Índice

1. Introducción a la modelización de series temporales **2**

2. Herramientas para desvelar propiedades de una serie temporal **6**
2.1. Análisis gráfico 6
2.1.1. Gráfico de la serie temporal 6
2.1.2. Gráfico rango-media 7
2.2. Determinación del orden de integración 7
2.2.1. Análisis gráfico 8
2.2.2. Contrastes formales sobre el orden de integración 10

3. Función de autocovarianzas y función de autocorrelación **10**

4. Otras herramientas estadísticas **CómoConR** **11**
4.1. Estadísticos descriptivos 11
4.2. Test de normalidad Jarque-Vera 11

Econometría Aplicada. Lección 5

Marcos Bujosa

24 de agosto de 2024

Resumen

En esta lección veremos algunas herramientas estadísticas.

Carga de algunas librerías de R

Primero cargamos la librería `tfarima` (Repositorio Cran: https://cran.r-project.org/web/packages/tfarima/index.html; repositorio GitHub: https://github.com/gallegoj/tfarima)

```
library(tfarima)      # librería de José Luis Gallego para Time Series
library(readr)        # para leer ficheros CSV
library(ggplot2)      # para el scatterplot (alternaticamente library(tidyverse))
library(ggfortify)    # para pintar series temporales
library(jtools)       # para representación resultados estimación
library(zoo)          # para generar objetos ts (time series)
```

y además fijamos los parámetros por defecto para las figuras en `png` del notebook

```
# fijamos el tamaño de las figuras que se generan en el notebook
options(repr.plot.width = 12, repr.plot.height = 4, repr.plot.res = 200)
```

1. Introducción a la modelización de series temporales

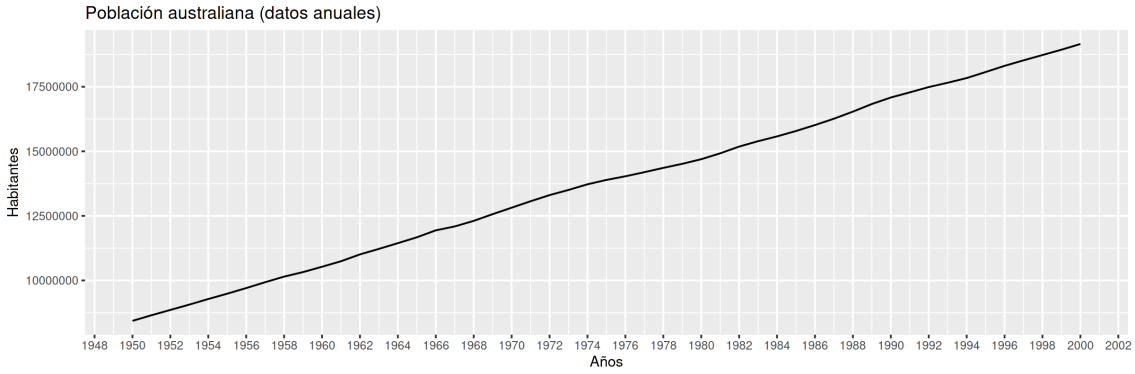
Corresponden a observaciones de un mismo objeto a lo largo del tiempo. El índice indica el instante de cada medición. *El orden cronológico puede ser crucial* al modelar los datos.

- El motivo es que frecuentemente el valor medido en un instante de tiempo está relacionado con otras mediciones próximas en el tiempo (*correlación serial*).
- Si es así, ya no deberíamos asumir que las variables aleatorias del proceso estocástico subyacente, $\boldsymbol{X} = (X_t \mid t \in \mathbb{Z})$, son independientes entre sí.

Esto tiene importantes implicaciones en las técnicas de análisis y los modelos a utilizar.

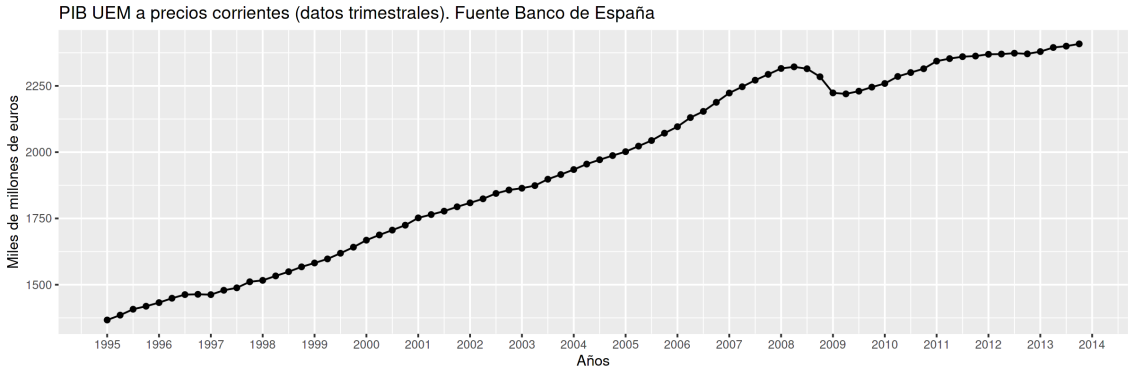
1. Población en Australia

```
PoblacionAustralia_ts = as.ts( read.zoo('datos/PoblacionAustralia.csv',
                                        header=TRUE,
                                        index.column = 1,
                                        sep=",",
                                        FUN = as.yearmon))
p <- autoplot(PoblacionAustralia_ts)
p <- p + labs(y = "Habitantes", x = "Años") + ggtitle("Población australiana (datos anuales)")
p <- p + scale_x_continuous(breaks = scales::pretty_breaks(n = 20))
p
```



2. PIB UEM

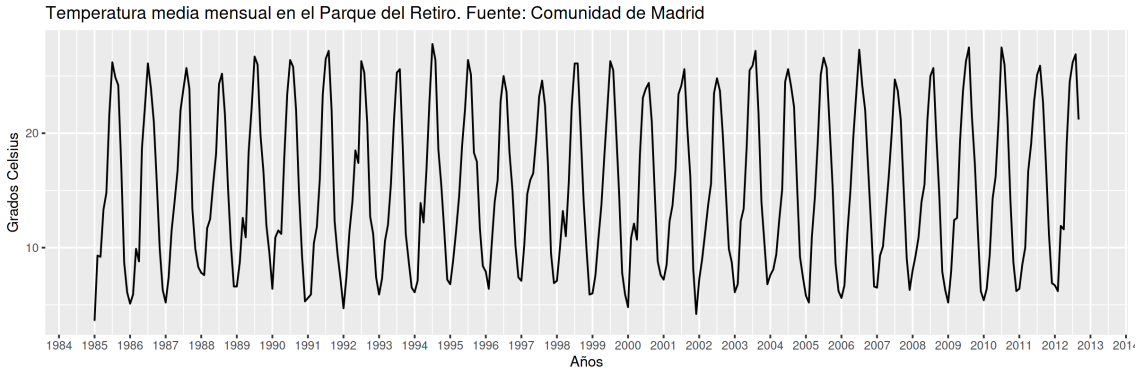
```
PIB_UEM_df <- read_csv("datos/PIB_UEM.csv",
                       show_col_types = FALSE)
fmt <- "%YQ%q"
PIB_UEM_df$Time <- as.yearqtr(PIB_UEM_df$obs, format = fmt)
# head(PIB_UEM_df,3)
P <- ggplot(PIB_UEM_df, aes(Time, PIB))
P <- P + geom_point() + geom_line()
P <- P + scale_x_continuous(breaks = scales::pretty_breaks(n = 15))
P <- P + labs(y = "Miles de millones de euros", x = "Años") + ggtitle("PIB UEM a precios corrientes (datos trimestrales).
P
```



3. Temperatura media en el Parque del Retiro. Madrid

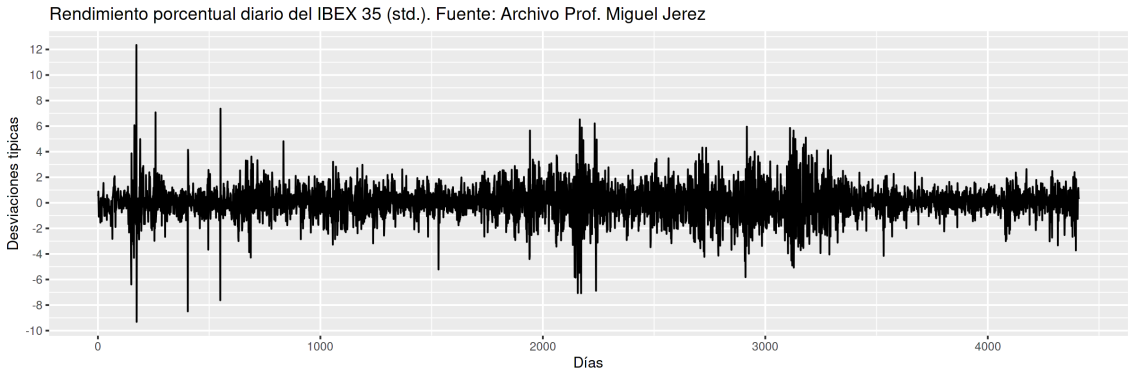
```
TemperaturaRetiro_df <- read_csv("datos/Retiro.txt", show_col_types = FALSE)
# Añadimos fechas
TemperaturaRetiro_df$Time <- as.yearmon(1985 + seq(0, nrow(TemperaturaRetiro_df)-1)/12)

P <- ggplot(TemperaturaRetiro_df, aes(Time, TemperaturaMedia))
P <- P + geom_line() # + geom_point()
P <- P + scale_x_continuous(breaks = scales::pretty_breaks(n = 25))
P <- P + labs(y = "Grados Celsius", x = "Años") + ggtitle("Temperatura media mensual en el Parque del Retiro. Fuente: Com
P
```



4. Rendimiento porcentual diario del IBEX 35 (std)

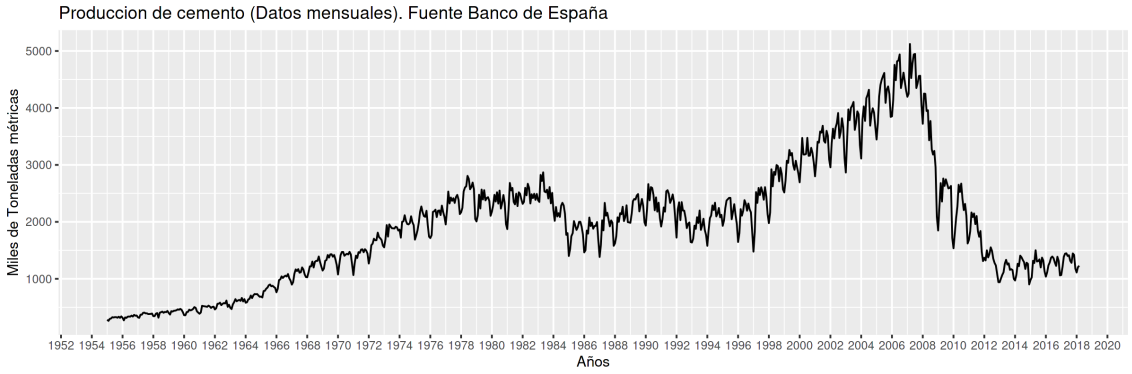
```
IBEX35_ts = as.ts( read.csv.zoo("datos/IBEX35.csv",
                                strip.white = TRUE))
P <- autoplot(IBEX35_ts) + scale_y_continuous(breaks = scales::pretty_breaks(n = 12))
p <- P + labs(y = "Desviaciones tipicas", x = "Días") + ggtitle("Rendimiento porcentual diario del IBEX 35 (std.). Fuente
p
```



- Datos centrados y estandarizados, i.e. el eje vertical está en desviaciones típicas.
- Los *volatility clustering* son característicos de series financieras de alta frecuencia.

5. Producción de cemento

```
ProduccionCemento_df <- read_csv("datos/ProduccionCemento.csv",
                      show_col_types = FALSE)
fmt <- "%YM%m"
ProduccionCemento_df$Time <- as.yearmon(ProduccionCemento_df$obs, format = fmt)
# head(ProduccionCemento_df,3)
P <- ggplot(ProduccionCemento_df, aes(Time, ProduccionCemento))
P <- P + geom_line() # + geom_point()
P <- P + scale_x_continuous(breaks = scales::pretty_breaks(n = 25))
P <- P + labs(y = "Miles de Toneladas métricas", x = "Años") + ggtitle("Produccion de cemento (Datos mensuales). Fuente B
P
```



Correlación serial vs muestreo aleatorio simple

Con datos de

sección cruzada solemos asumir que el muestreo es aleatorio simple

- i.e., los datos son realizaciones de variables aleatorias i.i.d.

series temporales dicha asunción resulta generalmente errónea

- con frecuencia el nivel esperado (o la volatilidad) parece cambiar con t
- con frecuencia hay dependencia temporal (correlación serial).

Ejemplo: dada la evolución de los datos, no parece aceptable asumir que $ProdCemento_{1960M01}$ tiene la misma distribución que $ProdCemento_{2000M04}$ (ni que sea independiente de $ProdCemento_{1959M01}$).

Veamos por qué esto genera dificultades...

Consideremos el proceso estocástico

$$\boldsymbol{X} = (X_t \mid t = 0, \pm 1, \pm 2, \ldots).$$

Caracterizar su distribución conjunta (todos los momentos) es demasiado ambicioso.

Así que, tentativamente, vamos a fijarnos *solo* en los dos primeros momentos:

$$E(X_t) = \mu_t \quad \text{y} \quad Cov(X_t, X_k) = E\big[(X_t - \mu_t)(X_k - \mu_k)\big] = \lambda_{t,k}; \quad t, k \in \mathbb{Z}$$

(si $k = t$ entonces $\lambda_{t,t} = Var(X_t) = \sigma_t^2$).

Si el proceso $\boldsymbol{X}$ fuera gaussiano, conocer estos *parámetros* bastaría para caracterizar la distribución conjunta. Pero aún así...

- necesitaríamos para cada X_t una muestra suficiente para estimar los parámetros
  - pero en una serie temporal tenemos una sola realización de cada X_t.
- Además... para cada variable aleatoria X_t hay infinitos parámetros.

Casos que simplifican el escenario

- Si el proceso es débilmente estacionario se reduce drásticamente el número de parámetros:

$$E(X_t) = \mu \tag{1}$$
$$Cov(X_t, X_{t-k}) = \gamma_k \tag{2}$$

- Si además pudiéramos asumir que el proceso es i.i.d. podríamos interpretar la serie temporal como una realización de un muestreo aleatorio simple (lo que habilita la inferencia estadística).

El desafío para el analista es (y nótese el abuso de lenguaje)

primero transformar los datos para lograr que sean "***estacionarios***".

- (Lo vimos en la lección 1))

después transformar los datos estacionarios en una secuencia de "**datos *i.i.d***"

- (Aún no visto)

Todo este proceso constituye la especificación y ajuste de un modelo a la serie temporal.

La especificación del modelo se escoge según las características de los datos.

£Es la serie

- "***estacionaria en media***"
  - (y si lo es, £cuál es su media?)
  - (y si no lo es, £cómo cambia o evoluciona su media?)
- "***estacionaria en varianza***"(*homocedástica*)
  - (y si lo es, £cuál es su varianza?)
  - (y si es *heterocedástica*, £cómo cambia o evoluciona su varianza?)?

£Están sus valores correlados con su historia pasada (autocorrelados)?
£Están correlados con los valores presentes o pasados de otras series?

Veamos algunas herramientas estadísticas para poder desvelar estas características.

2. Herramientas para desvelar propiedades de una serie temporal

2.1. Análisis gráfico

2.1.1. Gráfico de la serie temporal

Representa sus valores en el eje vertical (y) frente a una escala temporal en el horizontal (x). Es útil para detectar visualmente:

- tendencias y/o estacionalidad
- cambios de variabilidad

- valores atípicos (*outliers*)
  - el 95 % aprox. de una muestra de valores generados por una distribución normal debería estar comprendido entre $\mu \pm 2\sigma$
  - la probabilidad de que una variable normal genere un valor fuera de las bandas de $\mu \pm 3\sigma$ es 0,0023
    (véase gráfico IBEX 35)

En es importante escalar y rotular adecuadamente los ejes y asegurar la comparabilidad entre series y gráficos distintos (si los hubiere).

2.1.2. Gráfico rango-media

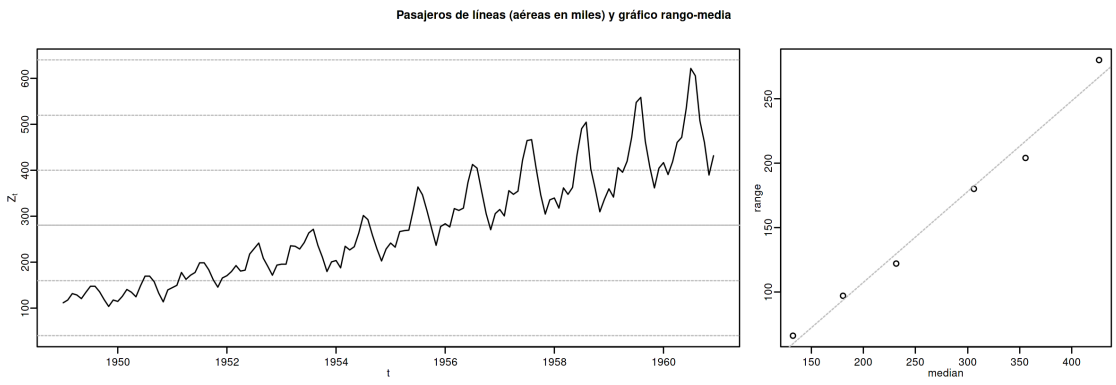
Cambios de variabilidad de una serie pueden evidenciarse en su gráfico temporal.
Pero también suelen verse bien en un gráfico rango-media, donde se representa:

en el eje x un indicador del nivel de la serie calculado para distintas submuestras no solapadas (normalmente la media).

en el eje y un indicador de la dispersión de la serie calculado para las mismas submuestras (normalmente el rango.)

Veamos el gráfico de la serie de pasajeros de líneas aéreas junto a su gráfico de rango media:

```
Z <- AirPassengers
ide(Z, graphs = c("plot", "rm"), main="Pasajeros de líneas (aéreas en miles) y gráfico rango-media")
```



El gráfico de rango media a veces se acompaña de una regresión de la dispersión sobre los niveles para medir la relación nivel-dispersión.

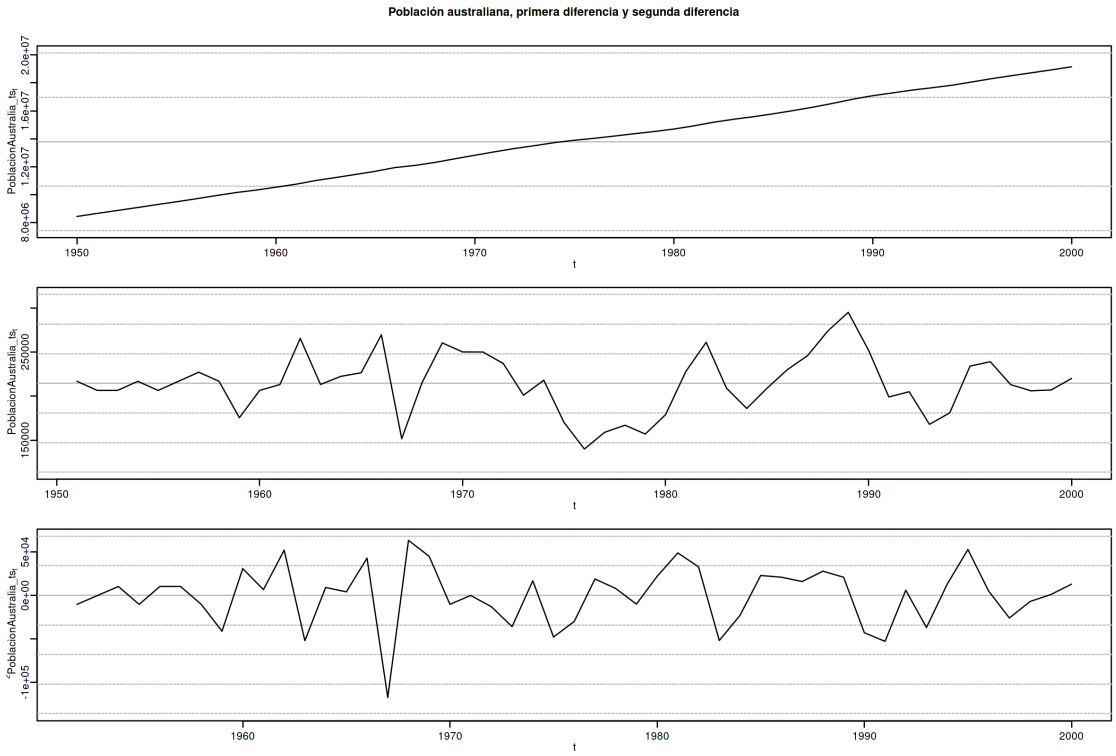
2.2. Determinación del orden de integración

Decidir adecuadamente el orden de integración es crucial en el análisis de series temporales.
Las herramientas utilizadas para tomar la decisión son

- el análisis gráfico
- los contrastes formales

2.2.1. Análisis gráfico

```
options(repr.plot.width = 12, repr.plot.height = 8, repr.plot.res = 200)
ide(PoblacionAustralia_ts,
    graphs = c("plot"),
    transf = list(list(bc = F), list(bc = F, d = 1), list(bc = F, d = 2)),
    main="Población australiana, primera diferencia y segunda diferencia" )
```



La serie de población $\boldsymbol{y}$ tiene una clara tendencia creciente (primer gráfico), que desaparece al tomar una diferencia ordinaria,

$$\nabla \boldsymbol{y} = (1 - \mathsf{B}) * \boldsymbol{y}$$

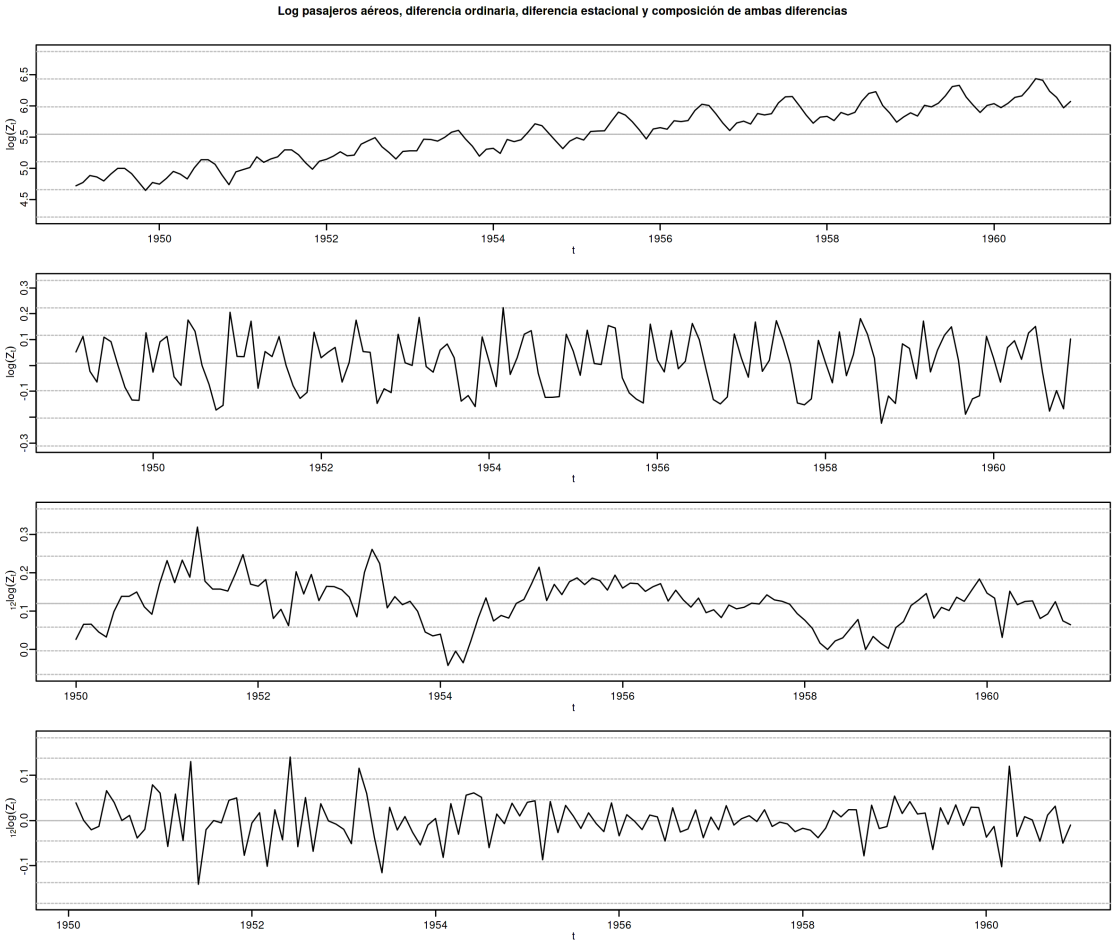
(segundo gráfico). Bastar con tomar una primera diferencia de la serie de población para obtener una nueva serie que se asemeja a la realización de un proceso estacionario.

No obstante, £qué pasa si tomamos una segunda diferencia ordinaria?

$$\nabla\nabla \boldsymbol{y} = \nabla^2 \boldsymbol{y} = (1 - \mathsf{B})^2 * \boldsymbol{y}$$

(segundo gráfico). Pues que la serie obtenida también es estacionaria, pero ojo, es un grave error tomar más diferencias de las necesarias al modelizar los datos. Se debe tomar el mínimo número de transformaciones que arrojen una serie "estacionaria" (recuerde que decir que una serie temporal es *estacionaria* es un abuso del lenguaje).

```
options(repr.plot.width = 12, repr.plot.height =  10, repr.plot.res = 200)
ide(Z,
    graphs = c("plot"),
    transf = list(list(bc=T), list(bc=T, d=1), list(bc=T, D=1), list(bc=T, D=1, d=1)),
    main = "Log pasajeros aéreos, diferencia ordinaria, diferencia estacional y composición de ambas diferencias" )
```



Como ya vimos, la serie pasajeros en logaritmos tiene tendencia y estacionalidad muy evidentes. No basta con tomar solo una diferencia ordinaria

$$\nabla \boldsymbol{y} = (1 - \mathsf{B}) * \boldsymbol{y};$$

pues el resultado muestra una pauta estacional. Ni tampoco basta con tomar solo una diferencia estacional

$$\nabla_{12}(\boldsymbol{y}) = (1 - \mathsf{B}^{12}) * \boldsymbol{y};$$

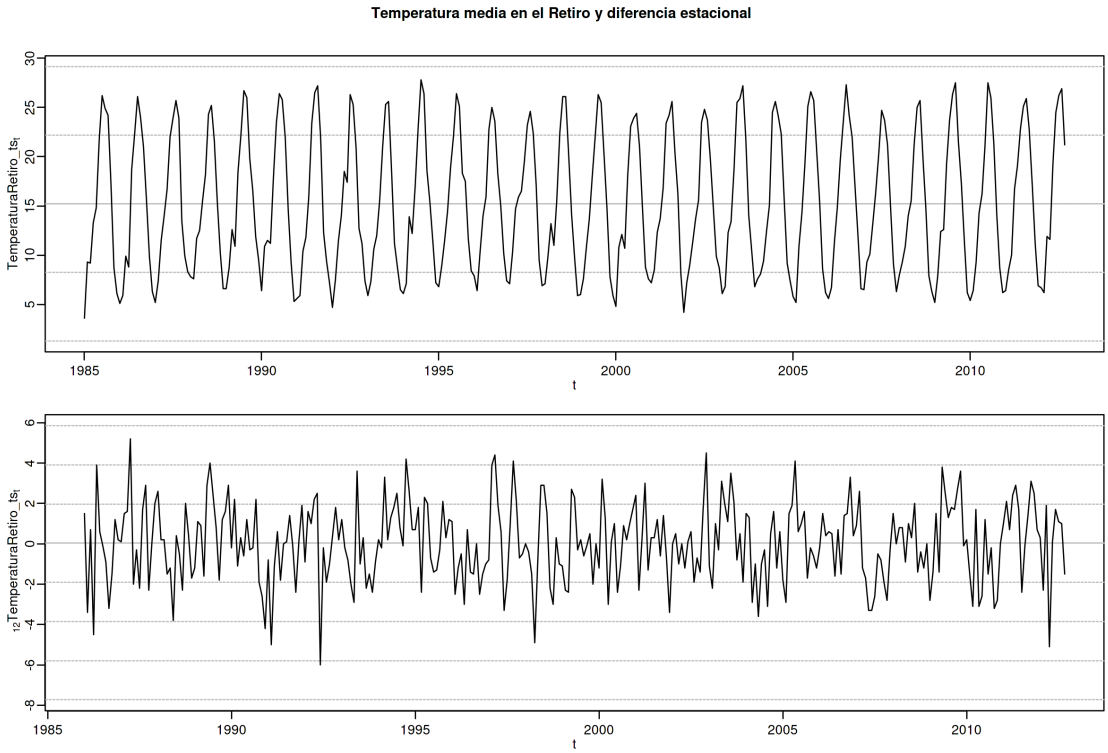
pues resulta una serie que "deambula", i.e., que no es *"estacionaria"* en media.

Tomar una diferencia ordinaria y otra estacional

$$\nabla\nabla_{12}(\boldsymbol{y}) = (1 - \mathsf{B}) * (1 - \mathsf{B}^{12}) * \boldsymbol{y}$$

arroja una serie que sí parece ser *"estacionaria"*.

```
options(repr.plot.width = 12, repr.plot.height =  8, repr.plot.res = 200)
TemperaturaRetiro_ts=ts(read.csv("datos/Retiro.txt"),start=c(1985, 1), end=c(2015,9), frequency=12)
ide(TemperaturaRetiro_ts,
    graphs = c("plot"),
    transf = list(list(), list(D = 1)),
    main="Temperatura media en el Retiro y diferencia estacional" )
```



En el caso de la serie de temperaturas en el Parque del Retiro, parece que es suficiente con tomar solo una diferencia estacional.

2.2.2. Contrastes formales sobre el orden de integración

1. Test de Dickey-Fuller (DF)

   H_0 la serie es $I(1)$

   H_1 la serie es $I(0)$.

   Consideremos el modelo

   $$y_t = \rho y_{t-1} + u_t,$$

   donde y_t es la variable de interés, ρ es un coeficiente, y u_t es un proceso de ruido blanco. Una raíz unitaria estará presente si $\rho = 1$. En tal caso el modelo será no-estacionario.

   El modelo de regresión se puede escribir como

   $$\nabla y_t = (\rho - 1) y_{t-1} + u_t = \delta y_{t-1} + u_t$$

2. Test de Dickey-Fuller aumentado (ADF)

3. Función de autocovarianzas y función de autocorrelación

- La secuencia (γ_k) con $k \in \mathbb{Z}$ se denomina *función de autocovarianzas*

- La secuencia $\{\rho_k\}$ con $k \in \mathbb{Z}$, donde

$$\rho_k = \frac{Cov(X_t, X_{t-k})}{\sqrt{Var(X_t)Var(X_{t-k})}} = \frac{\gamma_k}{\gamma_0}$$

  se denomina *función de autocorrelación* (ACF).

Debido a la estacionariedad, la correlación entre X_t y X_{t+k} no depende de t; tan solo depende de la distancia temporal k entre ambas variables.

$$\phi(\mathsf{B}) : \phi$$

$\phi(\mathsf{B})$

4. Otras herramientas estadísticas CómoConR

4.1. Estadísticos descriptivos

```
library(pastecs)      # resumen estadísticos descriptivos
# https://cran.r-project.org/web/packages/pastecs/index.html (stat.desc)
library(knitr)        # presentación de tabla resumen
# https://cran.r-project.org/web/packages/knitr/index.html (kable)
# https://bookdown.org/yihui/rmarkdown-cookbook/kable.html

# estadísticos principales y test de normalidad
kable(stat.desc(Z, basic=FALSE, norm=TRUE), 'rst')
```

```
============  =============
\                          x
============  =============
median            265.5000000
mean              280.2986111
SE.mean             9.9971931
CI.mean.0.95       19.7613736
var             14391.9172009
std.dev           119.9663169
coef.var            0.4279947
skewness            0.5710676
skew.2SE            1.4132515
kurtosis           -0.4298441
kurt.2SE           -0.5353818
normtest.W          0.9519577
normtest.p          0.0000683
============  =============
```

4.2. Test de normalidad Jarque-Vera

Jarque-Vera test (Wikipedia)
Podemos calcularlo con la librería momments:

```r
#install.packages("moments")
library(moments)
# Perform the Jarque-Bera test
jb_test <- jarque.test(as.numeric(Z))
# Print the test result
print(jb_test)
```

```
Jarque-Bera Normality Test

data:  as.numeric(Z)
JB = 8.9225, p-value = 0.01155
alternative hypothesis: greater
```

Otra librería alternativa para calcularlo: tseries

```r
library(tseries)
# Perform the Jarque-Bera test
jb_test <- jarque.bera.test(Z)
# Print the test result
print(jb_test)
```

```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo

Jarque Bera Test

data:  Z
X-squared = 8.9225, df = 2, p-value = 0.01155
```