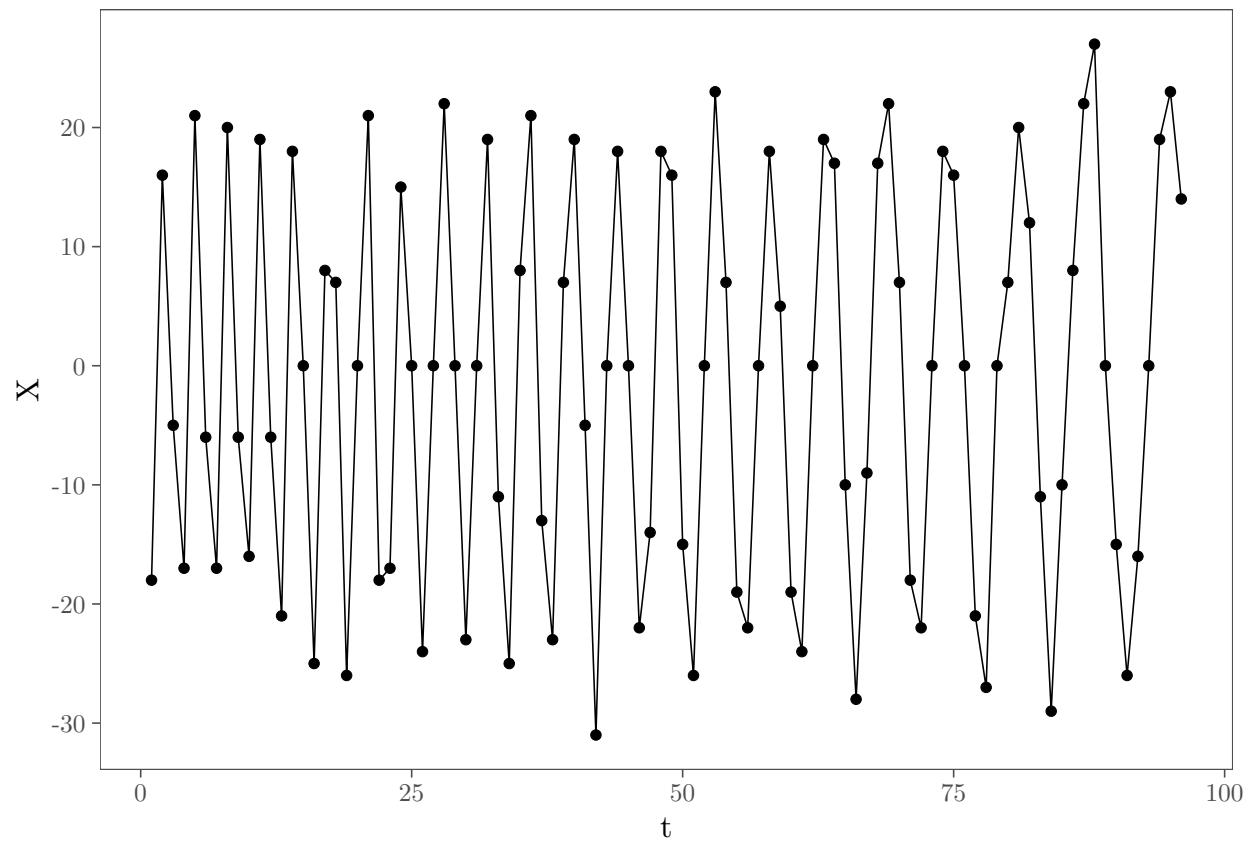


This looks pretty stationary to me! Let's try with some sample data:

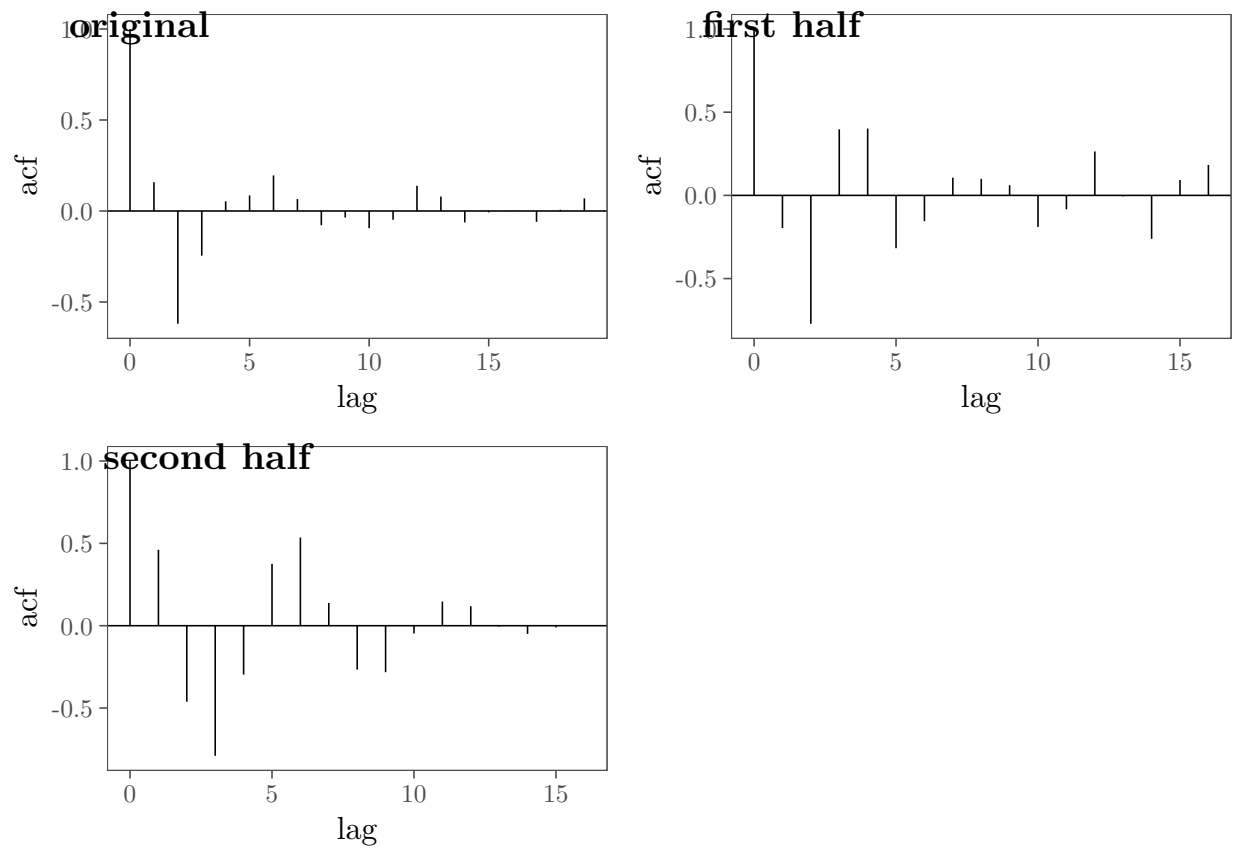
```

data("noctula")
tplot(noctula) + ggthemes::theme_few()

```

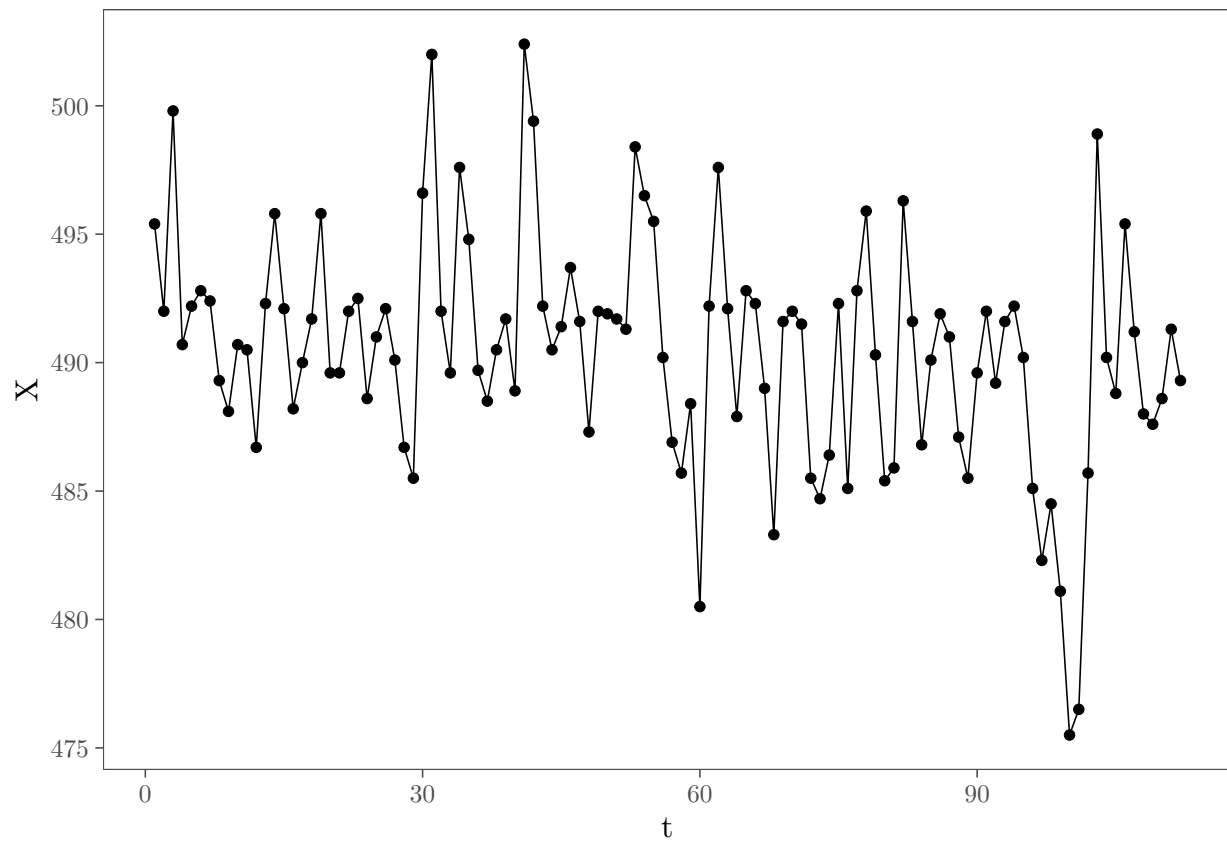


```
ggsplitacf(noctula)
```

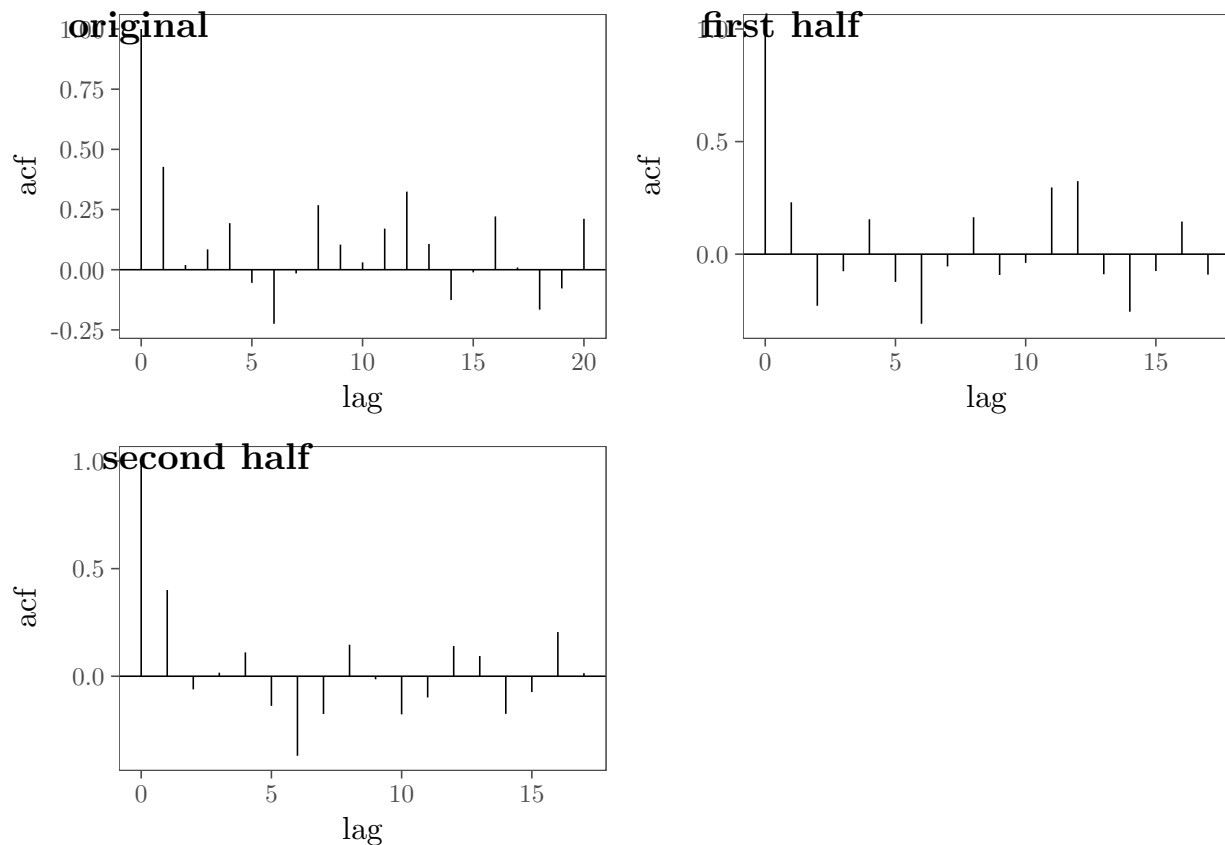


So in this case our mean probably looks constant, our variance could be constant as well, however the ACFs are clearly different. Let's look at a different dataset:

```
data(lavon)
tplot(lavon) + ggthemes::theme_few()
```



```
ggsplitacf(lavon)
```



In this case, our mean looks maybe to be constant around 495, our variance does not appear to be constant, and our ACF looks pretty good, however I would say overall this probably is not stationary, due to the variance and slightly off ACF.

4.5 Stationary Time Series: Putting it Together

4.5.1 Workflow:

- Step 1: Check condition 1
 - see if the assumption is more or less met
- Step 2: Check condition two
 - same thing
- Step 3: Check for condition three

If one of the things is iffy, and the other two arent, we are going to be just fine.

Chapter 5

Practical Autocorrelation

5.1 Estimation!

5.1.1 $\rho_k \rightarrow 0$

For a stationary time series, if the autocorrelation approaches zero, then :

A Single realization lets us estimate mean, variance, and autovariance!

Remember that $\rho_k = \frac{\gamma_k}{\gamma_0}$, and we use n-k pairs to calculate it (the summation)

5.1.1.1 Mean

Just calculate the mean normally for this case.

5.1.1.2 Variance

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \sum_{k=-(n-1)}^{n-1} \left(1 - \frac{|k|}{n}\right) \rho_k$$

σ^2 is calculated as normal, we will see ρ_k next!

remember !

Now it is time for some code!

```
library(glue)

##
## Attaching package: 'glue'
## The following object is masked from 'package:dplyr':
##
## collapse

xbar <- function(xs) {
  mean(xs)
}

ghat_zero <- function(xs) {
  summand <- (xs - xbar(xs))^2
  mean(summand)
}
```

```
}  
  
ghat_one <- function(xs) {  
  lhs <- xs[1:(length(xs) - 1)] - xbar(xs)  
  rhs <- xs[2:length(xs)] - xbar(xs)  
  summand <- lhs * rhs  
  summate <- sum(summand)  
  summate/length(xs)  
}  
  
rhohat_zero <- 1  
rhohat_one <- function(xs) {  
  ghat_one(xs)/ghat_zero(xs)  
}  
v <- c(76, 70, 66, 60, 70, 72, 76, 80)  
xbar(v)
```

```
## [1] 71.25
```

```
ghat_zero(v)
```

```
## [1] 34.9375
```

```
ghat_one(v)
```

```
## [1] 14.74219
```

```
rhohat_one(v)
```

```
## [1] 0.4219589
```

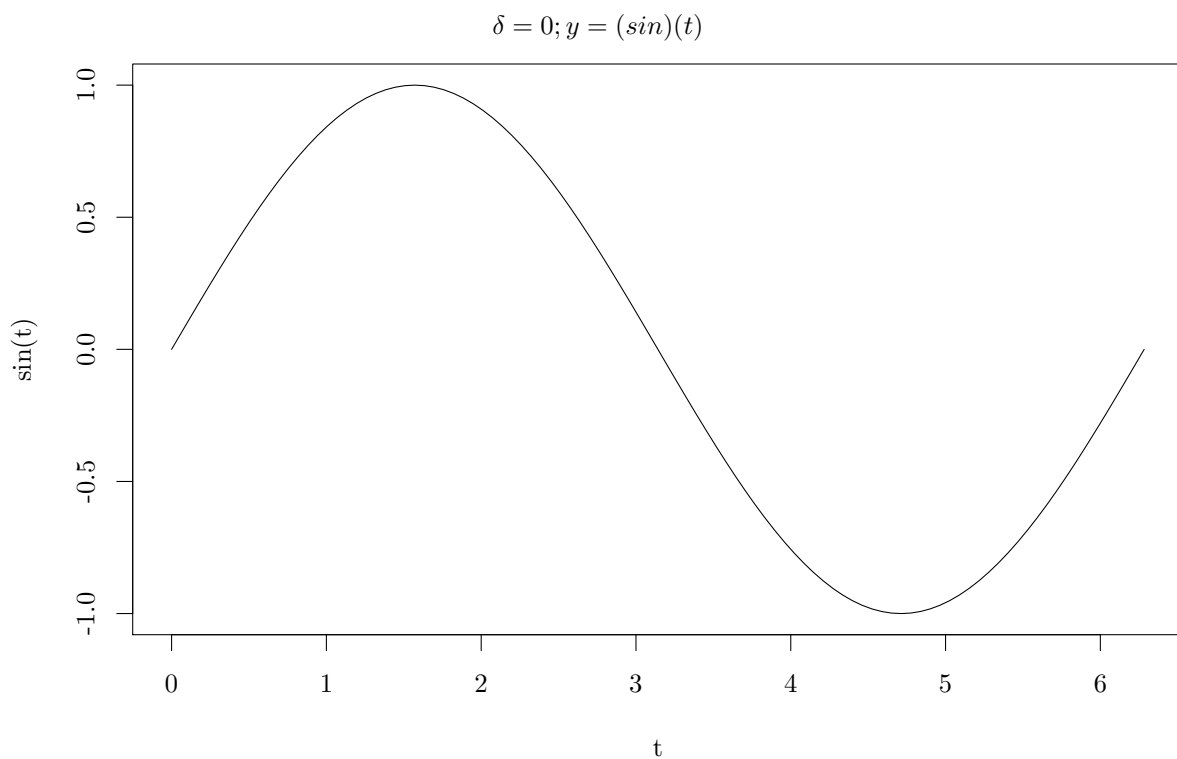
Chapter 6

The Frequency Domain

6.1 Review of sine and cosine

6.1.1 Phase shifts

This is what a sine wave looks like



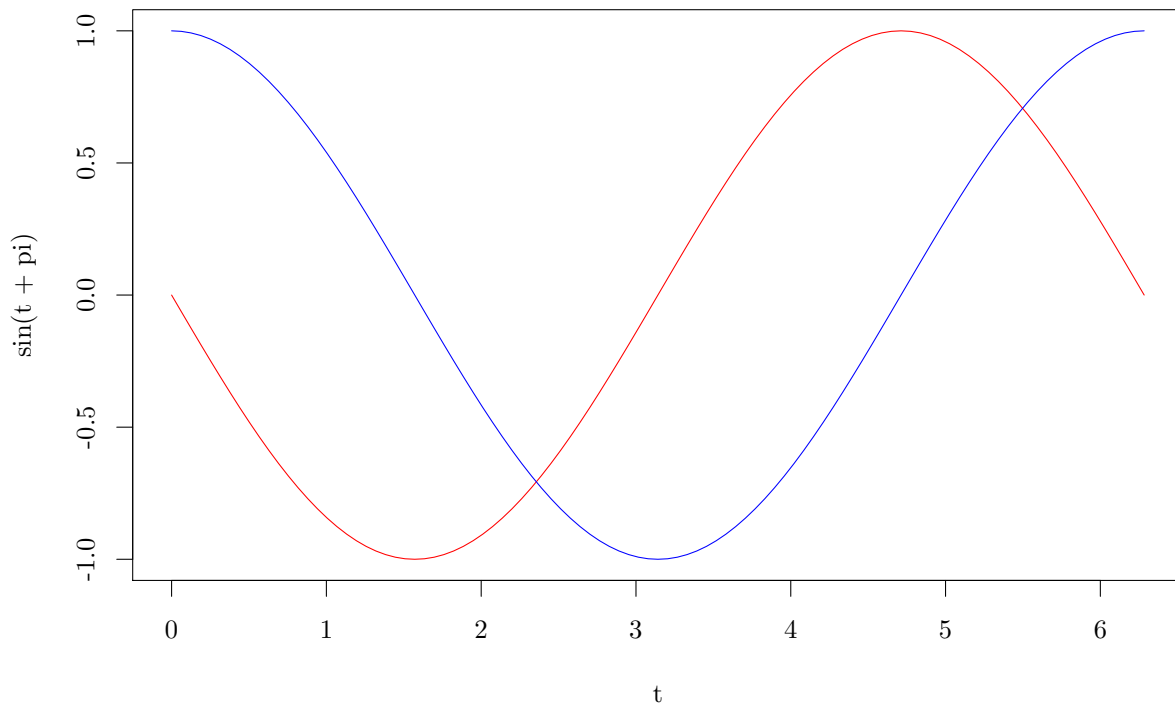
We can phase shift the sin wave by adding

RED:

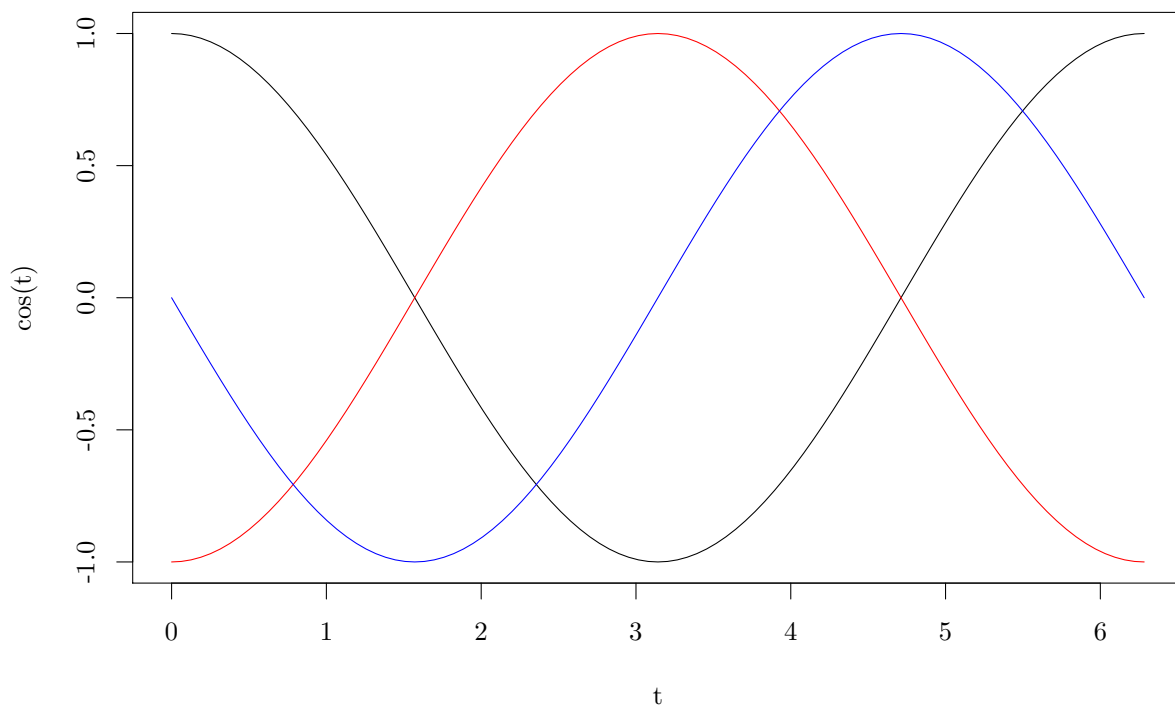
$$\delta = \pi; y = \sin(t + \pi)$$

BLUE:

$$\delta = \frac{\pi}{2}; y = \sin\left(t + \frac{\pi}{2}\right)$$



Now cosine



6.2 Frequency and Period

Let us say $f(x)$ is a periodic function. If $f(x)$ has period p , then p is the smallest value S.T.

$$f(x) = f(x + kp)$$

Cool

6.3 psueudo-periodicity

Data are *psuedo-periodic* with period p if p is the smallest value such that a cycle appears to repeat itself.

6.4 Aperiodicity

Aperiodic data is data in which no p exists.

6.5 Frequency

Frequency is calculated as : $\frac{1}{p}$

It symbolizes the number of cycles per unit(number of periods per unit)

6.6 Frequency Terminology

Frequency of a trigonometruc function:

Let $f(x) = \sin(2\pi ft)$

Then we have that :

$$p = 1/f$$

$$\nu = f$$

Consider:

$$\sin(Bt + C)$$

Then

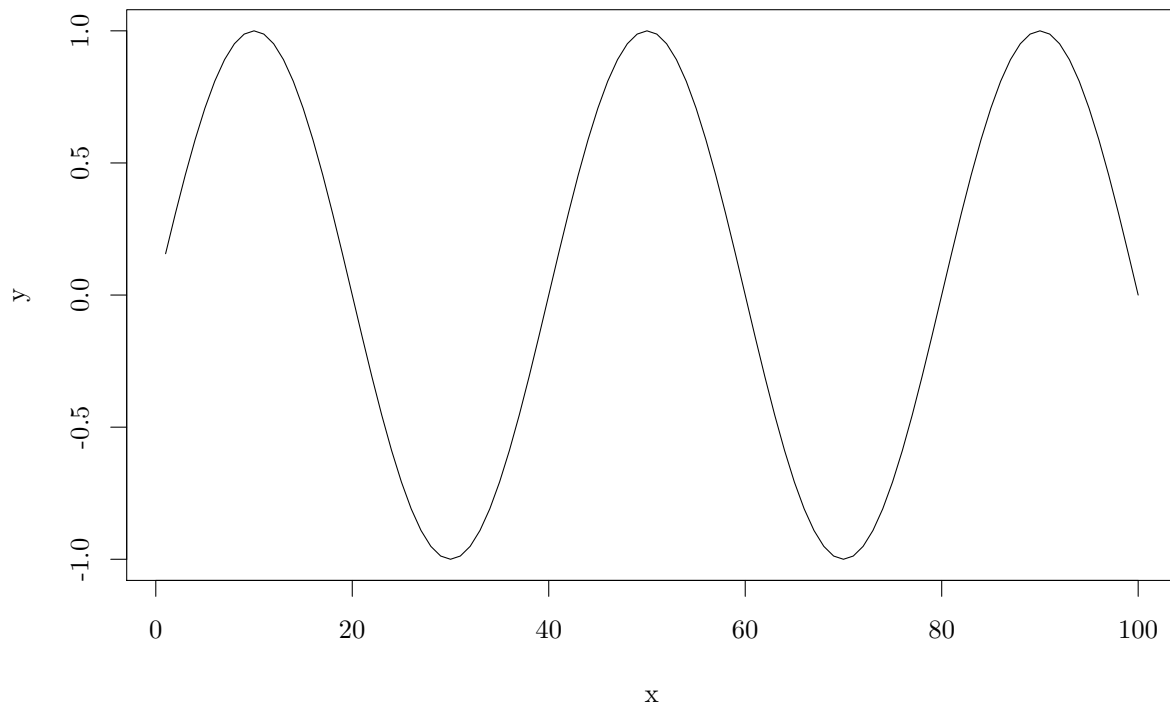
$$p = \frac{2\pi}{B}$$

$$\nu = \frac{B}{2\pi}$$

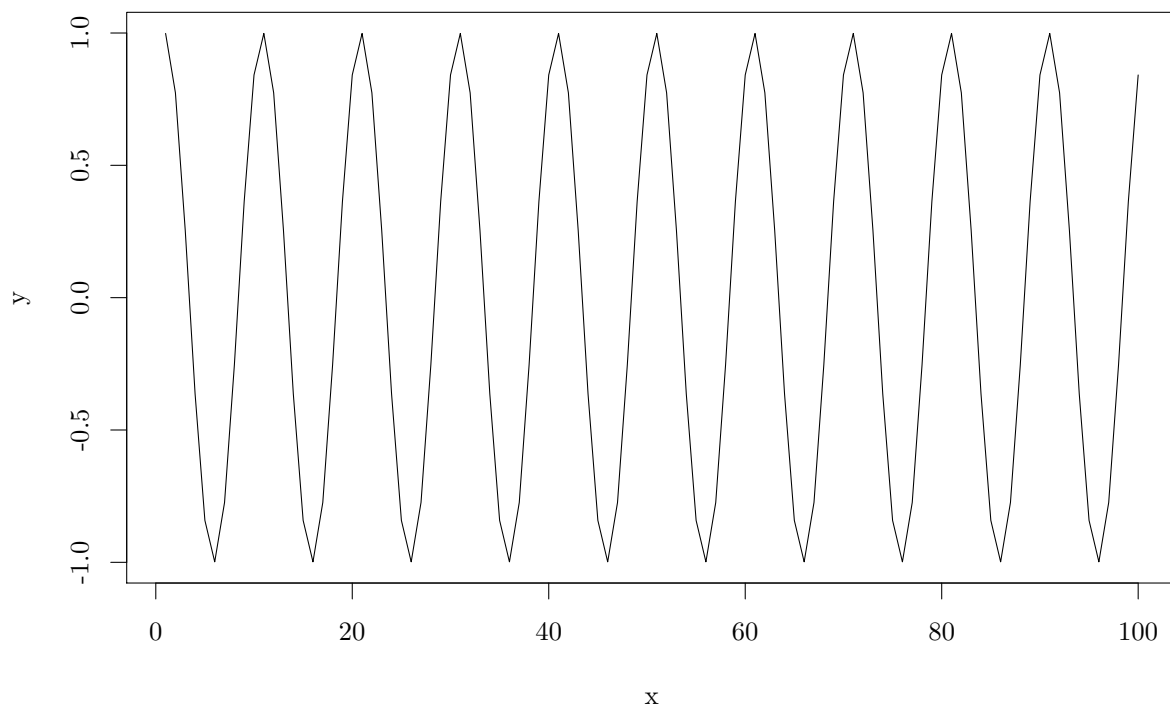
6.7 Composite sine function

```
t <- seq(1, 100, length = 100)
y1 <- sin(2 * pi * 0.025 * t)
y2 <- sin(2 * pi * 0.1 * t + 1)
y3 <- sin(2 * pi * 0.15 * t + 2.5)
lineplotter <- function(x, y) {
  plot(x, y, type = "l")
}

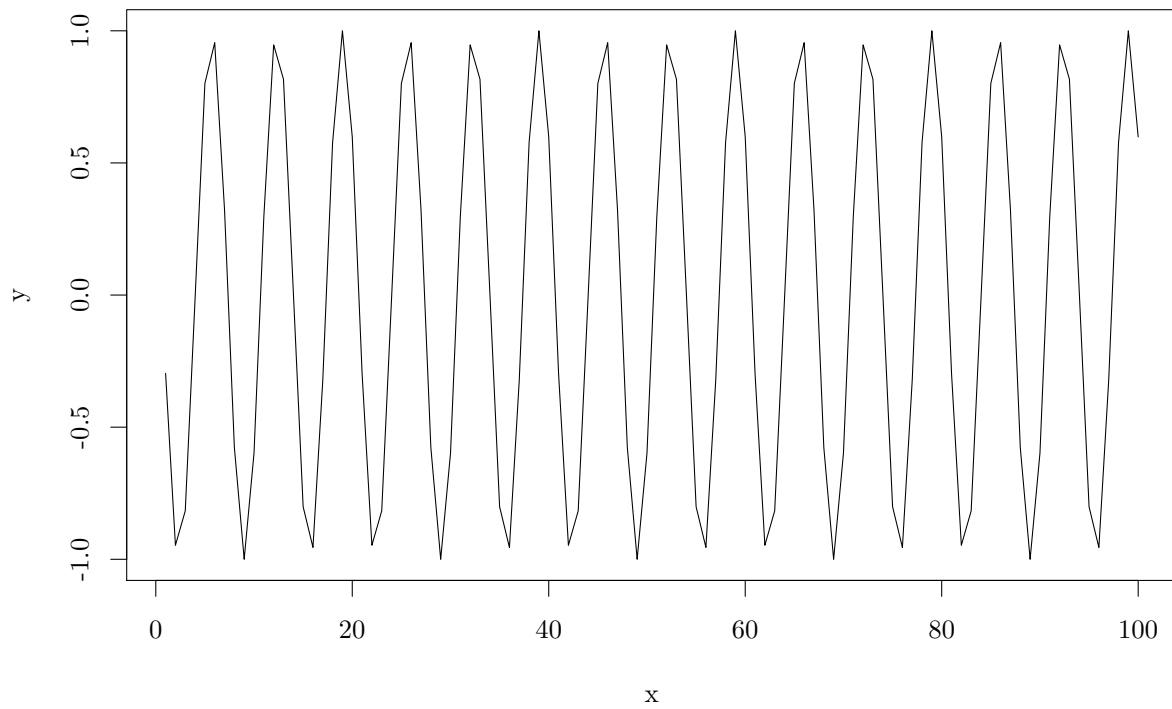
lineplotter(t, y1)
```



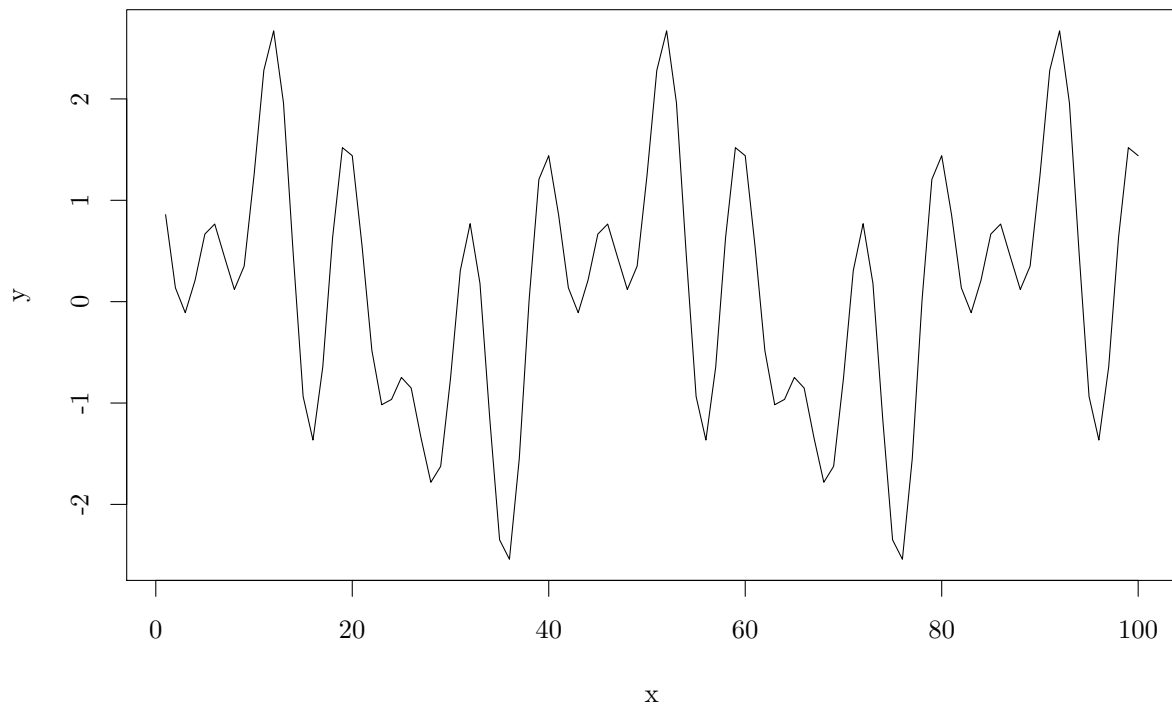
```
lineplotter(t, y2)
```



```
lineplotter(t, y3)
```



```
lineplotter(t, y1 + y2 + y3)
```



this is a signal with three frequencies!