Análisis de Series de Tiempo

Carrera de Especialización en Inteligencia Artificial

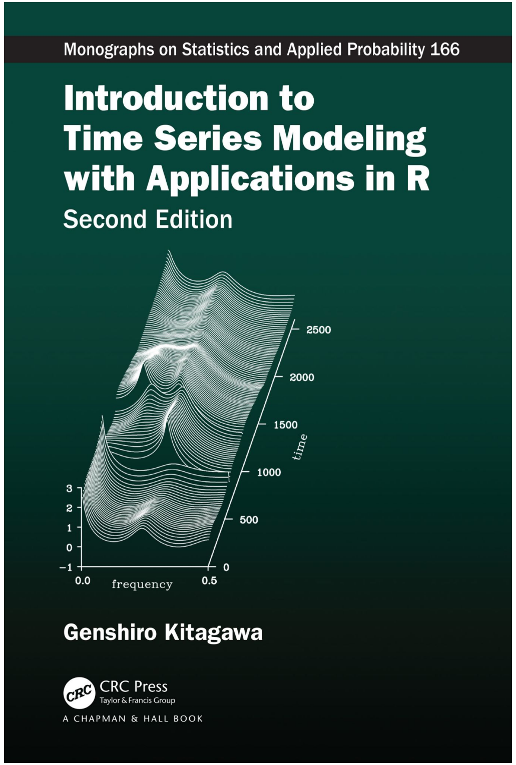
Clase 1

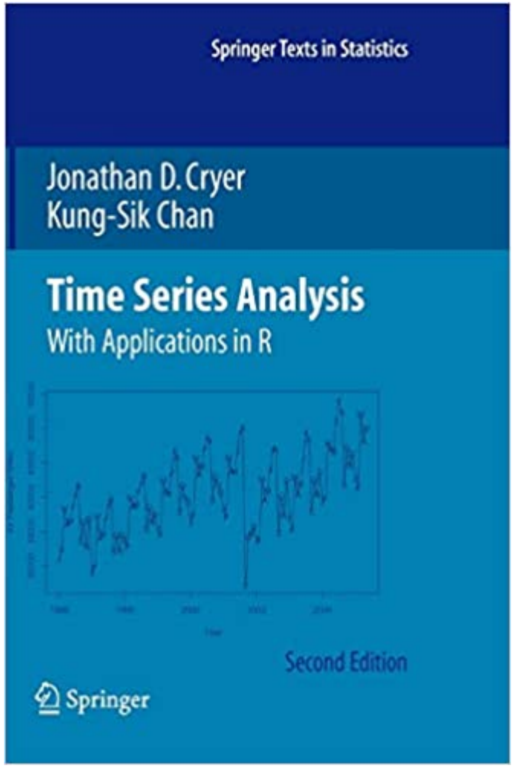
Ing. Magdalena Bouza, Ing. Carlos German Carreño Romano

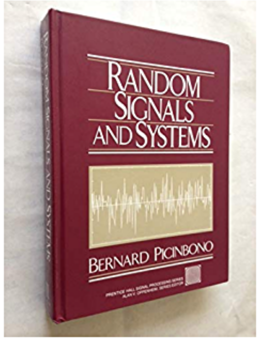
Cronograma

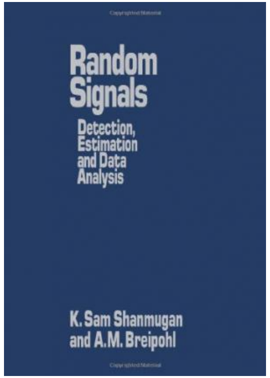
Clase 1	Introducción, nociones básicas de modelado, algunos ejemplos
Clase 2	Tendencia determinística, modelos estacionarios AR, MA, ARMA
Clase 3	Identificación y estimación de los modelos, presentación de TP, aplicaciones LSTM
Clase 4	Caso de estudio, tendencia estocástica, estacionalidad, modelos SARIMA
Clase 5	Predicciones, tratamiento de intervenciones y outliers
Clase 6	Aplicaciones. Análisis espectral. Markov*
Clase 7	Heteroscedasticidad. Frameworks.
Clase 8	Presentaciones trabajo final

Bibliografía









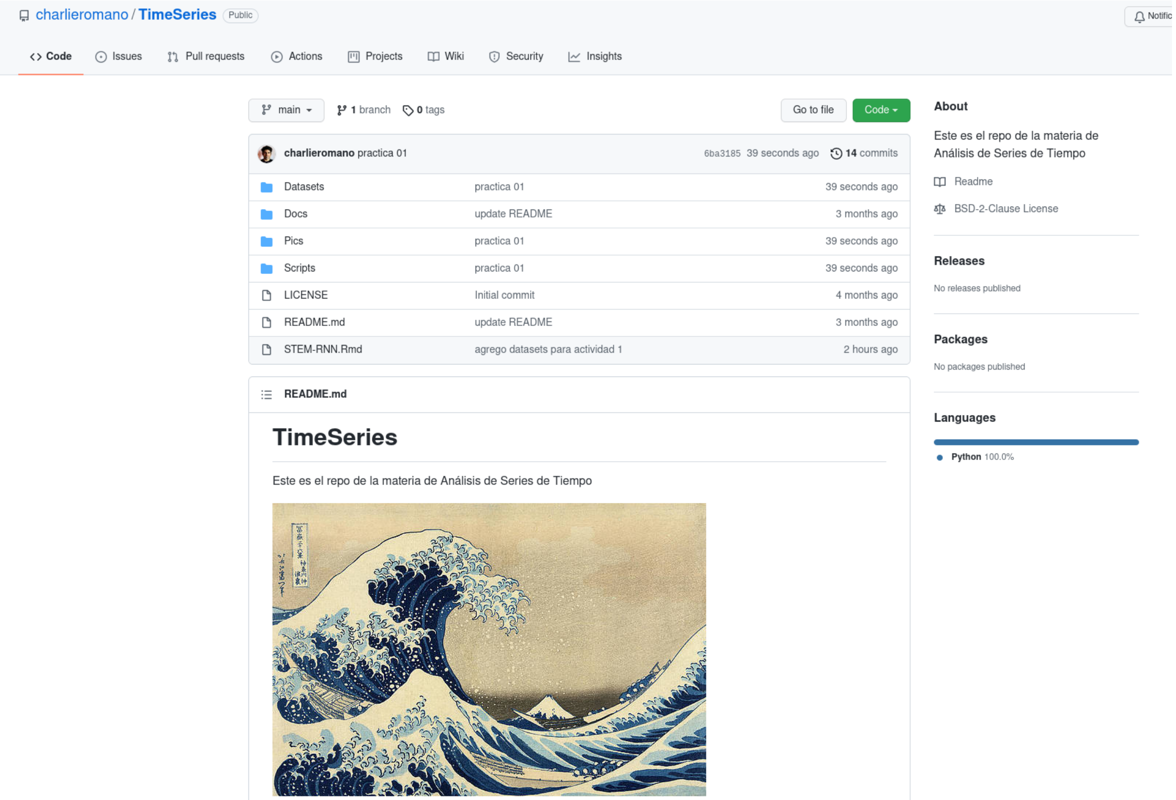
Acerca del curso

- Modelos clásicos
- Práctica matemática
- Análisis usando Python
- Aplicaciones con redes neuronales
- Aplicaciones en temas de interés de los alumnos
- Repo: www.github.com/charlieromano/timeseries

Repo

https://github.com/charlieromano/TimeSeries

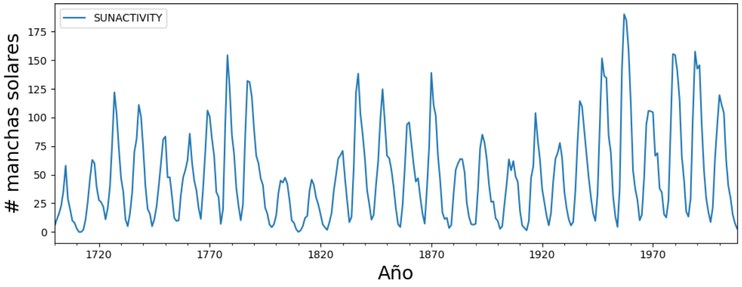
- Datasets
- Docs
- Pics
- Scripts
- README

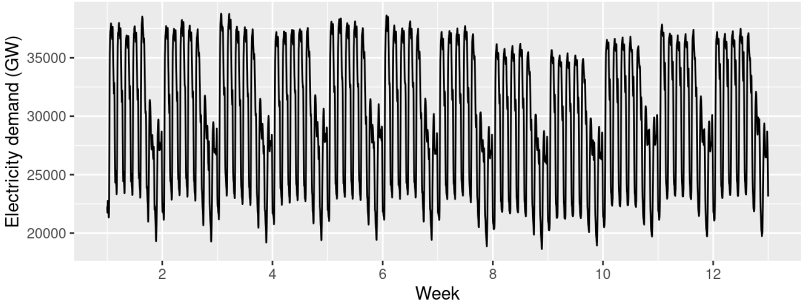


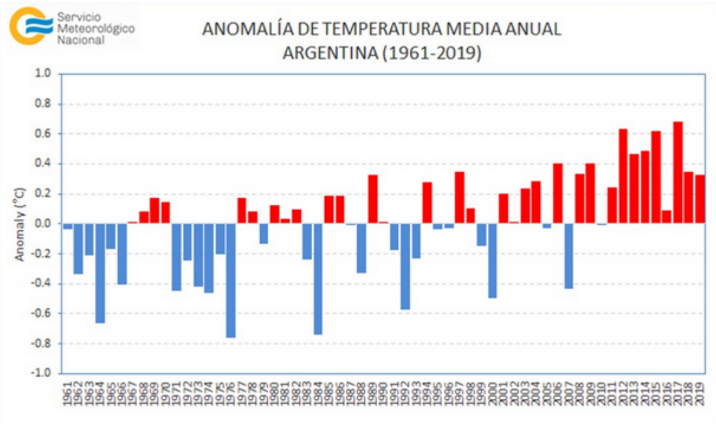
Introducción

¿Qué es una serie de tiempo?

Un registro de un fenómeno que varía en el tiempo de forma irregular es una serie de tiempo.







¿Para qué estudiar Series de Tiempo?

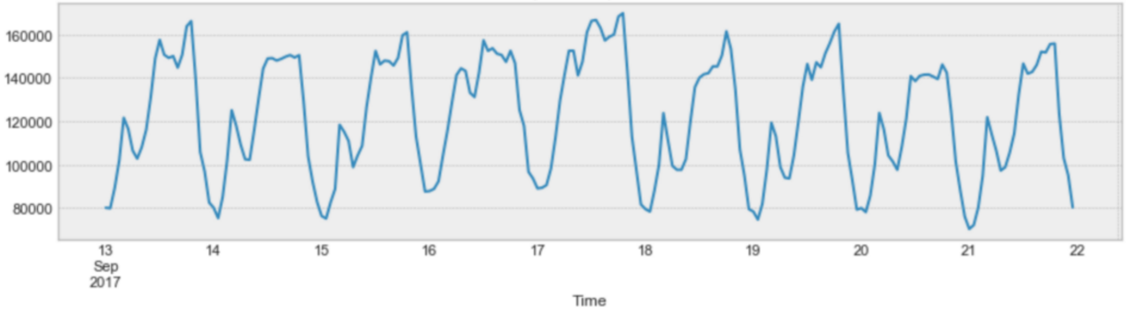
En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA. Para hacer esto solemos seguir un pipeline:

1. **Descripción**
2. **Modelado**
3. **Predicción**
4. **Extracción de señales**

¿Para qué estudiar Series de Tiempo?

En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA.

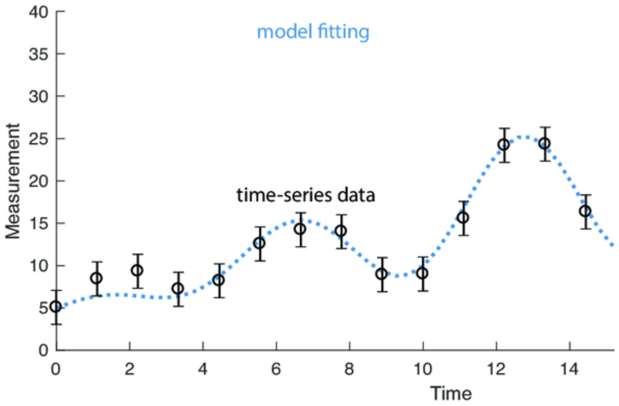
- **1: Descripción**: se desea resumir las características de una serie de tiempo.



¿Para qué estudiar Series de Tiempo?

En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA.

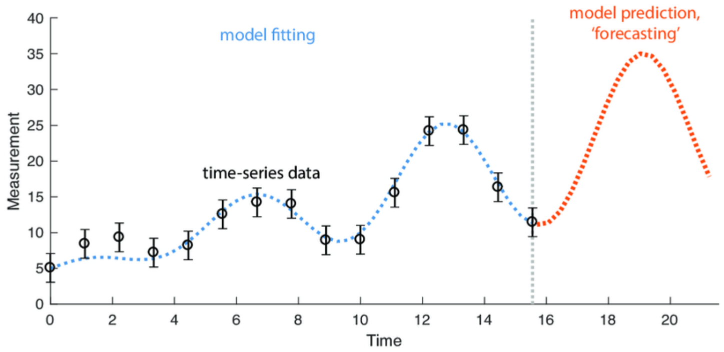
- **2: Modelado**: se busca identificar un modelo aproximado del problema. Es necesario identificar un modelo correcto, así como los parámetros asociados al mismo.



¿Para qué estudiar Series de Tiempo?

En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA.

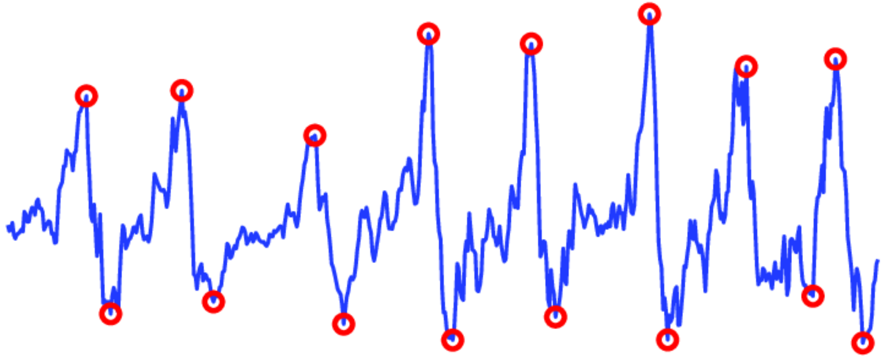
- **3: Predicción**: el objetivo es conocer, o predecir, el comportamiento futuro del sistema.



¿Para qué estudiar Series de Tiempo?

En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA.

- **4: Extracción de señales**: de todo el modelo, se busca extraer ciertas señales que resultan de interés para el problema en cuestión

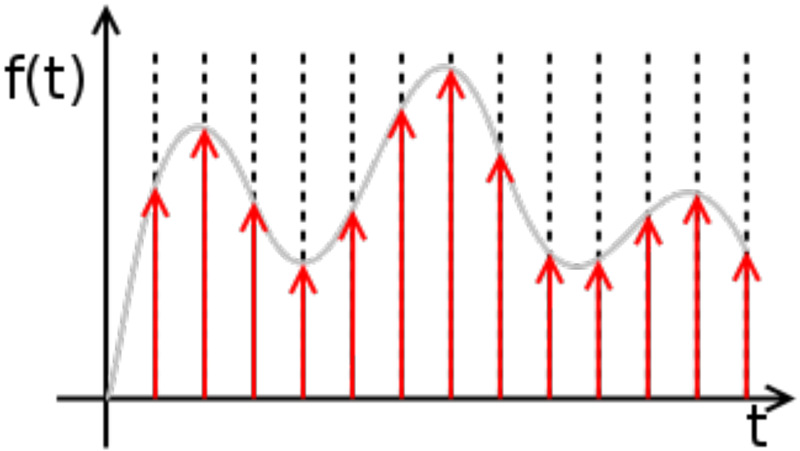


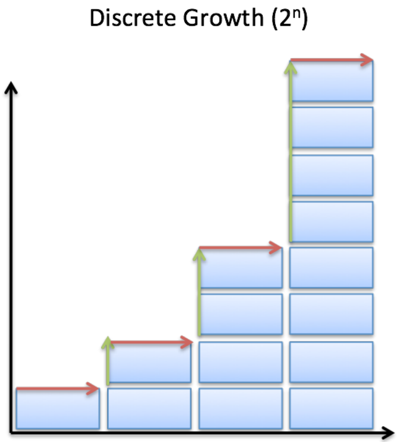
Cómo clasificar Series de Tiempo

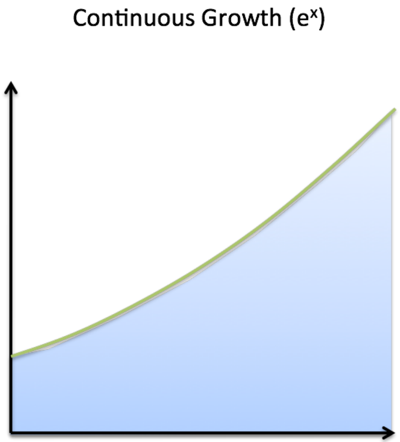
- Continuas o discretas
- Univariadas o multivariadas
- Estacionarias o no estacionarias
- Estacionales o no estacionales
- Gaussianas o no gaussianas
- Observaciones faltantes y outliers

Cómo clasificar Series de Tiempo

- **Continuas o discretas**: En general se consideran series de tiempo discretas, ya que los datos suelen medirse en intervalos de tiempo

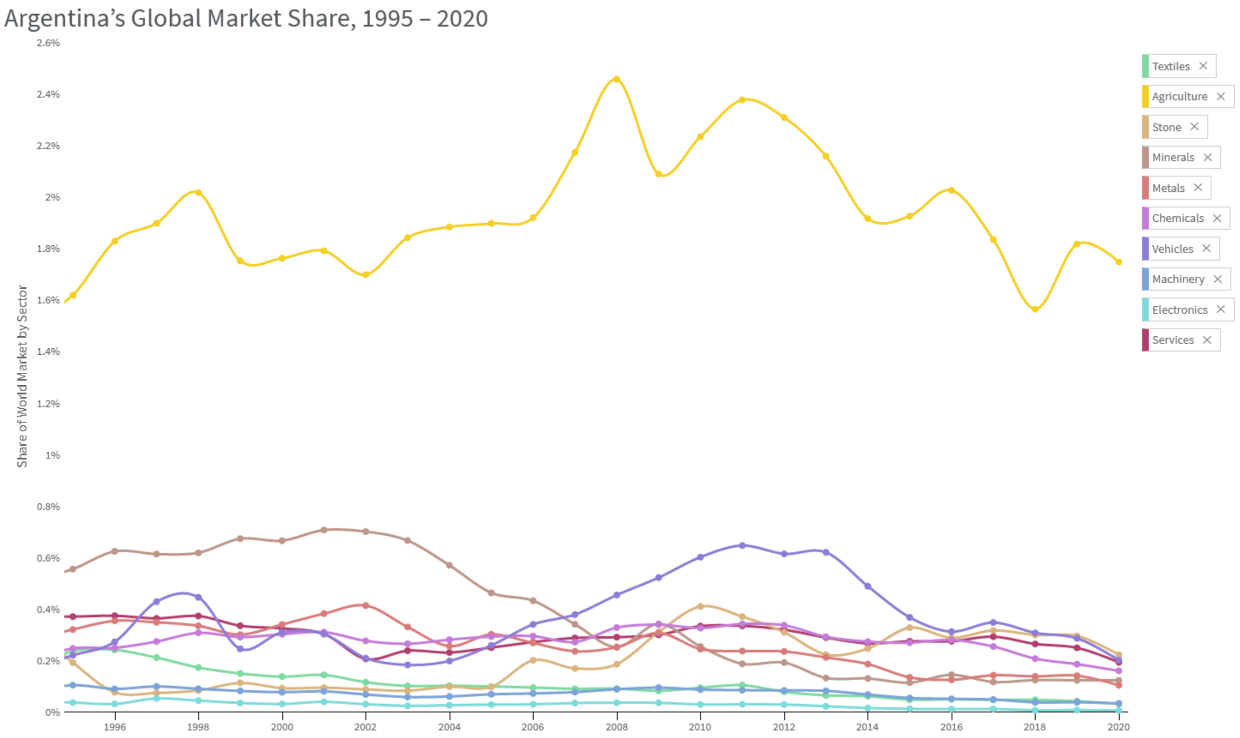






Cómo clasificar Series de Tiempo

- Univariadas vs. **multivariadas**

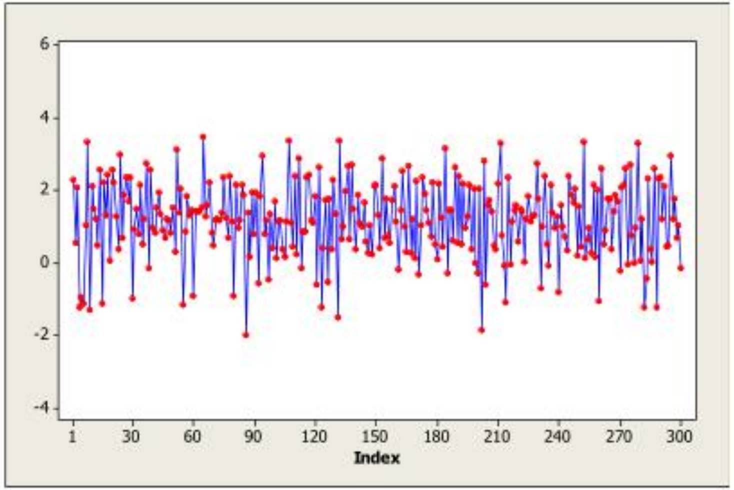


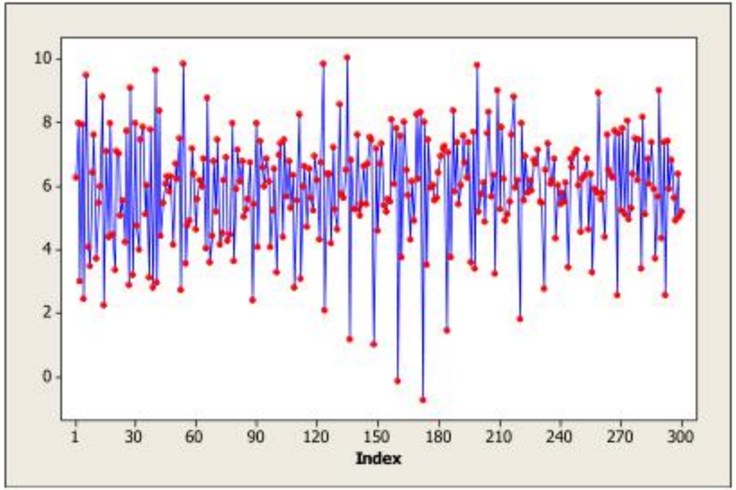
Ref: https://atlas.cid.harvard.edu/explore
Dashboard: https://atlas.cid.harvard.edu/explore/market?country=8&product=undefined&year=2020&queryLevel=location&productClass=HS&target=Product&partner=undefined&startYear=1995

Cómo clasificar Series de Tiempo

- **Estacionarias (no) estacionarias**

La serie de tiempo estudiada proviene de realizaciones de un proceso estocástico con una estructura invariante (media y desvio) en el tiempo se lo llama estacionario.

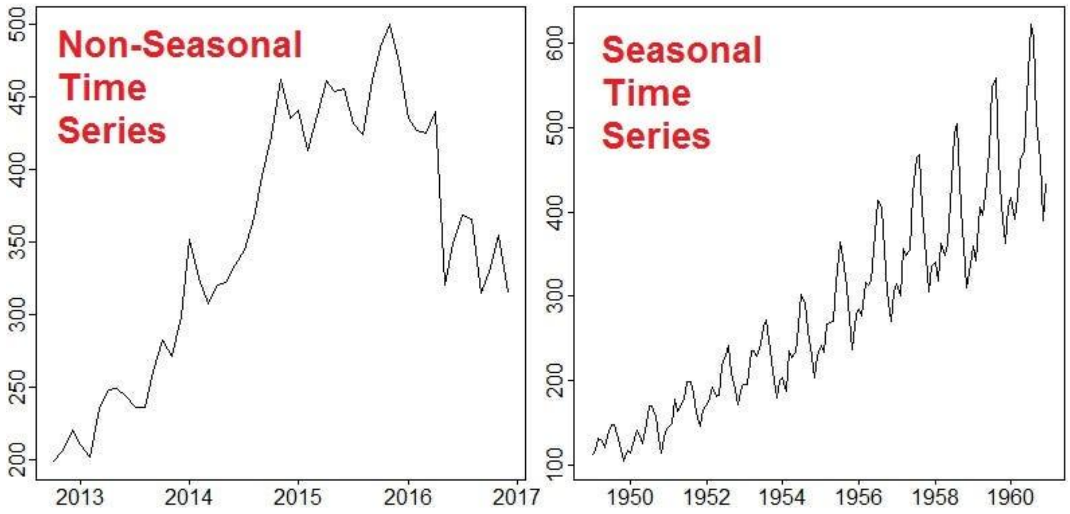




Cómo clasificar Series de Tiempo

- **Estacionales vs no estacionales**

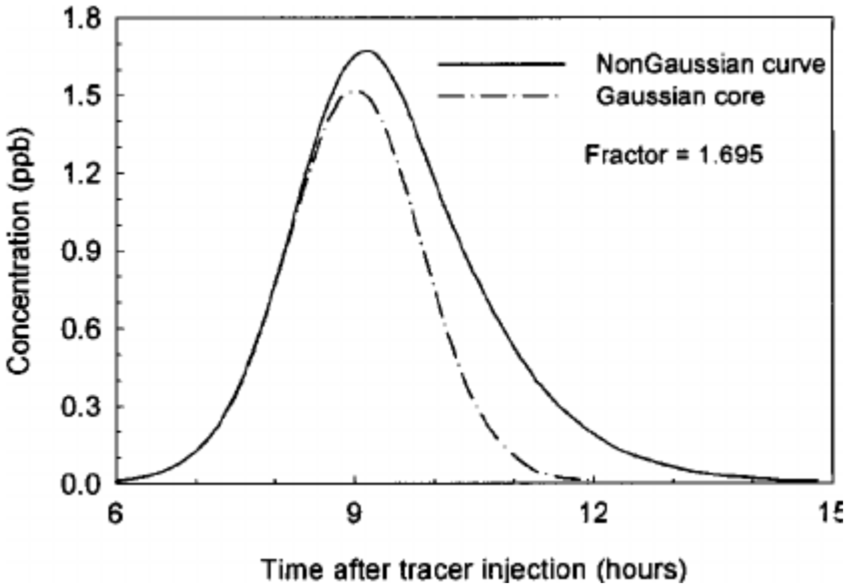
La serie presenta ciclos que tienen una frecuencia determinada, por ejemplo años o estaciones anuales (invierno, primavera, verano, otoño), mientras que la ausencia de ciclos a lo largo de toda la serie puede ser indicio de no estacionalidad.



Cómo clasificar Series de Tiempo

- **Gaussianas** o no gaussianas

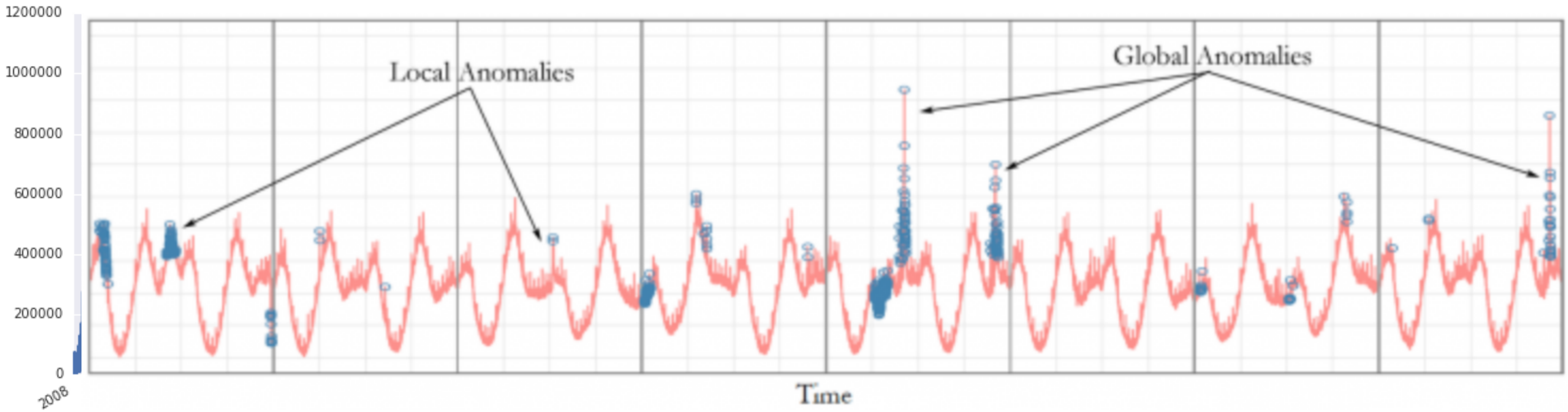
Si la serie de tiempo sigue una distribución Gaussiana



Cómo clasificar Series de Tiempo

- **Observaciones faltantes y outliers**

Es muy común en series de tiempo la aparición de valores fuera del rango esperado, o de anomalías locales o globales. Una aplicación general con técnicas de análisis de datos es la detección de anomalías.



Cómo clasificar Series de Tiempo

Hasta acá hemos visto cómo estudiar y clasificar series de tiempo. La clasificación es parte del primer paso del pipeline de análisis y modelado que corresponde a la descripción.

Describir una serie de tiempo tiene un grado muy fuerte de dependencia con la naturaleza de la serie. Siempre que sea posible debemos buscar y comprender las causas naturales de los movimientos de las series de tiempo para poder describirlas, y no quedarnos simplemente en una descripción técnica. En algunas disciplinas esto se diferencia entre análisis técnico y análisis fundamental.

Series de tiempo y
Procesos estocásticos

Procesos estocásticos

Las series de tiempo forman parte de lo que se conoce como procesos estocásticos.

Así como las variables aleatorias mapean los posibles resultados de un experimento aleatorio a un número (real), los procesos estocásticos mapean los resultados de un experimento aleatorio a un conjunto de funciones en el tiempo.

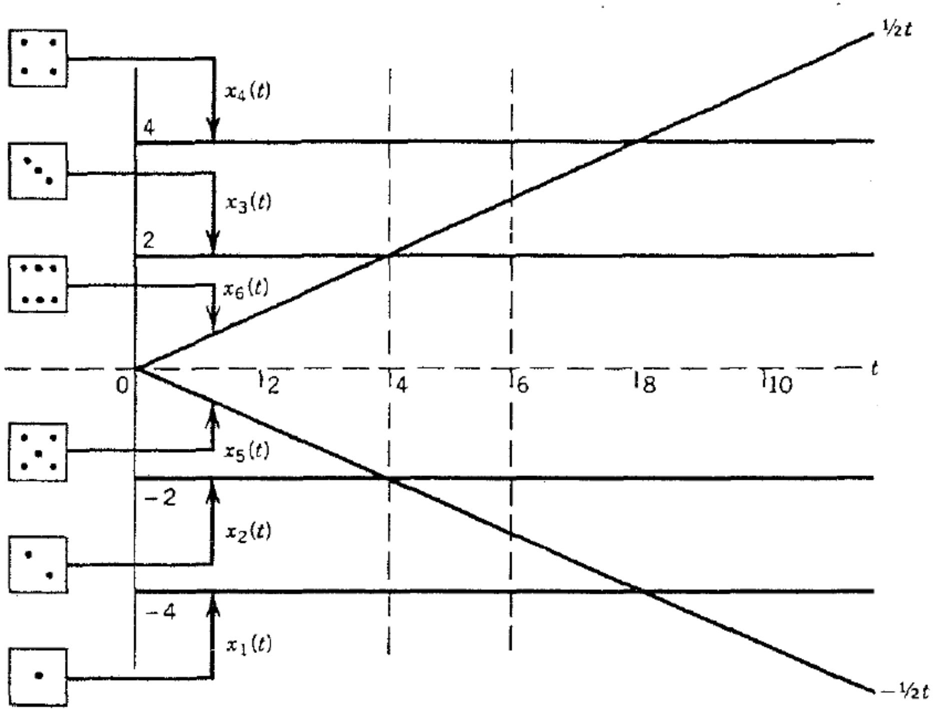


Figure 3.2 Example of a random process.

Ref.: Random Signals: Detection, Estimation and Data Analysis, *Shanmugan, Breipohl*

Procesos estocásticos

- Ensamble de funciones de tiempo

$$X(t,\Omega) = \{X(t,\omega_i) | \omega_i \in \Omega\} = \{x_1(t), x_2(t), \ldots\}$$

- Función de tiempo específica

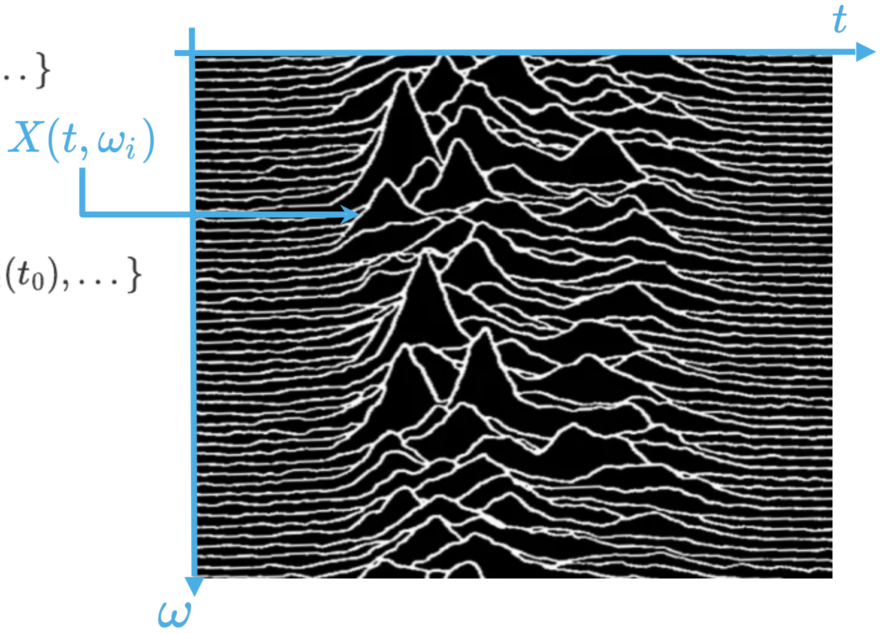
$$X(t,\omega_i) = x_i(t)$$

- Variable aleatoria

$$X(t_0,\Omega) = \{X(t,\omega_i) | \omega_i \in \Omega\} = \{x_1(t_0), x_2(t_0), \ldots\}$$

- Valor numérico de una función en un tiempo dado

$$X(t_0,\omega_i) = x_i(t_0)$$



Una serie de tiempo es una sucesión de variables aleatorias correlacionadas entre sí a lo largo del tiempo.

Estrategia de modelado

Hallar modelos para las series de tiempo no es algo trivial. En general el proceso consiste de 3 pasos:

- **Especificación** (o identificación) del modelo: se seleccionan algunos modelos que podrían ser apropiados para modelar la serie que estamos estudiando
- **Ajuste del modelo**: dado un modelo propuesto, se busca estimar los mejores parámetros de ese modelo
- **Diagnóstico** del modelo: analizar mediante diversos tests la bondad o calidad del modelo

En este curso nos vamos a encargar de presentar distintos modelos y estudiar cómo ejecutar cada uno de estos pasos.

Análisis de Series de Tiempo

Box and Jenkins plantean tres etapas de **modelado:**

1. **Especificaciones**
   a. Gráficos
   b. Estadísticas
   c. Contexto
2. **Ajuste**
   a. Parámetros
   b. Valores
3. **Diagnóstico**
   a. Testing

“everything should be made as simple as possible but no simpler”

Algunos modelos

Algunos Modelos

1. Ruido blanco (white noise)
2. Caminante aleatorio (Random walk)
3. ~~Coseno aleatorio (Random Cosine wave)~~

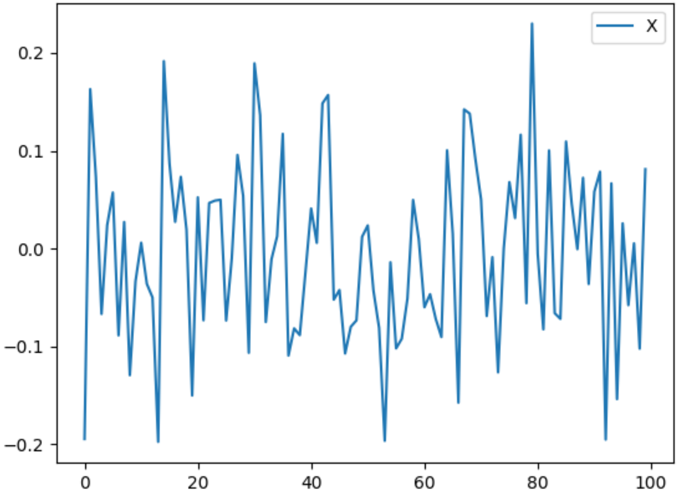
Ruido blanco

El proceso de **ruido blanco** se define como una secuencia de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) $\{e_t\}$

```
# random values data series
X = np.random.normal(mu, sigma, N)
```

$$X \sim \mathcal{N}(\mu, \sigma) \rightarrow f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

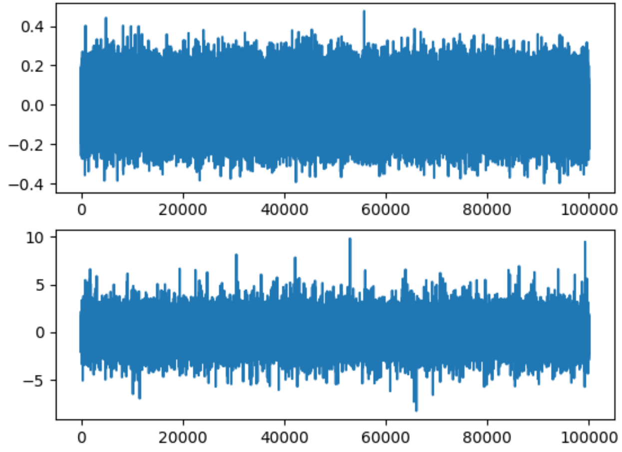
```
## stats
df.X.describe
df.Y.describe()
dt = pd.DataFrame(df.X.describe())
dt=pd.concat([df.X.describe(), df.Y.describe()],axis=1)
```



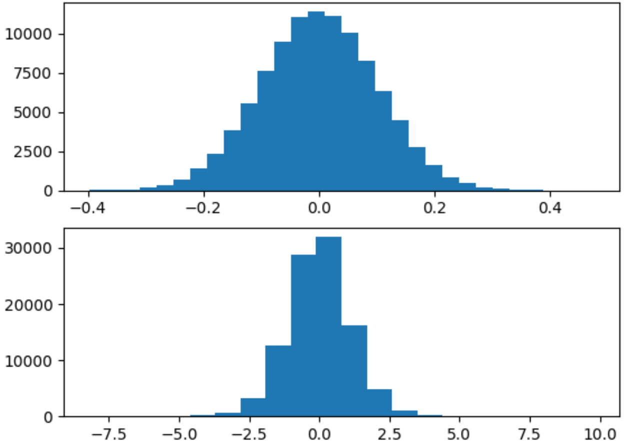
Stat	X	Y
count	100000	100000
mean	6.733E-05	6.608E-03
std	1.001E-01	1.129E+00
min	-3.980E-01	-8.216E+00
25%	-6.742E-02	-6.949E-01
50%	-3.210E-05	4.958E-03
75%	6.800E-02	7.056E-01
max	4.761E-01	9.797E+00

Gráficos

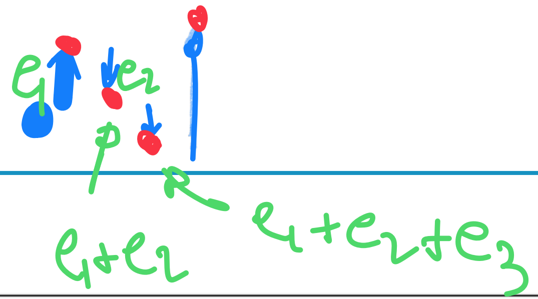
```
## data series
fig, (ax1, ax2) = plt.subplots(2)
fig.subtitle('X vs. Y timeseries')
ax1.plot(X)
ax2.plot(Y)
plt.show()
```




```
## histograms
fig, (ax1, ax2) = plt.subplots(2)
fig.subtitle('X vs. Y histograms')
ax1.hist(X,bins=30)
ax2.hist(Y,bins=20)
plt.show()
```



Caminata aleatoria

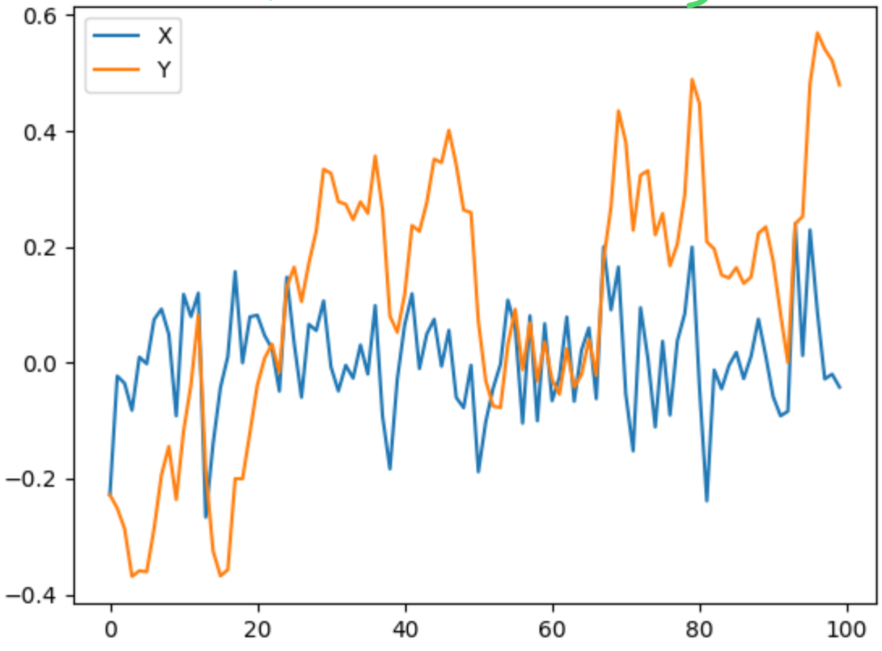


Sean $e_1, e_2, \ldots$ una secuencia de v.a. i.i.d., de media nula y varianza σ_e^2. Se construye la serie de tiempo como:

$$
\begin{aligned}
Y_1 &= e_1 \\
Y_2 &= e_1 + e_2 \\
&\vdots \\
Y_t &= e_1 + e_2 + \ldots + e_t
\end{aligned}
\qquad \Rightarrow Y_t = Y_{t-1} + e_t
$$

```python
# random values data series
X = np.random.normal(mu, sigma, N)

# Random walk
Y = np.cumsum(X)
```



https://github.com/charlieromano/TimeSeries/blob/main/Scripts/randomWalk.py

¿preguntas?

Momentos

Algunos momentos

Una serie de tiempo $Y(t)$, puede entonces modelarse a través de la secuencia de variables $\{Y(t), t = t_k, t_{k+1}, \ldots, t_{k+n}\}$ que definen un proceso estocástico. Gran parte de la información de la distribución de estos procesos está contenida en los momentos de dicha distribución: medias, varianzas y covarianzas. A diferencia de las variables aleatorias, estos momentos deberán definirse como función del tiempo.

- **Media** (o esperanza): $\mu_t = \mathbb{E}[Y_t]$
- **Función de autocovarianza**: $C_{t,s} = Cov(Y_t, Y_s) = \mathbb{E}\left[(Y_t - \mu_t)(Y_s - \mu_s)\right]$
- **Función de autocorrelación**: $R_{t,s} = Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{var(Y_t)var(Y_s)}}$

Algunos momentos (caso multivariado)

Análogamente, si se tiene una serie de tiempo multivariada, donde $Y_t = [Y_t^{(1)}, \ldots, Y_t^{(l)}]$, definimos

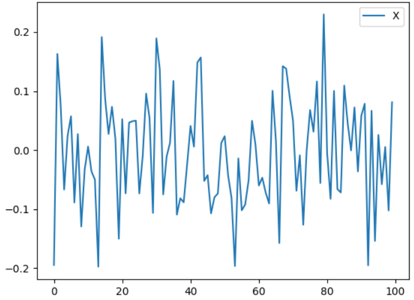
- **Media** (o esperanza): $\boldsymbol{\mu}_t = [\mu_t^{(1)}, \ldots, \mu_t^{(l)}] = [\mathbb{E}[Y_t^{(1)}], \ldots, \mathbb{E}[Y_t^{(l)}]]$
- **Matriz de cross-covarianza**: $C_{t,s} = \begin{bmatrix} C_{t,s}^{(1,1)} & \ldots & C_{t,s}^{(1,l)} \\ \vdots & \ddots & \vdots \\ C_{t,s}^{(l,1)} & \ldots & C_{t,s}^{(l,l)} \end{bmatrix}$, donde
$$C_{t,s}^{(i,j)} = Cov(Y_t^{(i)}, Y_s^{(j)}) = \mathbb{E}[(Y_t^{(i)} - \mu_t^{(i)})(Y_s^{(j)} - \mu_s^{(j)})]$$
- **Matriz de cross-correlación**: $R_{t,s} = \begin{bmatrix} R_{t,s}^{(1,1)} & \ldots & R_{t,s}^{(1,l)} \\ \vdots & \ddots & \vdots \\ R_{t,s}^{(l,1)} & \ldots & R_{t,s}^{(l,l)} \end{bmatrix}$

Ruido blanco

El proceso de **ruido blanco** se define como una secuencia de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) $\{e_t\}$.

Algunas propiedades:

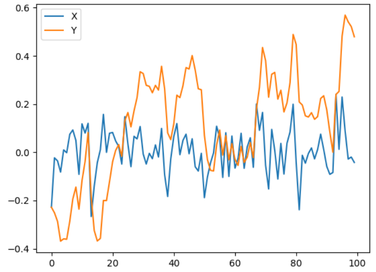
- $\mathbb{P}(e_{t_1} \leq x_1, \ldots, e_{t_n} \leq x_n) = \prod_{i=1}^{n} \mathbb{P}(t_{t_i} \leq x_i) = \mathbb{P}(e_{t_1 - k} \leq x_1, \ldots, e_{t_n - k} \leq x_n)$
- $\mu_t = \mu \quad \text{y} \quad var(e_t) = \sigma_e^2 \quad \forall t$
- $C_{t,s} = \sigma_e^2 \mathbf{1}\{t = s\}$



Caminante aleatorio (Random Walk)

Sean $e_1, e_2, \ldots$ una secuencia de v.a. i.i.d., de media nula y varianza σ_e^2. Se construye la serie de tiempo como:

$$\begin{aligned} Y_1 &= e_1 \\ Y_2 &= e_1 + e_2 \\ &\vdots \\ Y_t &= e_1 + e_2 + \ldots + e_t \end{aligned} \qquad \Rightarrow Y_t = Y_{t-1} + e_t$$



Podemos calcular distintos momentos para Yt:

$$\mu_t = \mathbb{E}[Y_t] = \mathbb{E}[e_1 + \ldots + e_t] = 0 \quad \forall t$$

$$\sigma_t^2 = var(Y_t) = var(e_1 + \ldots + e_t) = t\sigma_e^2 \quad \forall t$$

$$C_{t,s} = Cov(Y_t, Y_s) = t\sigma_e^2 \quad (1 \leq t \leq s)$$

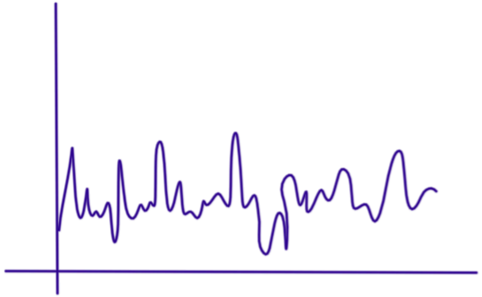
$$R_{t,s} = \sqrt{\frac{t}{s}}$$

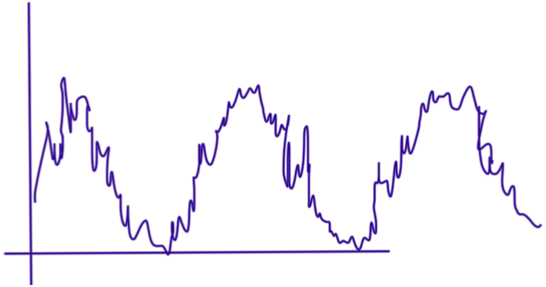
¿Qué podemos concluir?

Estacionariedad

Estacionariedad vs. Estacionalidad

- **Estacionariedad** es una propiedad de algunas series de tiempo que tiene que ver con **invarianza** a lo largo del tiempo.
- **Estacionalidad** se refiere a ciclos o períodos que se pueden observar en una serie de tiempo, por ejemplo las **estaciones** del año





Estacionariedad

Una suposición que suele hacerse a la hora de modelar series de tiempo es la estacionariedad.

Se dice que un proceso es **estacionario** (*stationary*) si la distribución del proceso no cambia a lo largo del tiempo. Matemáticamente, quiere decir que

$$f_{Y_{t_1},...,Y_{t_n}} = f_{Y_{t_{1+\delta}},...,Y_{t_{n+\delta}}}, \quad \forall t_1, \ldots t_n, \forall \delta$$

Una conclusión que se desprende de esta definición es que su los procesos son estacionarios, su media es constante en el tiempo y las funciones de autocorrelación y autocovarianza dependen sólo de la diferencia de tiempos:

$$\mu_t = \mu_s = \mu \quad \forall t, s$$

$$C_{t,s} = Cov(Y_t, Y_s) = Cov(Y_{t+\delta}, Y_{s+\delta}) = C_{s-t}, \quad \forall\, s, t, \delta$$

$$R_{t,s} = Corr(Y_t, Y_s) = Corr(Y_{t+\delta}, Y_{s+\delta}) = R_{s-t}, \quad \forall\, s, t, \delta$$

Estacionariedad débil

Se dice que un proceso es **débilmente estacionario (DE)** si sólo se cumple

$$\mu_t = \mu_s = \mu \quad \forall t, s$$

Es decir que sólo se pide que sean estacionarios los momentos hasta de segundo orden.

$$C_{t,s} = Cov(Y_t, Y_s) = Cov(Y_{t+\delta}, Y_{s+\delta}) = C_{s-t}, \quad \forall\, s, t, \delta$$
$$R_{t,s} = Corr(Y_t, Y_s) = Corr(Y_{t+\delta}, Y_{s+\delta}) = R_{s-t}, \quad \forall\, s, t, \delta$$

En general cuando hablemos de series de tiempo estacionarias nos estaremos refiriendo a este tipo de estacionariedad.

Estimación de momentos

Dadas N muestras de una serie de tiempo DE $\{y_1, \ldots, y_N\}$, entonces podemos estimar los momentos como:

$(x-\mu)^2$

$$\hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$\hat{C}_k = \frac{1}{N} \sum_{n=k+1}^{N} (y_n - \hat{\mu})(y_{n-k} - \hat{\mu})$$

$$\mathrm{Var}(x) = E[(x-\mu)^2]$$

$$\hat{R}_k = \frac{\hat{C}_k}{\hat{C}_0}$$

Resumen

¿Qué cubrimos en esta clase?

- Qué es una serie de tiempo, para qué sirve estudiarlas y cómo podemos modelarlas
- Momentos útiles para analizar series de tiempo: media, varianza, autocovarianza y autocorrelación
- Algunos modelos: ruido blanco, caminante aleatorio, coseno aleatorio
- **Estacionariedad**
- Algunas técnicas de preprocesamiento

Actividad

1. Generar una serie de tiempo con valores aleatorios y graficarla.
2. Graficar una serie de tiempo de los ejemplos del repositorio:
3. Redactar un informe describiendo tres series de tiempo distintas (máx. 3 páginas)