



Universidad
Tecmilenio®



CONACYT
Consejo Nacional de Ciencia y Tecnología

Estadística y pronósticos para la toma de decisiones

Profesor-investigador: Dr. Naím Manríquez

Consejo Nacional de Ciencia y Tecnología - CONACYT

Sistema Nacional de Investigadores

Profesor-investigador: Dr. Naim Manríquez

Sobre mí

- » Doctor en Economía Regional – Centro de Investigaciones Socioeconómicas (CISE).
- » Miembro del Sistema Nacional de Investigadores del CONACYT – Consejo Nacional de Ciencia y Tecnología
- » Temas de trabajo: ciencia de datos, estadística, medio ambiente, economía regional y urbana.

Especialización y proyectos de investigación

- » Especialización en **análisis de datos y estadística** – Laboratorio Nacional de Políticas Públicas.
- » Colaborador en Proyectos ProNacEs (Programas Nacionales Estratégicos) del CONACYT y Gobierno de México: Programas Nacionales Estratégicos: **vivienda y ciudades sostenibles**.

Estancias académicas e intercambios:

- » Centro de Investigación y Docencia Económicas (CIDE – Campus Aguascalientes).
- » Universidad Nacional de la Patagonia Austral (Rio Gallegos , Argentina).

Prerrequisitos del curso:

- Contar con un equipo de computo.
- Tener buenas bases y haber concluido asignaturas como fundamentos matemáticos y economía.
- Tener instalados **R** y **Rstudio** en el equipo de computo a utilizar, o en su defecto tener una cuenta en **Rstudio Cloud** con la cual ir trabajando el material de la clase. Se puede instalar R en este enlace: <https://cran.r-project.org/> y Rstudio en este enlace: <https://www.rstudio.com/products/rstudio/download/>
- Tener la disposición de aprender y superar sus límites.

Reglas del curso:

- Sobre el **pase de lista**; la asistencia se tomará a través de un cuestionario de **Google Forms** que les pasaré durante la clase.
- Los ejercicios, actividades y evidencias se realizarán en parejas.
- Los controles de lectura se realizarán de manera individual.
- Las fechas de entrega de los ejercicios y material de apoyo lo encuentran en la página de Github: https://github.com/naimmanriquez/LAE_estadistica_2022
- Se proponen uno o dos **descansos de 5 minutos** entre la clase para poder descansar un poco los ojos (favor de recordar)
- ...mas las reglas que se vayan sumando :P ...

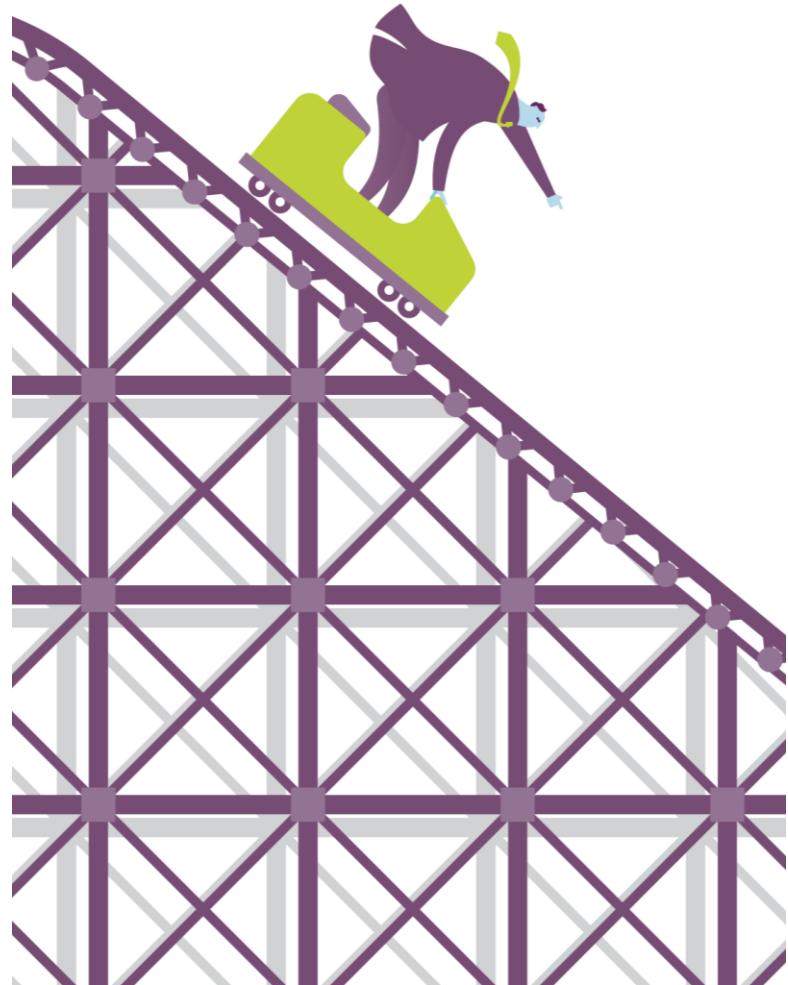
Dos sugerencias para el curso:

- 1. Tengan una cuenta de GitHub:** Una cuenta de GitHub siempre es un plus, para manejar nuestros archivos y bases de datos, tener control de versiones de nuestros proyectos y para presumirlos con la gente, entre otras cosas. Pueden registrarse en el siguiente enlace: <https://github.com/>
- 2. Trabajo en equipos de a dos:** Dado que a veces es pesado trabajar individual, vamos a agruparnos en parejas para trabajar durante todo el semestre, además se fomenta la colaboración y compañerismo.

Temario: estadística y pronósticos para la toma de decisiones...

Hay tres principales módulos que integran el temario del semestre:

- 1. Estadística y probabilidad:** Estadística descriptiva (representación gráfica de variables, detectar patrones en los datos), **modelos de probabilidad**, **estadística inferencial** (pruebas y contrastes de hipótesis).
- 2. Series de tiempo y regresión lineal:** modelos de series de tiempo, suavizamiento exponencial, regresión lineal simple, mínimos cuadrados ordinarios (MCO).
- 3. Regresión lineal múltiple:** análisis y predicción, efectos causales, variables dependientes y variables independientes, econometría, regresión logística.



Bibliografía del curso y bibliografía recomendada

Libro de texto del curso:

Rodríguez, J., Pierdant, A., y Rodríguez, E. (2016). *Estadística para administración* (2a ed.). México: Patria. ISBN: 978-6077443759

Libros de apoyo:

Hanke. J. E. y Wichern. D. W. (2010). *Pronósticos en los negocios* (9^a ed.). México: Pearson. ISBN: 9786074427004



Bibliografía del curso y bibliografía recomendada

Libros de texto recomendados:

- Heumann, Christian; Schomaker, Michael. (2016). **Introduction to Statistics and Data Analysis with exercises, solutions and applications in R.** (1st ed.). Springer, Editorial. ISBN 978-3-319-46160-1
- Wooldridge, Jeffrey (2019) **Introductory econometrics: a modern approach.** Cengage Editorial.
- Quintana Romero, Luis; Mendoza, Miguel. (2016). **Econometría aplicada usando R (1ra edición).** UNAM, Editorial
- Kopczewska, Katarzyna (2021) **Applied Spatial Statistics and Econometrics.** Routledge.

Recursos didácticos:

Calculadoras y notación matemática

- Symbolab. (2012). *Calculadora*. Recuperado de <https://www.symbolab.com/>
- Solve My Math. (2016). *Calculadora*. Recuperado de <http://www.solvemymath.com/>
- WolframAlpha. (2016). *Calculadora*. Recuperado de <http://www.wolframalpha.com/>

Bases de datos

Banco de Información Económica (INEGI)
<https://www.inegi.org.mx/sistemas/bie/>

DataMexico (INEGI)
<https://datamexico.org/>

Encuestas: ENOE, ENIGH, Censos Económicos, etc.
<https://www.inegi.org.mx/datos/?ps=Programas>

Programas y softwares a utilizar

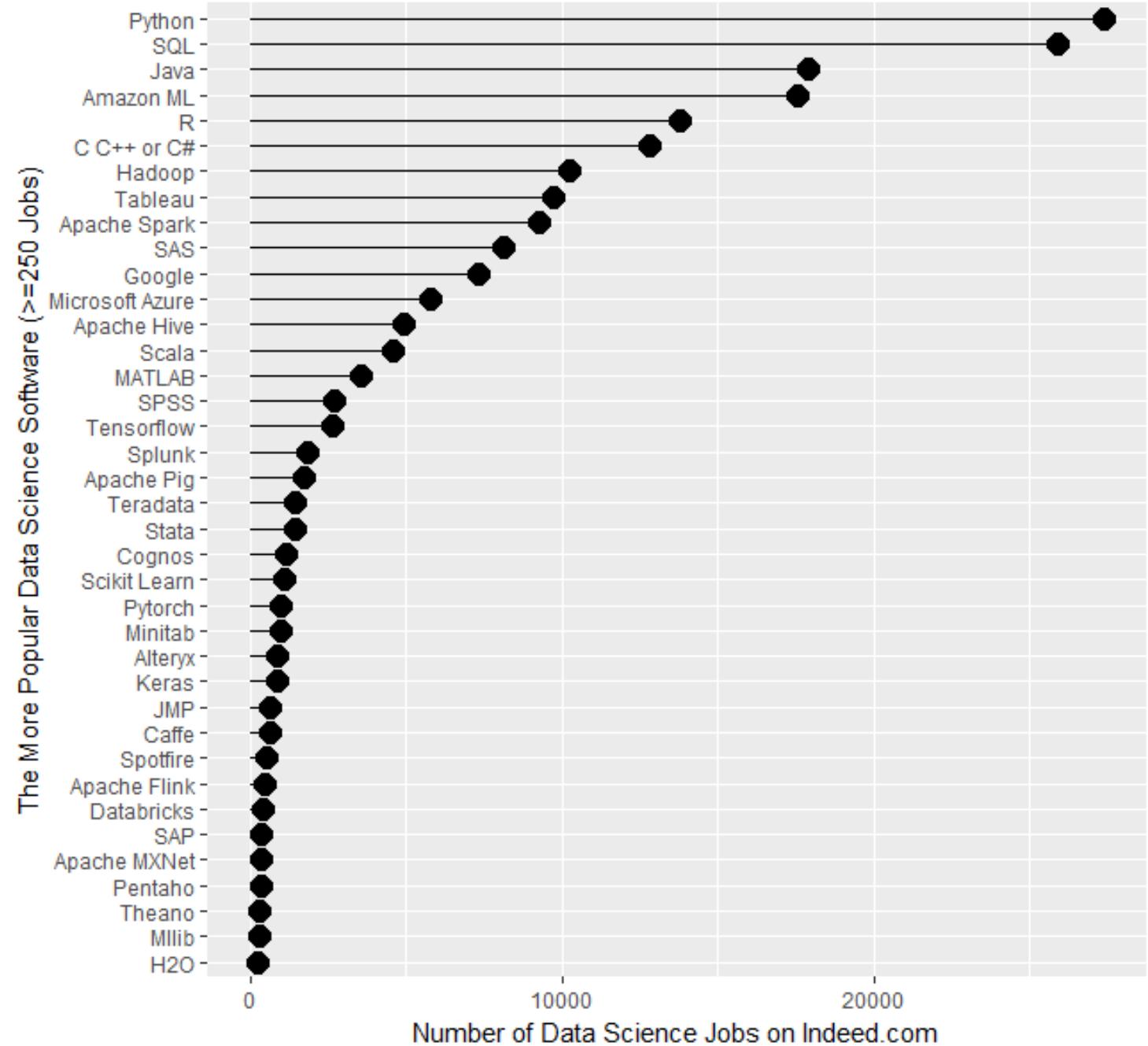
Software estadístico y lenguaje de programación



Hojas de cálculo



Lenguajes de programación para estadística y ciencia de datos...

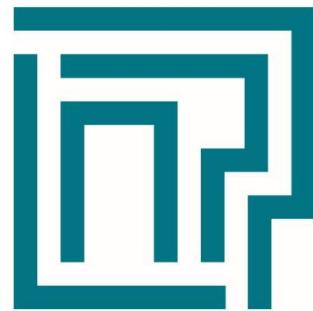
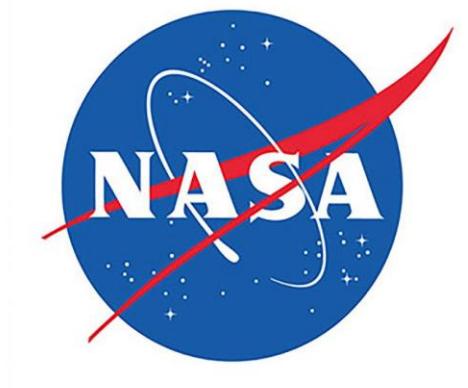


Algunas empresas que utilizan lenguaje de programación estadística...



INTELIGENCIA APLICADA
A DECISIONES

Algunas organizaciones públicas que utilizan lenguaje de programación estadística...



Universidades donde se enseñan lenguajes de programación estadística al menos en carreras de economía y negocios...



Tecnológico
de Monterrey



Massachusetts
Institute of
Technology



Berkeley
UNIVERSITY OF CALIFORNIA



HARVARD
UNIVERSITY



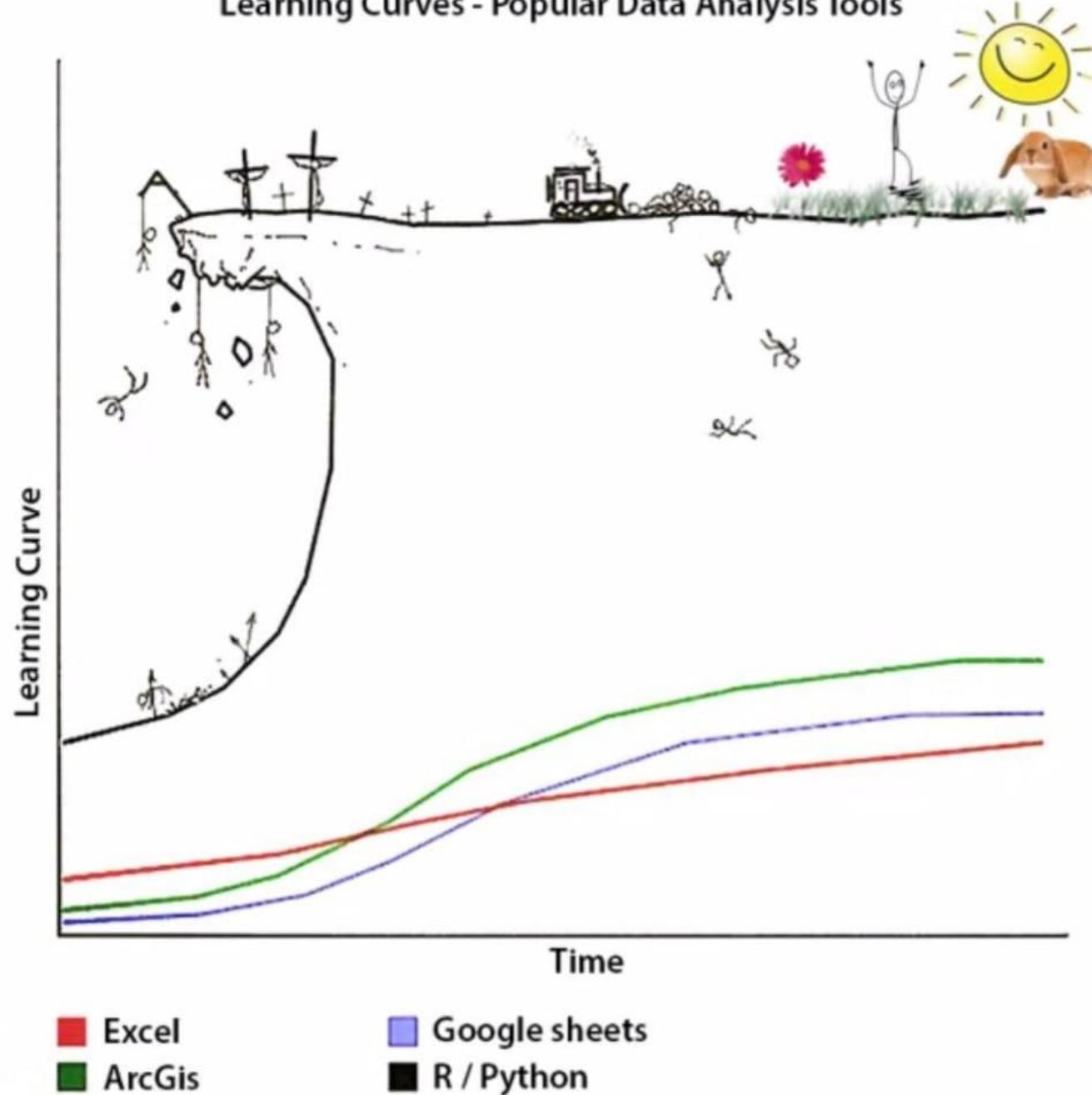
UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN



UNIVERSITY OF
CAMBRIDGE

Learning Curves - Popular Data Analysis Tools



Aprender estos lenguajes estadísticos tiene una curva de aprendizaje complicada pero no imposible de superar...

Sugerencias

No sufran en silencio

No acumules dudas por pena ni durante mucho tiempo, ya que los temas son acumulativos y una duda no resuelta en una clase puede hacer que no entiendas las clases posteriores.

¿Qué se puede hacer con R?

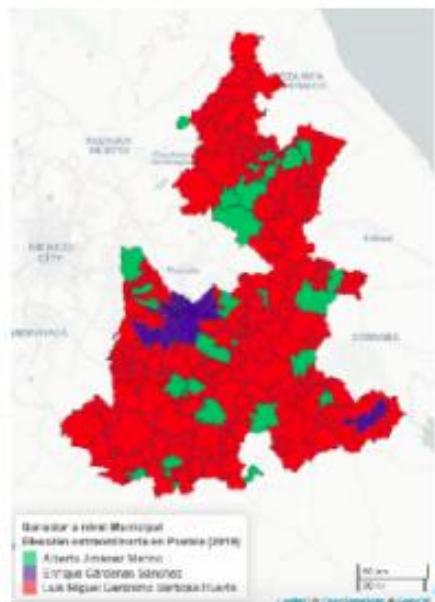
*Esto lo vamos a ver
en clase*



1. Manejo y visualización de datos.

```
library(tidyverse)
```

```
72 datos <- prep %>%
73   select(ECS, AJM, LMGBH, TOTAL_VOTOS, LISTA_NOMINAL, MUNICIPIO, DISTRITO) %>%
74   filter(!is.na(MUNICIPIO)) %>%
75   group_by(MUNICIPIO) %>%
76   summarise(ECS = sum(ECS, na.rm = TRUE),
77             AJM = sum(AJM, na.rm = TRUE),
78             LMGBH = sum(LMGBH, na.rm = TRUE),
79             Total_Votos = sum(TOTAL_VOTOS, na.rm = TRUE),
80             ListaNominal = sum(LISTA_NOMINAL, na.rm = TRUE)
81           )
```



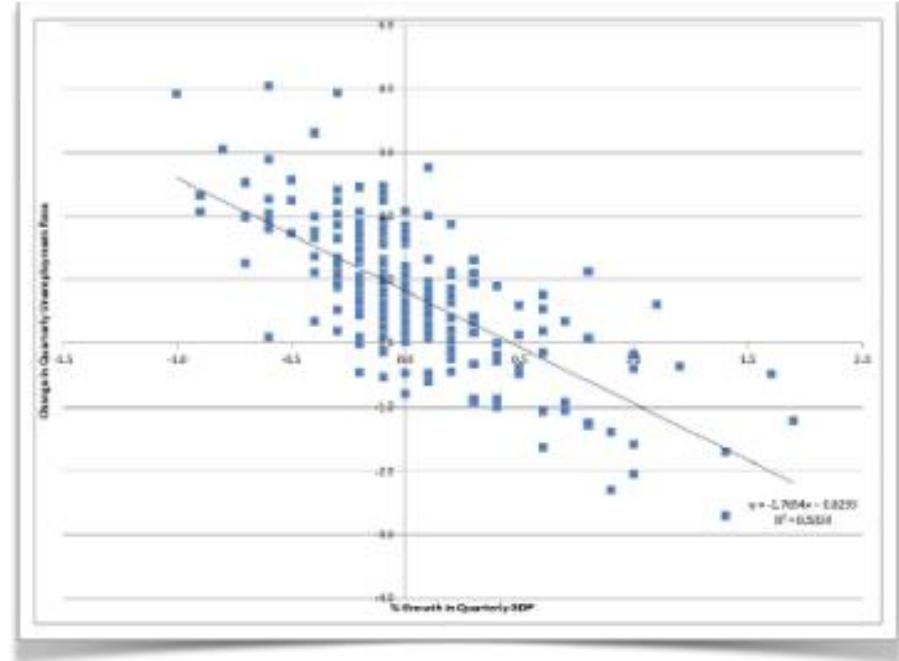
¿Qué se puede hacer con R?

*Esto lo vamos a ver
en clase* 😊

2. Análisis estadístico y econometría.

`library(base)`

`library(MASS)`



¿Qué se puede hacer con R?

*Esto no lo vamos a
ver en clase*



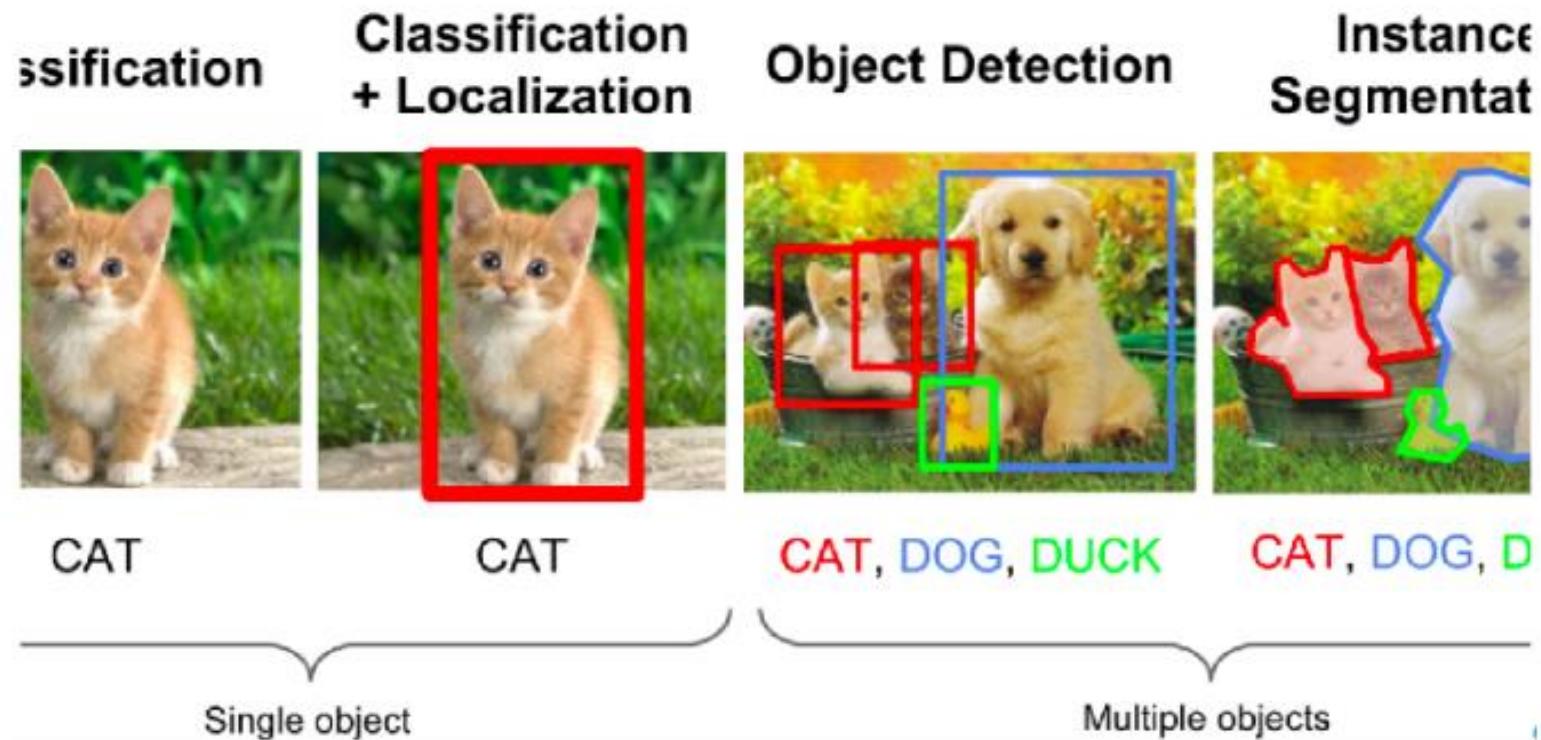
3. Machine Learning y Deep Learning.

```
library(e1071)
```

```
library(tensorflow)
```

```
library(caret)
```

```
library(rpart)
```



¿Qué se puede hacer con R?

*Esto lo vamos a ver
en clase*



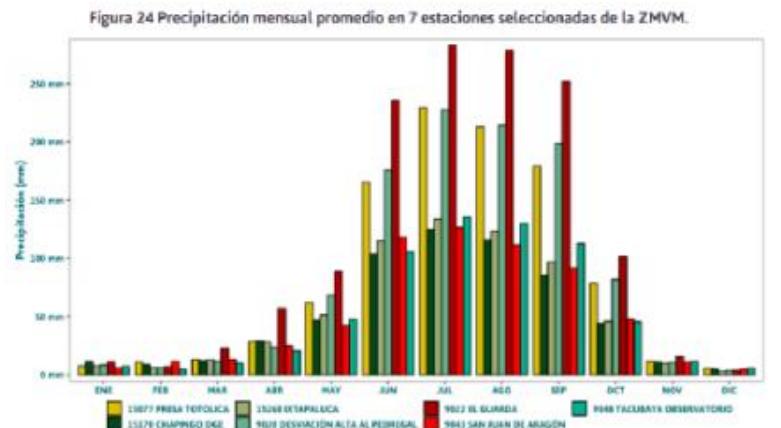
4. Visualización de datos.

```
library(ggplot2)
```

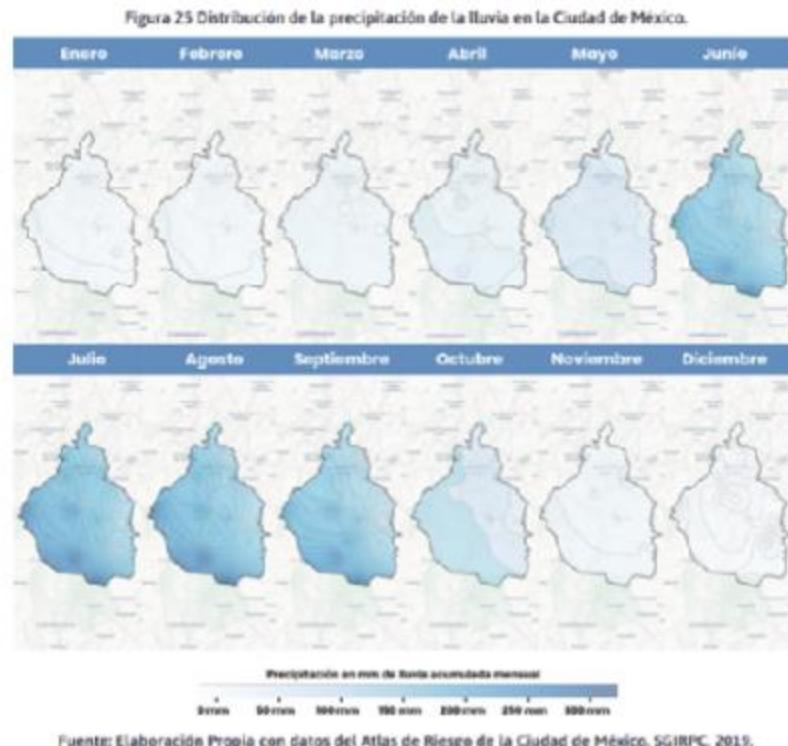
```
library(plotly)
```

```
library(leaflet)
```

```
library(htmlwidgets)
```



Gráfica de barras de la distribución de la lluvia en la CDMX, en el tiempo.



Mapas de la distribución de la lluvia en la CDMX, en el espacio y tiempo.

¿Qué se puede hacer con R?

5. Análisis de texto.

library(tm)

```
library(stringr)
```



Nube de palabras.

Solicitudes de Acceso a información realizadas en el Estado de Morelos.



Nube de palabras.

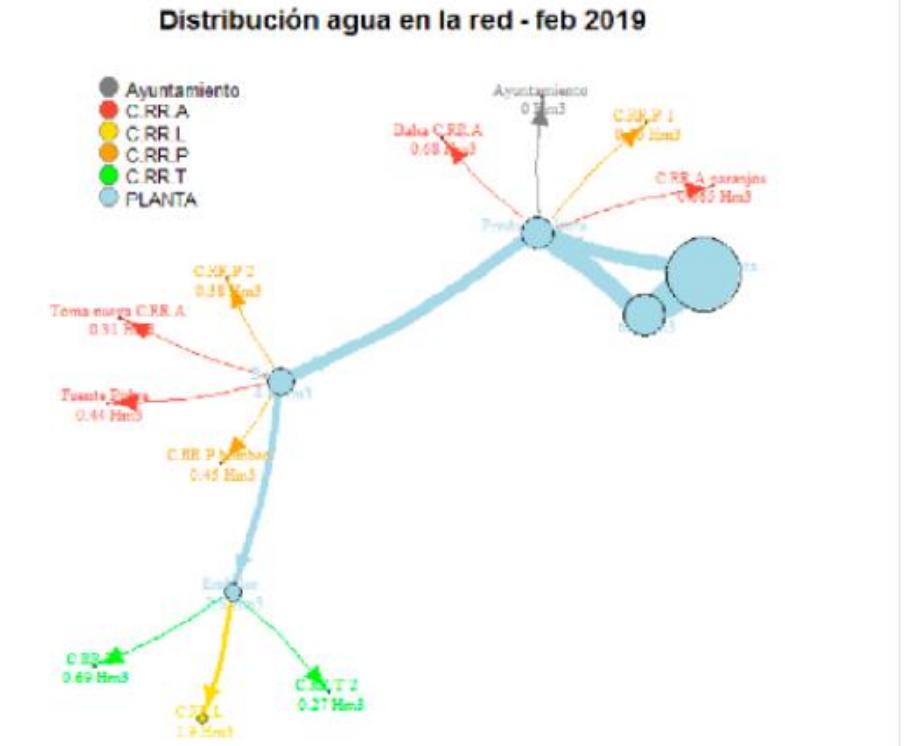
Plan Nacional de Desarrollo, 2019.

*Esto lo vamos a ver
en clase*



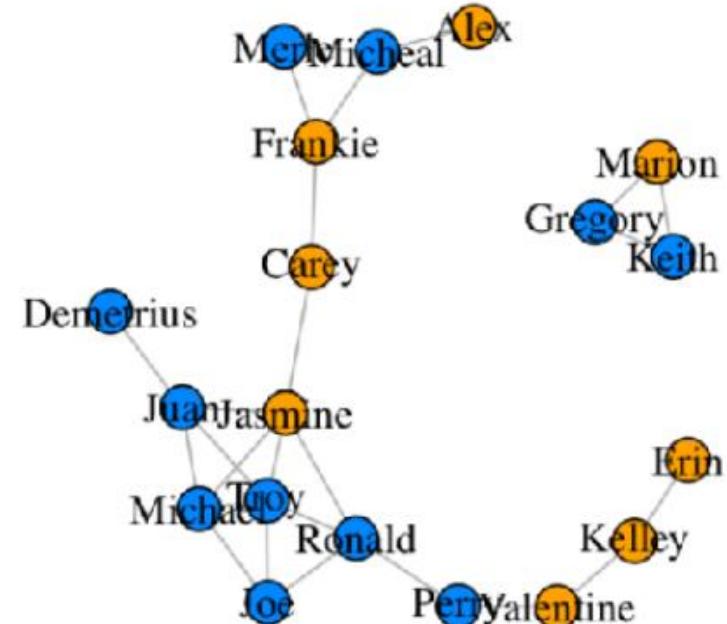
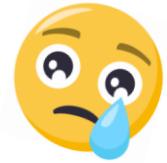
¿Qué se puede hacer con R?

6. Análisis de redes. library(igraph)



Red de distribución de Agua

*Esto no lo vamos a
ver en clase*



Red social de amigos en una prepa

¿Qué se puede hacer con R?

Esto no lo vamos a ver en clase



7. Recolección automática de información (Web Scrapping, Data Crawling).

```
library(rvest)  
library(xml)
```

Ejemplo: Extraer precios e información de vuelos desde Google Flights

Flight Details	Duration	Stops	Price
06:05 - 14:35 ¹ United - ANA	18 h 30 m MEX-NRT	1 stop 2 h 35 m SFO	MX\$20,089 round trip
07:30 - 15:20 ¹ United, ANA	17 h 50 m MEX-NRT	1 stop 1 h 35 m IAH	MX\$20,089 round trip
02:20 - 06:45 ¹ ANA	14 h 25 m MEX-NRT	Non-stop	MX\$21,689 round trip
01:30 - 06:20 ¹ Aeroméxico - ANA	14 h 50 m MEX-NRT	Non-stop	MX\$31,471 round trip
17:30 - 14:20 ² United, ANA	30 h 50 m MEX-NRT	2 stops ▲ IAU, LORB	MX\$20,089 round trip
17:30 - 14:20 ² United, ANA	30 h 50 m MEX-NRT	2 stops ▲ IAU, LORB	MX\$20,089 round trip

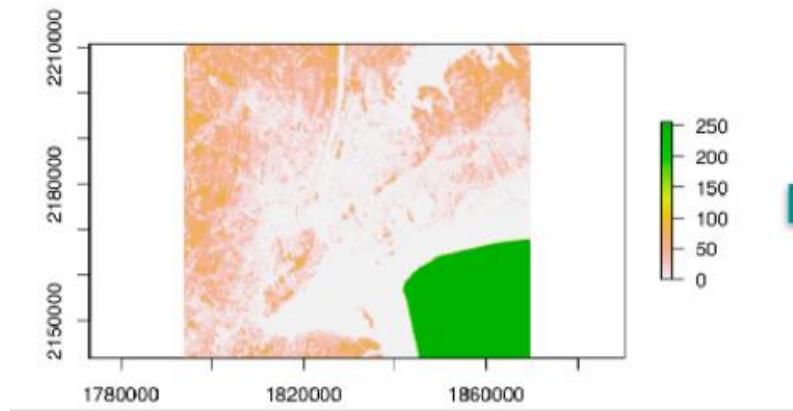
https://www.google.com/flights?lite=0#flt=MEX.NRT.2019-10-21*NRT.MEX.2019-11-05;c:MXN;e:1;sd:1;t:f

¿Qué se puede hacer con R?

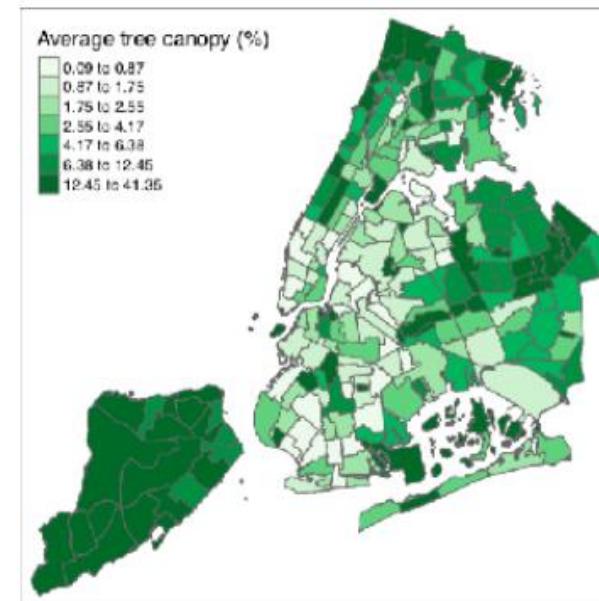
*Esto lo vamos a ver
en clase* 😊

8. Análisis Geoespacial. `library(sf)`

Abrir información geográfica, modificarla y visualizarla, así como realizar análisis a partir de esta.



Datos Crudos
(Raw Data)



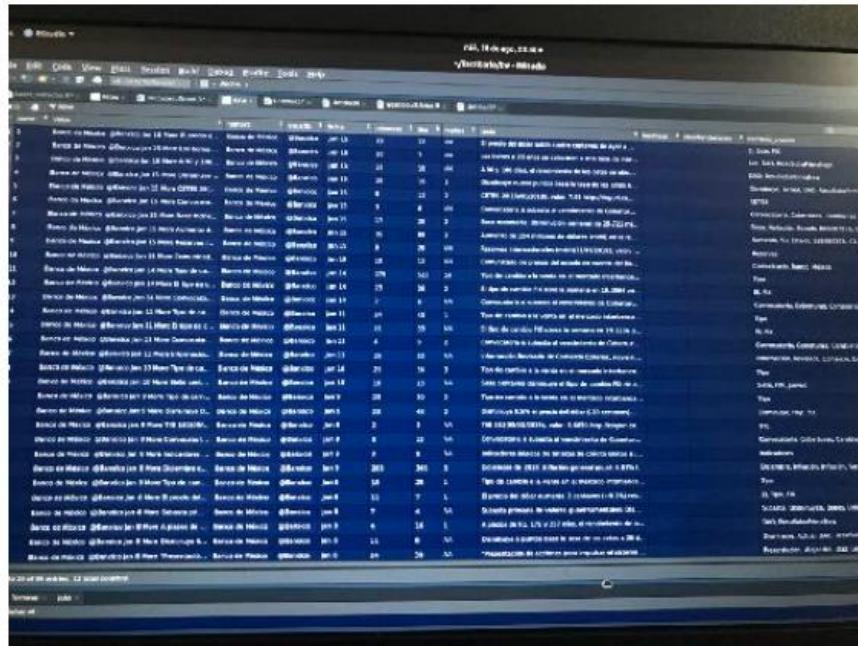
Datos Procesados que permiten llegar a conclusiones

¿Qué se puede hacer con R?

9. Automatización de tareas. library(Rselenium)

RSelenium permite programar el navegador para que replique cosas que nosotros podríamos hacer manualmente (p. ej. Descargar archivos, revisar Twitter, mandar correos, etc.).

*Esto no lo vamos a
ver en clase*



¿Qué se puede hacer con R?

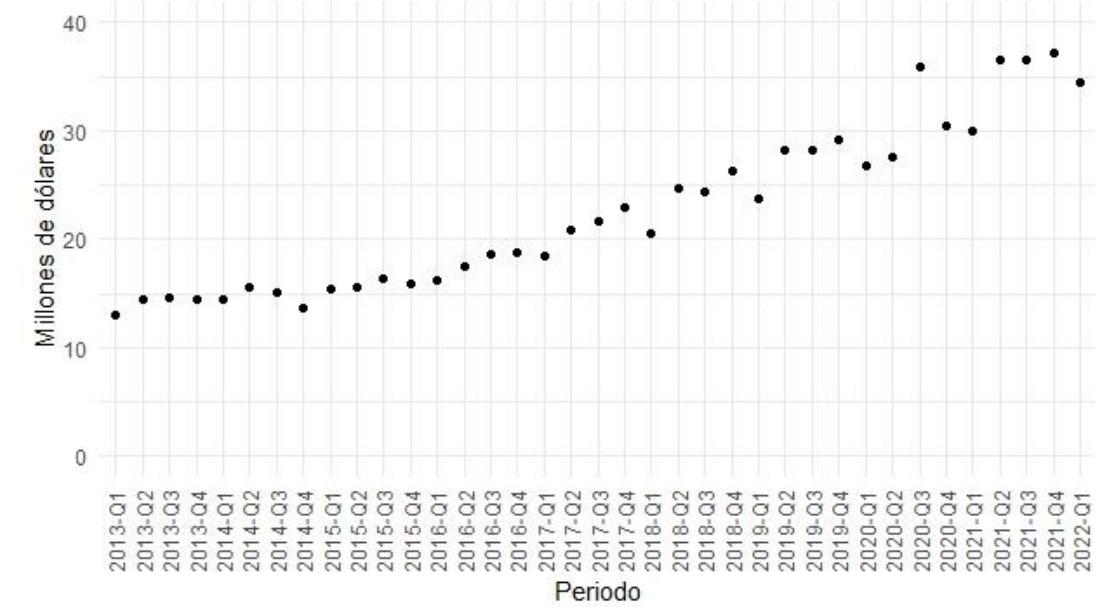
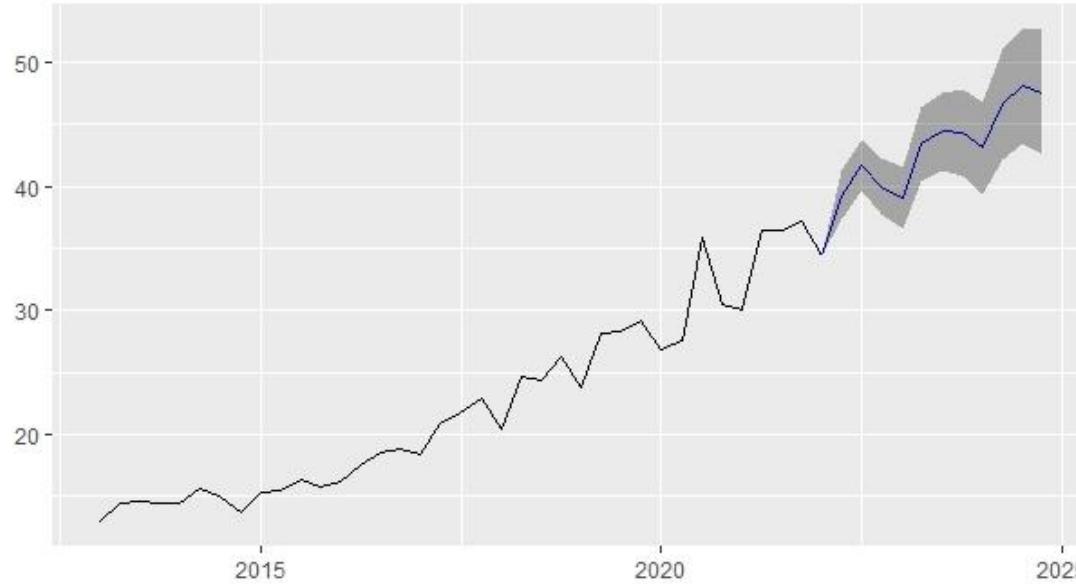
*Esto lo vamos a ver
en clase*



10. Pronósticos en series de tiempo.

`library(forcats)`

R nos permite hacer proyecciones y análisis de series de tiempo de alguna variable de interés.



Recursos extras para los estudiantes

Grupo de **Facebook**: Ciencia de datos con R.
<https://www.facebook.com/groups/1059429834256215>

Para dudas en tiempo real sobre códigos, comandos y técnicas utilizadas.



R-Ladies es una organización mundial cuya misión es promover la diversidad de género en la comunidad de R.

<https://rladies.org/>



Github:
Cuenta para compartir datos y tutoriales
<https://github.com/naimmanriquez>



Algunos grupos de R-Ladies en el mundo ...



Horario y fechas importantes

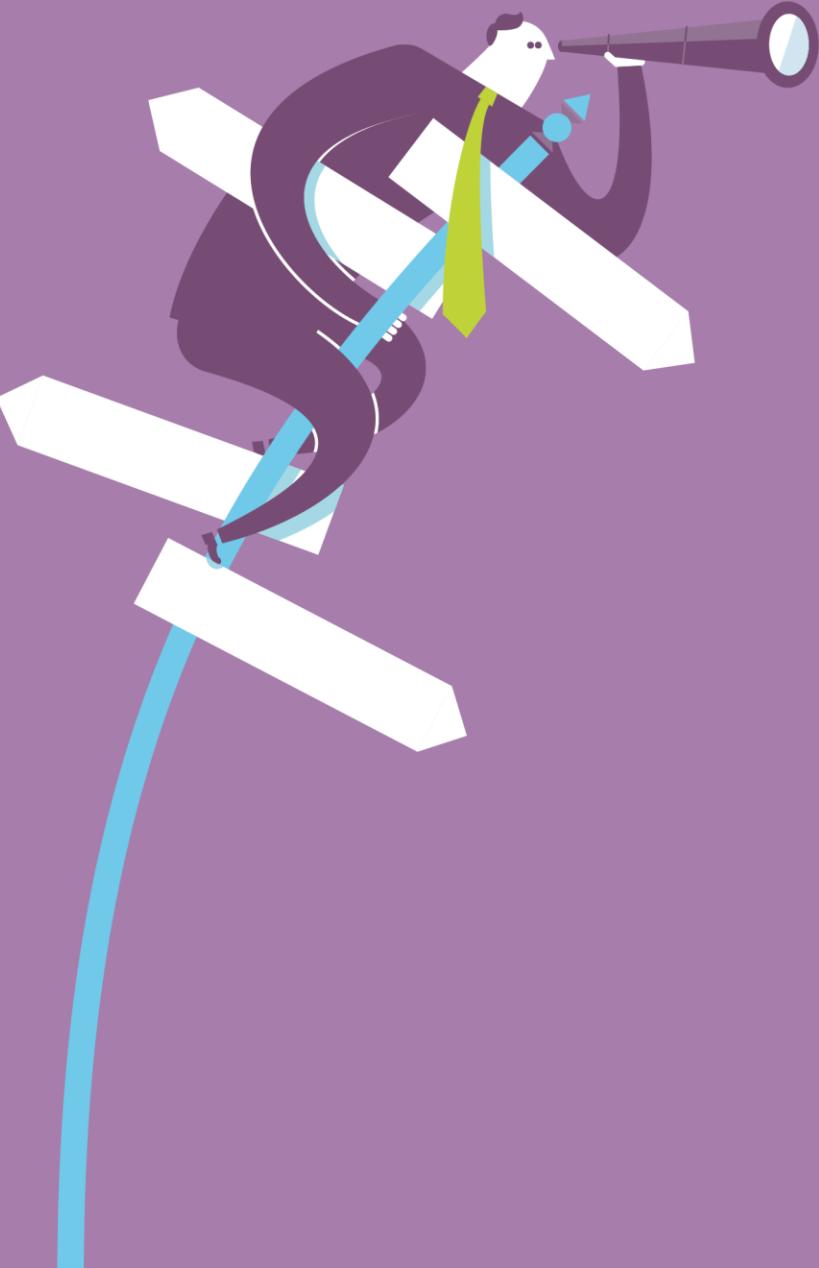
- **Horario:** martes, jueves de 7:30am a 8:55am y viernes de 7:00am a 8:55am
- **Inicio de clases:** 8 de agosto de 2022
- **Asuertos:** 16 de septiembre y 21 de noviembre
- **Primer parcial:** 9 de septiembre
- **Segundo parcial:** 14 de octubre
- **Último día de clases:** 29 de noviembre
- **Exámenes finales:** 30 noviembre – 8 de diciembre

Otras fechas importantes

- **Clase en modo virtual:** 23, 27, 29 y 30 de septiembre. Motivo: Profesor viaja a Colombia a presentar un proyecto sobre “Ciudad, planificación, ordenamiento territorial y técnicas estadísticas para el análisis de entornos urbanos”. Pontificia Universidad Javeriana, Departamento Administrativo Nacional de Estadística, y Alcaldía de Bogotá.
- **Clase en modo virtual:** 4, 6 y 7 de octubre. Motivo: Profesor presenta proyecto en Ciudad de México sobre: “Vivienda y acceso justo al hábitat”. (Fecha tentativa).

Modulo 1: Estadística y modelos de probabilidad.





Tema 1. Estadística descriptiva, representación gráfica y descripción matemática de la información.

Definición de estadística - RAE

estadística.

(Del al. Statistik).

- **1. f.** Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas.
- **2. f.** Conjunto de estos datos.
- **3. f.** Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades.



Definición #1: estadística

La estadística es parte del método que permite organizar, sintetizar, presentar, analizar, cuantificar e interpretar gran cantidad de datos, de tal forma que se puedan tomar decisiones y obtener conclusiones acerca de los fenómenos o líneas de investigación en estudio. (Rodríguez, Pierdant y Rodríguez, 2016).



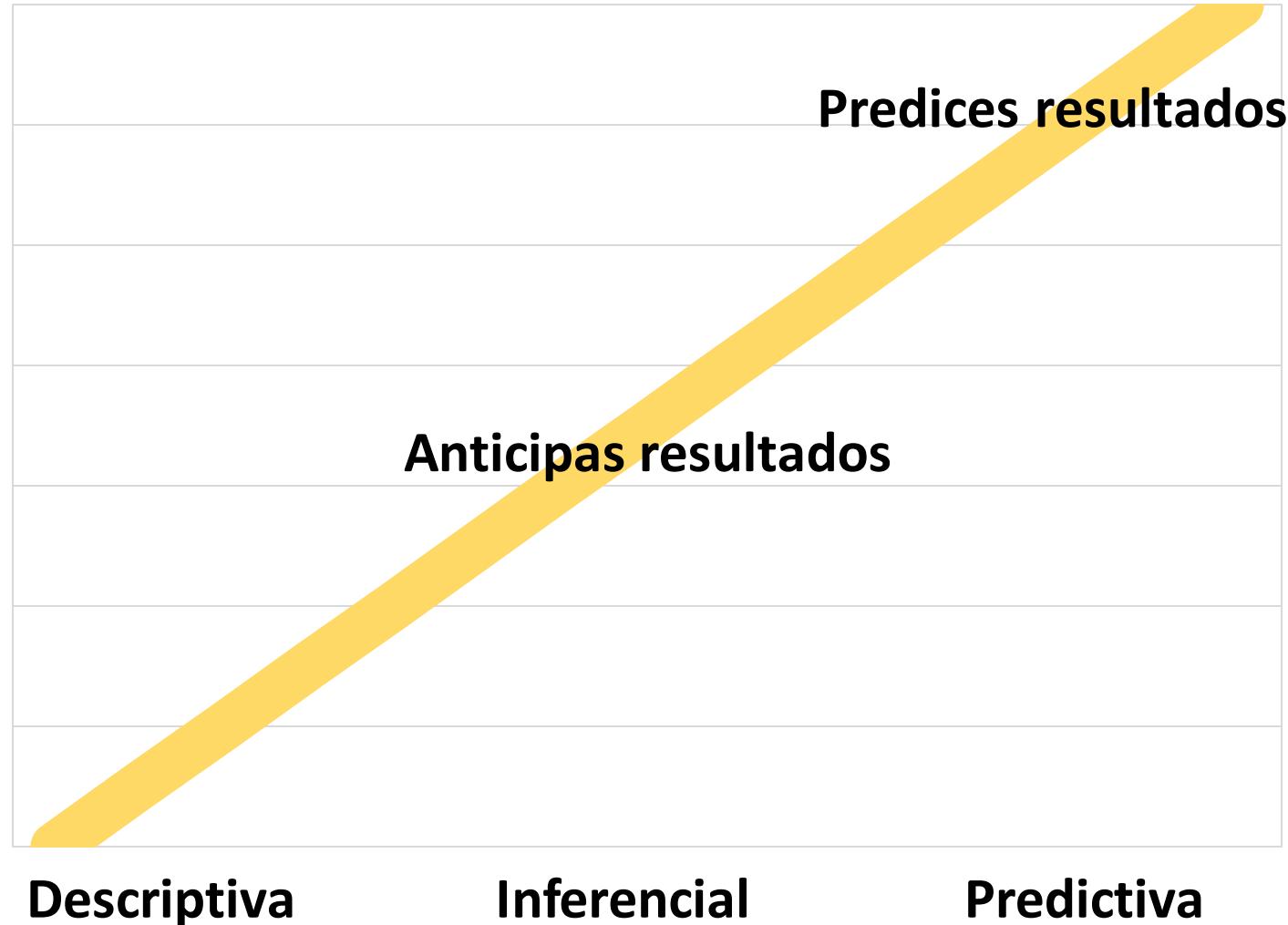
Definición #2: econometría

Entre el agregado de la estadística existe el término llamado “econometría”, el cuál es la aplicación de métodos estadísticos y matemáticos al análisis de los datos económicos, con el propósito de dar un contenido empírico a las teorías económicas y verificarlas o refutarlas.

Evolución en la estadística y analítica de datos

La estadística puede dividirse en dos grandes apartados, descriptiva e inferencial pero con la ciencia de datos y la econometría se puede lograr mejores predicciones...

- Descriptiva**
¿Cómo se están comportando los datos, qué patrones existen...?
- Inferencial**
Efectos causales y contraste de hipótesis, ¿cuál es la causa de que suceda ese patrón de datos...?
- Predictiva**
¿Qué va a pasar...? ¿A qué nos vamos a enfrentar?





Rama del conocimiento

Estadística

Tipos de estadística

Descriptiva

Inferencial

Predictiva

Técnicas y herramientas

Media, moda, mediana

Varianza, desviación estándar

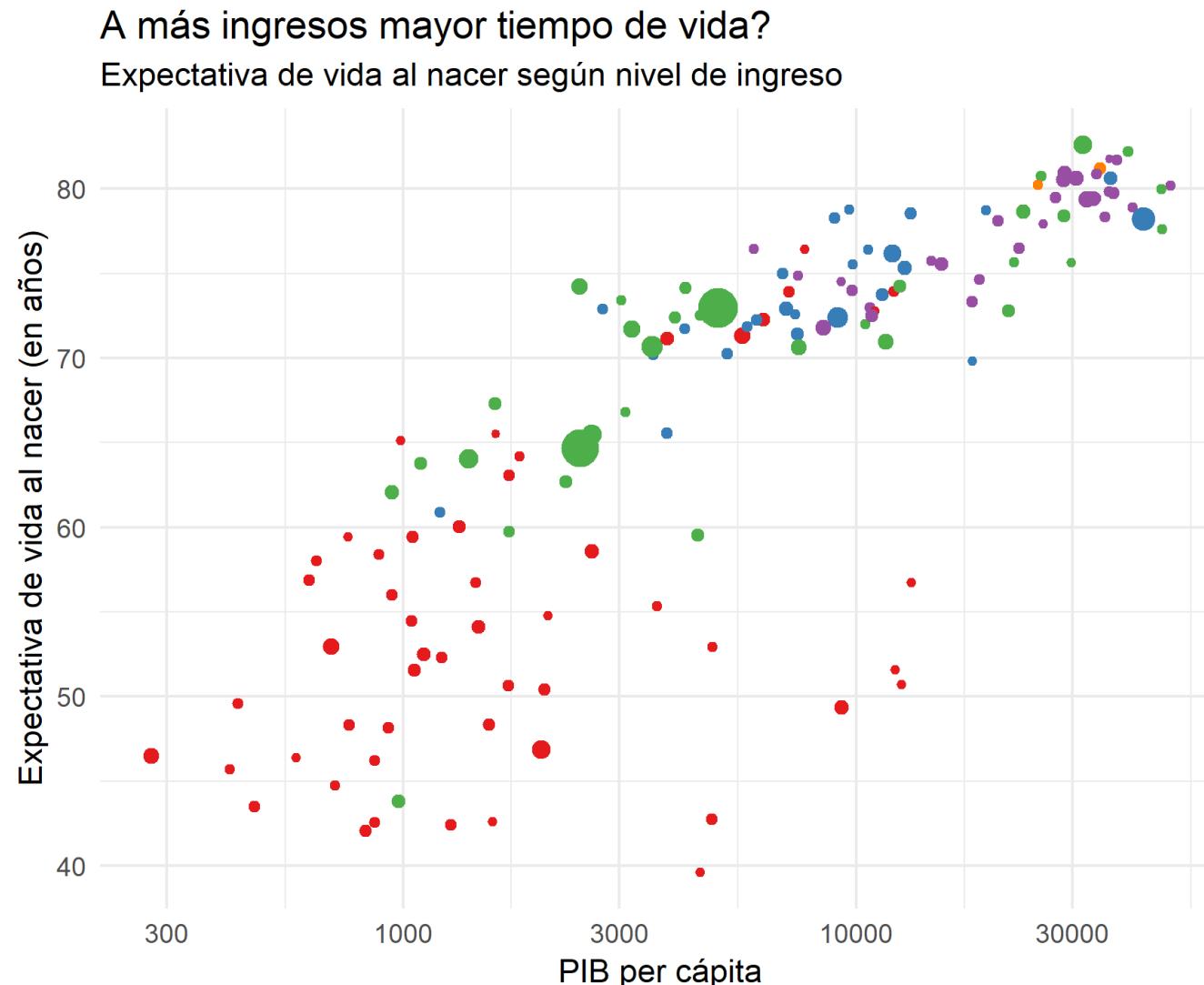
Contraste de hipótesis, probabilidad

Modelos econométricos: regresión, probit, logit

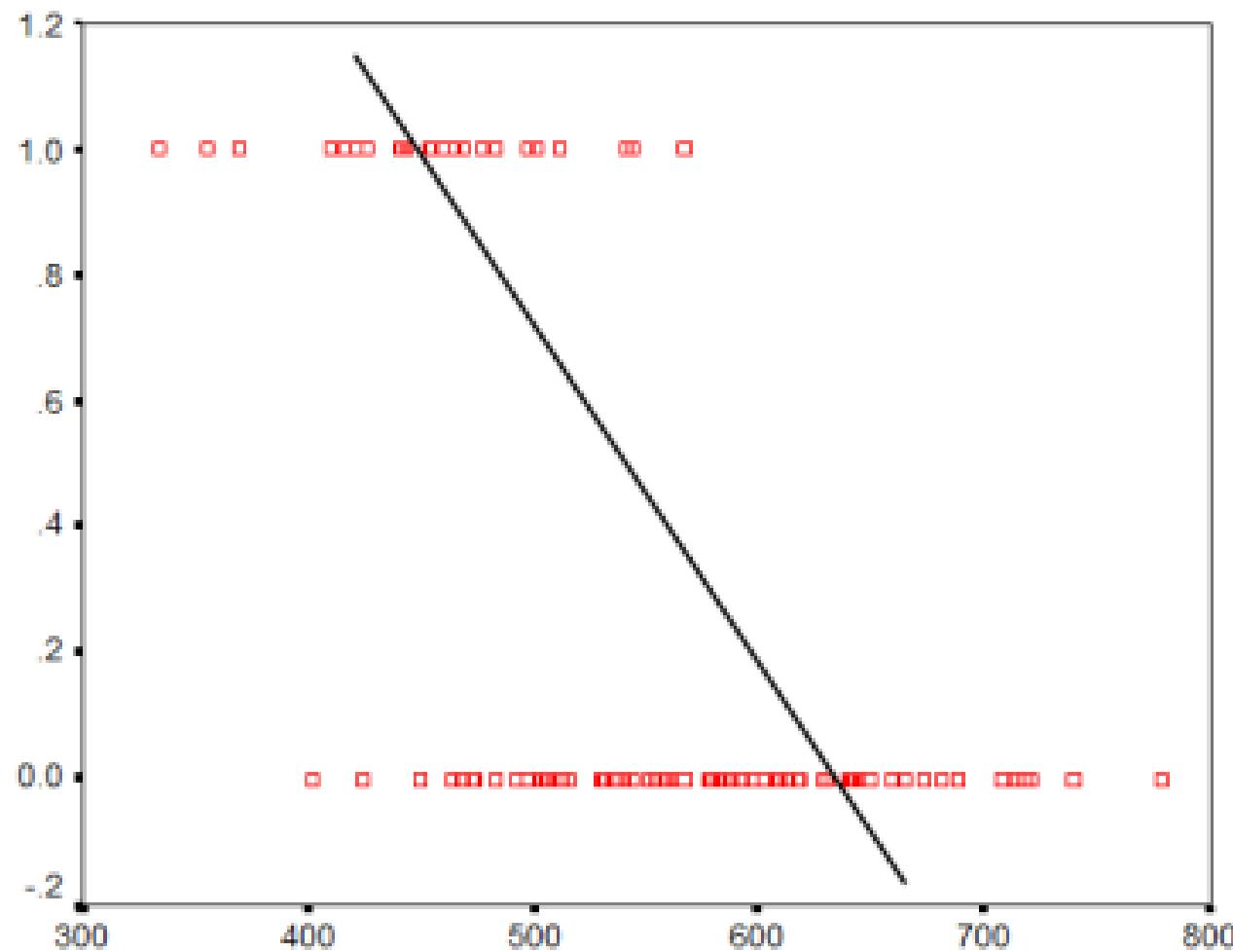
Tipo de variables en la estadística

#	Tipo	Descripción
1	Cuantitativas.	Se refiere a exclusivamente cantidades numéricas: ventas, producción total, gasto, número de delitos, etc.
2	Cualitativas.	Expresan cualidades, atributos, categorías o características de algo. Pueden capturarse por ejemplo como 0 y 1, las llamadas variables dicotómicas: 1, si se presenta una característica, 0 si no la presenta.
#	Georreferenciar.	Es una parte en la estadística y econometría espacial donde a cualquier variable cualitativa o cuantitativa se le asigna a un espacio o territorio.

Variables cuantitativas: ejemplo de gráfica de dispersión variable X (PIB per cápita) vs variable Y (esperanza de vida)

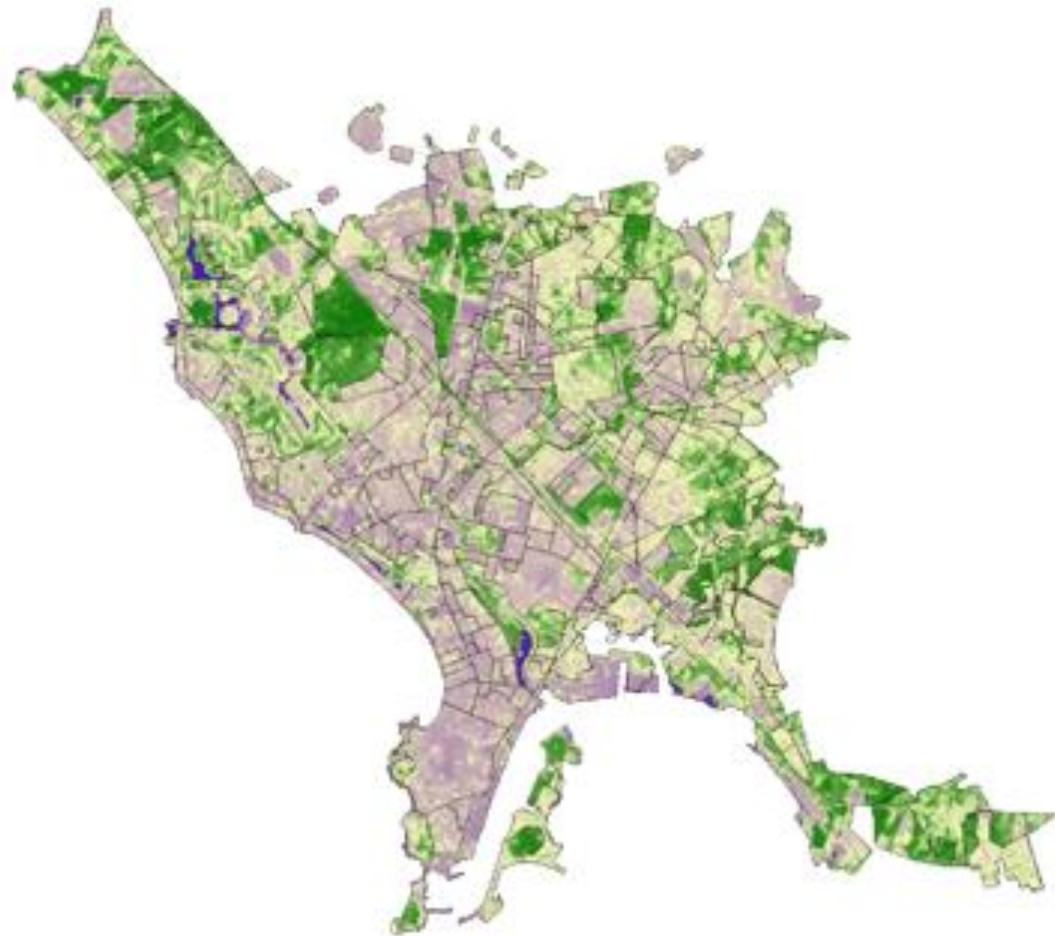


Variables cualitativa con cuantitativa: se contrasta la variable cualitativa de si una persona ha contratado un servicio, 1 = si contrata, y 0 = si no contrata vs variable cuantitativa: ingreso.

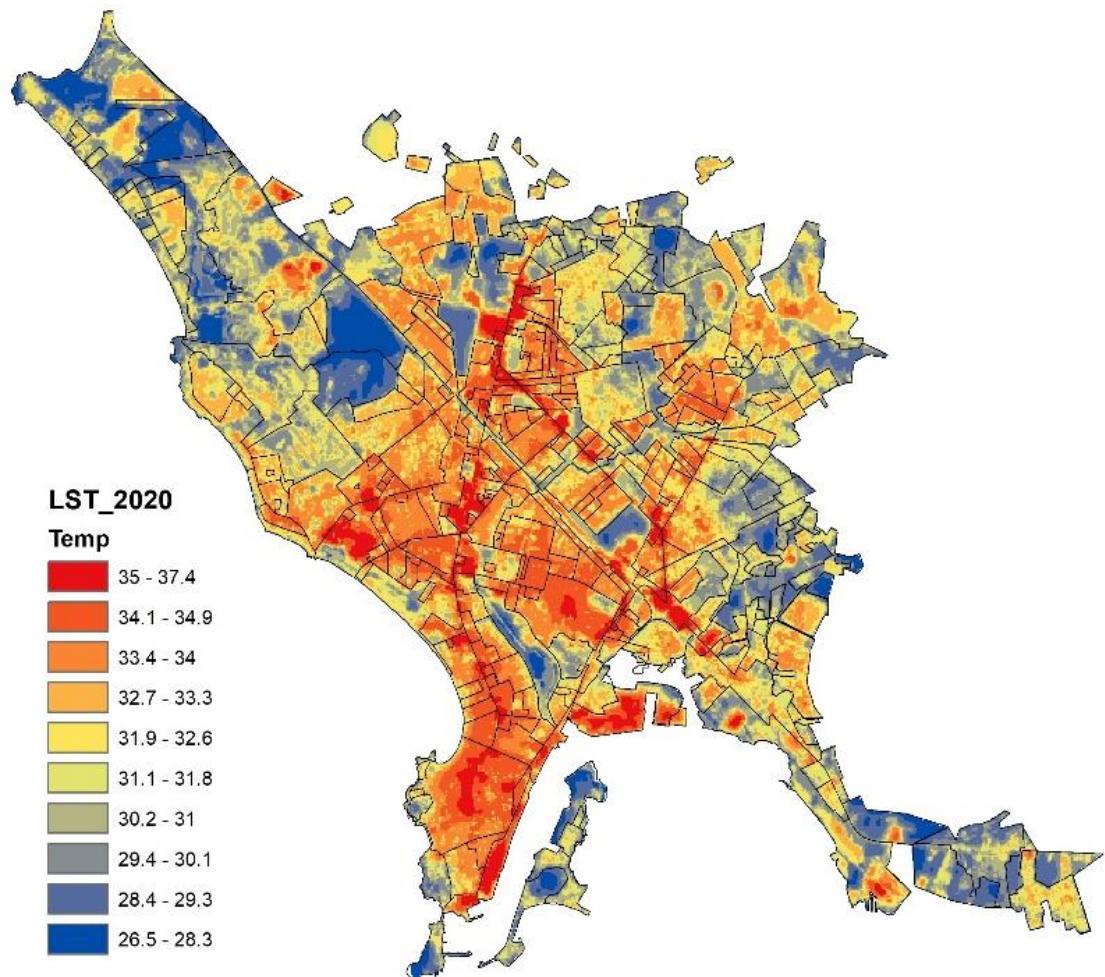


Variable georreferenciada: áreas verdes en Mazatlán y evidencia de isla de calor

Índice de Vegetación de Diferencia Normalizada



Isla de calor en la ciudad de Mazatlán



Elaboración propia en Rstudio con datos de la Administración Nacional de Aeronáutica y el Espacio. La isla de calor se refiere a la presencia de aire más caliente en ciertas zonas de ciudad,

Operadores matemáticos en estadística.

#	simbolo	Descripción
1	Σ	Este símbolo (llamado sigma) significa "sumatoria". Por lo tanto, si ves este simbolo " Σx_i " solo significa "sumar todos los valores recopilados"..
2	Π	Este símbolo (pi) significa "multiplicar". Entonces, si ves algo como " Πx_i " solo significa "multiplicar todos los valores recopilados"..
3	\sqrt{x}	Significa sacar la raíz cuadrada de x.

Símbolos griegos: ejemplo de algunos.

#	simbolo	Descripción
1	σ	Significa la desviación estándar de un conjunto de datos..
2	β_i	Coeficiente asociado a variable en el análisis de regresión..
3	ρ	Significa el nivel de correlación entre dos variables. Va entre -1 y 1. Puede interpretarse como una correlación positiva fuerte cuando el numero es mayor a 0.50, y negativa fuerte cuando el valor es mayor de -0.50.

Sumatoria: Sigma, Σ .

$$\sum_{i=1}^n x_i$$

debe leerse como “la suma de los números x_i desde x_1 hasta x_n ”.

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

El valor del índice i en la parte inferior de la letra griega sigma indica cuál es el primer término de la suma, mientras que el último de la parte superior indica el último término de la misma.

Si $x_1 = 2$, $x_2 = 3$, $x_3 = 2$, $x_4 = 0$

$$\sum_{i=1}^4 x_i = x_1 + x_2 + x_3 + x_4 = 2 + 3 + 2 + 0 = 7$$

Media muestral: \bar{X} .

La media aritmética de n observaciones de la variable x se denotará con el símbolo \bar{X} y se define como la suma de ellas dividida por n . Simbólicamente, se representa de la siguiente manera:

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

Con los datos del ejemplo anterior, tenemos lo siguiente

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{4} \sum_{i=1}^4 x_i = \frac{2 + 3 + 2 + 0}{4} = 1.7500$$

Mediana

La mediana de un conjunto de n números ordenados de menor a mayor es el número central en el arreglo. Es un valor que divide a los datos en mitades, una con todas las observaciones mayores o iguales a la mediana y otra con aquellas menores o iguales a ella. Si n es un número non, solo hay un valor central. Si n es un número par, hay dos valores centrales, y la mediana debe tomarse como la media aritmética de estos dos valores.

Mediana para datos impares

Datos sin ordenar	46	47	30	17	43	48	21
-------------------	----	----	----	----	----	----	----

Datos ordenados	17	21	30	43	46	47	48
-----------------	----	----	----	----	----	----	----



Mediana para datos pares

Datos sin ordenar	46	47	30	17	42	48	21	36
-------------------	----	----	----	----	----	----	----	----

Datos ordenados	17	21	30	36	42	46	47	48
-----------------	----	----	----	----	----	----	----	----

$$36+42=78$$

$$78/2 = 39$$

Mediana = 39



Moda

Otra medida de tendencia central es la moda. Se define como el valor que se presenta con mayor frecuencia en una serie de datos.

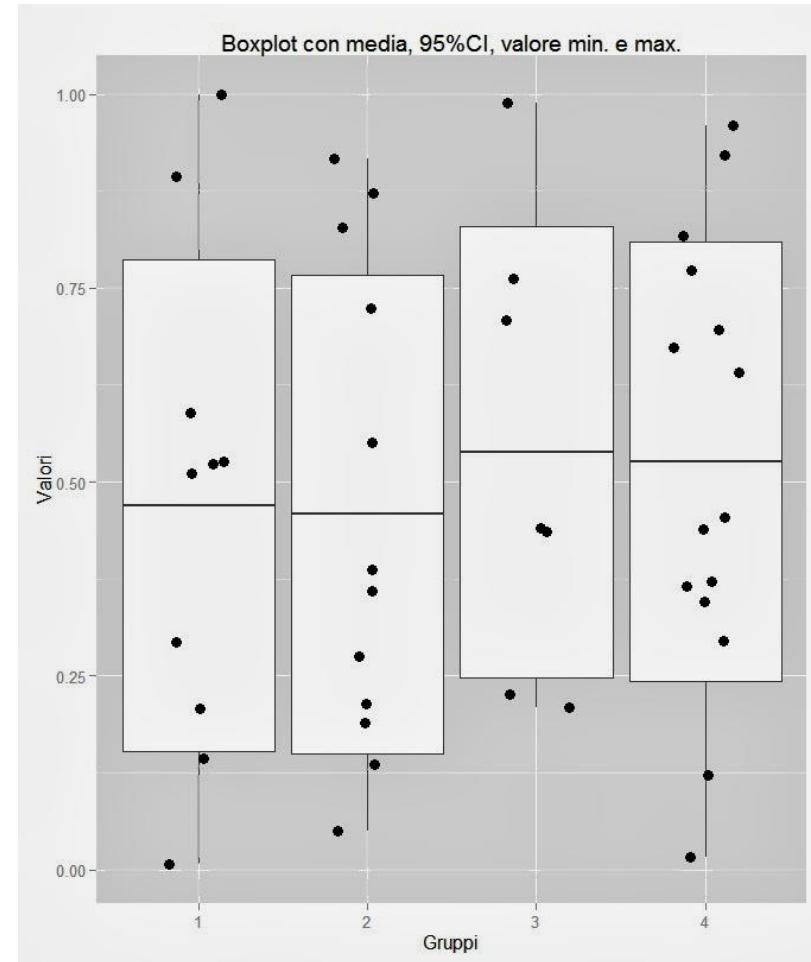
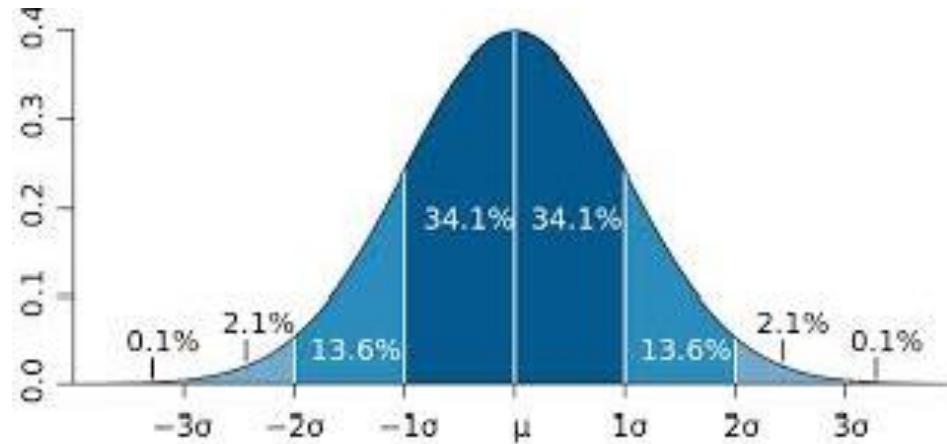
Ejemplo

Las calificaciones obtenidas por un alumno en ocho exámenes del curso de Estadística son:

100	85	80	85	90	80	85	90
-----	----	----	----	----	----	----	----

La moda de este conjunto es 85, puesto que tiene frecuencia 3, mientras que los otros números tienen frecuencia de 1, 2 y 2, respectivamente.

Medidas de dispersión



La dispersión se refiere a la separación de los datos en una distribución, es decir, al grado en que las observaciones se separan. Aquí por ejemplo el símbolo μ significa la media muestral de los datos que tenemos, y σ significa la desviación estándar.

Las medidas de dispersión más comunes son rango, desviación estándar y varianza.

Rango

Es el intervalo que existe entre el valor máximo y el valor mínimo de una serie de datos. Nos da una idea de la dispersión de los datos, de tal forma que cuanto más grande es el rango, es más probable que los datos se encuentren más dispersos entre sí.

Datos	
Límite inferior	Límite superior
436	868
510	
520	
562	
591	
658	
665	
678	
680	
708	
718	
727	
728	
741	
762	
799	
813	
831	
834	

Rango

$$R = L_s - L_i$$

$$\text{Rango} = 868 - 436$$

$$\text{Rango} = 432$$

Varianza y desviación est醖ar

La varianza (s^2) de un conjunto de datos se define como la suma de cuadrados de las desviaciones de las observaciones con respecto a la media y dividida por el n閞mero de observaciones menos uno. Su ecuaci髇 es la siguiente:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

	Datos	Media o promedio	$x - \bar{x}$	$(x - \bar{x})^2$
L韗ite inferior	436	691.45	-255.45	65254.7025
	510	691.45	-181.45	32924.1025
	520	691.45	-171.45	29395.1025
	562	691.45	-129.45	16757.3025
	591	691.45	-100.45	10090.2025
	658	691.45	-33.45	1118.9025
	665	691.45	-26.45	699.6025
	678	691.45	-13.45	180.9025
	680	691.45	-11.45	131.1025
	708	691.45	16.55	273.9025
	718	691.45	26.55	704.9025
	727	691.45	35.55	1263.8025
	728	691.45	36.55	1335.9025
	741	691.45	49.55	2455.2025
L韗ite superior	762	691.45	70.55	4977.3025
	799	691.45	107.55	11567.0025
	813	691.45	121.55	14774.4025
	831	691.45	139.55	19474.2025
	834	691.45	142.55	20320.5025
	868	691.45	176.55	31169.9025
			Suma $(x - \bar{x})^2$	264868.95
			Varianza	13940.47105

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

Desviación estándar muestral

La desviación estándar de un grupo de observaciones es la raíz cuadrada positiva de la varianza de las observaciones.

	Datos	Media o promedio	$x - \bar{x}$	$(x - \bar{x})^2$
Límite inferior	436	691.45	-255.45	65254.7025
	510	691.45	-181.45	32924.1025
	520	691.45	-171.45	29395.1025
	562	691.45	-129.45	16757.3025
	591	691.45	-100.45	10090.2025
	658	691.45	-33.45	1118.9025
	665	691.45	-26.45	699.6025
	678	691.45	-13.45	180.9025
	680	691.45	-11.45	131.1025
	708	691.45	16.55	273.9025
	718	691.45	26.55	704.9025
	727	691.45	35.55	1263.8025
	728	691.45	36.55	1335.9025
	741	691.45	49.55	2455.2025
Límite superior	762	691.45	70.55	4977.3025
	799	691.45	107.55	11567.0025
	813	691.45	121.55	14774.4025
	831	691.45	139.55	19474.2025
	834	691.45	142.55	20320.5025
	868	691.45	176.55	31169.9025
	Suma $(x - \bar{x})^2$		264868.95	
	Varianza		13940.47105	
Desviación Estándar		118.069772		

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

$$\sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Tabla de frecuencias

Es una tabla que agrupa datos en intervalos no traslapados llamados clases y que registra el número de datos en cada clase. Ejemplo rango de estatura en los jugadores del FIFA 2022.

Estatura de los jugadores del FIFA 2022.

value	N	Raw %	Valid %	Cum. %
160-	54	0.28	0.28	0.28
161 a 170	1716	8.74	8.74	9.02
171 a 180	7617	38.80	38.80	47.82
181 a 189	8493	43.27	43.27	91.09
190 a 206	1750	8.91	8.91	100.00

Fuente: datos obtenidos de EA Sports.

Ejemplo de como elaborar una tabla de frecuencias

En el siguiente cuadro se presentan 40 valores aleatorios sobre los gastos en pesos de diferentes personas:

405	648	876	1082
465	680	885	1099
502	697	887	1130
537	707	905	1131
538	745	908	1147
559	749	917	1163
577	764	953	1164
598	768	982	1178
617	815	1009	1189
622	824	1058	1198

Primero, determinamos el rango:

Límite superior	1198
Límite inferior	405
Rango	793

$R = L_s - L_i$

Determinación del número de clases

Para determinar en cuántas clases dividiremos los datos para su estudio, emplearemos la siguiente relación:

$$k \geq \frac{\log N}{\log 2}$$

Donde:

N = número de datos

2 = límites superior e inferior de cada clase

k = número de clases buscado

$$k \geq \frac{\log 40}{\log 2} \geq 5.32$$

Como obtenemos un valor mixto, subimos al siguiente valor entero.

Clases	5.32	6.00
--------	------	------

Tamaño de clase

Para el tamaño de clase, empleamos la siguiente relación:

$$T_c \geq \frac{R}{k}$$

donde:

R = rango de los datos

K = número de clases entero que se obtuvo en el punto anterior

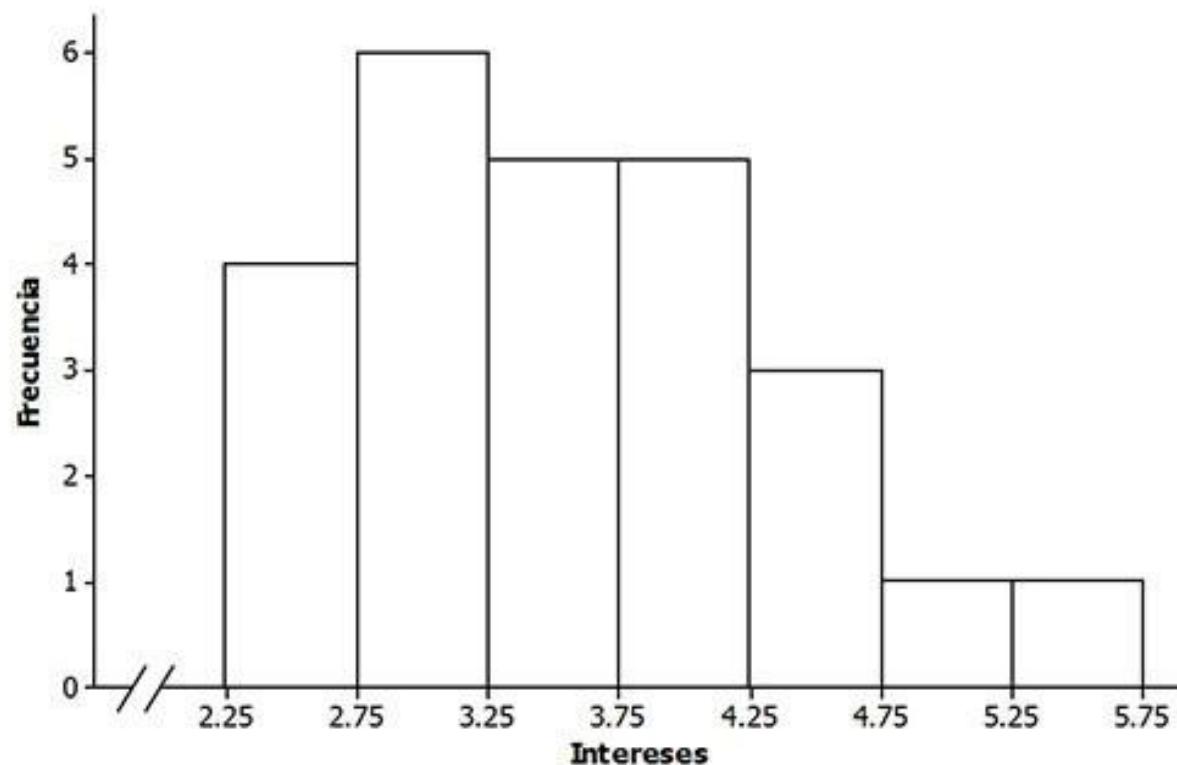
$$T_c \geq \frac{793}{6} \geq 132.1 \geq 133$$

Ahora, procedemos a llenar la siguiente tabla:

k	Lím. inf.	Lím. sup.	Frec. abs	Frec. relativa	Frec. Relativa acumulada	MC	MCxFa	Med. arit.	MC-Med. arit.	(MC-Ma)^2
1	405	537	4	0.1	0.1	471	1884	840.075	-369.075	136216.3556
2	538	670	7	0.175	0.275	604	4228	840.075	-236.075	55731.40563
3	671	803	7	0.175	0.45	737	5159	840.075	-103.075	10624.45563
4	804	936	8	0.2	0.65	870	6960	840.075	29.925	895.505625
5	937	1069	4	0.1	0.75	1003	4012	840.075	162.925	26544.55563
6	1070	1202	10	0.25	1	1136	11360	840.075	295.925	87571.60563
			40	1			33603			317583.8838

Histograma

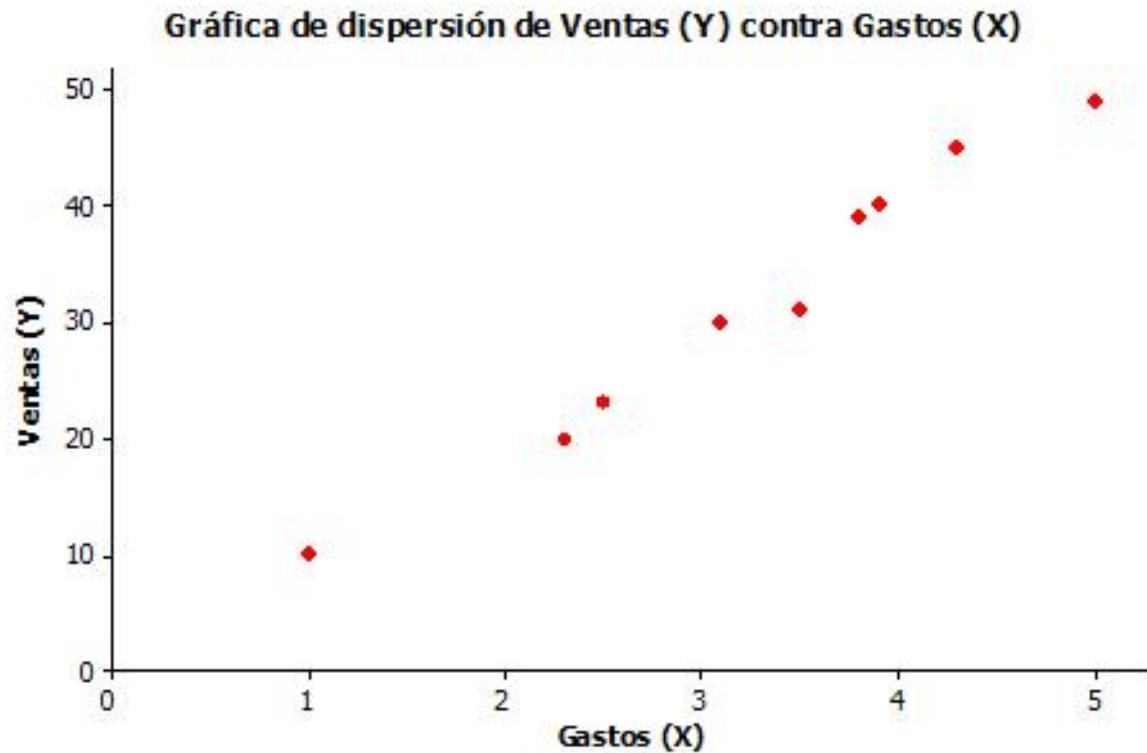
“El histograma condensa los datos, agrupando valores similares en clases. Se puede construir un histograma colocando la variable de interés en el eje horizontal, y la frecuencia, frecuencia relativa o frecuencia porcentual, en el eje vertical” (Hanke y Wichern, 2010).



Diagramas de dispersión

Los diagramas de dispersión se utilizan para visualizar la relación entre dos variables. En el siguiente gráfico de dispersión se presentan 10 pares de datos para el gasto en publicidad y las ventas. Puede apreciarse que las ventas tienden a aumentar cuando se incrementan los gastos de publicidad.

Gastos en publicidad (miles de \$)	Ventas (miles de \$)
X	Y
1.0	10
2.3	20
2.5	23
3.1	30
3.5	31
3.9	40
3.8	39
4.3	45
5.0	49



Tema extra: Coeficiente de correlación

A menudo estamos interesados en **observar y medir la relación entre 2 variables numéricas**. Por ejemplo, si queremos evaluar la relación entre:

1. Las horas que se dedican a estudiar una asignatura y la calificación obtenida en el examen correspondiente.
2. La relación entre los niveles de educación y los ingresos de un grupo de individuos.
3. Los niveles de contaminación en un lugar y los niveles educativos de una población.

Lo que **nos interesa es identificar el tipo de relación o asociación entre ambas variables, su dispersión y si existen datos que se comportan de manera atípica (también llamados outliers)**.

Este coeficiente de correlación toma valores entre -1 y 1:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n(\sum x^2) - (\sum x)^2)(n(\sum y^2) - (\sum y)^2)}}$$

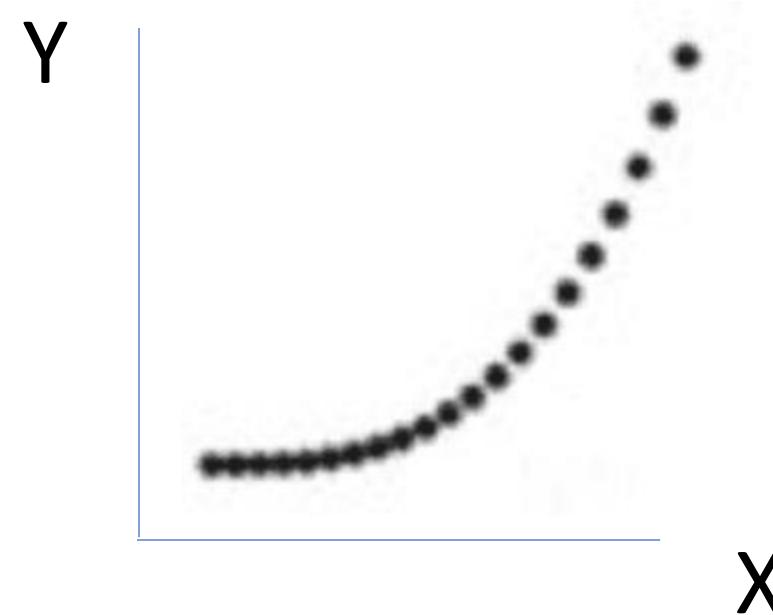
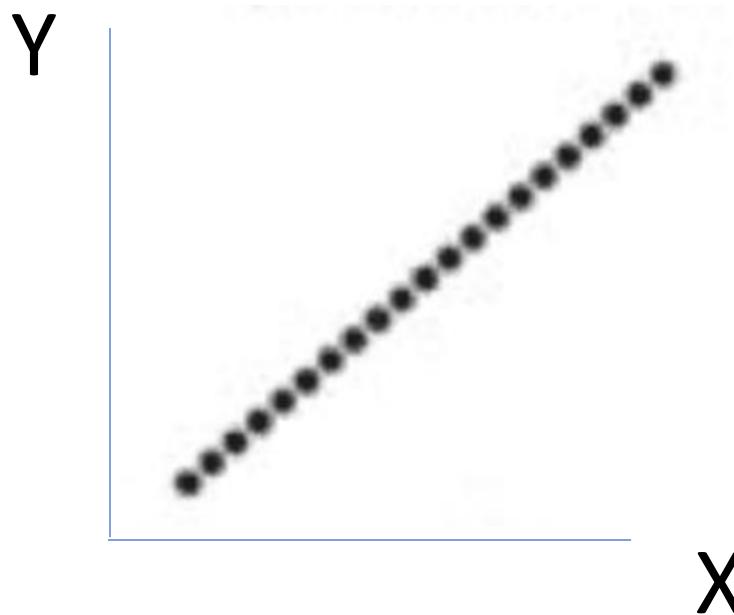
Dependiendo de su valor, nos dirá si hay una relación positiva o negativa. Existe una clasificación para medir su intensidad.

Resultado			Coeficiente de correlación lineal (positivo)
0.00	a	0.09	Nula
0.10	a	0.19	Muy débil
0.20	a	0.49	Débil
0.50	a	0.69	Moderada
0.70	a	0.84	Significativa
0.85	a	0.95	Fuerte
0.96	a	1.00	Perfecta

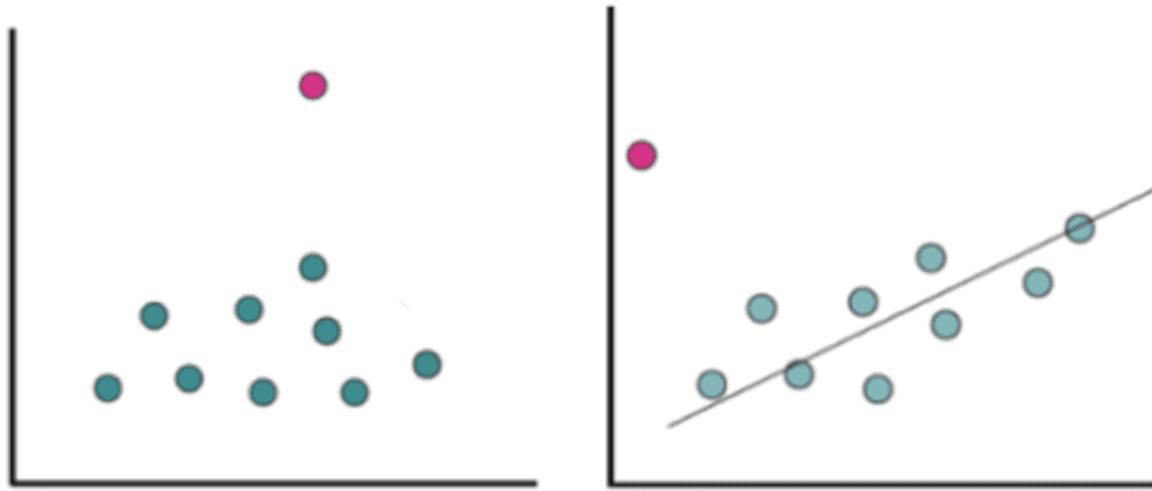
Resultado			Coeficiente de correlación lineal (negativo)
0.00	a	0.09	Nula
-0.10	a	-0.19	Muy débil
-0.20	a	-0.49	Débil
-0.50	a	-0.69	Moderada
-0.70	a	-0.84	Significativa
-0.85	a	-0.95	Fuerte
-0.96	a	-1.00	Perfecta

El diagrama de dispersión nos permite también observar características importantes de la correlación.

Forma: lineal o no lineal



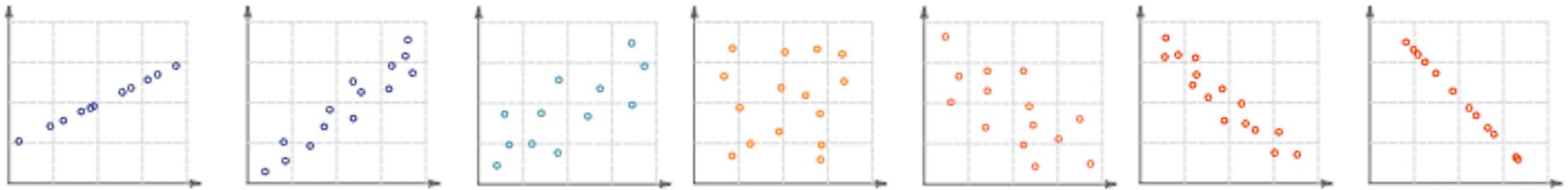
La **Presencia o no de datos atípicos (Outliers)**, puntos que no se ajustan al comportamiento del resto de la nube.



Los outliers pueden afectar los análisis de correlación y/o regresión (que veremos en temas mas adelante) debido a que la relación entre las dos variables cambia con la presencia de estos valores.

Dirección: Positiva o negativa

Fuerza: Qué tanta dispersión existe.



Si existe poca dispersión a lo largo de la tendencia diremos que la relación es fuerte, mientras que si la dispersión es grande o la nube de puntos es circular, diremos que la relación es débil.

Preguntas de seguimiento

Todas los viernes iniciaremos clase con las siguientes preguntas:

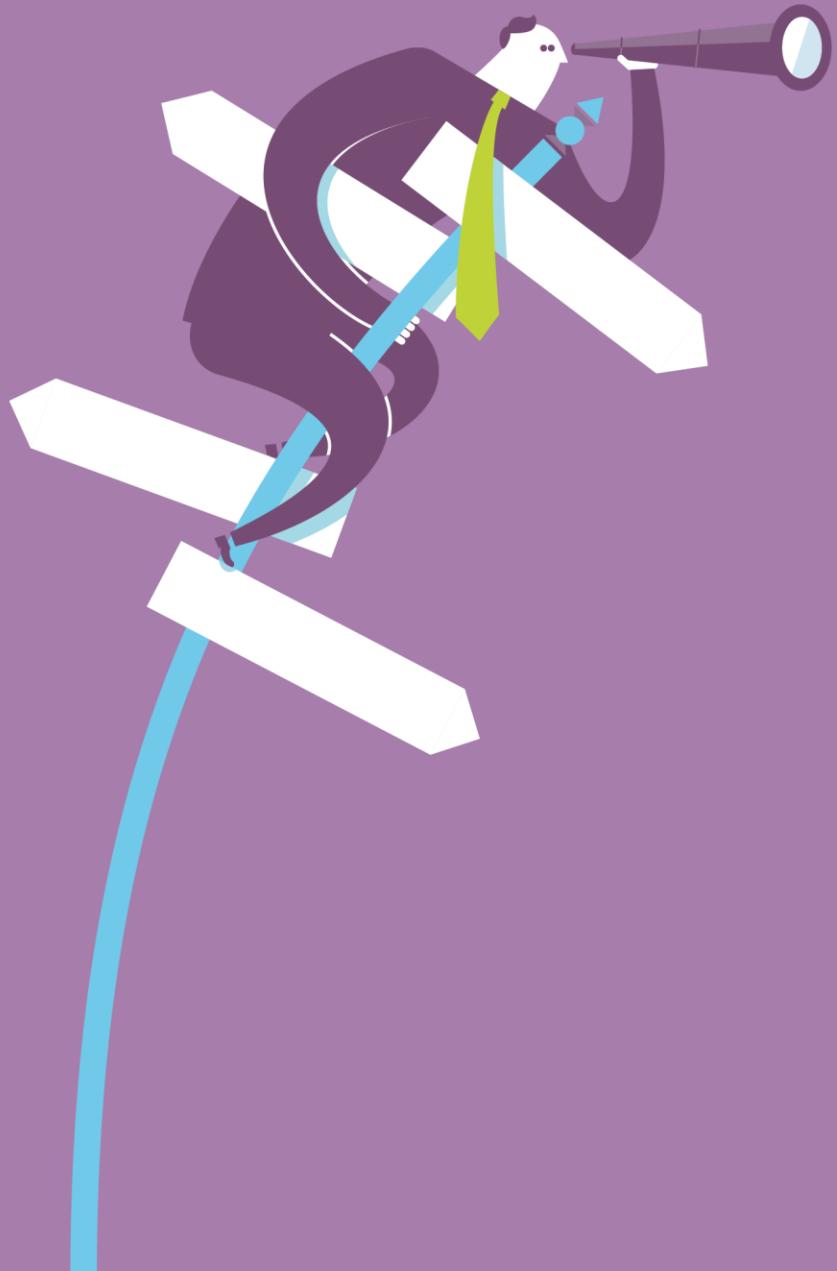
- a. ¿Qué han hecho durante la semana, referente a la materia?**
- b. ¿Que piensan hacer respecto a la materia durante la siguiente semana?**
- c. ¿En qué les podemos ayudar para que logren sus objetivos con el menor sufrimiento posible?**

Preguntas de repaso previo al examen.

1. La siguiente tabla muestra el nivel de pelea de pokemones tipo agua.

Obtenga la desviación estándar de la muestra.

Pokemon	Ataque
Squirtle	48
Poliwag	50
Dewgong	45
Kingler	55
Goldeen	67
Starmie	75
Magikarp	10
Gyarados	125



Tema 2. Teoría de la probabilidad,
conteo, independencia de eventos y
medición de incertidumbre.

Definición de probabilidad - RAE

Probabilidad.

(Del lat. *probabilitas, -ātis*).

1. Es un proceso aleatorio, razón entre el número de casos favorables y el número de casos posibles.
2. Cualidad de probable (que se verificará o sucederá)

Probabilidad y teoría de la probabilidad

Habitualmente cuando hablamos de la teoría de la probabilidad, estamos hablando de una serie de conceptos teóricos y matemáticos para entender el comportamiento de eventos que están sujetos a **incertidumbre**.

La probabilidad es un concepto fundamental del análisis cuantitativo:

Produce una “medición de la incertidumbre” de un evento:

- O sea, le pone un número a la incertidumbre: un número con el que podemos trabajar. Es un número que sigue reglas muy estrictas, aunque:
- No podemos saber con precisión qué pasará, pero sí podemos hacernos de una idea de qué tan probable es que suceda si repetimos el experimento muchas veces

Probabilidad - Introducción

Habitualmente usamos frases como:

- Es probable que el Monterrey (fútbol) pierda la final del Torneo Apertura 2022.
- Se espera que la inflación no alcance el 3 %.
- El Banco de México espera que el próximo semestre se tenga una tasa de crecimiento del 0.2%

Todas estas frases, contienen un sentido de incertidumbre sobre sucesos cuyos resultados finales no pueden predecirse exactamente.

De estos sucesos conocemos todos los resultados posibles y algunos resultados nos parece que son más probables que otros.

Tres conceptos fundamentales

- **Experimento** es una acción o grupo de acciones que producen eventos de forma aleatoria (o estocástica o impredecible)
- El **espacio muestral Ω** es el conjunto de todos los resultados posibles.
- El **evento (A)** es un subconjunto del espacio muestral.

$$p(A) = \frac{\text{número de elementos en } A}{\text{número de elementos en } \Omega}$$

Es un concepto constraintuitivo... este es un debate matemático, filosófico y metodológico.



Trading In The Zone @Tradingindzone · 23 ago.

“The central idea in The Black Swan is that: rare events cannot be estimated from empirical observation since they are rare.”
- Nassim Nicholas Taleb

Probabilidad - Conceptos básicos

En primer lugar, definimos el concepto de un experimento aleatorio y sus posibles resultados.

Definición 1. Un **experimento aleatorio** es el proceso de observar un fenómeno cuyos posibles resultados son inciertos. Se supone que se saben todos los posibles resultados del experimento de antemano y que se puede repetir el experimento en condiciones idénticas.

Ejemplo 1. Lanzar una moneda y observar si sale cara o cruz.

Ejemplo 2. Los valores, al final del año, de la inflación, la tasa de desempleo, etcétera..

Definición 2. El **espacio muestral**, que denotamos por Ω (omega), es el conjunto de todos los posibles resultados del experimento.

Ejemplo 3. Si el experimento es lanzar la moneda una vez, el espacio muestral es

$\Omega = \{C, X\}$ donde C denota cara y X denota cruz.

Si el experimento es lanzar la moneda dos veces, el espacio muestral es $\Omega = \{(C, C), (C, X), (X, C), (X, X)\}$ donde, por ejemplo, (C, X) es el suceso de que la primera tirada sea cara y la segunda cruz.

Definición 3. Los posibles resultados del experimento o componentes del espacio muestral, que denotaremos por e_i , se llaman **sucesos (eventos) elementales** y $\Omega = \{e_1, \dots, e_k\}$.

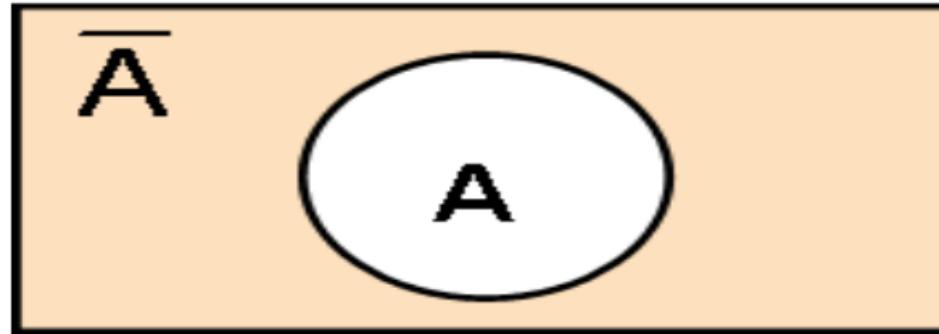
Ejemplo 4. En el caso de lanzar la moneda dos veces, los sucesos elementales son $e_1 = (C, C)$, $e_2 = (C, X)$, $e_3 = (X, C)$ y $e_4 = (X, X)$.

Definición 4. Un **suceso** es un conjunto de sucesos elementales.

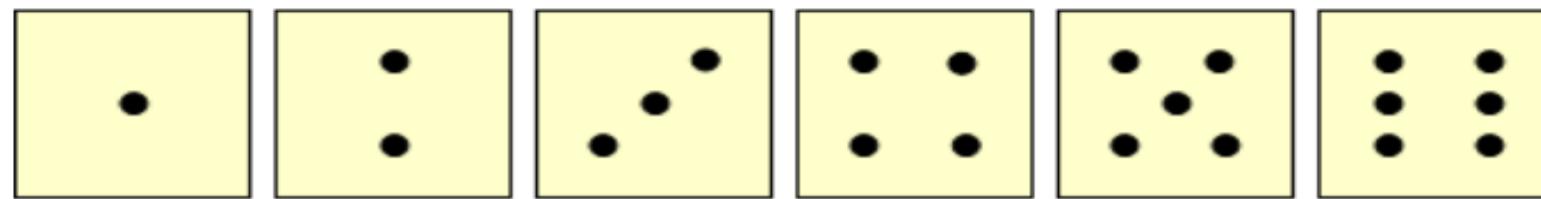
Ejemplo 5. En el caso de lanzar la moneda dos veces, el suceso A = “sale una cara y una cruz” es $A = \{(C, X)\}$.

Suceso seguro: El espacio muestral completo Ω . Siempre ocurre. **Suceso imposible:** El conjunto vacío \emptyset . Nunca ocurre.

Suceso complementario o contrario a un suceso A: suceso que ocurre cuando no lo hace A. Se compone de todos los sucesos elementales de Ω que no están en A. Se denota por A^c o por \bar{A} .



Ejemplo dados:



Espacio muestral: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Sucesos elementales: $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$

Suceso complementario o contrario a un suceso A: necesito que salga un $\{1\}$,
 suceso complementario: $\{2, 3, 4, 5, 6\}$

Suceso imposible: que te salga un $\{7\}$

Probabilidad. Intuición

La **probabilidad** de un suceso es una medida de la confianza que tenemos a priori en que el suceso ocurra cuando se realice el experimento aleatorio (a mayor probabilidad de un suceso, más cabe esperar que ocurra).

Al tirar un dado : Intuitivamente,

- La probabilidad de que salga un 1 es menor que la de que salga un numero mayor que uno
- La probabilidad de que salga un 4 es igual que la de que salga un 6.
- La probabilidad de que salga un 7 es mínima, ya que es un suceso imposible.
- La probabilidad de que salga un numero positivo es máxima, ya que es un suceso seguro.

Tres enfoques/interpretaciones

Probabilidad clásica (regla de Laplace): Considera un experimento en el que los sucesos elementales son equiprobables. Si el suceso A tiene $n(A)$ puntos muestrales, entonces se define la probabilidad de A como:

$$P(A) = \frac{\text{número de casos favorables a } A}{\text{número de casos posibles}} = \frac{n(A)}{n(\Omega)}.$$

Enfoque frecuentista: Si repitiéramos el experimento muchas veces, la frecuencia relativa con que ocurriría el suceso A convergería a su probabilidad.

$$P(A) = \text{valor límite de la frecuencia del suceso } A$$

Probabilidad subjetiva: Depende de la información de que dispongamos.

$$P(A) = \text{grado de creencia o certeza de que ocurra el suceso } A$$

Interpretación clásica de la probabilidad: En algunas situaciones, la definición del experimento asegura que todos los sucesos elementales tienen la misma probabilidad de ocurrir. En este caso, se dice que el espacio muestral es equiprobable.

Ejemplo:

Se clasifica un grupo de 100 ejecutivos en acuerdo con su peso y si tienen hipertensión.

La tabla de doble entrada muestra el número de ejecutivos en cada categoría.

	Insuficiente	Normal	Sobrepeso	Total
Hipertenso	2	8	10	20
Normal	20	45	15	80
Total	22	53	25	100

Si se elige un ejecutivo al azar, ¿Cuál es la probabilidad de que tenga hipertensión? Hay 20 ejecutivos con hipertensión, por tanto,

$$\Pr(H) = \frac{20}{100} = 0,2.$$

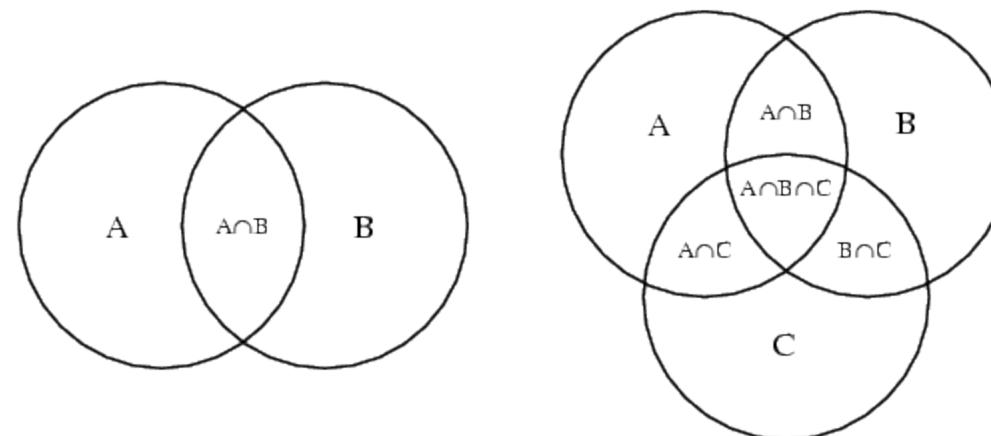
Reglas de la probabilidad

Teorema 1:

La probabilidad del conjunto $A \cup B$ (probabilidad de la unión de dos eventos A y B, se obtiene mediante la expresión):

\cup = Unión
 \cap = Intersección

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Ejemplo

Los empleados de cierta compañía han elegido a cinco de ellos para que los representen en el consejo administrativo y de personal sobre productividad.

Los perfiles de los cinco elegidos son:

- Hombre de 35 años.
- Hombre de 32 años.
- Mujer de 45 años.
- Mujer de 20 años.
- Hombre de 40 años.

Este grupo decide elegir un vocero sacando de un sombrero uno de los nombres impresos.

¿Cuál es la probabilidad de que el vocero seleccionado sea mujer o cuya edad esté por arriba de 35 años?

$$P_{(mujer \text{ o mayor de } 35 \text{ años})} = \\ P_{(mujer)} + P_{(mayor de 35 años)} - P_{(mujer \text{ y mayor de } 35 \text{ años})}$$

Hombres	3
Mujeres	2
Total	5

Mayor de 35 años	2
Menor de 35 años	3
Total	5

	Mayor de 35 años	Menor o de 35 años
Hombre	1	2
Mujer	1	1

$$P_{(mujer \text{ o mayor de } 35 \text{ años})} = \frac{2}{5} + \frac{2}{5} - \frac{1}{5} = \frac{3}{5} = 0.6$$

Probabilidad condicional

A menudo la ocurrencia de un evento depende de la ocurrencia de otros. Por ejemplo, considera las calificaciones de un estudiante en dos cursos, uno preliminar y otro avanzado. Es razonable suponer que la calificación que obtenga en el curso avanzado depende en cierta medida de la que haya obtenido en el curso preliminar.

Esta dependencia de unos eventos con respecto a otros lleva a formular el concepto de **probabilidad condicional**:

Sean A y B dos eventos en un espacio muestral S . Si $P(B) \neq 0$, se define la probabilidad condicional de un evento A dado un evento B como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Donde la línea vertical “|” debe leerse “dado que”.

Repaso de formulas

- Probabilidad de la unión de los eventos (sucesos):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Probabilidad condicionada:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Probabilidad clásica de ocurrencia de un evento:

$$P(A) = \frac{\text{Número de resultados del evento}}{\text{Número total de resultados posibles}}$$

Ejemplo de tarea:

Una caja contiene 8 bolas blancas y 4 bolas rojas. El experimento consiste en extraer 2 bolas de la caja, sin reemplazamiento.

Encuentra la probabilidad de que las 2 bolas sean blancas.

Solucion: Para calcular la probabilidad de que la primera bola extraída sea blanca utilizamos la definición clásica de la probabilidad; es decir, dividimos el número de casos favorables entre el número de casos posibles.

El número de casos favorables es 8, ya que hay 8 bolas blancas; el número de casos posibles es de 12, el total de bolas en la caja.

Entonces

$$P(\text{primera bola es blanca}) = \frac{8}{12} = \frac{2}{3}$$

Ahora hay que calcular la probabilidad que la segunda bola sea blanca, sabiendo que la primera extraída fue blanca. Dado que no hay reemplazamiento, al sacar una bola blanca nos quedan en la urna 7 bolas blancas y 4 bolas rojas, así que ahora la probabilidad de sacar otra vez bola blanca es el número de casos favorables, 7, entre el número de casos totales, 11; es decir, la probabilidad es 7/11.

Ahora la definición de la probabilidad condicional nos dice que:

$P(\text{segunda bola es blanca, sabiendo que la primera es blanca}) =$

$$\frac{P(\text{ambas son blancas})}{P(\text{primera es blanca})}$$

Así que, $P(\text{ambas son blancas}) = P(\text{Primera bola es blanca}) \times P(\text{segunda bola es blanca, sabiendo que la primera es blanca})$ y por tanto,

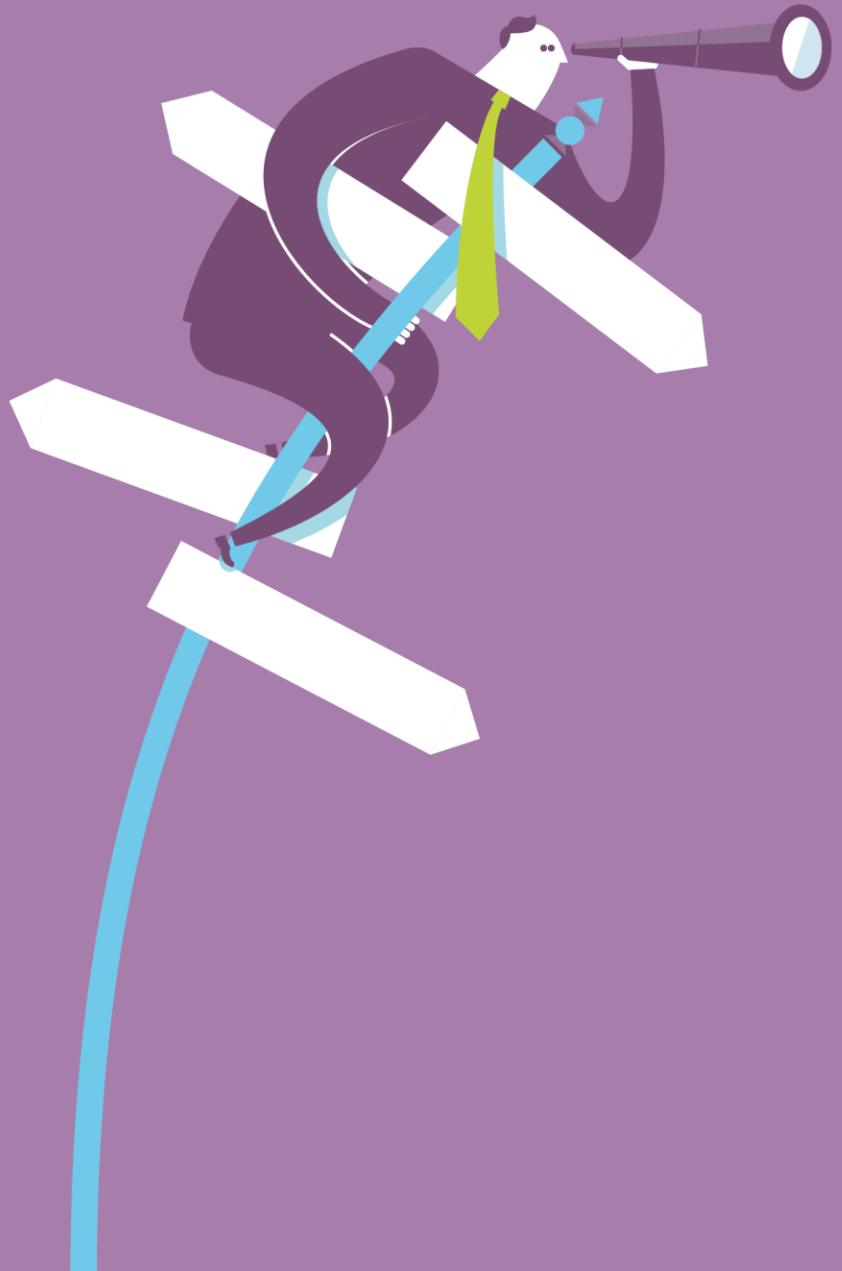
$$P(\text{ambas son blancas}) = \frac{2}{3} \times \frac{7}{11} = \frac{14}{33} = 0.424$$

Tarea para repasar previo a examen: 9 de septiembre

1. Se saca al azar una bola de una caja que contiene 6 bolas amarillas, 4 negras y 5 verdes. Encuentra la probabilidad de que la bola extraída sea amarilla.
2. En la clase de estadística para la toma de decisiones en Universidad Tecmilenio, todos practican un deporte. El 60% juega fútbol, el 10% juega basquetbol, el 10% juega Badminton y el resto montañismo. ¿Cuál es la probabilidad de que escogido un alumno de la clase?: 1. Uno juegue fútbol, 2. Uno juegue al basquetbol, 3. Uno juegue Badminton o montañismo.
3. En una estantería hay 60 novelas y 20 mangas de Dragon Ball. Una persona A elige un manga al azar de la estantería y se lo lleva. A continuación una persona B elige otro manga. ¿Cuál es la probabilidad de que lo seleccionado por una persona C sea una novela?

4. La siguiente tabla muestra el rango de estatura y sus frecuencias de los jugadores del FIFA 2022. ¿Cuál de las siguientes probabilidades representa la probabilidad de que si se eligiera un jugador de entre los 19,630 futbolistas midiera entre 171 a 180?

value	N	Raw %	Valid %	Cum. %
160-	54	0.28	0.28	0.28
161 a 170	1716	8.74	8.74	9.02
171 a 180	7617	38.80	38.80	47.82
181 a 189	8493	43.27	43.27	91.09
190 a 206	1750	8.91	8.91	100.00



Tema 3. Modelos de probabilidad, funciones y distribuciones de probabilidad.

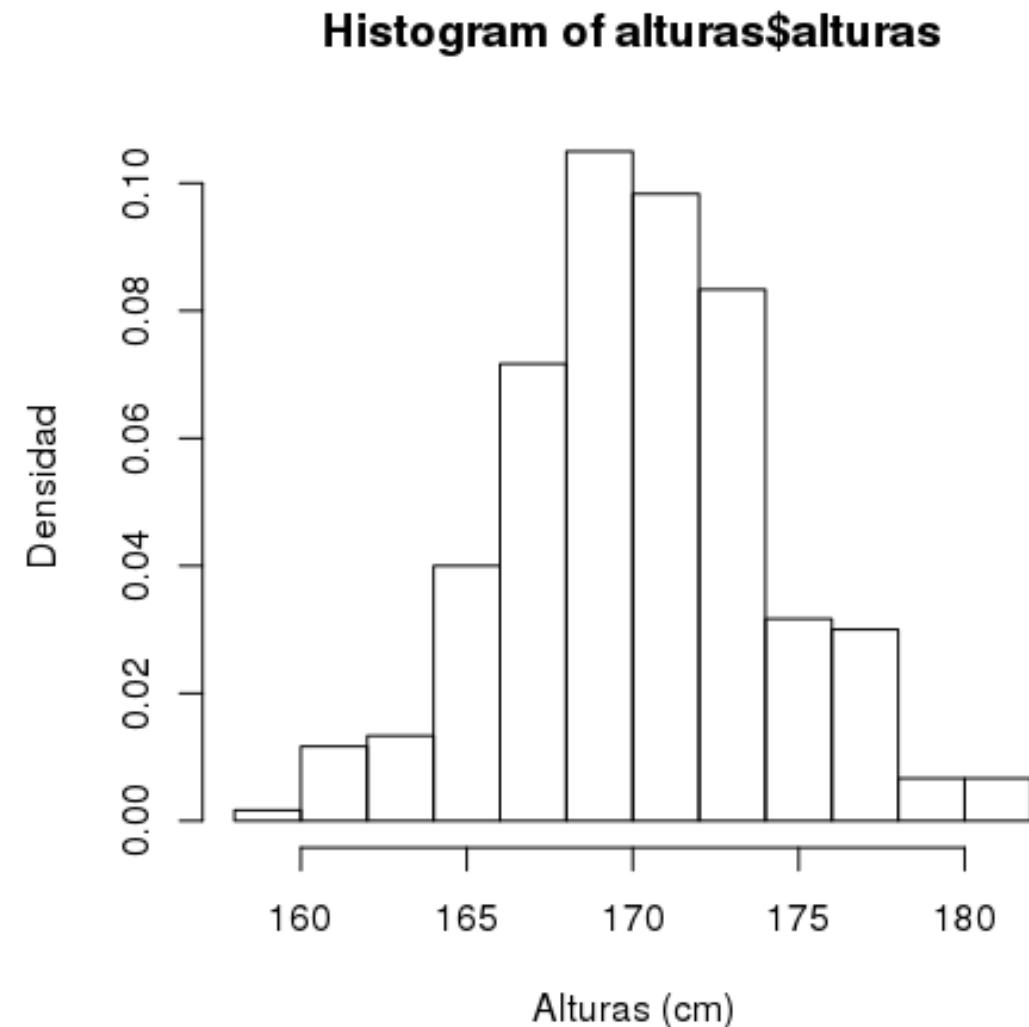
Distribución normal: una pequeña intuición.

Generalmente para hacer modelos estadísticos una de las claves para construirlos es saber la distribución de los datos. En particular se debe conocer si la distribución se ajusta a una **distribución normal**, **distribución de poisson** o **distribución bernoulli**.

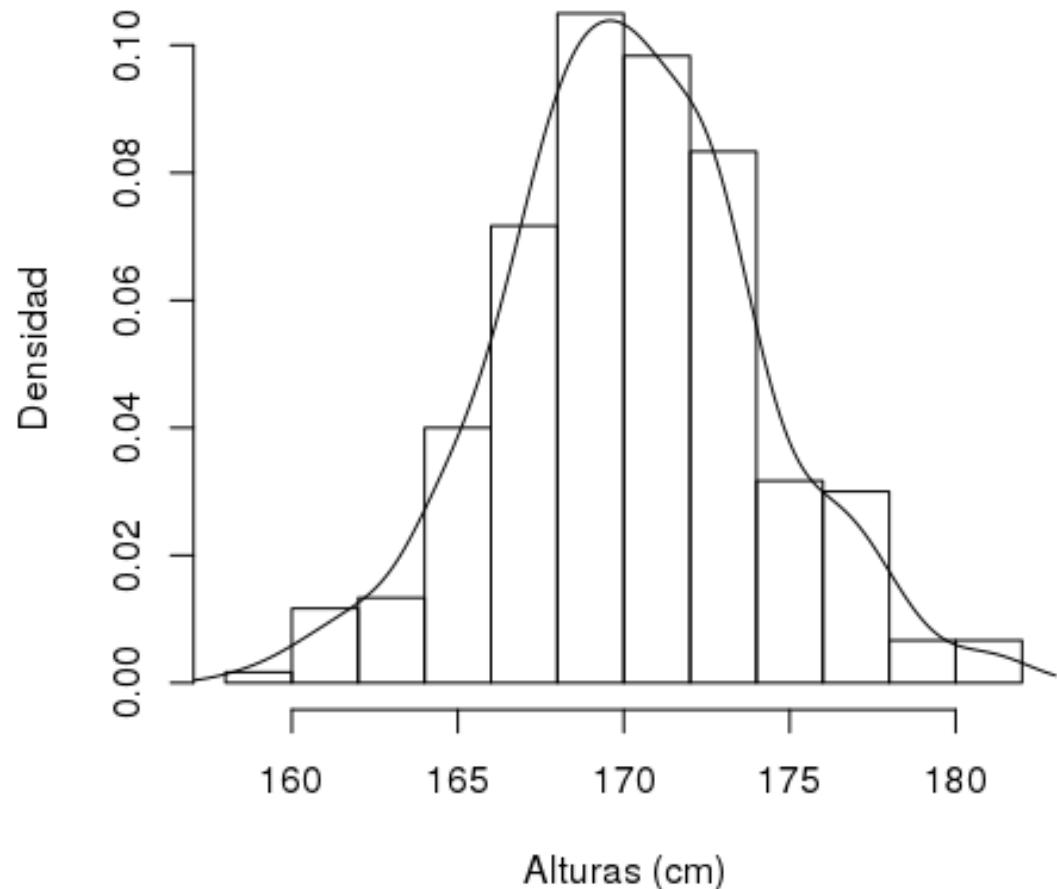
Para entender la distribución normal pensemos el siguiente ejemplo:

Pensemos en la altura de las personas. Si obtenemos la altura de las personas que estudian en el Tecmilenio veremos que existe un espectro de valores entre bajos, intermedios y altos. También podrás observar que la mayoría de personas tiene alturas intermedias mientras que la minoría de ellas tiene alturas bajas o altas.

Si dibujamos un histograma de las alturas tenemos:



Histogram of alturas\$alturas

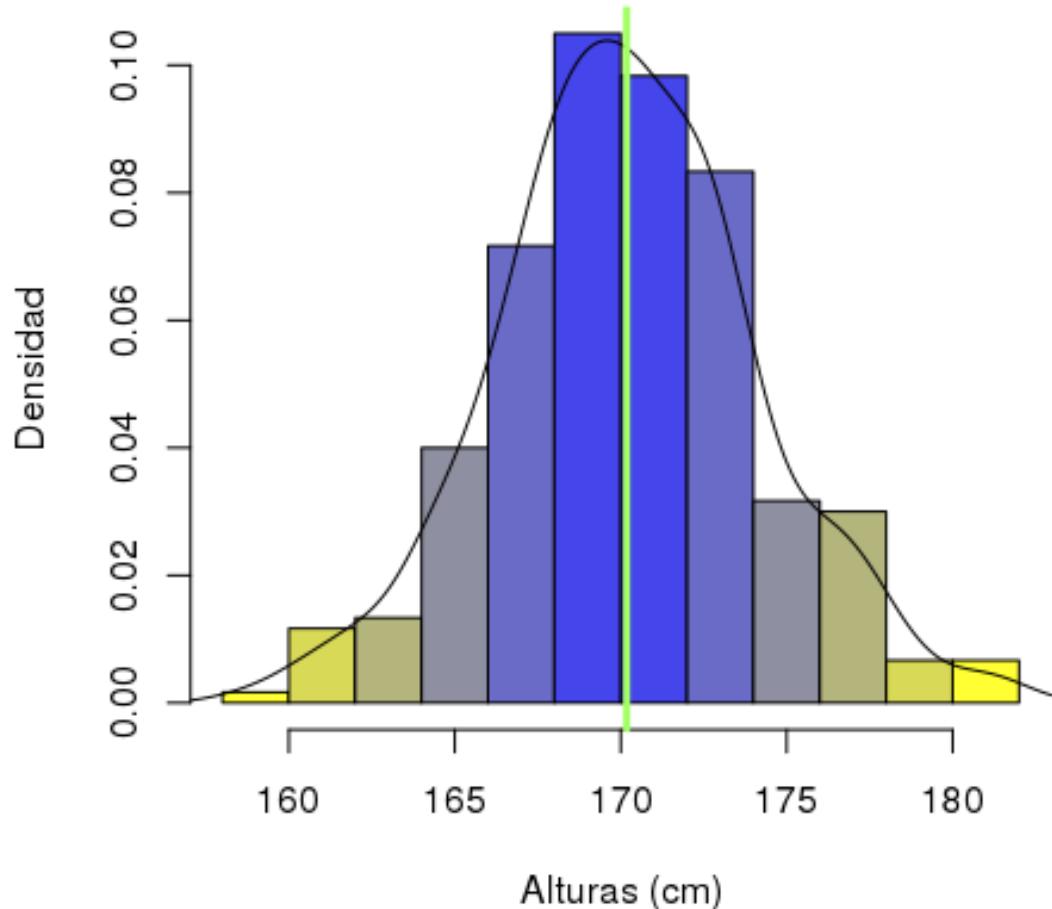


Adicionalmente si dibujamos una curva de densidad en el histograma ...

Pues resulta que la distribución de *muchas* variables numéricas tienen este tipo de *figura*, la mayoría tiende al valor central (la media) mientras que la minoría se aleja de este valor.

En muchas situaciones se asume que, a nivel poblacional, una variable numérica x tiene esta distribución, sin embargo, vale tener cierto cuidado ya que no siempre esto es así.

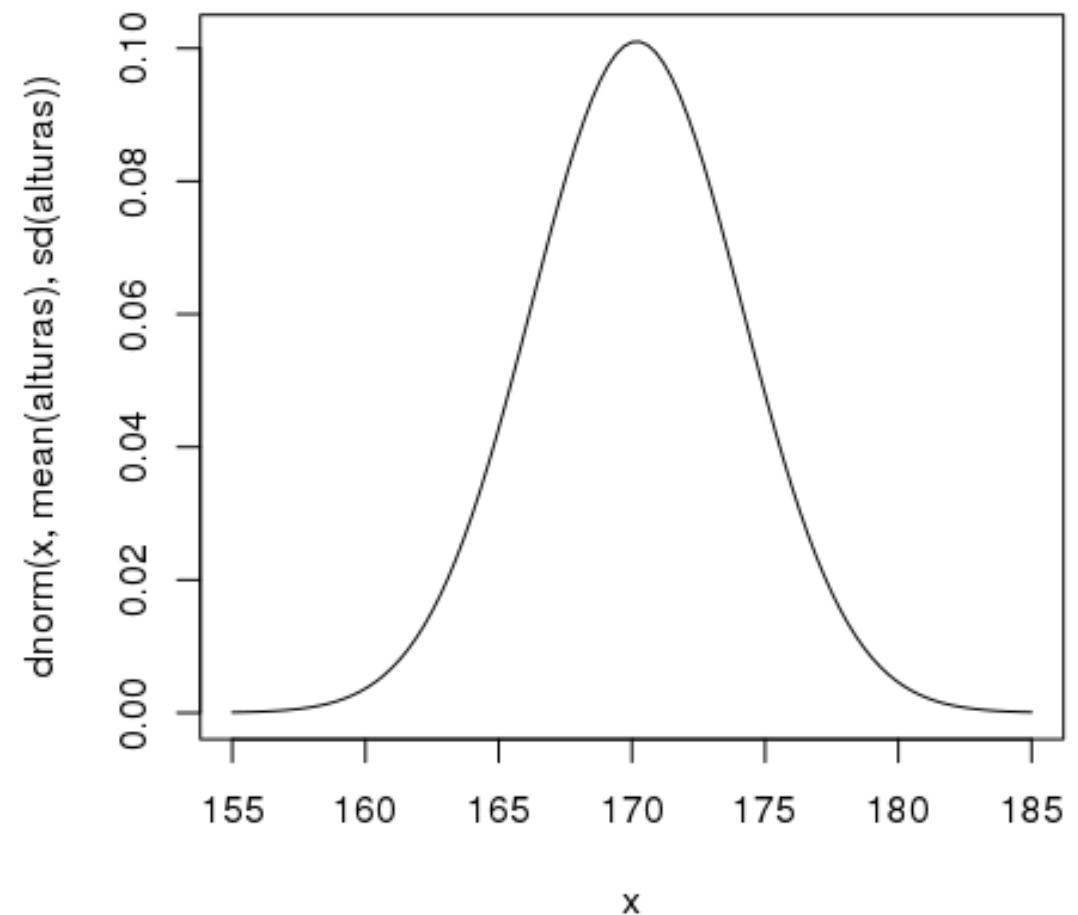
Esta distribución también es simétrica, es decir, si trazamos una línea recta vertical (la media), la parte izquierda será similar a la parte derecha dividiendo el conjunto de valores en 50%|50%.



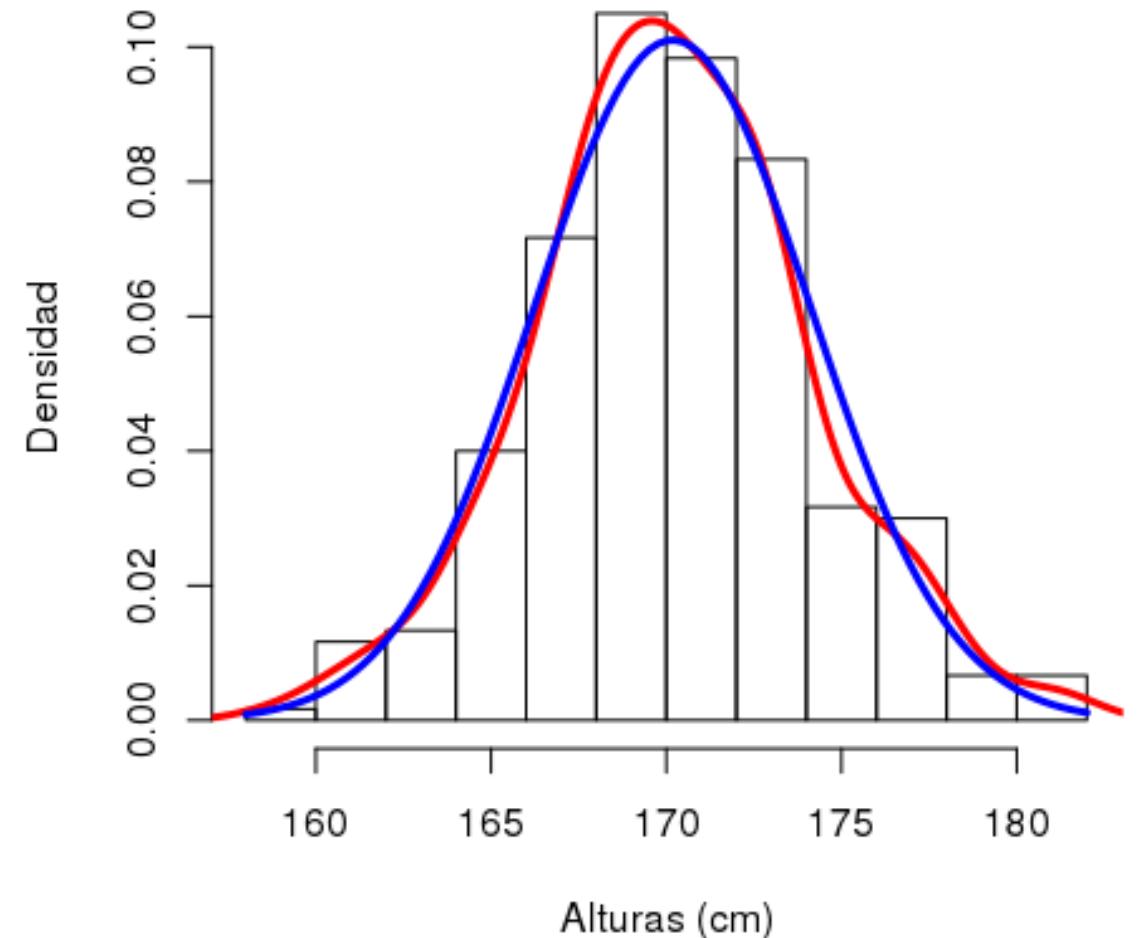
Se dice que la
distribución normal
es simétrica ...

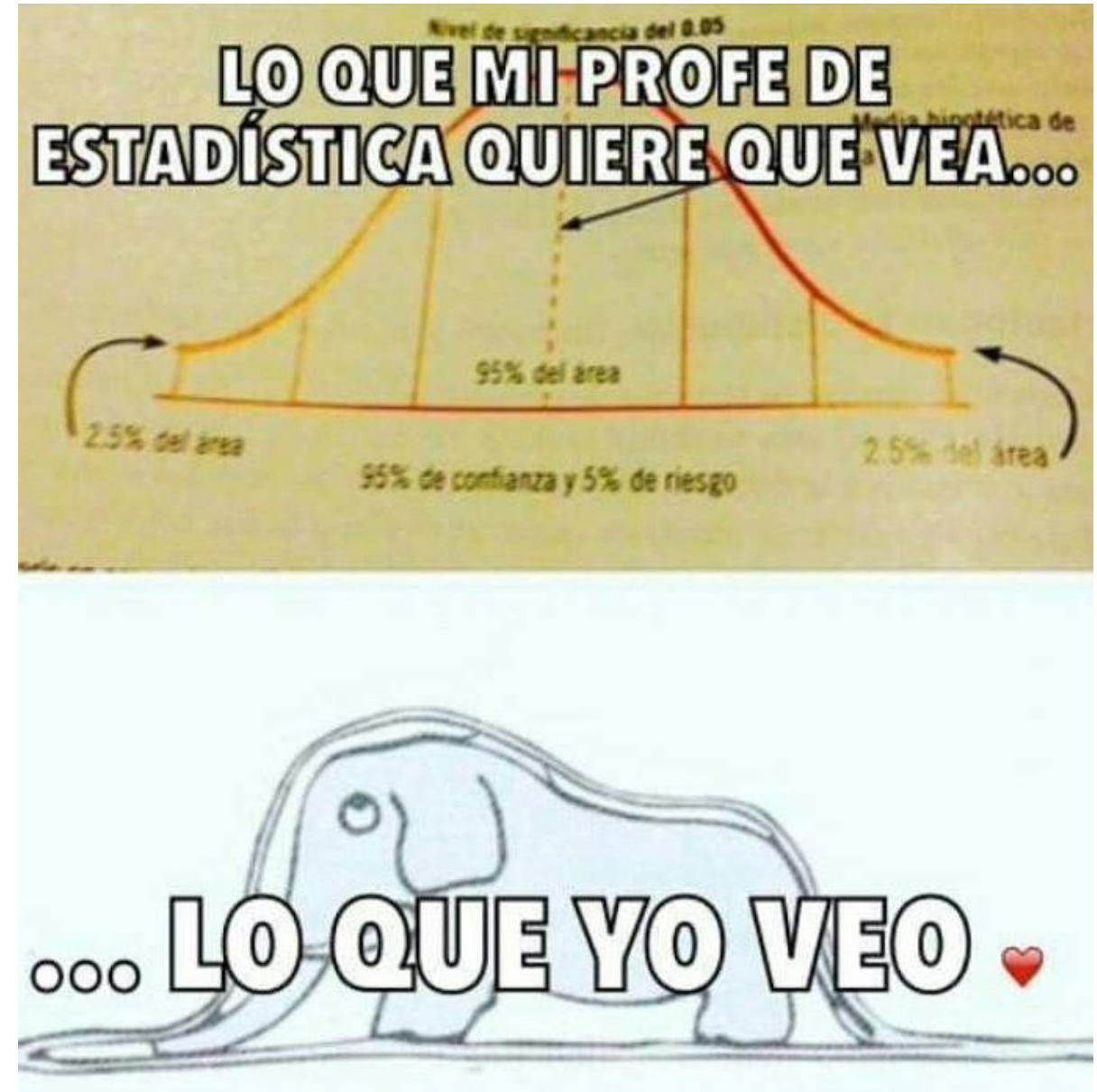
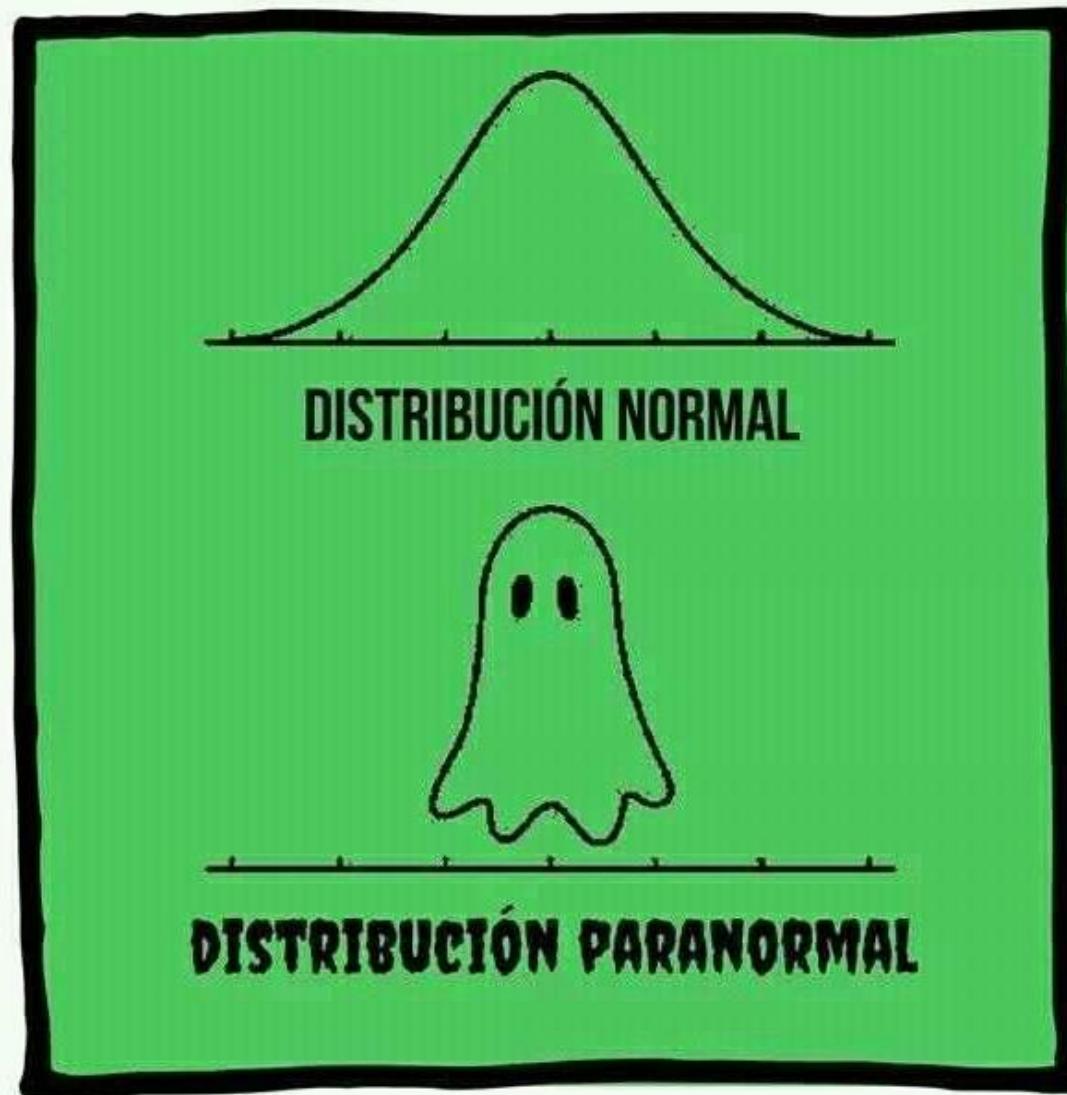
Una ventaja de la distribución normal es que puede explicarse con la media y la desviación estándar, por ejemplo, podemos generar una curva teórica a partir de los valores de la media y la desviación estándar de nuestro conjunto de alturas.

En otras palabras, podemos tomar una sub muestra poblacional y graficar su distribución.



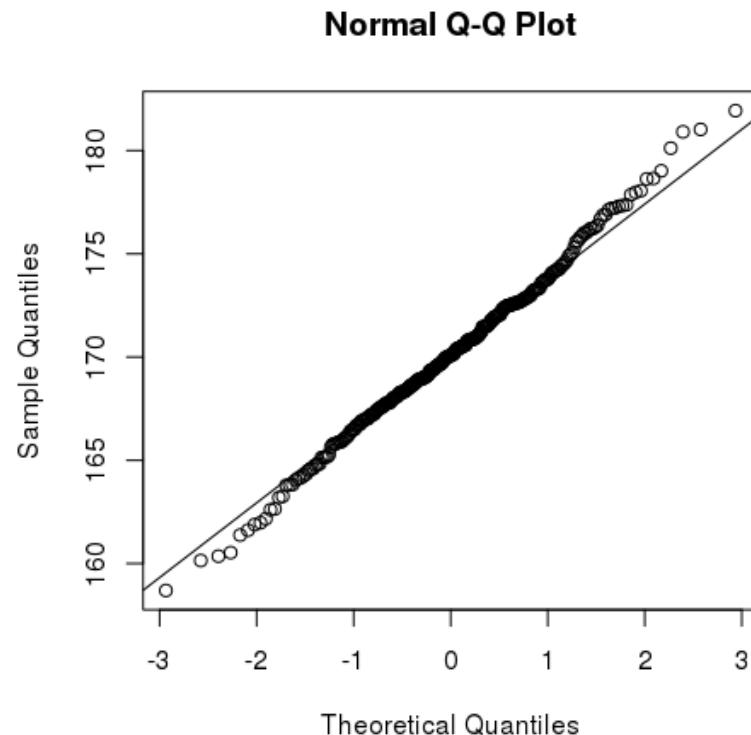
Como se puede ver la curva roja (observada) y la azul (teórica) se parecen mucho. Cuando esto ocurre decimos que la curva de nuestra muestra se ajusta a una distribución normal (posiblemente porque la variable poblacional también sigue la misma distribución).





Tema extra: *Quantile – Quantile Plot*

Otro tipo de gráficos que podemos hacer para conocer el ajuste a una distribución normal es el *Quantile-Quantile Plot* más conocido como *q-q plot*.

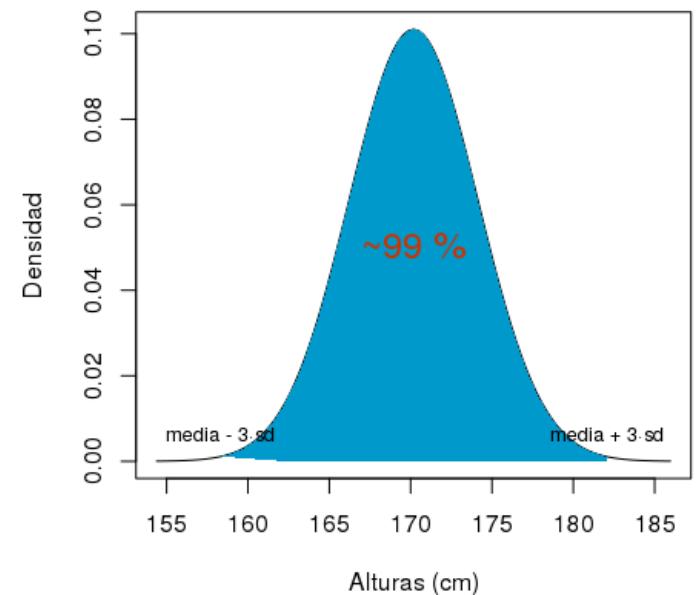
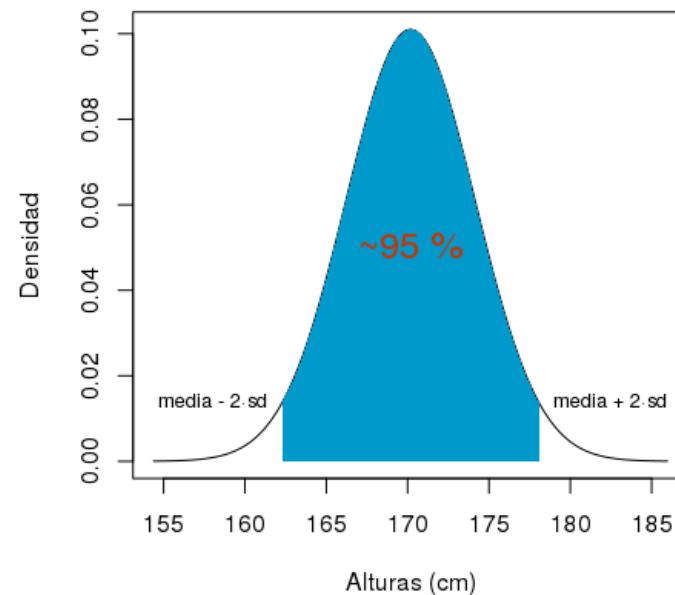
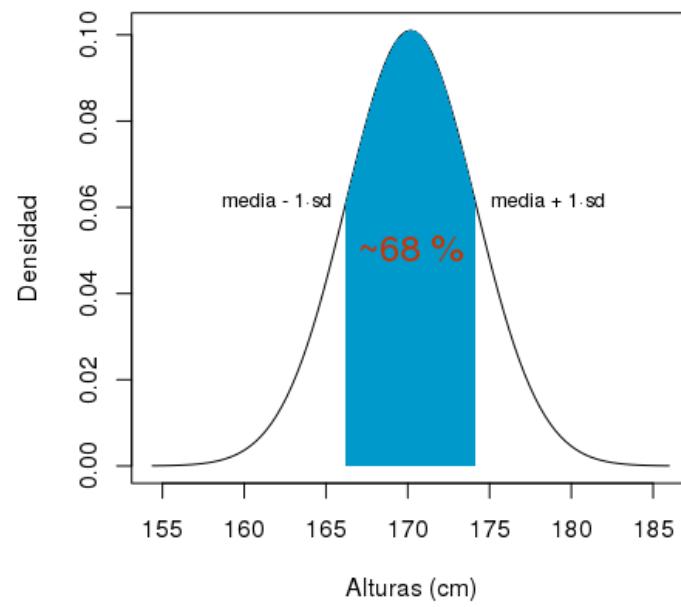


Si los puntos se ajustan a la línea diagonal, diremos que nuestros datos siguen una distribución normal.

La línea diagonal es como si fuera la curva teórica y el conjunto de puntos los valores observados.

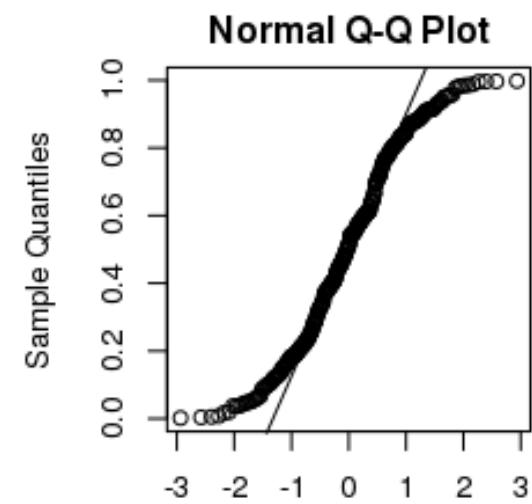
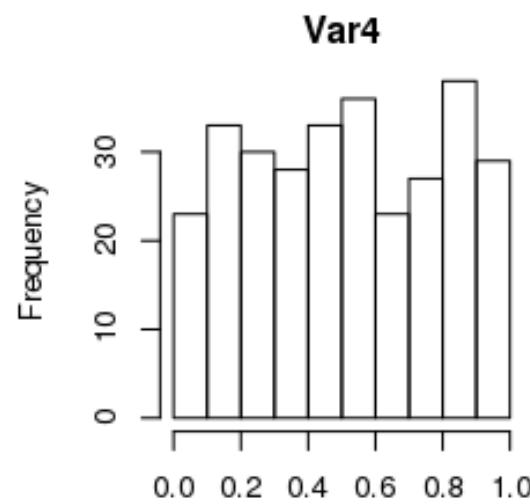
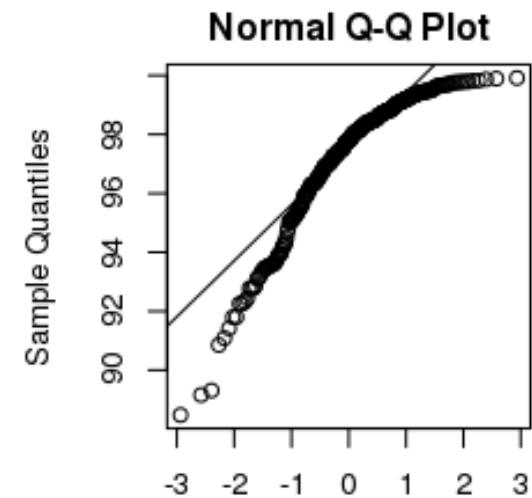
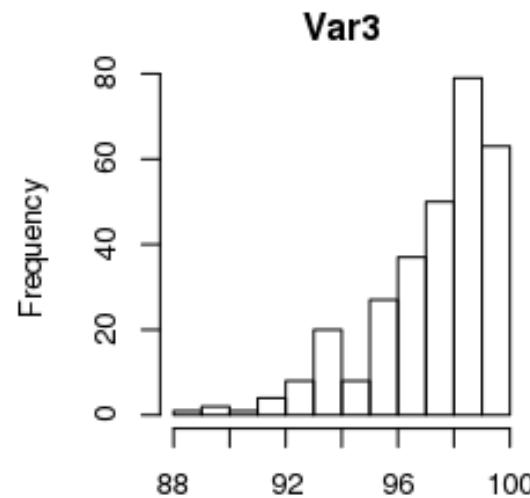
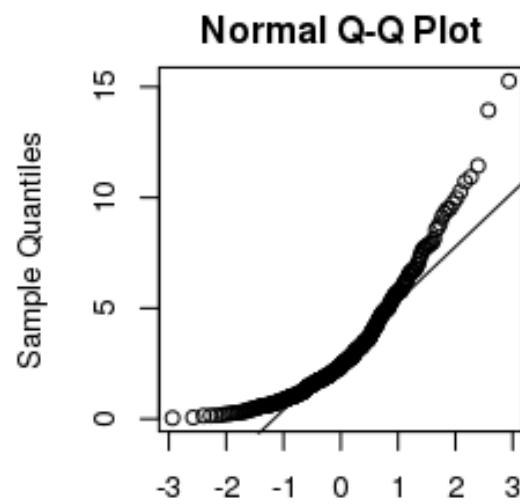
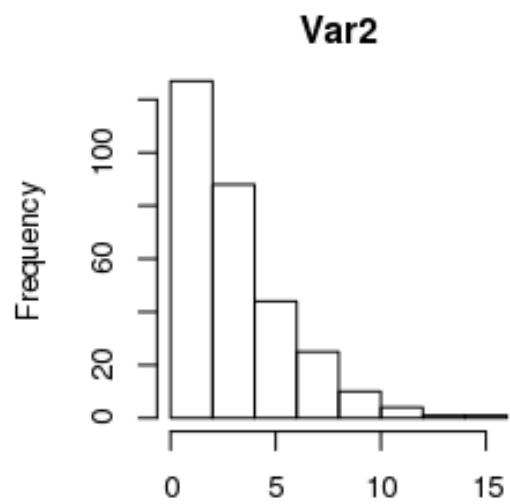
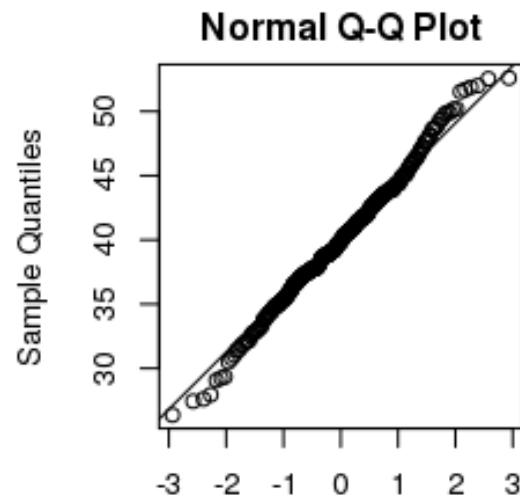
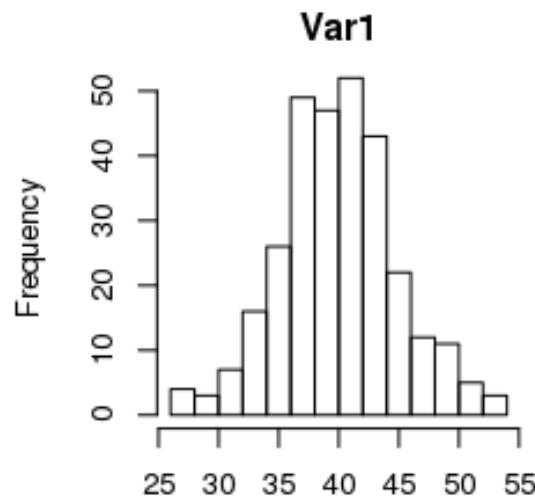
Volviendo a la curva normal, existe una característica muy interesante cuando consideramos la media y la desviación estándar.

- El 68 % (aprox.) de los valores se encuentra entre la media-1*sd y la media+1*sd
- El 95 % (aprox.) de los valores se encuentra entre la media-2*sd y la media+2*sd
- El 99 % (aprox.) de los valores se encuentra entre la media-3*sd y la media+3*sd



Esta es la razón por la que se suele acompañar el valor de la media con el de la desviación estándar y la distribución normal.

Sin embargo, ¡cuidado!, esto tiene sentido cuando se trata de una curva con distribución normal. Así, si tu conjunto de datos se aleja de la distribución normal describirlos con la media y la desviación estándar no sería adecuado.



No siempre nos encontramos con la distribución normal ...

Nivel de ingresos entre los individuos.

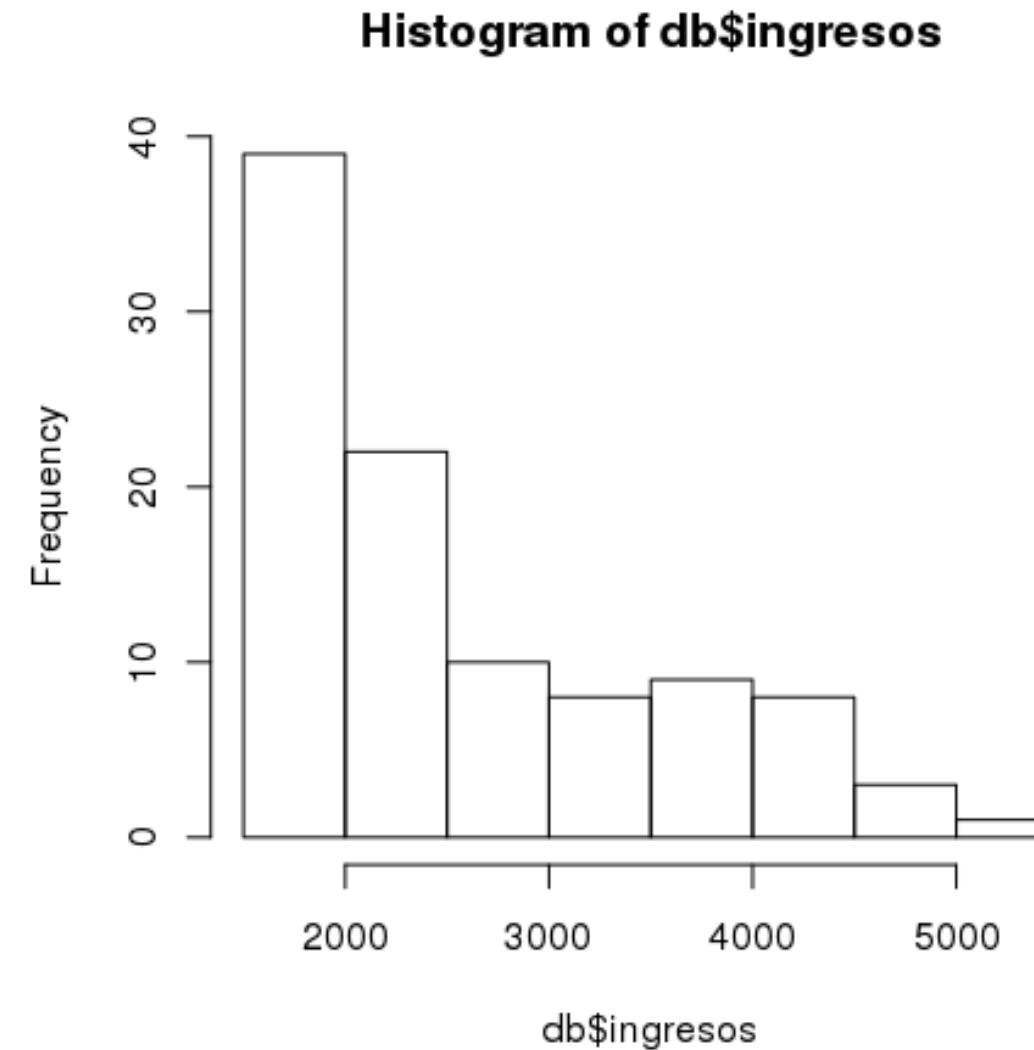
Si tenemos una variable que sigue una distribución normal, podemos usar la media y la desviación estándar para describirla pero qué hacemos si tenemos una que no lo hace, por ejemplo, los ingresos económicos.

Utilizaremos otro tipo de modelos.

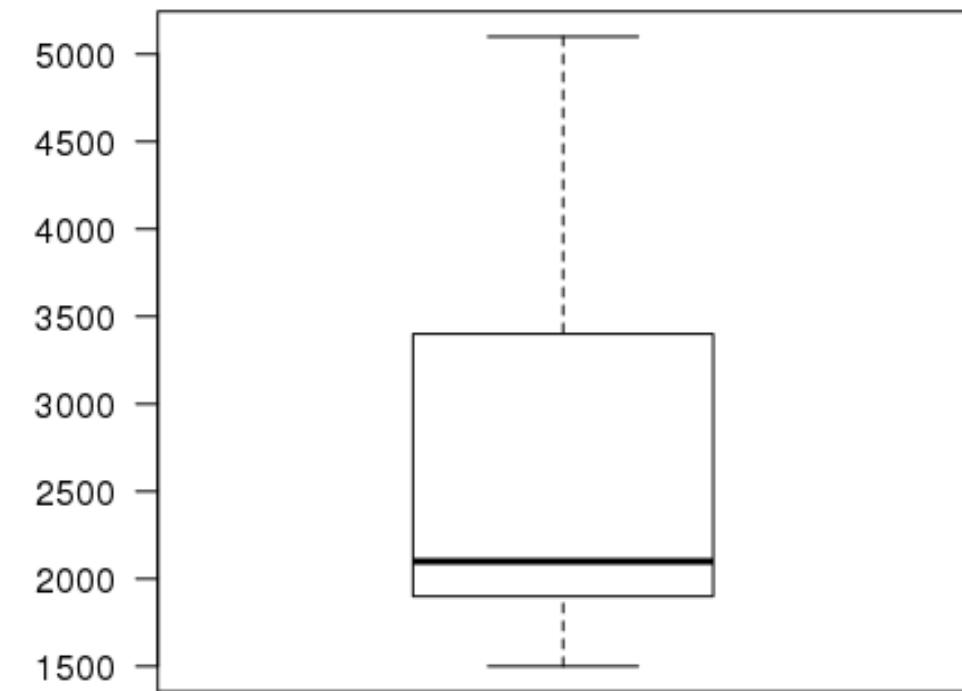
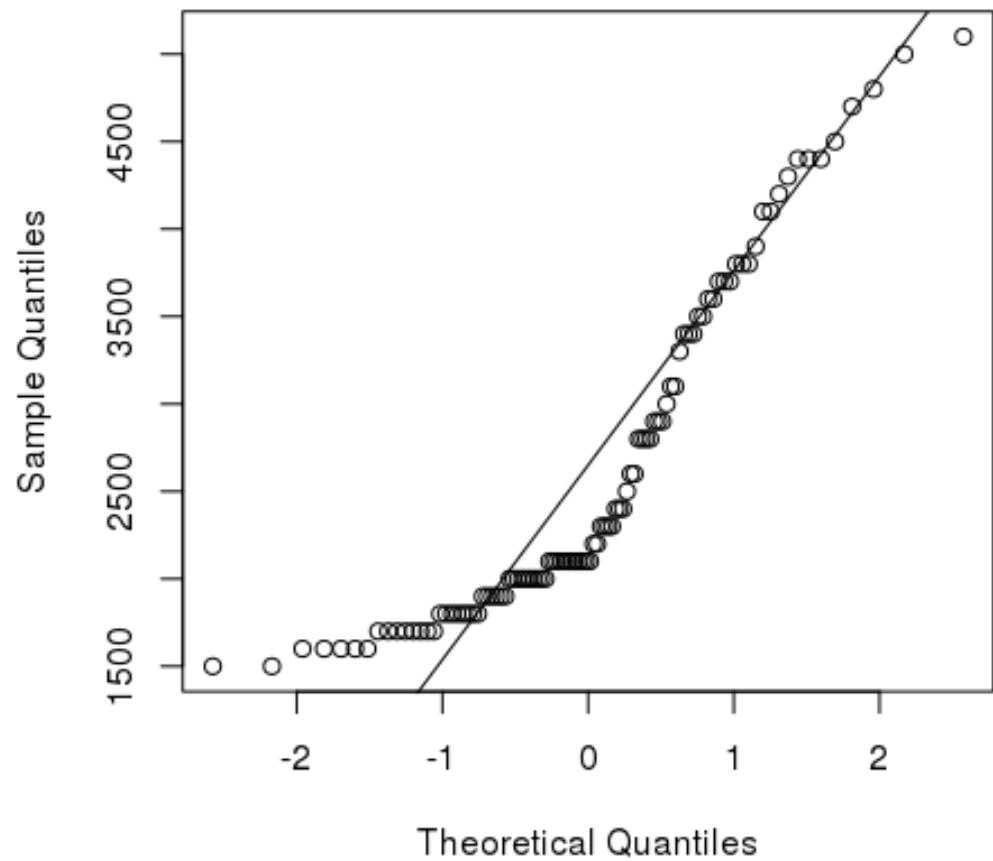
Por ejemplo, veamos los ingresos económicos de personas en una ciudad X.

Puede que la mayor parte de su población sea de ingresos bajos, y la minoría sea de ingresos altos.

**Esta distribución
evidentemente se
aleja de una
distribución
normal ...**

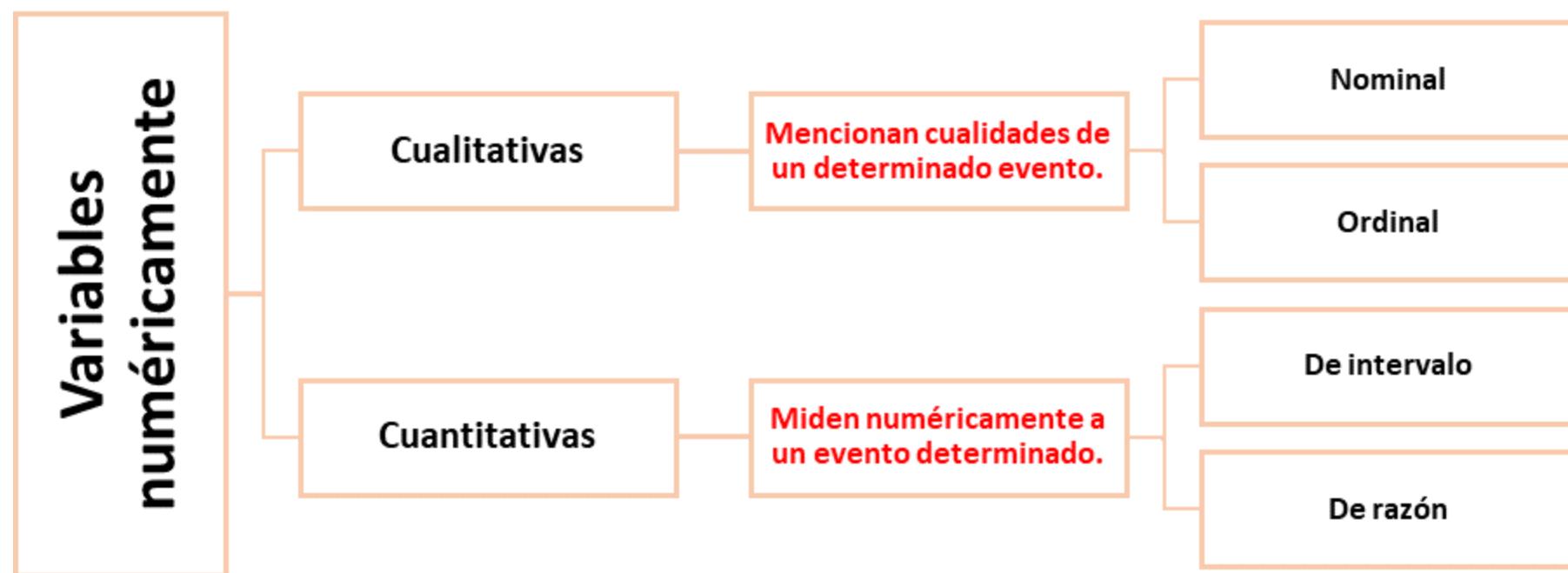


Normal Q-Q Plot



Modelos de probabilidad, funciones y distribuciones de probabilidad: variables aleatorias.

Una **variable aleatoria** (VA) es la que toma diferentes valores numéricos mediante un proceso de contar o medir, como producto de un experimento aleatorio (Rodríguez, Pierdant y Rodríguez, 2016).



Conceptos: Variable y aleatoriedad

Variable:

- Adjetivo.** 1. Que varía o puede variar.
2. Que está sujeto a cambios frecuentes o probables.

Aleatoriedad:

1. Se asocia a todo proceso cuyo resultado no es previsible.

Una secuencia numérica se dice que es aleatoria cuando no contiene patrones reconocibles.

Ejemplo: son procesos aleatorios cuando lanzamos un dado, no sabemos con exactitud el resultado, el número de delitos en una ciudad, el tiempo que puede durar una llamada.

Variable aleatoria

Definición 1. Una **variable aleatoria** es una función que asocia un valor numérico a todos los posibles resultados de un experimento aleatorio.

Ejemplo: Consideramos el experimento de lanzar un dado dos veces. Sea X = suma de las dos tiradas

¿Cuantos y qué valores puede tomar la variable X ? ¿Cuantos sucesos elementales tiene este experimento?

Variable aleatoria

Ejemplo 1. La tabla muestra los sucesos elementales asociados con cada posible valor de X .

x	Sucesos elementales				
2	(1, 1)				
3	(1, 2)	(2, 1)			
4	(1, 3)	(2, 2)	(3, 1)		
5	(1, 4)	(2, 3)	(3, 2)	(4, 1)	
6	(1, 5)	(2, 4)	(3, 3)	(4, 2)	(5, 1)
7	(1, 6)	(2, 5)	(3, 4)	(4, 3)	(5, 2)
8	(2, 6)	(3, 5)	(4, 4)	(5, 3)	(6, 2)
9	(3, 6)	(4, 5)	(5, 4)	(6, 3)	
10	(4, 6)	(5, 5)	(6, 4)		
11	(5, 6)	(6, 5)			
12	(6, 6)				

Variable aleatoria discreta

Variable aleatoria

- Las variables pueden ser **discretas**, como en el ejemplo 1, y tomar valores en un conjunto finito numerable de valores.
- También pueden ser **continuas**, por ejemplo, el tiempo que dure una llamada telefónica, y tomar valores en un intervalo de los números reales.
- El tratamiento de estos dos tipos de variables es distinto pero ambos comparten algunos de los conceptos claves: distribución, media, varianza, etc.

¿Función de probabilidad?

Lógicamente, una vez tenemos un suceso, nos preocupa saber si hay muchas o pocas posibilidades de que este suceso ocurra. Por lo tanto, sería interesante el tener alguna función que midiera el *grado de confianza* a depositar en que se verifique el suceso.

A esta función la denominaremos *función de probabilidad*.

La función de probabilidad será, pues, una aplicación entre el conjunto de resultados y el conjunto de números reales, que asignará a cada suceso la probabilidad de que se verifique.

Función de probabilidad de una v.a. discreta

Definición 2. Sea X una variable aleatoria discreta con posibles valores $\{x_1, x_2, \dots\}$. Sean $p_i = \Pr(X = x_i)$ para $i = 1, 2, \dots$ las correspondientes probabilidades. Este conjunto de probabilidades se llama **función de probabilidad** o **función de masa** de la variable.

x	$\Pr(X = x)$
2	$\frac{1}{36}$
3	$\frac{2}{36}$
4	$\frac{3}{36}$
5	$\frac{4}{36}$
6	$\frac{5}{36}$
7	$\frac{6}{36}$
8	$\frac{5}{36}$
9	$\frac{4}{36}$
10	$\frac{3}{36}$
11	$\frac{2}{36}$
12	$\frac{1}{36}$

Ejemplo 1. La función de probabilidad de la variable $X =$ suma de las dos tiradas es la siguiente:

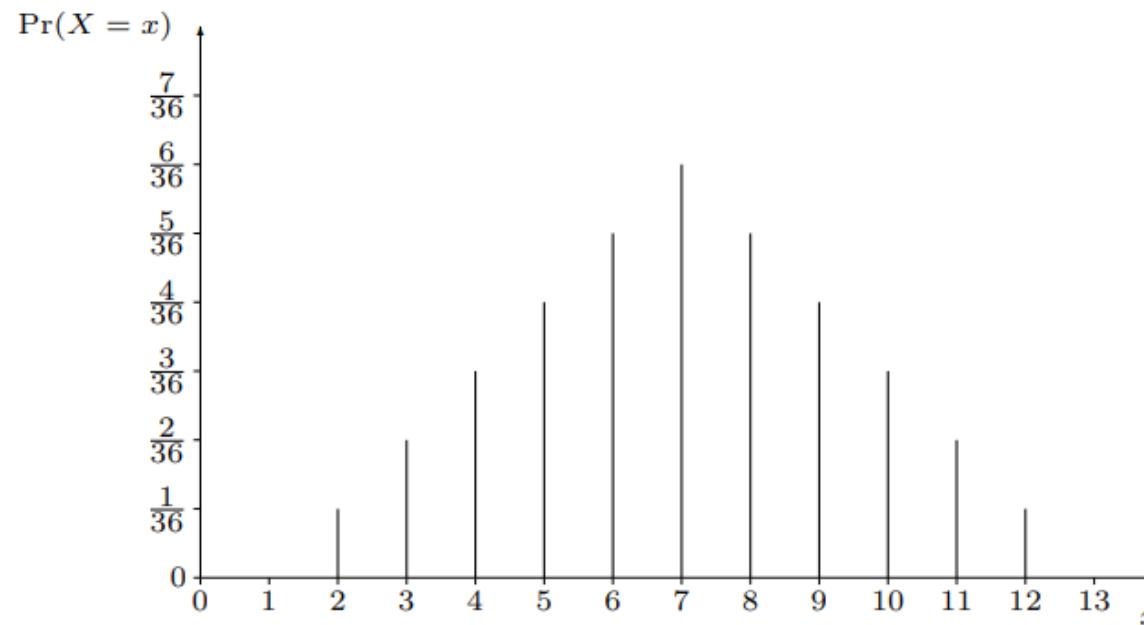
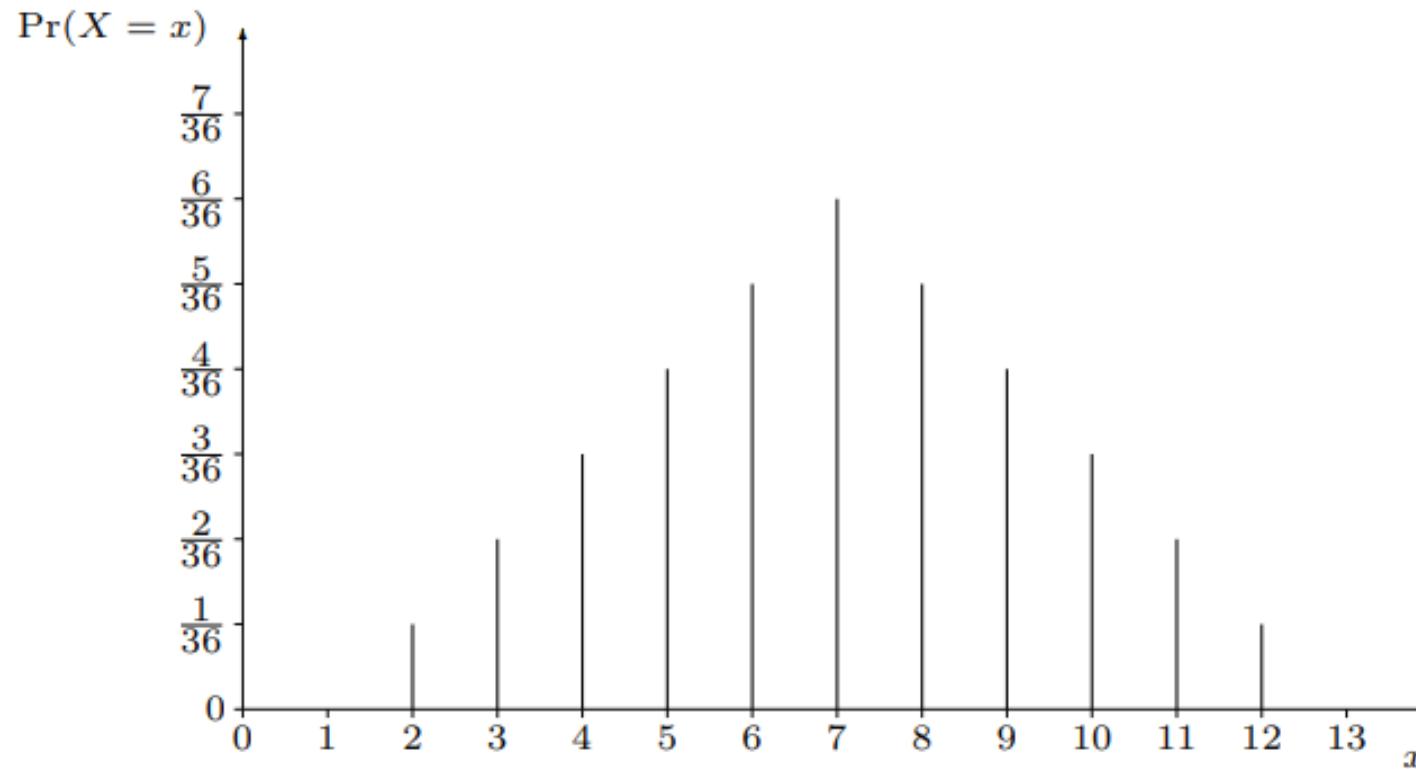


Gráfico de la función de probabilidad de X



Vemos que, en este caso, la distribución es simétrica y unimodal.

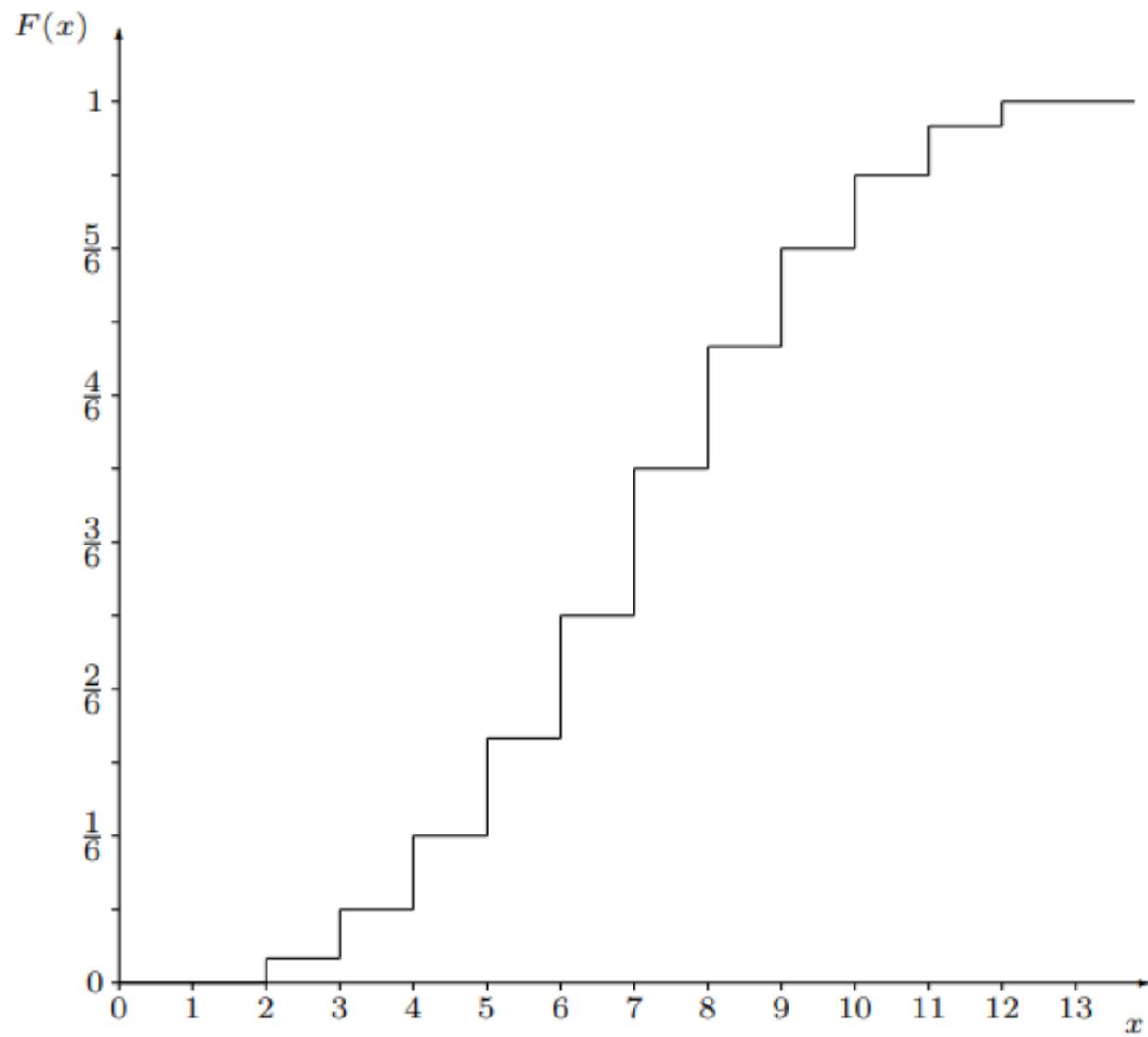
Función de distribución de una v.a. discreta

Definición 3. La función de distribución (acumulada) de una variable aleatoria X es la función $F(X) = \Pr(X \leq x)$.

Ejemplo. Volviendo al ejemplo 1, obtenemos la función de distribución de X = “la suma de los dos dados”

x	$\Pr(X = x)$	$F(x)$
2	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{2}{36}$	$\frac{3}{36}$
4	$\frac{3}{36}$	$\frac{6}{36}$
5	$\frac{4}{36}$	$\frac{10}{36}$
6	$\frac{5}{36}$	$\frac{15}{36}$
7	$\frac{6}{36}$	$\frac{21}{36}$
8	$\frac{5}{36}$	$\frac{26}{36}$
9	$\frac{4}{36}$	$\frac{30}{36}$
10	$\frac{3}{36}$	$\frac{33}{36}$
11	$\frac{2}{36}$	$\frac{35}{36}$
12	$\frac{1}{36}$	1

Grafico de la función de distribución de X



Esperanza de una v.a. discreta

Supongamos que se repite un experimento (tirar dos dados) n veces y que se observan los resultados (suma de las dos tiradas) cada vez.

Definición 4. La esperanza o media de una variable aleatoria discreta X es

Volvemos al Ejemplo 1. La media de la suma de dos dados, X , es

$$E[X] = \frac{1}{36} \times 2 + \frac{2}{36} \times 3 + \frac{3}{36} \times 4 + \dots + \frac{2}{36} \times 11 + \frac{1}{36} \times 12 = \boxed{7}.$$

Varianza y desviación típica

Recordamos que la varianza y la desviación típica muestral son medidas de la desviación de los datos en torno a la media. Podemos definir de manera semejante estas medidas para una variable.

Definición 6. La varianza de una variable X que tiene media $E[X] = \mu$ es

$$V[X] = E[(X - \mu)^2] = \sum_i \Pr(X = x_i) \times (x_i - \mu)^2.$$

La desviación típica es

$$DT[X] = \sqrt{V[X]}.$$

- **Ejemplo.** Retomamos el Ejemplo 1, sobre los dados. Tenemos

$$V[X] = \frac{1}{36} \times (2 - 7)^2 + \frac{2}{36} \times (3 - 7)^2 + \dots + \frac{1}{36} \times (12 - 7)^2 \approx 6,389$$

- La desviación típica es

$$DT[X] = \sqrt{6,389} \approx 2,53.$$

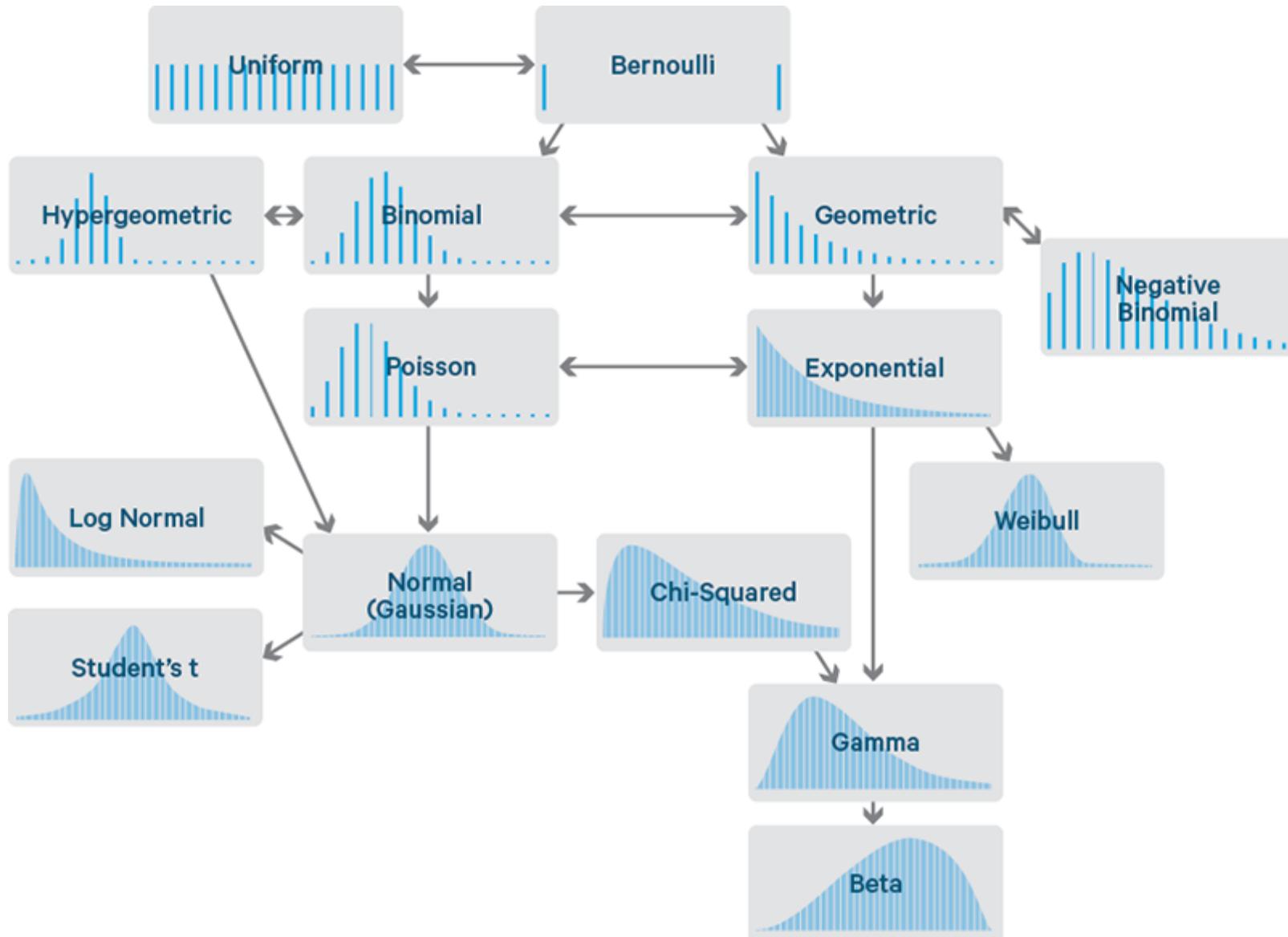
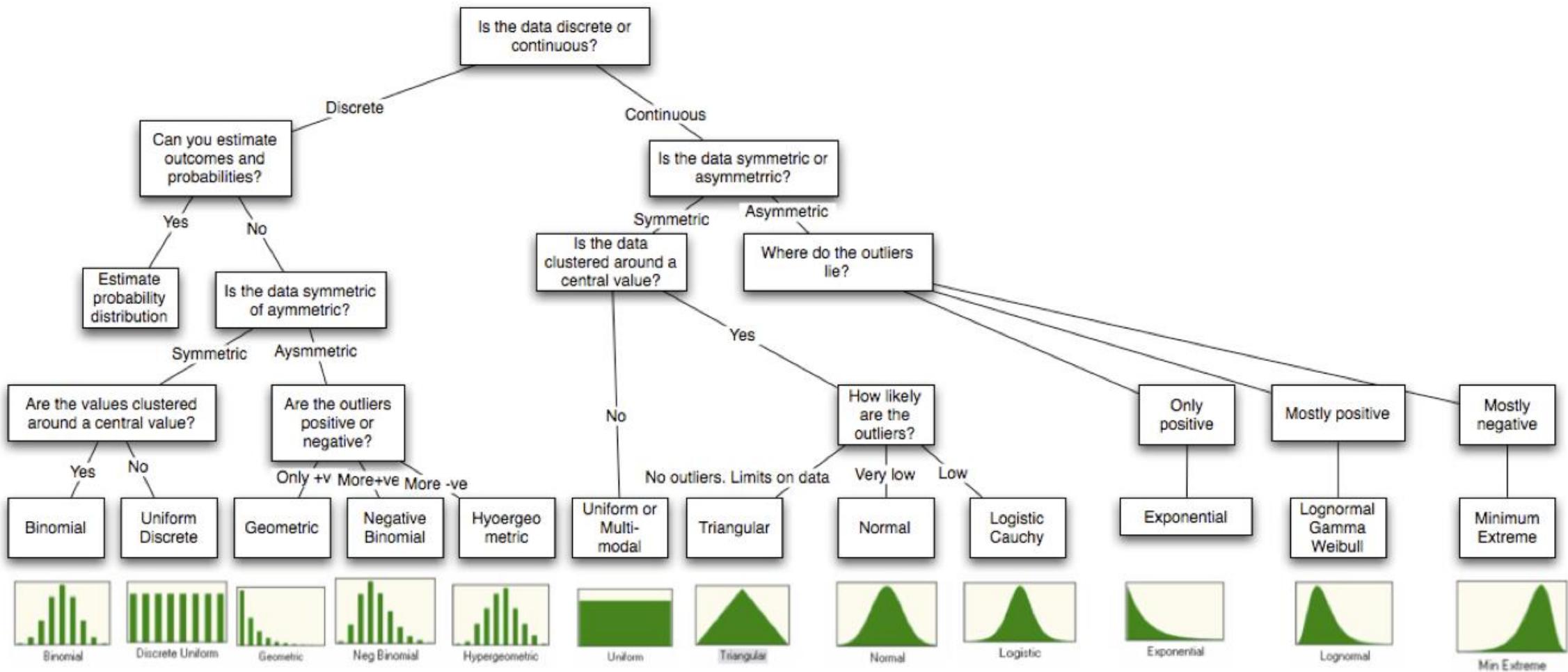


Figure 6A.15: Distributional Choices



¿Lo real y lo observado?

¿Cuál debe ser el valor que adoptará una variable?, ¿cómo aproximarnos para intentar conocer el valor “verdadero” de una variable que, así, nos permita caracterizar a una población?

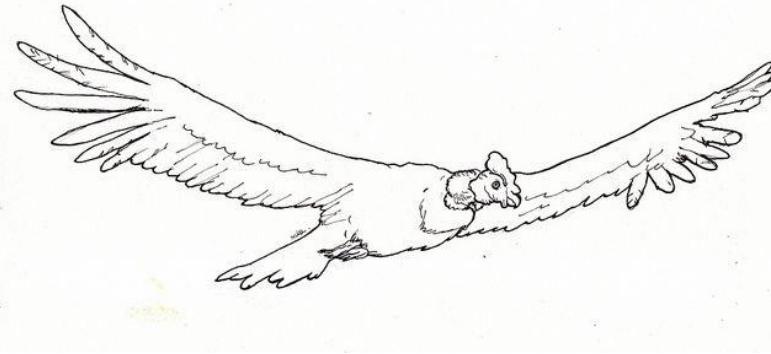
Inicialmente para conocer una característica o variable de la realidad, requerimos información completa sobre la población que analizamos. En ocasiones esto se busca lograr mediante el levantamiento de información censal o registros administrativos (por ej. censos poblacionales, reportes de resultados electorales, registros de votación legislativa, entre otros).

Sin embargo, es muy frecuente que este tipo de información sea difícil o costosa de generar y, también, corre el riesgo de desactualizarse rápidamente.

Una solución para intentar conocer dichos valores de variables es a través de recurrir a información muestral. Una muestra es un subconjunto de individuos o casos que, a su vez, forman parte de la población objetivo que se desea conocer.

$$Z = \frac{X - \mu}{\sigma}$$

La ecuación anterior, es una forma de acercarnos a la población real es mediante inferencias y probabilidades. Esta formula se llama estandarización de la distribución normal.



Ejemplo

Se sabe que la longitud de las alas de un Condor es una variable aleatoria que sigue una distribución normal, de media 120 cm. y desviación típica 8 cm.

1. Calcúlese la probabilidad de que la longitud de un ave elegida al azar sea:
 - a.- Mayor de 130 cm

Solución

$$\begin{aligned}a. \quad P(X > 130) &= P\left(Z > \frac{x-\mu}{\sigma}\right) \\P(X > 130) &= P\left(Z > \frac{130-120}{8}\right) \\&P(Z > 1.25).\end{aligned}$$

Checamos en tablas de distribución de Z

$$P(Z > 1.25) = 1 - 0.8944 = \mathbf{0.1056}$$

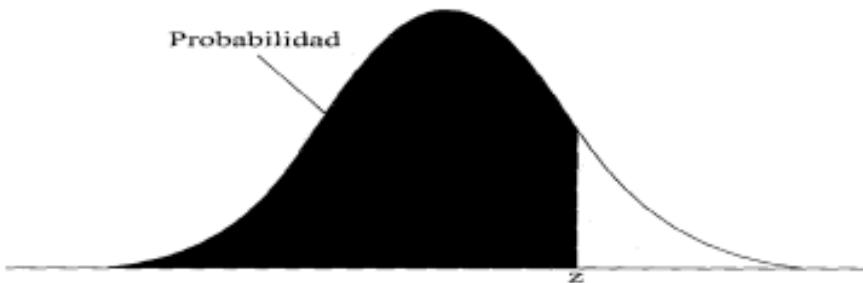
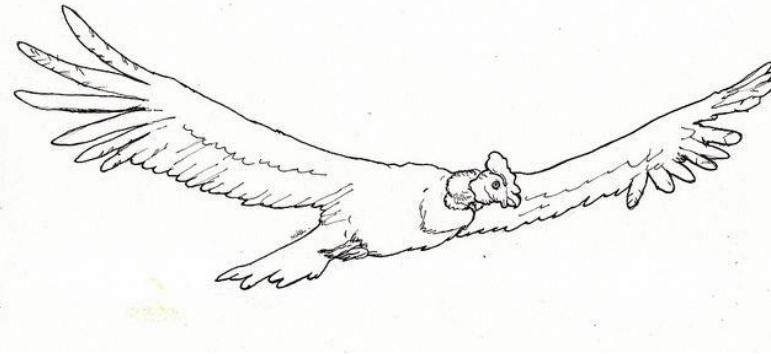


Tabla 3. (continuación) Probabilidad de que una variable normal de media cero y desviación típica uno tome un valor menor que z

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Tabla 3. Probabilidad de que una variable normal de media cero y desviación típica uno tome un valor menor que z

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
−3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
−3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
−3,2	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
−3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
−3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
−2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
−2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
−2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
−2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
−2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
−2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
−2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
−2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
−2,1	0,0179	0,0174	0,0170	0,0166	0,016	0,0158	0,0154	0,0150	0,0146	0,0143
−2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
−1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
−1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
−1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
−1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
−1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
−1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
−1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0855	0,0869	0,0853	0,0838	0,0823
−1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
−1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
−1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
−0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
−0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
−0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
−0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
−0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
−0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
−0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
−0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
−0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
−0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,721	0,4681	0,4641



Ejemplo en R

Se sabe que la longitud de las alas de un Condor es una variable aleatoria que sigue una distribución normal, de media 120 cm. y desviación típica 8 cm.

1. Calcúlese la probabilidad de que la longitud de un ave elegida al azar sea:
a.- Mayor de 130 cm

En R podemos utilizar la función `pnorm()` para poder calcularlo.

```
pnorm(130, mean = 120, sd = 8, lower.tail = TRUE)
```

0.8943502

Distribución binomial

Dentro de las distribuciones de probabilidad discretas, una de las más populares es la distribución binomial. Esta distribución de probabilidad se ocupa de experimentos en donde su resultado solo puede tomar un solo valor de dos posibles: “éxito” o “fracaso”.

Propiedades de un experimento binomial

1. El experimento consiste de una serie de n ensayos idénticos.
2. En cada ensayo hay dos resultados posibles: éxito y fracaso.
3. La probabilidad de éxito, denotada por p , no cambia de un ensayo a otro.
4. Los ensayos son independientes.

Así:

Probabilidad de éxito = p

Probabilidad de fracaso = $1 - p = q$

En este caso es de interés encontrar la probabilidad de x éxitos en n pruebas. Los diversos valores de x , junto con sus probabilidades, forman la **distribución binomial**. Estas probabilidades pueden encontrarse a partir de la siguiente expresión:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

Donde:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

p = probabilidad de éxito en cada prueba

$q = 1 - p$ probabilidad de fracaso

n = número de pruebas

Ejemplo:

Se afirma que una nueva dieta es exitosa el 85 por ciento de las veces. Si la dieta la realizan cinco personas y se puede suponer que los resultados son independientes entre sí, entonces:

- ¿Cuál es la probabilidad de que cuatro personas tengan éxito en la dieta?

Solución:

Si $x = 4$ es el número de personas que tienen éxito con la dieta, entonces $n = 5$ y la probabilidad de que tengan éxito es $p = 0.85$. Entonces, utilizando la ecuación de la ecuación binomial dada anteriormente:

$$n = 5$$

$$X = 4$$

$$p = 0.85$$

$$q = 1 - 0.85 = 0.15$$

y efectuando los cálculos en la ecuación:

$$\begin{aligned}P(X = x) &= \binom{n}{x} p^x q^{n-x} \\P(X = 4) &= \binom{5}{4} (0.85)^4 (0.15)^{5-4} \\&= \frac{5!}{4! (5-4)!} (0.85)^4 (0.15)^1 \\&= \frac{5 \times 4 \times 3 \times 2 \times 1}{(4 \times 3 \times 2 \times 1)(1)} (0.85)^4 (0.15)^1 \\&= 5(0.5220)(0.15) = 0.3915\end{aligned}$$

Ejercicio 3: entrega 28/08/2022

1. Identifica las siguientes variables como discretas o continuas:

- a. _____ Altura del agua en una presa.
- b. _____ Cantidad de dinero concedida a un demandante por un tribunal.
- c. _____ Número de personas esperando ser atendidas en la sala de emergencias.
- d. _____ Cantidad de lluvia acumulada en la presa San Juan.
- e. _____ El tiempo de reacción de un conductor de automóvil.
- f. _____ El número de accidentes aéreos observados por una torre de control.

Tips para resolver este problema en siguiente diapositiva.

a. Para resolver las pruebas de hipótesis y validarlas estadísticamente, es necesario estandarizar los datos. Emplea la fórmula estandarización de la distribución normal y las tablas de probabilidad de valores de z . **Calcula lo siguiente y represéntalo en la curva normal.**

a. $P(Z \leq 1.17) =$ _____

b. $P(0 \leq Z \leq 1.17) =$ _____

c. $P(Z \geq 1.17) =$ _____

d. $P(Z \leq -1.17) =$ _____

Recordando la distribución normal

Una **curva de densidad normal** (o de Gauss) describe la densidad de probabilidades en la distribución de valores de observaciones (muestra) de una variable aleatoria, cuando el número de observaciones es bastante grande.

La curva de distribución de valores con $\mu = 0$ y $\sigma = 1$ se conoce como la **curva normal estandarizada**, y su función de densidad de probabilidades es:

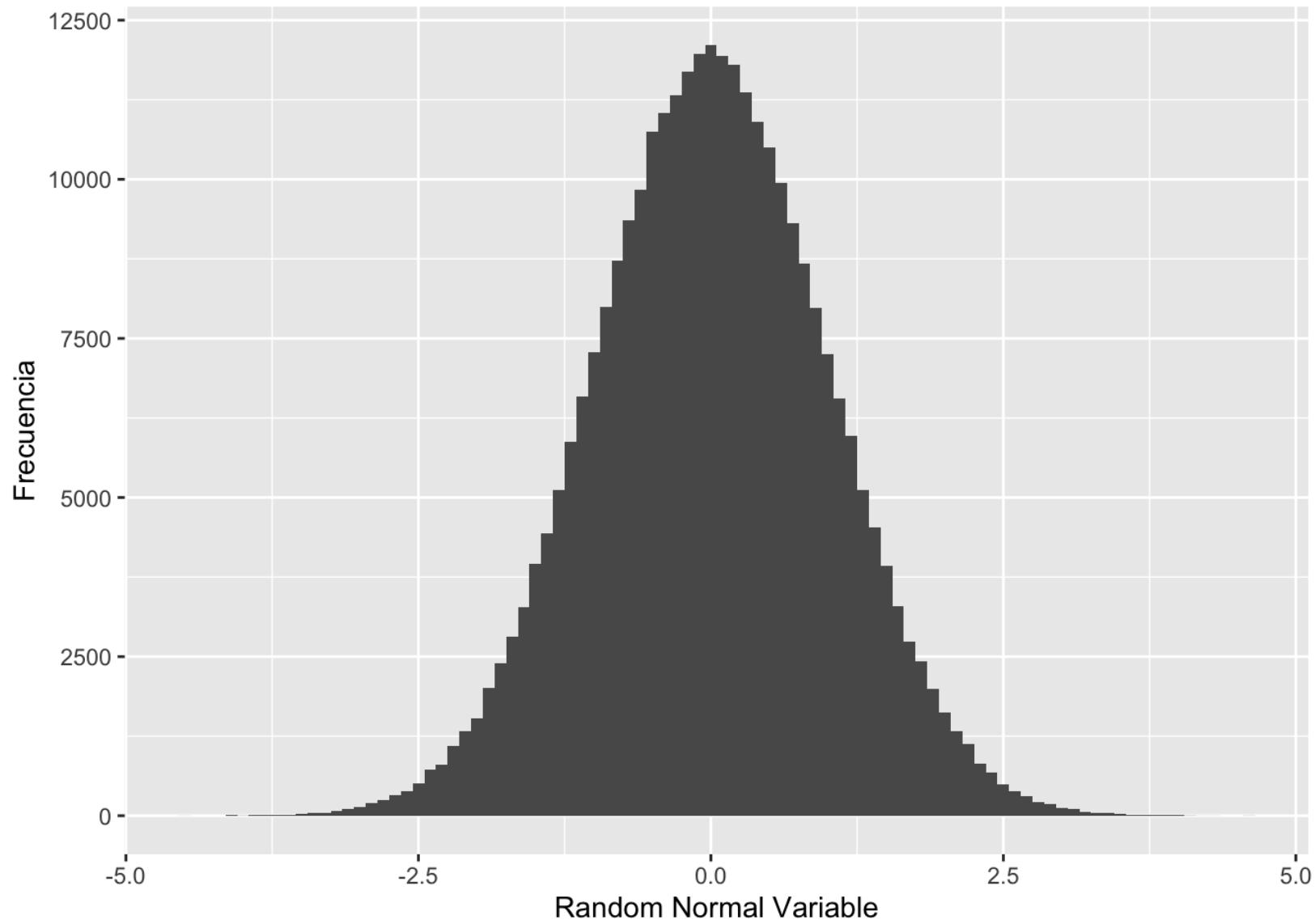
$$Y_i = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-x_i^2}{2}}$$

Para obtener valores que se basen en la distribución normal, R, dispone de cuatro funciones:

<i>dnorm()</i>	<i>pnorm()</i>
<i>qnorm()</i>	<i>rnorm()</i>

Imaginemos el siguiente problema: Sea Z una variable aleatoria normal con una media de 0 y una desviación estándar igual a 1. Determinar:

- a)** $P(Z > 2)$.
- b)** $P(-2 \leq Z \leq 2)$.
- c)** $P(0 \leq Z \leq 1.73)$.



Apartado a)

Para resolver este apartado, necesitamos resolver: $P(Z > 2)$, por lo tanto, usamos la función acumulada de distribución indicando que la probabilidad de cola es hacia la derecha:

```
> pnorm(2, mean = 0, sd = 1, lower.tail = F)  
[1] 0.02275013
```

Apartado b)

Necesitamos resolver: $P(-2 \leq z \leq 2)$, volvemos a emplear la función de densidad acumulada, esta vez, con la probabilidad de cola por defecto, hacia la izquierda:

```
> pnorm(c(2), mean = 0, sd = 1) - pnorm(c(-2), mean = 0, sd = 1)  
[1] 0.9544997
```

Apartado c)

Necesitamos resolver: $P(0 \leq z \leq 1.73)$, este ejercicio se resuelve con el mismo procedimiento que el apartado anterior, por lo tanto, volvemos a emplear la función de densidad acumulada:

```
> pnorm(c(1.73), mean = 0, sd = 1) - pnorm(c(0), mean = 0, sd = 1)
[1] 0.4581849
```

2. Para resolver las pruebas de hipótesis y validarlas estadísticamente, es necesario estandarizar los datos. Elabora una tabla de distribución normal e identifica los valores de acuerdo con el valor de z :

a. $P(Z = 1.17) = \underline{\hspace{2cm}}$

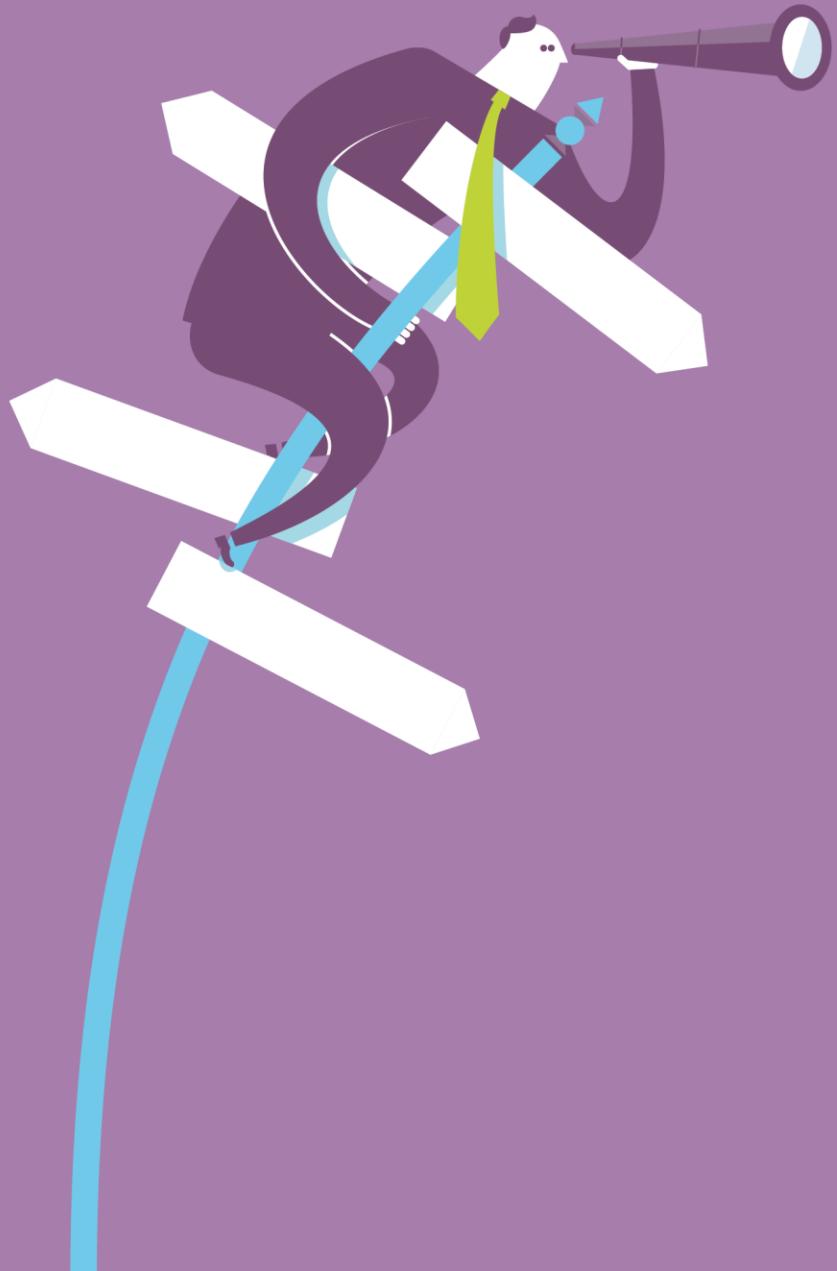
b. $P(Z = -1.46) = \underline{\hspace{2cm}}$

c. $P(Z = 1.04) = \underline{\hspace{2cm}}$

d. $P(Z = 2.66) = \underline{\hspace{2cm}}$

```
library(tigerstats)
```

```
pnormGC(1.17, region="above", mean=0,  
sd=1,graph=TRUE)
```



Tema 4. Inferencia estadística, intervalos de confianza y pruebas de hipótesis.

Definiciones clave: inferir e inferencia

Inferir

Del lat. *inferre* 'llevar a'.

1. tr. Deducir algo o sacarlo como conclusión de otra cosa.
2. Extraer un juicio o conclusión a partir de hechos, proposiciones o principios, sean generales o particulares.

La **inferencia estadística** es el conjunto de métodos que permiten inducir, a través de una muestra, el comportamiento de una determinada población.

Inferencia estadística

Objetivo: Obtener información de interés sobre una **población** a partir de una **muestra** de la misma.

Ejemplo: La ENIGH (Encuesta Nacional de Ingreso y Gasto de los Hogares) busca obtener información sobre en qué gastan las familias mexicanas. Ya que preguntar casa por casa es muy costoso, lo que hace es tomar una muestra representativa de los hogares.

La **inferencia estadística** se ocupa principalmente de **inferir los valores desconocidos** de los parámetros poblacionales de una v.a. a partir de la información proporcionada por una muestra.

Muestreo

Una **muestra** es un subconjunto finito de una población. El número de individuos que la componen se denomina **tamaño muestral**. Entre más representativa sea la muestra, mayor probabilidad de obtener un dato más preciso del resultado.

Entre los motivos para considerar una muestra en lugar de toda la población destacamos los siguientes:

- Puede ser inviable económicamente estudiar a toda la población.
- El estudio de la población puede llevar un tiempo excesivo.
- Además, sus características podrían cambiar con el tiempo.

Ejemplo de muestreo e inferencia

Un medico atendió cierto día a 24 pacientes en su consulta. Para ese día, tenemos una población finita de $N = 24$ individuos, siendo la variable de interés X = “Duración (min.) de una consulta”.

Los valores de X en la población fueron:

5.1	1	0.9	3.8	10.2	2.1	9.5	4.5
1	2.2	1.5	4.8	1.6	8.8	4.3	1
9	5.1	0.2	2.3	0.8	7.8	7.7	1.5

Por tanto, la **media** de X es $E[X] = 4$. Nota: en este caso el término E hace referencia a la esperanza. Recordar el ejemplo de los dados que su esperanza era 7.

Ejemplo de muestreo e inferencia

Seleccionamos una muestra aleatoria de tamaño 7 dada por:

3.8	9.5	4.8	1.6	0.2	0.8	1.5
-----	-----	-----	-----	-----	-----	-----

La **media muestral** de estos valores que denotaremos por

$$\bar{X} = 3.17.$$

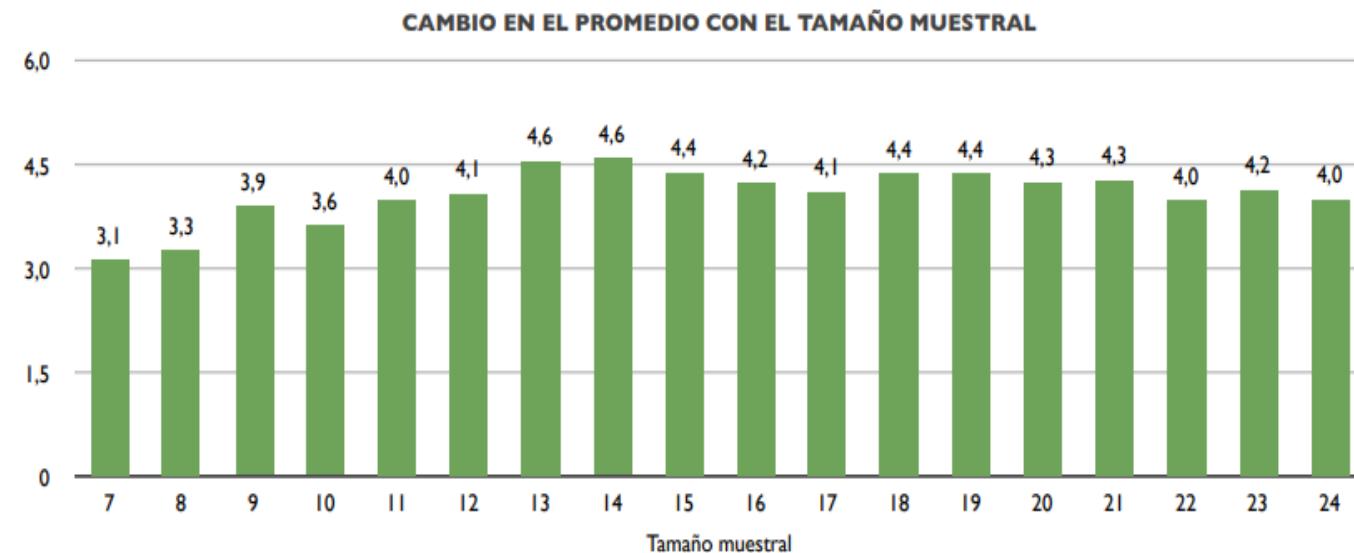
Por lo tanto, el **error (sesgo) relativo** es $(4 - 3.17)/4 = 0.207$.

Nota: En estadística se le llama sesgo de un estimador a la diferencia entre su esperanza matemática y el valor numérico del parámetro que estima. Un estimador cuyo sesgo es nulo se llama insesgado o centrado.

El **sesgo estadístico** es un error que se detecta en los resultados de un estudio y que se debe a factores en la recolección, análisis, interpretación o revisión de los datos.

Recapitulando ...

Si a la muestra anterior le añadimos nuevos elementos, la media muestral cambia. De hecho, vemos que el aumento reiterado de elementos hace que la media muestral converja a la media poblacional.



Estimación puntual

Lo que hemos hecho en la sección anterior: utilizar la media muestral para asignar un valor aproximado a la media poblacional se denomina **estimación**.

Propiedades de estimadores puntuales:

Sesgo: diferencia entre la media de un estimador y el valor del parámetro.

Estimadores insesgados: los que tienen sesgo igual a cero

Criterios para seleccionar un buen estimador:

- Inseguido Eficiente
 - Consistente Suficiente

Criterios	Argumento
Insesgado	Las medias de la distribución muestral tomadas de la misma población son iguales a la media de la población misma. Es lo opuesto a sesgado, que se refiere a tomar una muestra de información tendenciosa. Por ejemplo, si quiero saber qué opina un país de su presidente y tomo una muestra solo de las zonas económicas en donde sé que tiene preferencia, se dice que la información está sesgada.
Eficiente	La eficiencia se refiere al tamaño del error estándar del estadístico.
Consistente	Una estadística es un estimador consistente de un parámetro de población si al aumentar el tamaño de la muestra, se tiene casi la certeza de que el valor de la estadística se aproxima bastante al valor del parámetro poblacional .
Suficiente	Un estimador es suficiente si utiliza tanta información de la muestra que ningún otro estimador puede extraer información adicional acerca del parámetro de población que se está estimando.

Intervalos de confianza ... una pequeña intuición

Este intervalo es muy importante (*muy importante* quiere decir *muy importante*) para la toma de decisiones. La definición más básica que se puede tener es que “el intervalo de confianza es un intervalo de posibles valores (basados en la evidencia) dentro del cual *podría* estar el valor de un parámetro poblacional”.

La confianza NO ES probabilidad sino que hace referencia a que si repetimos muchas veces el mismo experimento (en nuestro caso serían muestreos), se llega a un número X de intervalos donde se puede incluir el valor poblacional.

Definición de intervalo y confianza según la RAE.

Intervalo

Del lat. intervallum

1. m. Espacio o distancia que hay de un tiempo a otro o de un lugar a otro.
2. m. Conjunto de los valores que toma una magnitud entre dos límites dados.
Intervalo de temperaturas, de energías, de frecuencias.

Confianza

De confiar

1. f. Esperanza firme que se tiene de alguien o algo.
2. f. Seguridad que alguien tiene en sí mismo.

Estimación por intervalos

Motivación

En muchos casos prácticos la información de una estimación puntual no es suficiente.

En ciertos casos nos interesa conocer también información sobre la fiabilidad del estimador puntual.

La manera más habitual de proporcionar esa información es calcular un estimador por intervalos.

Intervalo de confianza: permite acotar un par o varios pares de valores, dentro de los cuales se encontrará la estimación buscada

Por ejemplo, dada una muestra nos gustaría conocer un intervalo de valores que con toda seguridad incluya al verdadero valor de la media poblacional.

Consideramos un procedimiento de construcción de intervalos, contega el verdadero valor de la media poblacional. Aquí, a es el **nivel de confianza** y el intervalo obtenido se llamará **intervalo de confianza**.

El **nivel de confianza del intervalo** lo fijamos nosotros., se suele trabajar con 95% y a veces con 99% o el 90%.

Por lo tanto, para una muestra particular x_1, \dots, x_n , un intervalo de confianza está dado por:

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

Donde:

$z_{\alpha/2}$ = valor de z que corresponde al nivel de confianza

n = tamaño muestral

σ = desviación estándar de la población

Coeficiente de Confianza ($1 - \alpha$)	α	$\alpha/2$	$z_{\alpha/2}$
0.90	0.10	0.05	1.645
0.95	0.05	0.025	1.96
0.99	0.01	0.005	2.58

Ejemplo



Supongamos que los rendimientos de las acciones de la empresa MONTALVO INNOVATION & TECHNOLOGY siguen una distribución normal de media μ en euros y desviación $\sigma = 1$. Se toma una muestra de $n = 20$ rendimientos, obteniendo los valores:

5.29	3.66	5.71	6.62	4.30
5.85	6.25	3.40	3.55	5.57
4.60	5.69	5.81	5.71	6.29
5.66	6.19	3.79	4.98	4.84

La media muestral para estos 20 datos es $\bar{X} = 5.18$. Por lo tanto, el intervalo de confianza al 90 % para el rendimiento medio de esta empresa es:

$$\left(5,188 - 1,645 \frac{1}{\sqrt{20}}, 5,188 + 1,645 \frac{1}{\sqrt{20}} \right) = (4,6678, 5,7082)$$

Cálculo de intervalos de confianza

Procedimiento general

Pasos a seguir:

1. Identificar la variable con distribución conocida y la distribución sobre la que construiremos el intervalo de confianza
2. Buscar los percentiles de esa distribución que cubran el nivel de confianza elegido
3. Construir el intervalo para la variable con distribución conocida
4. Sustituir los valores muestrales

Ejemplo: ejercicio de practica

Se ha realizado una encuesta a 60 personas en la que se pedía a los encuestados que valorasen de 0 a 5 la calidad de un servicio. La valoración media en la muestra fue de 2.8 puntos y la desviación es de 0.7 puntos. Calcula un intervalo de confianza al 90 % para la valoración media en la población.

Resultados

$$\begin{aligned}\alpha &= 1 - 0,9 = 0,1, & z_{\alpha/2} &= z_{0,05} = 1,645 \\ -1,645 &\leq \frac{2,8 - \mu}{0,7 / \sqrt{60}} \leq 1,645 \\ 2,65 &\leq \mu \leq 2,95\end{aligned}$$

Distribución de t : cuando no se conoce la desviación σ poblacional

Si no se conoce la desviación estándar de la población, para una estimación por intervalo de la media poblacional cuando σ no se conoce es la siguiente.

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

donde s es la desviación estándar de la muestra, y $t_{\alpha/2}$ es el valor de t que proporciona un área de $1 - \alpha/2$ en la cola superior de la distribución t para $n - 1$ grados de libertad.

Tabla t-Student



Grados de libertad	0.25	0.1	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3137	12.7062	31.8210	63.6559
2	0.8165	1.8856	2.9200	4.3027	6.9645	9.9250
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8408
4	0.7407	1.5332	2.1318	2.7765	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9979	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7970
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500

Ejemplo



La empresa LERMA S.A. desea estimar el precio medio de la venta de abrigos, una muestra aleatoria de $n = 25$ tiene un precio promedio de $\bar{x} = 50$, euros y $S = 8$. Construya un intervalo de confianza del 95% para μ .

Datos: $n = 25$, $s = 8$, $\bar{x} = 50$, g.l. $n - 1 = 24$, $t_{\alpha/2} = 0.025 = 2.0636$

Solución:

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 50 \pm (2.0636) \frac{8}{\sqrt{25}}$$

$$= 46.70 \leq \mu \leq 53.30$$

Interpretación: Con 95% de conf. el precio promedio está entre 46.7 y 53

Recapitulando

Para los **intervalos de confianza** se usan dos formulas:

1. Cuando conocemos la desviación de la población: usamos intervalo de confianza para

z

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

2. Cuando no conocemos la de la población pero si tenemos la desviación típica de la muestra: **usamos t - Student**

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Contraste de hipótesis

Un **Contraste de hipótesis** es un conjunto de reglas para tomar una decisión acerca de una hipótesis, falsa o no falsa, en base a una probabilidad.

Las etapas a seguir son:

- Plantear una hipótesis y utilizarla como premisa
- Deducir de lo anterior una situación esperable
- Usar lo observado en los datos como prueba
- Cuantificar la discrepancia entre lo observado y lo esperado: Obs vs Esp
- Tomar una decisión a favor o en contra de la hipótesis

Para hacer estos **contrastes de hipótesis**, generalmente se utilizan algunos test estadísticos:

test Z:

$$z = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$$

test T:

$$t = \frac{\bar{x} - \mu_0}{(s/\sqrt{n})},$$

$$df = n - 1$$

Contraste de hipótesis

Hipótesis.

(Del lat. *hypothēsis*).

1. f. Suposición de algo posible o imposible para sacar de ello una consecuencia.

Hipótesis estadística: Una **hipótesis estadística** es una afirmación respecto a una característica de una población. En muchas situaciones, el interés está en probar cierta afirmación sobre la población más que en estimar o predecir alguno de sus parámetros. Toda investigación debe tener definida una **hipótesis**, la cual es una declaración acerca del tema de investigación que puede ser verdadera o falsa.

Contraste o test de hipótesis

Un **contraste de hipótesis** es un procedimiento formal para rechazar o no una hipótesis estadística planteada sobre una población utilizando para ello una muestra de observaciones.

Un contraste de hipótesis es un procedimiento que:

- se basa en datos muestrales
- para proporcionarnos información cara a tomar una decisión

Esta hipótesis a confrontar se conoce como la hipótesis nula (H_0):

Será mantenida a menos que la muestra aporte suficiente evidencia contraria

Elementos de la prueba de hipótesis

La **hipótesis estadística nula**, simbolizada como H_0 , es la hipótesis que se somete a prueba. Por lo general, es una afirmación acerca de que un parámetro poblacional tiene un valor específico.

La **hipótesis estadística alternativa**, simbolizada como H_1 , es una afirmación sobre el mismo parámetro poblacional considerado en la hipótesis nula, que especifica que el mismo tiene un valor diferente, de alguna manera, al postulado en la hipótesis nula.



La hipótesis nula: ejemplos

Cereales dinorah

La empresa fabricante de paquetes de cereales DINORAH S.A. afirma que, en promedio, cada paquete pesa al menos 400 g. Quieres contrastar esta información a partir de los pesos de los paquetes en una muestra aleatoria.

Población: X = 'peso de un paquete de cereales (en g)'

Hipótesis nula, H_0 : $\mu \geq 400$

Si la hipótesis nula no es válida, alguna alternativa debe ser correcta. Para realizar el contraste, el investigador debe especificar una hipótesis alternativa H_1 frente a la que se contrasta la hipótesis nula.



La empresa fabricante de paquetes de cereales Dinorah S.A. afirma que, en promedio, cada paquete pesa al menos 400 g. Quieres contrastar esta información a partir de los pesos de los paquetes en una muestra aleatoria.

Población: X = 'peso de un paquete de cereales (en g)'

Hipótesis nula, $H_0 : \mu > 400$

Hipótesis alternativa $H_1 : \mu \leq 400$

Si no es razonable, rechazamos la hipótesis nula en favor de la alternativa.

Formulas para el contraste de hipótesis

Para el contraste de hipótesis tenemos dos formulas (test), la diferencia se encuentra en que en el *Z-test* se supone que la desviación σ poblacional es conocida, mientras que en el *T-test* es desconocida y se toma la de la muestra:

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma}{n}}} \quad t = \frac{\bar{X} - \mu}{\sqrt{\frac{s}{n}}}$$

También uno es para muestras mayores a 30 (z) o menores o igual a 30 (t).

Aquí \bar{x} es la media muestral, μ se refiere a la hipótesis nula, n es el tamaño de la muestra. La única diferencia es σ poblacional y s de la muestra.

Volviendo al ejemplo de los Cereales Dinorah



Cereales dinorah

La empresa fabricante de paquetes de cereales Dinorah S.A. afirma que, en promedio, cada paquete pesa al menos 400 g. Quieres contrastar esta información a partir de los pesos de los paquetes en una muestra aleatoria. La muestra fue de 20 cajas de cereales (5 frutti lupis, 10 chocodino, 5 zukadino). La media de la muestra es de 402, y la desviación de la muestra es de 2.1581.

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}}$$

$$H_0: \mu = 400 \quad vs \quad H_1: \mu \neq 400$$

Luego,

$$\text{Se calcula } t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{402 - 400}{2.1581/\sqrt{20}} = \frac{2}{2.1581/\sqrt{20}} = 4.15$$

$$\text{Si } \alpha = 0.05, \text{ entonces } t_{\alpha/2}(19) = t0.025(19) = 2.0930$$

Si el valor absoluto del valor t es mayor que el valor crítico (tablas), se rechaza la **hipótesis nula**. En este caso 4.15 es mayor que el valor de tablas 2.0939. Por lo tanto se rechaza la hipótesis nula, y se acepta la alternativa.

Conclusión: Cuando hacemos test de hipótesis en estadística las decisiones son dos: a) rechazamos la **hipótesis nula** b) no rechazamos la **hipótesis nula**.

Errores en el contraste de hipótesis

A la hora de hacer un contraste, es posible que la muestra que seleccionemos no necesariamente nos dé evidencias sobre lo que ocurre en la población, por lo que podríamos rechazar la hipótesis nula cuando en realidad es cierta o lo contrario:

Tipos de errores		
Errores tipo I y II		
	La H_0 es verdadera.	La H_0 es falsa.
Rechazamos H_0	Error tipo I: rechazar la hipótesis nula cuando esta es verdadera.	Decisión correcta.
Aceptamos H_0	Decisión correcta.	Error tipo II: rechazar la hipótesis nula cuando esta es falsa.

Error tipo I

El **error tipo I** consiste en rechazar la hipótesis nula, cuando ésta es cierta. A la probabilidad de cometer este error se le llama nivel de significancia (α) y se debe fijar antes del inicio del estudio:

$$\alpha = P(\text{Rechazar } H_0 | H_0 \text{ es cierta})$$

Ejemplo-problema de investigación

Supongamos que el nivel de significancia se fijó en 10%, lo cuál significaría, que en el 10% de los casos podríamos decir que hay diferencias entre los tratamientos cuando en realidad no las hay.

Error tipo II

El error tipo II consiste en no rechazar la hipótesis nula, cuando ésta es falsa. A la probabilidad de cometer este error se le llama **función característica**.

$$\beta_\tau(\theta) = P_\theta(\text{No rechazar } H_0 | H_0 \text{ es falsa})$$

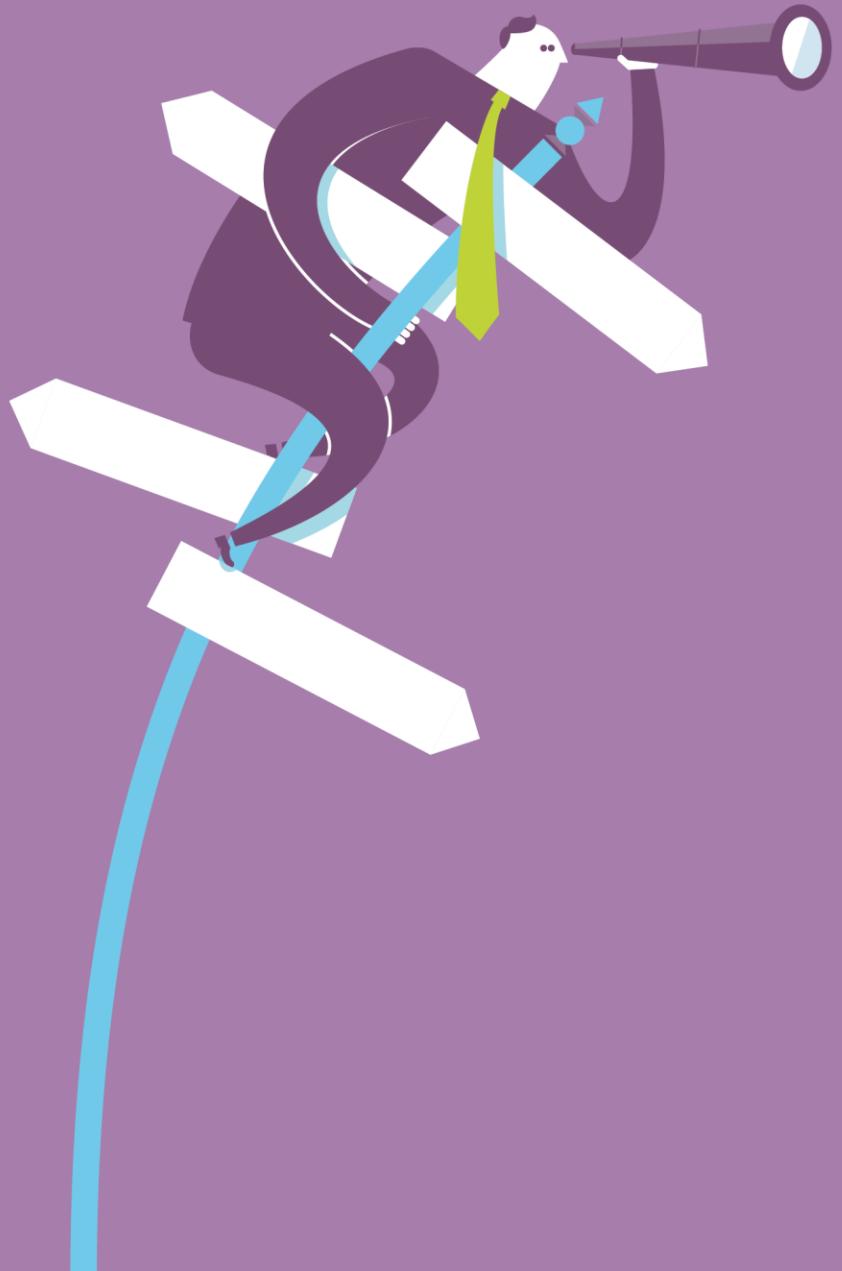
Es una función, pues dado que no conocemos el verdadero valor del parámetro, se debe calcular con los posibles valores de éste.

Type I Error (false-positive)



Type II Error (false-negative)





Tema 5. Métodos de pronósticos para series de tiempo y datos de corte transversal.

Introducción

Definición 1. Una **serie de tiempo** es una sucesión de observaciones de una variable tomadas en varios instantes de tiempo.

- Nos interesa estudiar los cambios en esa variable con respecto al tiempo.
- Predecir sus valores futuros.

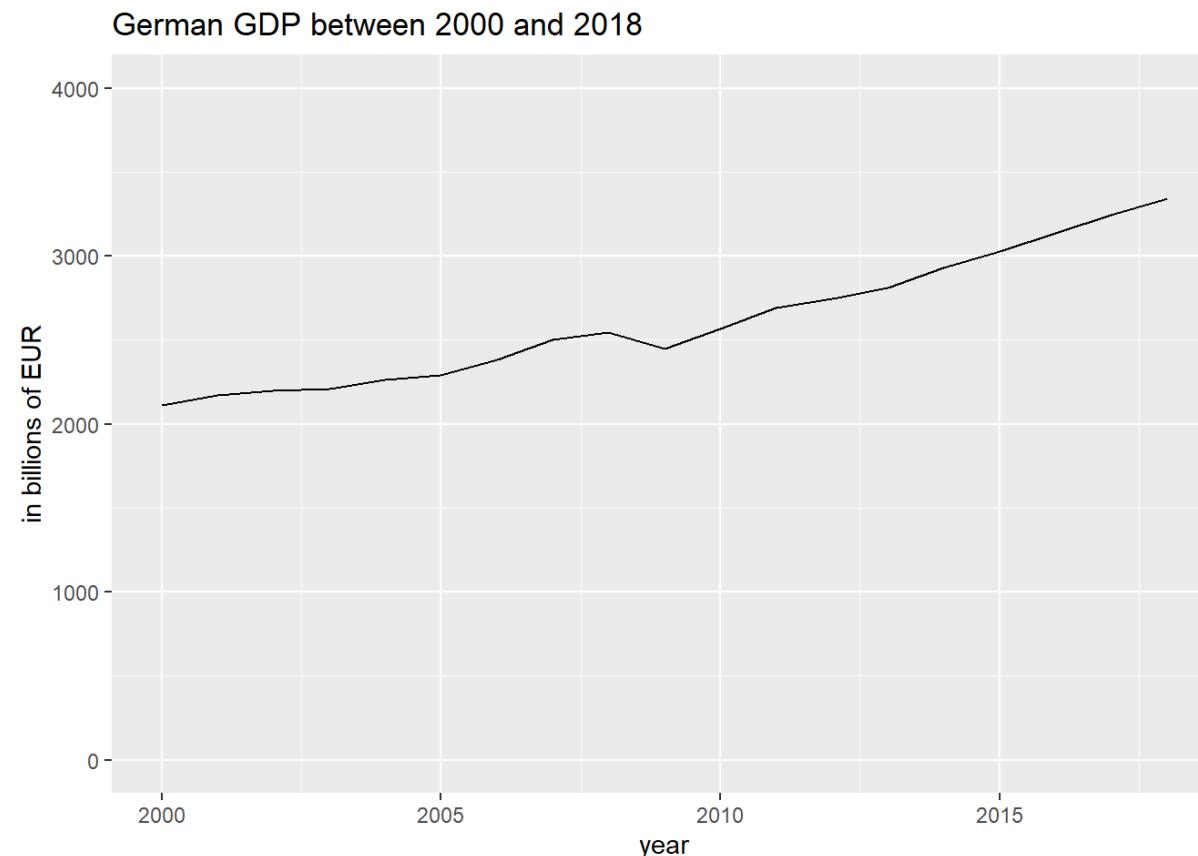
Ejemplos de series temporales podemos encontrarlos en muchos campos de conocimiento:

- Economía: Producto Interno Bruto anual, tasa de inflación, tasa de desempleo, etc.
- Teoría del crimen: tasa de delincuencia, número de robos.
- Economía ambiental: emisiones de CO₂ en un periodo de tiempo, producción de plásticos por año, etc.

Representación gráfica de una serie de tiempo

A menudo, se representa la serie en un gráfico temporal, con el valor de la serie en el eje de ordenadas (y) y los tiempos en el eje de abscisas (x).

Ejemplo 1. Producto Interno Bruto de Alemania entre 2000 a 2018



Periodicidad

- Anual
- Mensual
- Semanal
- Diaria

Claramente, existen otros tipos de periodicidad: semestral, trimestral, horaria. El tipo de periodicidad es una característica muy importante en una serie y aparecerán determinadas pautas debida a ella.

Clasificación de series temporales

Definición 2. Una serie temporal es una sucesión de observaciones de una variable tomadas en varios instantes de tiempo. Estas observaciones provienen de una distribución que puede ser diferente en cada instante del tiempo.

Una serie es **estacionaria** si la media y la variabilidad se mantienen constantes a lo largo del tiempo.

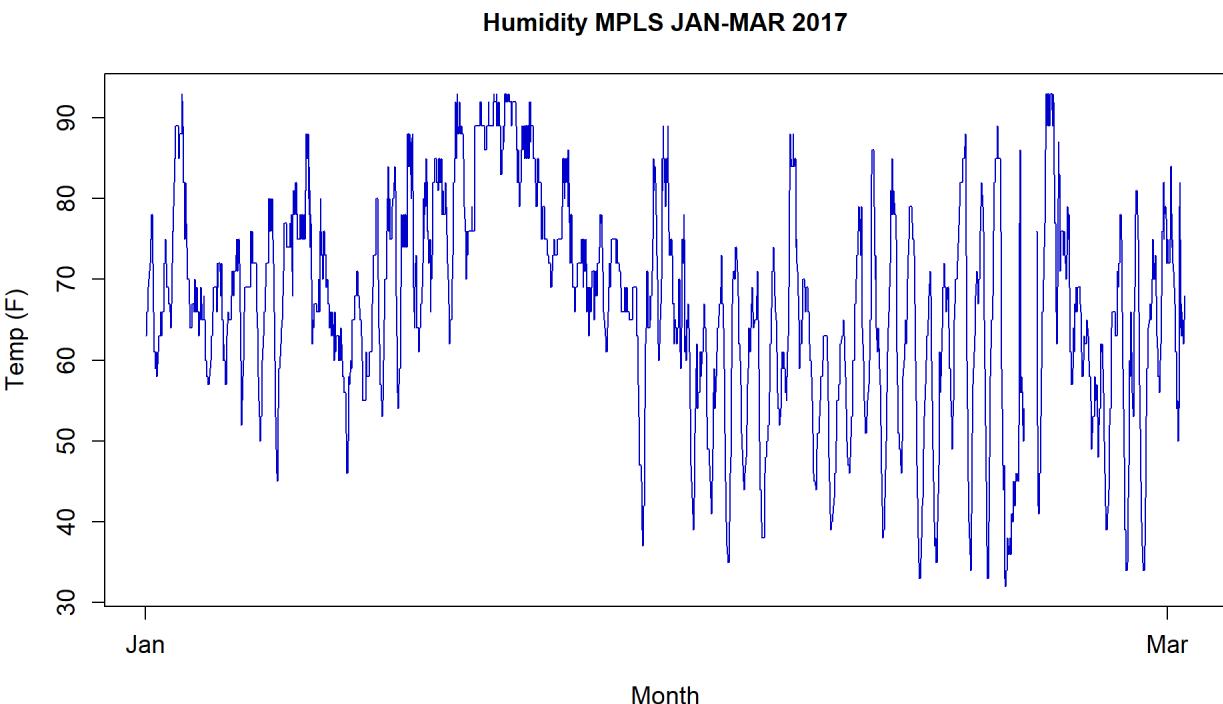
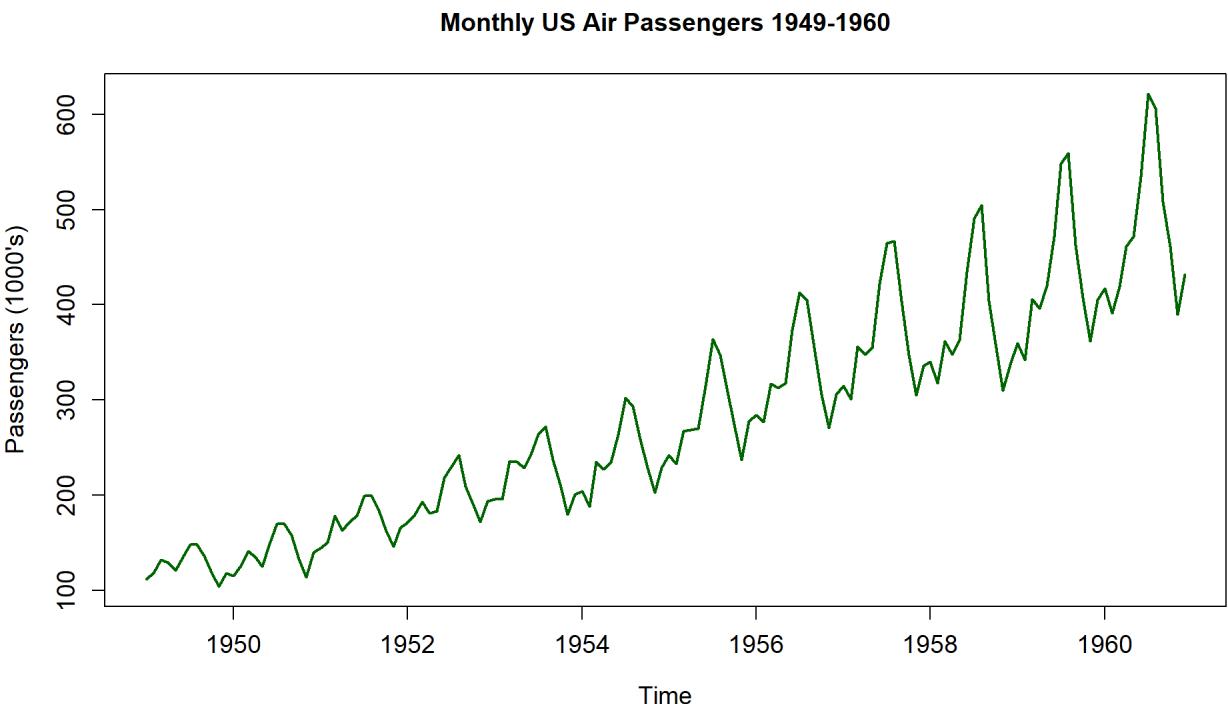
Una serie es **no estacionaria** si la media y/o la variabilidad cambian a lo largo del tiempo.

- Series no estacionarias pueden mostrar cambios de varianza.
- Series no estacionarias pueden mostrar una tendencia, es decir que la media crece o baja a lo largo del tiempo.

¿Por qué es bueno que las series sean estacionarias?

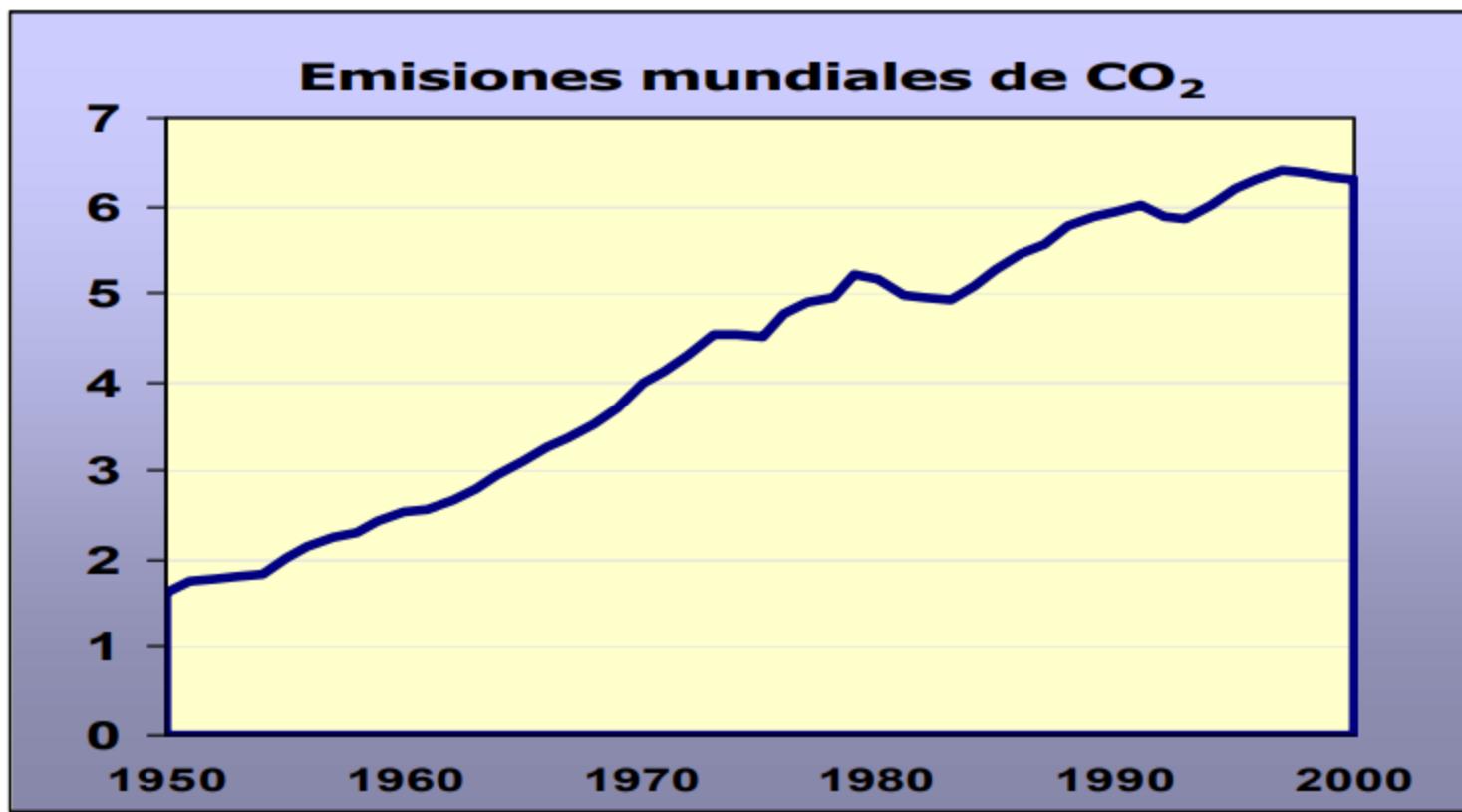
- Con series estacionarias podemos obtener predicciones fácilmente
- Como la media es constante, podemos estimarla con todos los datos, y utilizar este valor para predecir una nueva observación.
- También se pueden obtener intervalos de predicción (confianza) para las predicciones asumiendo que X_t sigue una distribución conocida, por ejemplo, normal.

Ejemplo: Serie no estacionaria vs serie estacionaria



Clasificación de series temporales

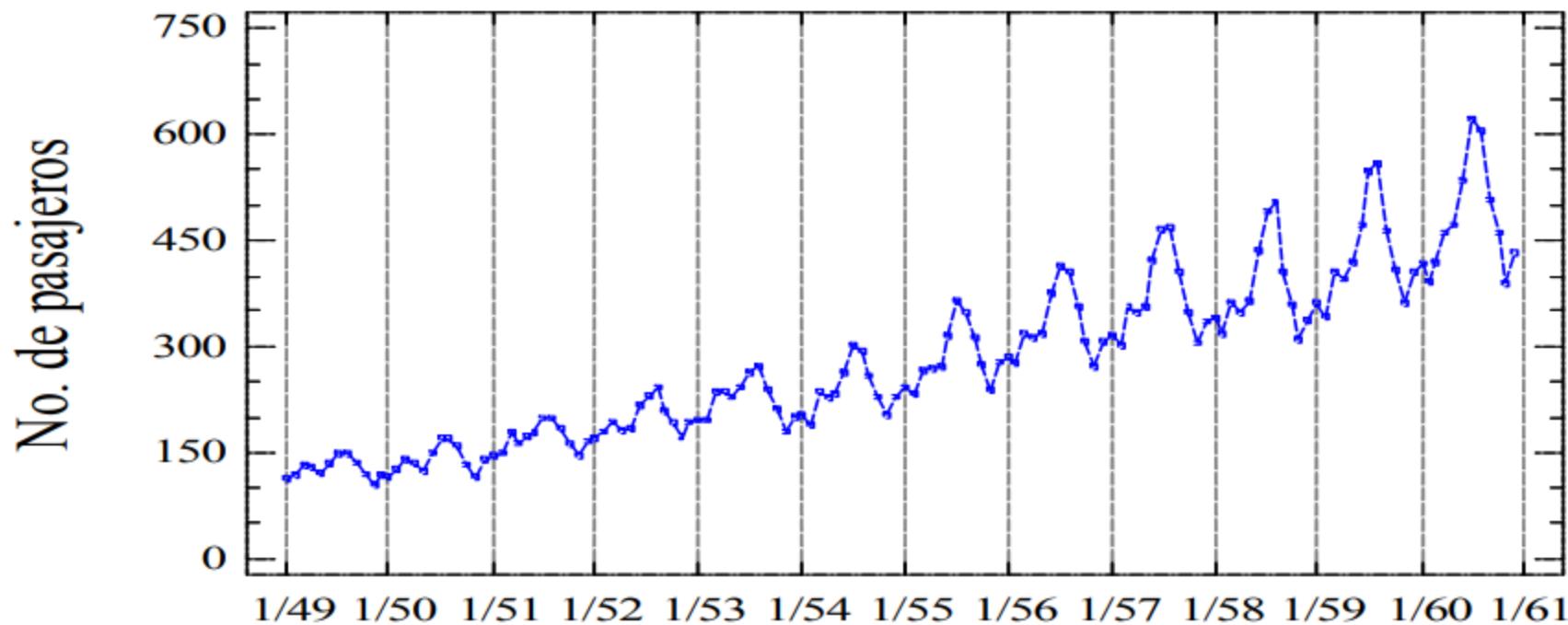
Serie no estacionaria: Emisiones mundiales de CO_2 .



Tendencia

Clasificación de series temporales

Serie no estacionaria: Número mensual de pasajeros de avión, USA,
Enero:1949 a Diciembre:1960



Fuente de datos: Box, G. & Jenkins, G. (1976) Time Series Analysis:
Forecasting and Control.

Componentes de una serie temporal

En muchos casos, se supone que la serie temporal es la suma de varias componentes:

$$X_t = T_t + S_t + I_t$$

Valor esperado = Tendencia + Estacionalidad + Irregular

Tendencia: comportamiento o movimiento suave de la serie a largo plazo.

Estacionalidad: movimientos de oscilación dentro del año. **Irregular:** variaciones aleatorias alrededor de los componentes anteriores.

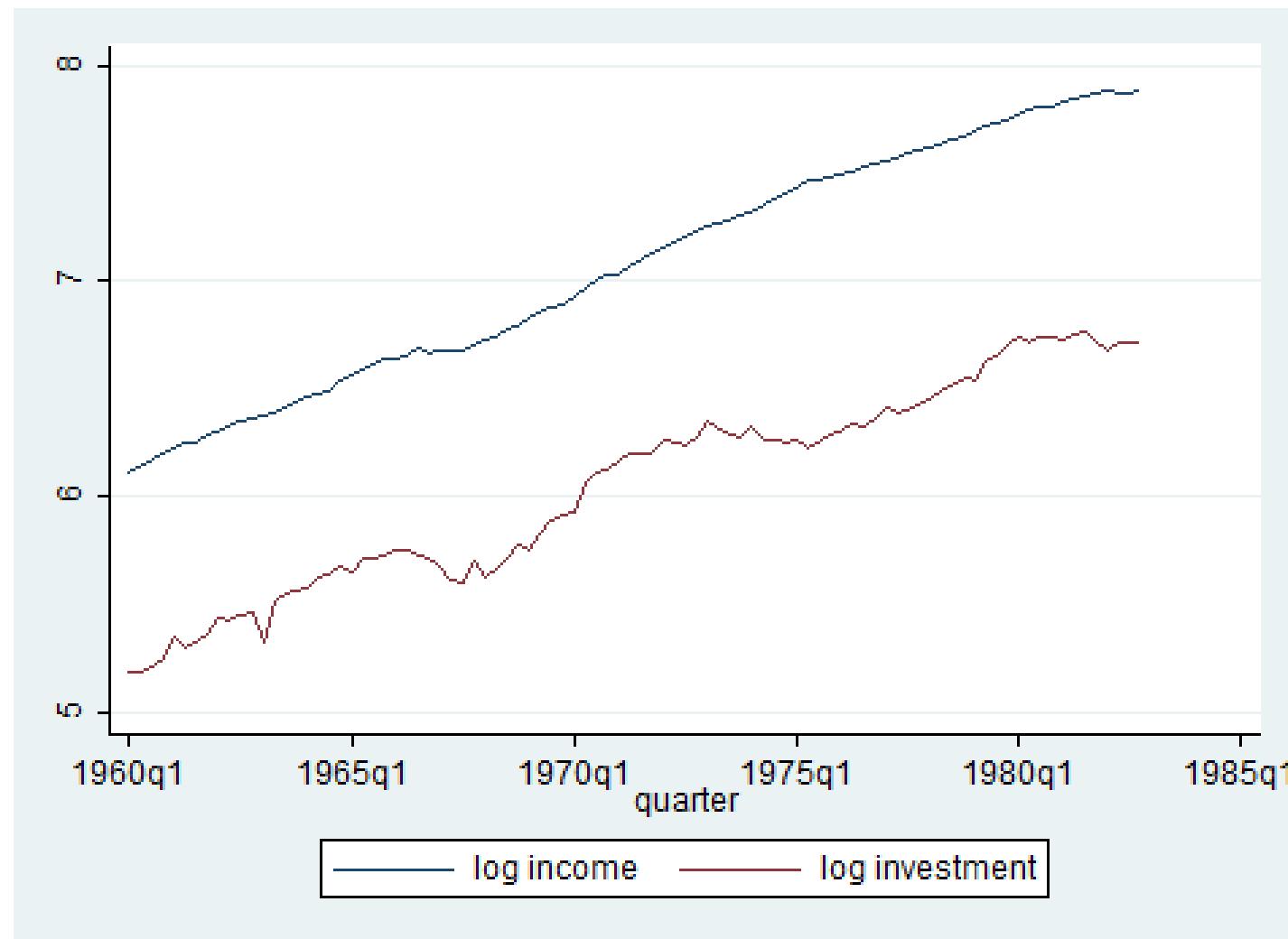
En esos casos, es interesante obtener o “aislar” los distintos componentes.

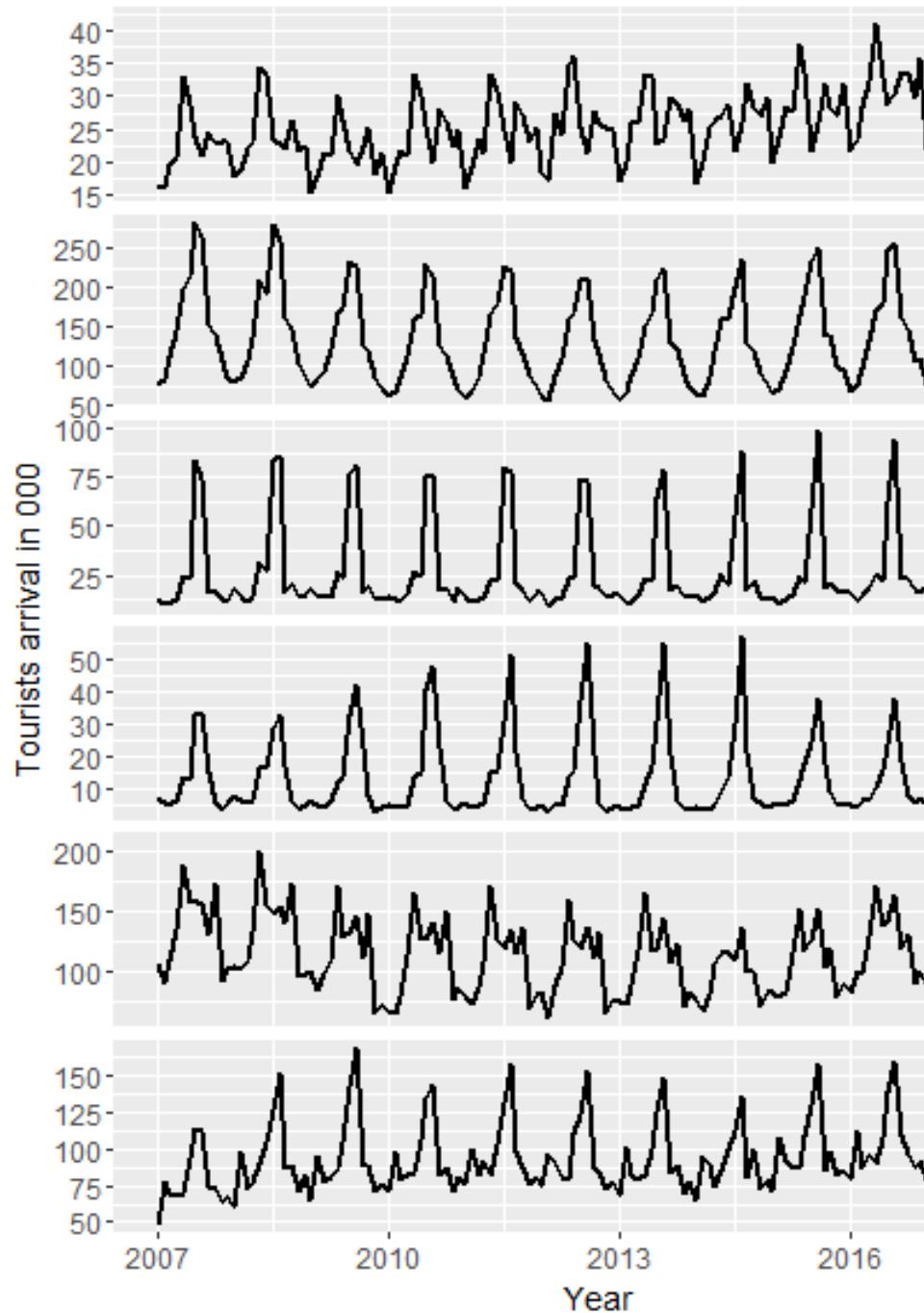
Tendencias

La **tendencia** es el componente de largo plazo que representa el crecimiento o descenso en la serie de tiempo durante un periodo extenso.

Las técnicas que deben considerarse al considerar series con tendencia son los modelos de promedios móviles, suavizamiento exponencial y modelos autorregresivos integrados de promedio móvil (ARIMA, del inglés *Autorregressive Integrated Moving Average*).

Tendencias en series de tiempo



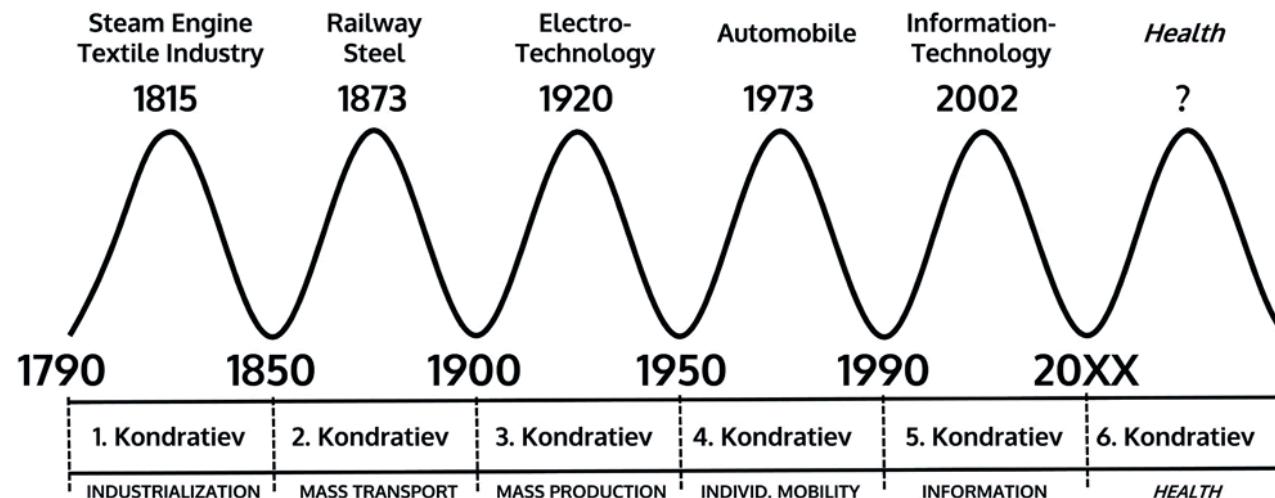


Estacionales

- El componente **estacional** es un patrón que se repite año tras año. Como habría de esperarse, este tipo de variación implica patrones de cambio en el lapso de un año que tienden a repetirse anualmente.

Cíclico

El componente **cíclico** es la oscilación alrededor de la tendencia. El ejemplo más común de fluctuación cíclica es el ciclo económico a través del tiempo.



Nota: Para **Kondratieff** la economía vive **ciclos** de cincuenta años que se repiten y repetirán indefinidamente.

Tipos de datos con series temporales

Building	Year	Rental cost per square foot (US\$)	Occupancy (%)	Building age (years)
1	2000	2.0	90	10
1	2001	2.0	90	11
1	2002	2.2	100	12
2	2000	0.8	60	15
2	2001	0.9	70	16
2	2002	1.2	90	17
3	2000	3.7	80	2
3	2001	3.8	85	3
3	2002	4.0	100	4

 Cross-sectional data  Time series data  Pooled data  Panel data

Temario para examen parcial

1. Tipos de errores en prueba de hipótesis (error tipo I y error tipo II)
2. Identificar los componentes de una serie de tiempo (estacional, cíclico, irregular)
3. Encontrar la probabilidad (tips en diapositiva 87)
4. Formulas para calcular la desviación estándar y la media muestral
5. Calculo de la desviación estándar y calculo de la media muestral
6. Calcular e identificar la frecuencia relativa y la frecuencia absoluta
7. Formulas del intervalo de confianza (estudiar las dos formulas y sus diferencias)
8. Calculo e interpretar la formula estandarizada de la distribución normal.
9. Calculo de los intervalos de confianza para muestras grandes y pequeñas
10. Realizar diagrama de dispersión

Temario para examen parcial

11. Identificar los grados de correlación
12. Realizar histograma
13. Identificar probabilidad en las tablas Z.
14. Técnicas para series de tiempo
15. Diagramas de Venn, realizar unión e intersección
16. Identificar variables aleatorias y continuas
17. Probabilidad condicional, ejercicio.
18. Probabilidad de la unión de eventos, ejercicio.
19. Definición de estimador insesgado

Ejercicios de repaso

En un examen de econometría, la calificación media fue de 70 y la desviación típica de 12. Determinar en cuántas desviaciones estándar de la media están los alumnos que obtuvieron 65.

Solución:

$$Z = \frac{X - \mu}{\sigma}$$

μ : 70

σ : 12

X = 65

$$z = \frac{65 - 70}{12}$$

Ejercicio

Ximena y Nathalia discuten sobre el nivel de crecimiento económico a nivel mundial. Ximena en su posición optimista, cree que este es de 3.2%. Nathalia es un poco pesimista, cree que es del 1.9%. En su afán de estar en la razón, Nathalia toma una muestra de 20 países. Construye un intervalo de confianza del 95%. Considera una desviación estándar de la muestra de 1.75.

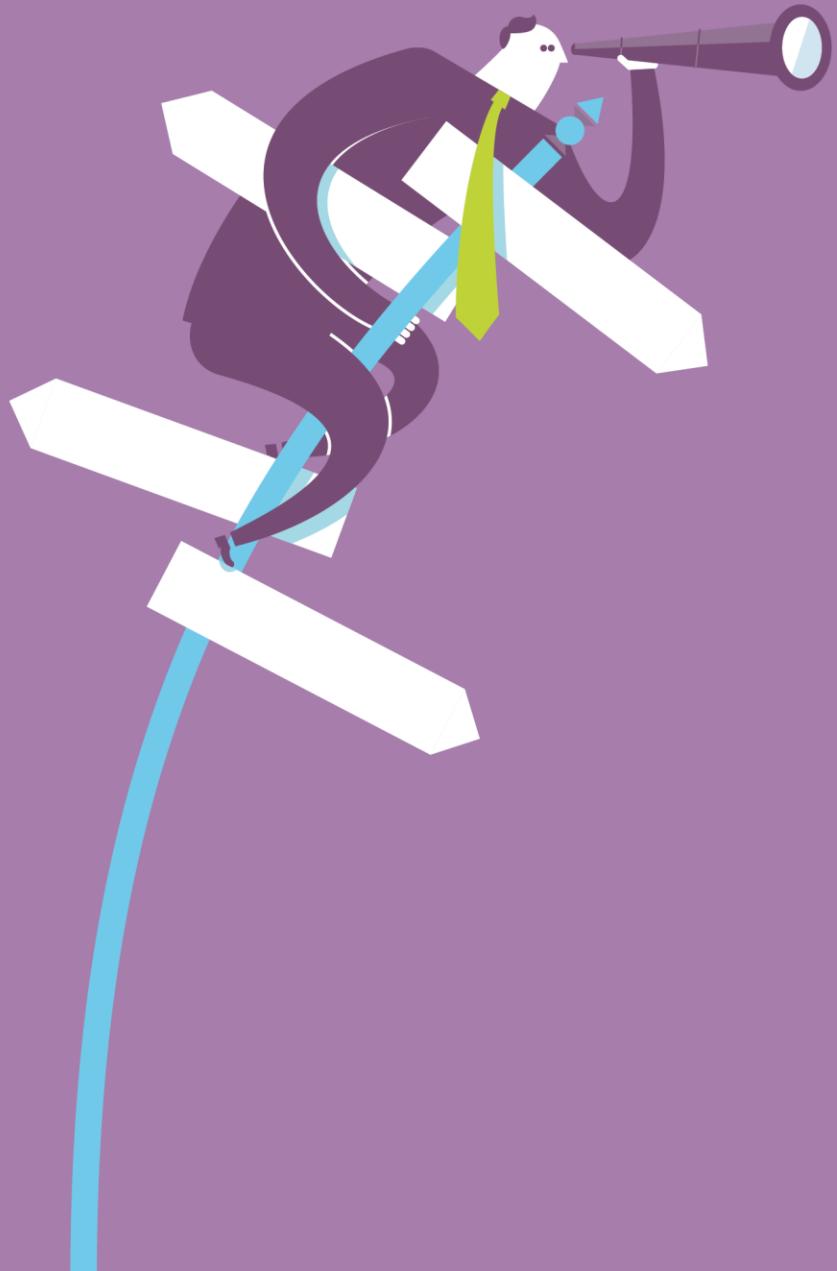
Azerbaiyán = 1.4	Colombia = 2.7	Japón = 0.8	Rwanda = 8.7
Bolivia = 4.2	Dinamarca = 1.4	México = 2.0	Sudáfrica = 0.6
Brasil = 1.1	Bélgica = 1.4	Noruega = 1.4	Uruguay = 1.6
Burundi = 1.6	Francia = 1.7	Portugal = 2.1	Ucrania = 3.3
Canadá = 1.9	Guatemala = 3.1	Rep. Checa = 2.9	Turquía = 2.6

Fuente: Datos originales de Banco Mundial (2018)

¿Quién estaba en la razón?

Modulo 2: Series de tiempo y regresión lineal simple.





Tema 6. Patrón de datos en las series de tiempo y análisis de autocorrelación.

Recapitulando temas anteriores ...

Antes de continuar con las series de tiempo tendremos que revisar primero qué es la **correlación** entre dos variables, que es el primer concepto que debemos profundizar antes de comenzar con las series de tiempo. Antes de definir lo que es una autocorrelación, primero debes saber exponer lo que significa la correlación.

La **correlación** es una medida estadística que expresa hasta qué punto dos variables están relacionadas linealmente (esto es, cambian conjuntamente a una tasa constante).

A menudo estamos interesados en **observar y medir la relación entre 2 variables numéricas**. Por ejemplo, si queremos evaluar la relación entre:

1. Las horas que se dedican a estudiar una asignatura y la calificación obtenida en el examen correspondiente.
2. La relación entre los niveles de educación y los ingresos de un grupo de individuos.

Lo que **nos interesa es identificar el tipo de relación o asociación entre ambas variables. Para ello, se usa el coeficiente de correlación**.

Este coeficiente de correlación toma valores entre -1 y 1:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n(\sum x^2) - (\sum x)^2)(n(\sum y^2) - (\sum y)^2)}}$$

Dependiendo de su valor, nos dirá si hay una relación positiva o negativa. Existe una clasificación para medir su intensidad.

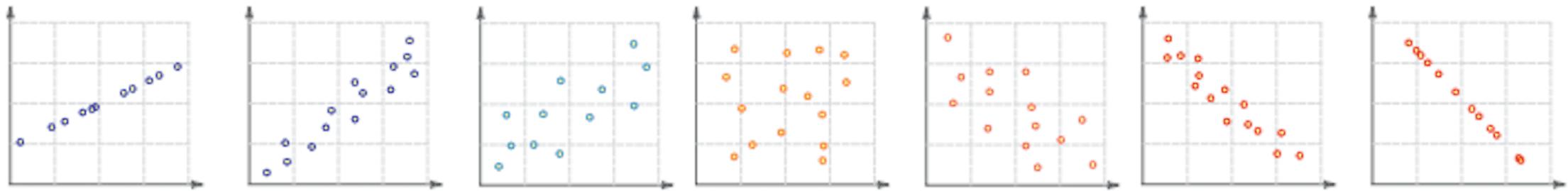
Resultado			Coeficiente de correlación lineal (positivo)
0.00	a	0.09	Nula
0.10	a	0.19	Muy débil
0.20	a	0.49	Débil
0.50	a	0.69	Moderada
0.70	a	0.84	Significativa
0.85	a	0.95	Fuerte
0.96	a	1.00	Perfecta

Resultado			Coeficiente de correlación lineal (negativo)
0.00	a	0.09	Nula
-0.10	a	-0.19	Muy débil
-0.20	a	-0.49	Débil
-0.50	a	-0.69	Moderada
-0.70	a	-0.84	Significativa
-0.85	a	-0.95	Fuerte
-0.96	a	-1.00	Perfecta

La correlación puede mostrar dos cosas: dirección y fuerza.

Dirección: Positiva o negativa

Fuerza: Qué tanta dispersión existe.



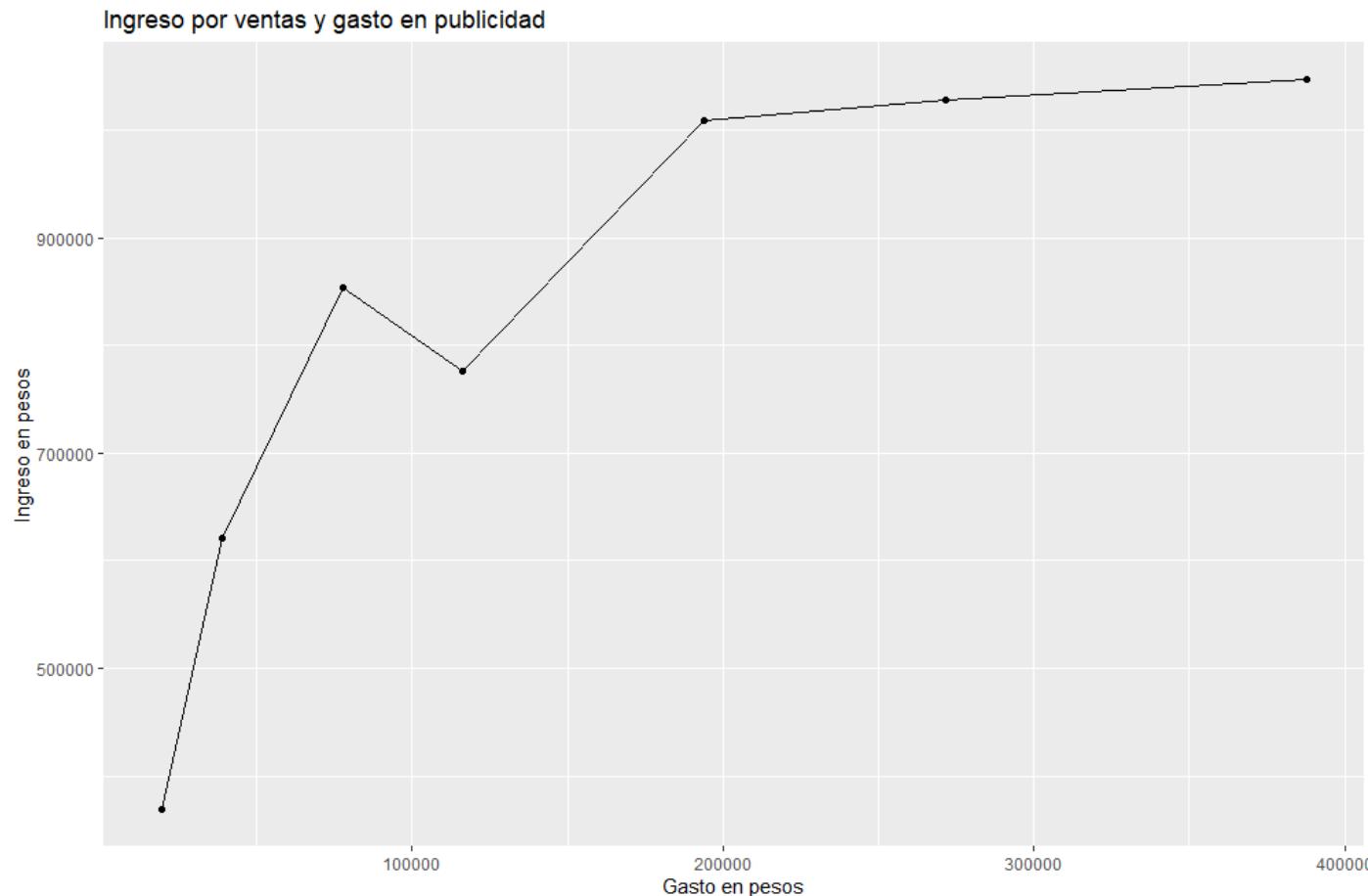
Si existe poca dispersión a lo largo de la tendencia diremos que la relación es fuerte, mientras que si la dispersión es grande o la nube de puntos es circular, diremos que la relación es débil.

Ejemplo de la correlación ...

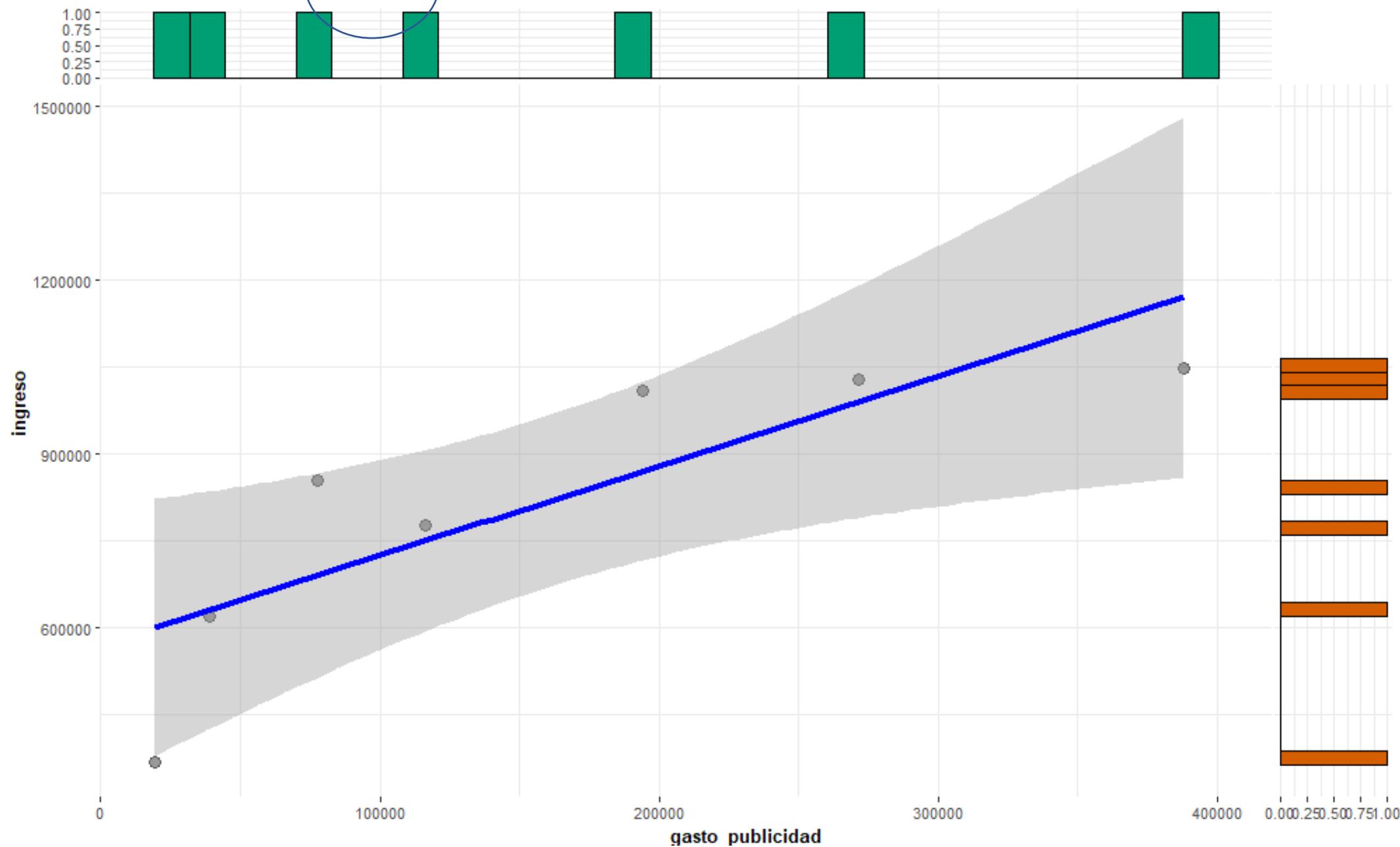
A continuación, se presentan los gastos anuales en publicidad y los ingresos de un restaurante de la localidad:

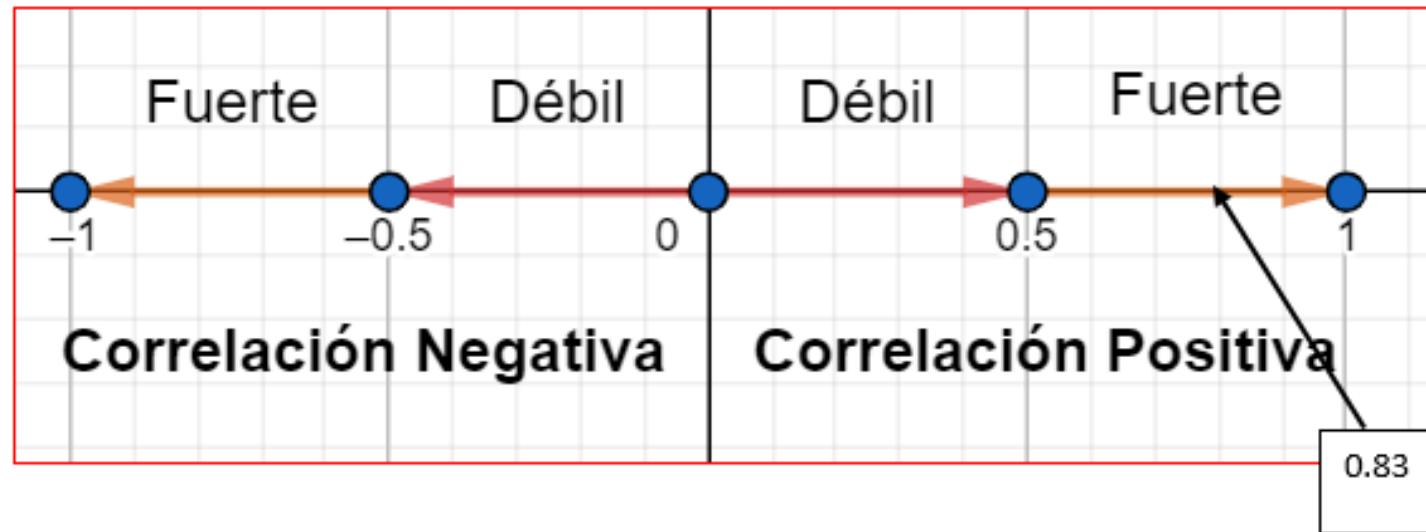
Año	Ingresos	Gastos en publicidad
2012	\$ 368,600.00	\$ 19,400.00
2013	\$ 620,800.00	\$ 38,800.00
2014	\$ 853,600.00	\$ 77,600.00
2015	\$ 776,000.00	\$ 116,400.00
2016	\$ 1,008,800.00	\$ 194,000.00
2017	\$ 1,028,200.00	\$ 271,600.00
2018	\$ 1,047,600.00	\$ 388,000.00

De manera **visual**, podemos realizar una simple exploración y observar qué pasa con los ingresos de la empresa a medida que se incrementa el gasto en publicidad. ¿Qué tipo de relación observas? ¿Una relación positiva, negativa o no hay relación?



$t_{Student}(5) = 3.34, p = 0.02$ $\hat{r}_{Pearson} = 0.83, \text{CI}_{95\%} [0.21, 0.97], n_{pairs} = 7$





Conclusión sobre los valores de correlación: En la medida que los valores se acercan a -1 (correlación negativa) o a 1 (Correlación positiva). La relación entre las dos variables tenderá a ser más fuerte. Es decir, deseamos que el coeficiente de correlación esté cercano a -1 o a 1. Entre más cercano la correlación será más fuerte. Y entre más cercano sea a 0 la correlación será más débil o prácticamente nula cuando el coeficiente de correlación es 0.

Autocorrelación

Bien pues ahora es momento de hablar de la autocorrelación.

Autocorrelación es la correlación que existe entre una variable cuando se retarda uno o más periodos consigo misma.

Es muy importante tomarlo en cuenta, porque, por ejemplo, volviendo al cuadro anterior de los gastos de publicidad, **muchas de las veces, el gasto de publicidad en enero traerá resultados de ventas hasta febrero y el éxito de la publicidad en febrero, probablemente se reflejará en las ventas de marzo. Así, generalmente, en todos los meses del año es probable que exista un desfase de tiempo.**

Ejemplo

Considérese el siguiente ejemplo:

Periodo t	Serie, \hat{Y}_t	Serie con un retardo Y_{t-1}
1	86	-
2	81	86
3	82	81
4	85	82
5	90	85
6	93	90
7	89	93
8	86	89
9	-	86

Al omitir los periodos 1 y 9 (debido a que una serie o la otra no tienen valor en ellos), puede calcularse la correlación entre Y_t y Y_{t-1} mediante la expresión:

$$r_k = \frac{\sum_{t=k+1}^n (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2} \quad k = 0, 1, 2, \dots$$

En donde:

r_k = coeficiente de autocorrelación para un retardo de k periodos

\bar{Y} = media de los valores de la serie

Y_t = observación en el periodo t

Y_{t-k} = observación en k periodos anteriores o durante el periodo $t-k$

El coeficiente de autocorrelación de retardo 1 (r_1), es decir, la autocorrelación entre $y Y_t$ y Y_{t-1} , se calcula con la tabla que se presenta enseguida.

Periodo t	Serie						
	Serie, Y_t con un retardo Y_{t-1}	$(Y_t - \bar{Y})$	$(Y_{t-1} - \bar{Y})$	$(Y_t - \bar{Y})^2$	$(Y_t - \bar{Y})(Y_{t-1} - \bar{Y})$		
1	86	-	-0.5	-	0.25	-	
2	81	86	-5.5	-0.5	30.25	2.75	
3	82	81	-4.5	-5.5	20.25	24.75	
4	85	82	-1.5	-4.5	2.25	6.75	
5	90	85	3.5	-1.5	12.25	-5.25	
6	93	90	6.5	3.5	42.25	22.75	
7	89	93	2.5	6.5	6.25	16.25	
8	86	89	-0.5	2.5	0.25	-1.25	
9	-	86					
Total	692			114	66.75		

$$\bar{Y} = \frac{692}{8} = 86.5$$

$$r_1 = \frac{\sum_{t=1+1}^n (Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2} = \frac{66.75}{114} = 0.5855 \quad k = 1$$

Es un término estadístico la autocorrelación se utiliza para describir la presencia o ausencia de correlación en los datos de las series temporales, indicando, si las observaciones pasadas influyen en las actuales.

Por tanto, se puede decir que la autocorrelación hace referencia cuando los valores que toman una variable en el tiempo no son independientes entre sí, sino que un valor determinado depende de los valores anteriores.

Para interpretar la función de autocorrelación se necesita un método para probar cuáles de las autocorrelaciones r_k son estadísticamente diferentes de cero. Una prueba aproximada se presenta enseguida.

Prueba de hipótesis para los coeficientes de autocorrelación

Hipótesis:

Hipótesis nula: $H_0 : \rho_k = 0$ (la autocorrelación es igual a cero)

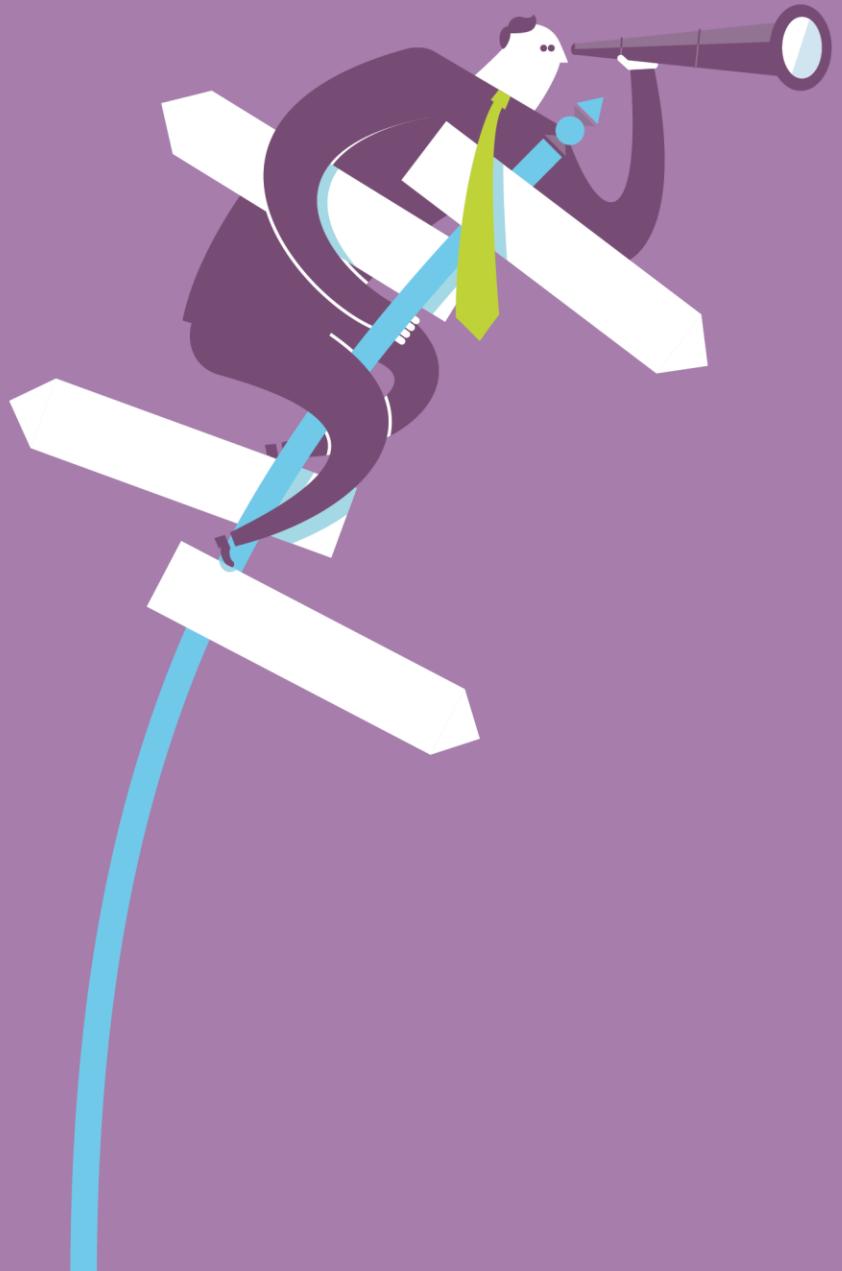
Hipótesis alternativa: $H_a : \rho_k \neq 0$ (la autocorrelación es diferente de cero)

Donde ρ_k es el coeficiente de autocorrelacional poblacional del lapso k .

Estadística de prueba : r_k , el coeficiente de autocorrelación muestral del lapso k .

Región de rechazo: A un nivel aproximado de significancia de 0.05 ($\alpha \approx 0.05$), se rechaza H_0 si $|r_k| > 2 / \sqrt{n}$

Conclusión: En el presente caso, puesto que 0.5855 es menor que $2 / \sqrt{n} = 2 / \sqrt{8} = 0.7071$, no hay razón para descartar la H_0 , es decir, puede concluirse que no existe suficiente evidencia para indicar que la autocorrelación sea diferente de cero.



Tema 7. Métodos de pronósticos basados en promedios, suavización y descomposición.

Series de tiempo: recapitulando

Una serie de tiempo es una serie de registros realizados en diversos periodos de tiempo (días, semanas, meses, trimestres, años). Los registros son valores numéricos que varían en el tiempo. Un aspecto básico del estudio de las series de tiempo requiere analizar la naturaleza de estas variaciones y hacer predicciones.

Estacionaria: fáciles de predecir debido a que tienden a ser constantes en el tiempo.

No estacionaria: presentan cierta tendencia y es mas difícil pronosticar con este tipo de series de tiempo.

Nomenclatura de las series temporales

Esquema de elaboración de un pronóstico

Datos pasados

..... $Y_{t-3}, Y_{t-2}, Y_{t-1}$

Usted está aquí
t

Y_t

Periodos por pronosticar

$Y_{t+1} Y_{t+2} Y_{t+3} \dots$

Donde

Y_t

Es la observación más reciente

Y_{t+1}

Es el pronóstico para el siguiente periodo en el futuro

Antes de profundizar en el análisis de las series temporales es necesario señalar que, para llevarlo a cabo, hay que tener en cuenta los siguientes supuestos:

1. Los datos deben ser homogéneos en el tiempo, o lo que es lo mismo, se debe mantener la definición y la medición de la magnitud objeto de estudio.
2. Se considera que existe una cierta estabilidad en la estructura del fenómeno estudiado.

El objetivo fundamental del estudio de las series temporales es el conocimiento del comportamiento de una variable a través del tiempo para, a partir de dicho conocimiento, y bajo el supuesto de que no van a producirse cambios estructurales, **poder realizar predicciones.**

Indudablemente, la calidad de las previsiones realizadas dependerán, en buena medida, del proceso generador de la serie: así, **si la variable observada sigue algún tipo de esquema o patrón de comportamiento más o menos fijo** (serie determinista) **seguramente obtengamos predicciones más o menos fiables**, con un grado de error bajo. Por el contrario, **si la serie no sigue ningún patrón de comportamiento específico** (serie aleatoria), **seguramente nuestras predicciones carecerán de validez por completo**.

Dentro de los métodos de predicción, se pueden distinguir dos grandes enfoques alternativos:

1. Por un lado, **el análisis univariante de series temporales** mediante el cual se intenta realizar previsiones de valores futuros de una variable, utilizando como información la contenida en los valores pasados de la propia serie temporal. Dentro de esta metodología se incluyen los métodos de descomposición y la familia de modelos ARIMA univariantes.
2. El otro gran bloque dentro de los métodos cuantitativos estaría integrado por el **análisis multivariante o de tipo causal**, denominado así porque en la explicación de la variable o variables objeto de estudio intervienen otras adicionales a ella o ellas mismas.

Enfoques para pronosticar una serie de tiempo

Promedios /media móviles

Suavización exponencial

Descomposición

Promedios/media móviles

Las **medias móviles** (MA, por sus siglas en inglés “Moving Averages”) son, tal como su nombre lo indica, el valor promedio de los datos en un período de tiempo. El término “Móvil” se debe a que sólo son utilizados los datos más recientes en el cálculo.

La **media móvil** es un indicador de tendencias. Se puede decir que este indicador tiene como propósito indicar el comienzo de una tendencia nueva o la finalización de una ya existente o de cambios de dirección. Se utiliza para seguir tendencias, sin embargo no es un indicador que intente predecir o delimitar un pronóstico.

Formula: promedios móviles

$$Y_{t+1} = \frac{Y_t + Y_{t-1} + \cdots + Y_{t-k+1}}{K}$$

En donde:

Y_{t+1} = es el valor pronosticado en el siguiente periodo

Y_t = es el valor real en el periodo t

k = número de términos en el promedio móvil

El método de **pronóstico móvil simple** se utiliza cuando se quiere dar más importancia a conjuntos de datos más recientes para obtener la previsión.

Ejemplo de aplicación de un pronóstico de promedio móvil

Una compañía presenta en el siguiente tabulado el reporte de ventas correspondiente al año 2009:

MES	VENTAS REALES (2009)
Enero	80
Febrero	90
Marzo	85
Abril	70
Mayo	80
Junio	105

El pronóstico restante al ser un pronóstico con un período móvil de 6 meses, este deberá efectuarse a partir del mes de Julio, es decir que para su cálculo tendrá en cuenta seis períodos, es decir, enero, febrero, marzo, abril, mayo y junio.

$$\hat{X}_{7(\text{Julio})} = \frac{80 + 90 + 85 + 70 + 80 + 105}{6}$$

$$\hat{X}_{7(\text{Julio})} = 85$$

Suavización exponencial

La suavización exponencial es una técnica útil para suavizar una serie de tiempo, además es utilizada para conseguir predicciones a corto plazo. **Se caracteriza por dar un mayor peso a los últimos valores de la serie y un menor valor a los primeros. Para ello se da cuenta de un parámetro α** es un parámetro que toma valores comprendidos entre 0 y 1.

Formula: suavizamiento exponencial

$$\hat{x}_t = \hat{x}_{t-1} + (\alpha \cdot (x_{t-1} - \hat{x}_{t-1}))$$

Para efectos académicos siempre suele proporcionarse el factor de suavización

Ejemplo de aplicación de un pronóstico de Suavización Exponencial.

En enero un vendedor de vehículos estimó unas ventas de 142 automóviles para el mes siguiente. En febrero las ventas reales fueron de 153 automóviles. Utilizando una constante de suavización exponencial de 0.20 presupueste las ventas del mes de marzo.

Solución

$$\hat{x}_3 = 142 + (0,2 \cdot (153 - 142))$$

$$\hat{x}_3 = 144$$

Podemos así determinar que el pronóstico de ventas para el período 3 correspondiente a marzo es equivalente a 144 automóviles.

En la ecuación del suavizamiento exponencial, la constante de suavizamiento α sirve como factor de ponderación. El valor de α determina la medida en que la observación actual influye en el pronóstico de la siguiente observación.

Cuando α se acerca a uno, básicamente el nuevo pronóstico será la última observación real. (De forma equivalente, el nuevo pronóstico será el pronóstico antiguo más un ajuste sustancial por el error que haya ocurrido en el pronóstico anterior). A la inversa, cuando α se acerca a cero, el pronóstico nuevo será muy similar al pronóstico anterior y la última observación real tendrá poca importancia.

Descomposición

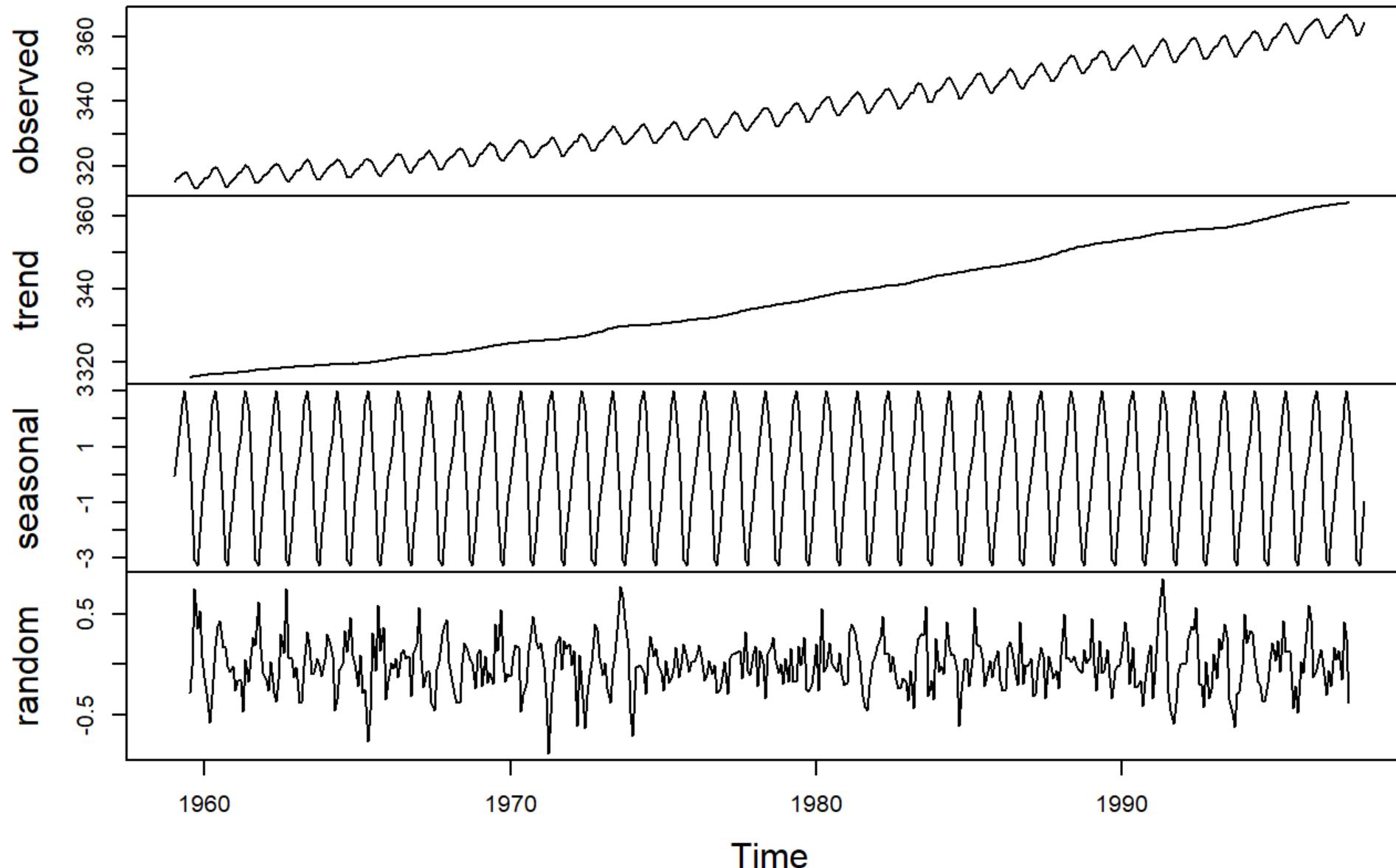
Tradicionalmente, en los métodos de descomposición de series temporales, se parte de la idea de que la serie temporal se puede descomponer en todos o algunos de los siguientes componentes:

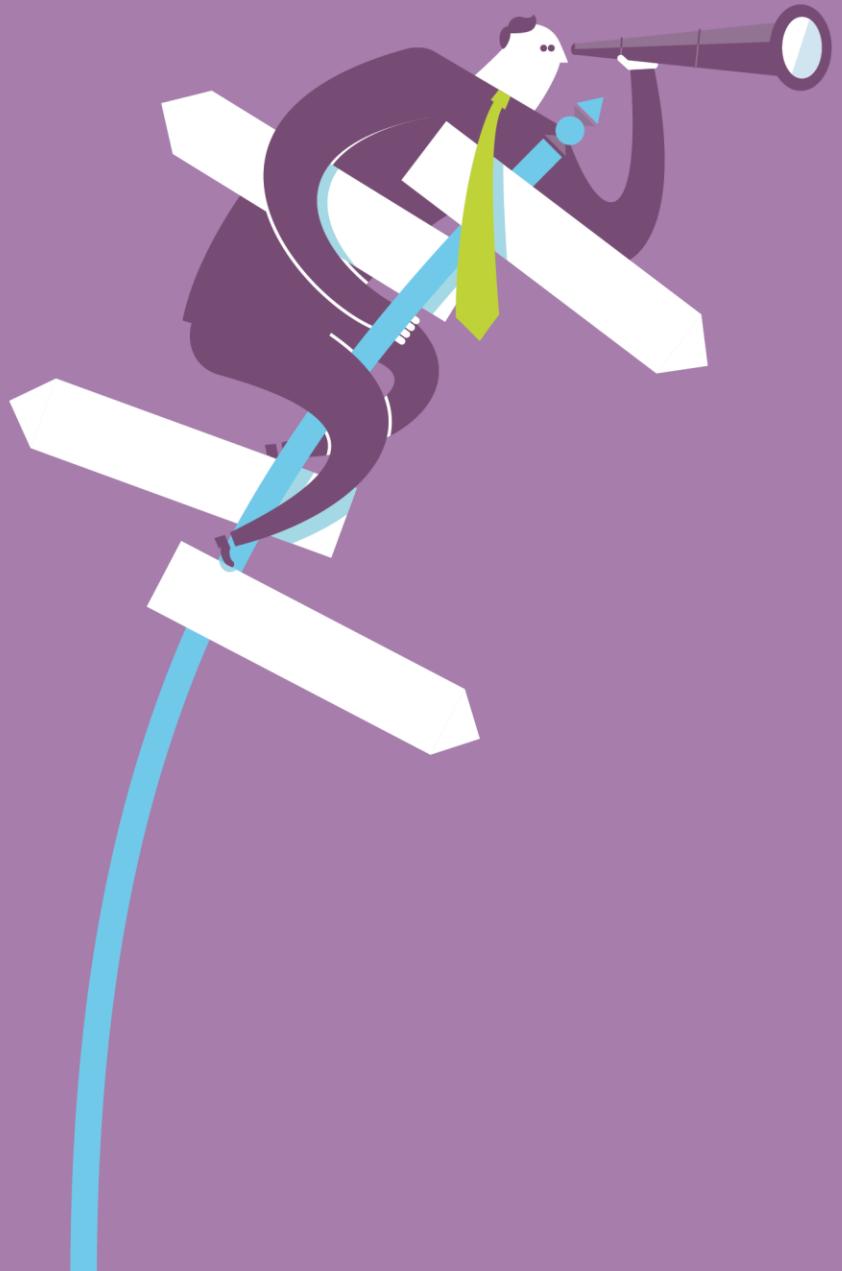
1. **Tendencia (T)**, que representa la evolución de la serie en el largo plazo
2. **Fluctuación cíclica (C)**, que refleja las fluctuaciones de carácter periódico, pero no necesariamente regular, a medio plazo en torno a la tendencia. Este componente es frecuente hallarlo en las series económicas, y se debe a los cambios en la actividad económica.

3. Variación Estacional (S): recoge aquellos comportamientos de tipo regular y repetitivo que se dan a lo largo de un período de tiempo, generalmente igual o inferior a un año, y que son producidos por factores tales como las variaciones climatológicas, las vacaciones, las fiestas, etc.

4. Movimientos Irregulares (I), que pueden ser aleatorios, la cual recoge los pequeños efectos accidentales, o erráticos, como resultado de hechos no previsibles, pero identificables a posteriori (huelgas, catástrofes, etc.)

Decomposition of additive time series





Tema 8. Criterios de estimación de la precisión del pronóstico.

Antes de empezar... ¿qué es un pronóstico?

Según la Real Academia de la Lengua ...

pronóstico

Del lat. *prognosticum*, y este del gr. *προγνωστικόν* *prognōstikón*.

1. m. Acción y efecto de pronosticar.
2. m. Señal por donde se conjetura o adivina algo futuro.

Notación básica para pronósticos

La notación básica para pronósticos se resume a continuación:

Y_t = *Valor de una serie de tiempo en el periodo t*

\hat{Y}_t = *Valor pronosticado del periodo Y_t*

$e_t = Y_t - \hat{Y}_t$

El símbolo $\hat{}$ (circunflejo) se coloca arriba de la letra **Y** para indicar que se trata de un pronóstico o estimación.

Visión general de los problemas de estimación.

El diseño de un pronóstico consiste en **construir una función real** que, a partir del valor de unas determinadas variables de observación, **proporcione predicciones acerca de una variable objetivo**.

A modo de ejemplo, considérese **la producción de energía en una planta nuclear**. Con el fin de maximizar el beneficio de explotación **resulta muy deseable adecuar la generación de energía a la demanda real**, ya que la **capacidad de almacenamiento de la energía no consumida es muy limitada**.

Visión general de los problemas de estimación

Por lo tanto, en este contexto **es muy importante considerar estos problemas de estimación**, pues en el caso anterior, **los errores en que se incurre al realizar el pronóstico acarrean determinadas penalizaciones**.

En este sentido, **el objetivo perseguido en el diseño de un pronóstico suele ser la minimización de dicha penalización**.

Criterios de estimación de la precisión del pronóstico

Existen varios métodos, cuya finalidad es estimar los errores generados por una técnica específica de pronósticos. La mayoría de estas medidas requieren calcular el promedio de las diferencias entre el valor real de una función y su valor pronosticado; tales diferencias se conocen como **residuos**, dicho de otra forma, estimar los errores de una técnica de pronóstico requiere determinar el promedio de los residuos.

Criterios de estimación de la precisión del pronóstico

En la actualidad hay algunas formas para estimar el rendimiento y evaluar el ajuste de un pronóstico modelo, algunas de ellas son: el error cuadrático medio (RMSE, por sus siglas en inglés, root mean squared error), error absoluto medio (MAE, mean absolute error), R-cuadrado.

Ejemplo

Por ejemplo, supongamos que en diciembre de 2015 se pronosticó la inflación para enero de 2016, y supongamos que este pronóstico fue de $\hat{I}_{E2016} = 2.98\%$ luego en enero de 2016, una vez que todos los agentes económicos han afectado el precio de los bienes y servicios se puede determinar el valor real de la inflación, supongamos que en este caso fue de $I_{E2016} = 3.01\%$, de lo anterior tenemos que el residuo o error de pronóstico es:

$$e_{E2016} = I_{E2016} - \hat{I}_{E2016} = 3.01 - 2.98 = 0.03\%$$

Desviación Absoluta Media (DAM)

La **desviación absoluta media** (DAM) mide la exactitud de un pronóstico, promediando las magnitudes de los errores de pronóstico (los valores absolutos de los errores). La siguiente ecuación muestra cómo se calcula DAM:

$$DAM = \frac{1}{n} \sum_{i=1}^n |Y_t - \hat{Y}_t|$$

Error Cuadrático Medio (ECM)

Cada error o residuo se eleva al cuadrado, luego se suman y se dividen entre el número de observaciones. Este enfoque sanciona errores grandes en la elaboración de pronósticos, ya que los errores se elevan al cuadrado, lo cual es importante, porque una técnica que produce errores moderados quizá sea preferible a una que usualmente tenga pequeños errores.

La siguiente ecuación muestra cómo se calcula el ECM:

$$ECM = \frac{1}{n} \sum_{i=1}^n (Y_y - \hat{Y}_t)^2$$

Los valores más bajos de RMSE indican un mejor ajuste.

Error Absoluto Medio (EPAM)

El **error porcentual absoluto medio** (EPAM) se calcula obteniendo el valor absoluto de los residuos de cada periodo y dividiendo éste entre el valor real observado en dicho periodo y promediando estos errores porcentuales absolutos. El resultado final se multiplica después por 100 y se expresa como porcentaje. El EPAM es especialmente útil cuando los valores de Y_t son grandes.

$$EPAM = \frac{1}{n} \sum_{i=1}^n \frac{|Y_t - \hat{Y}_t|}{Y_t}$$

Error Porcentual Medio (EPM)

El **error porcentual medio (EPM)**, se calcula obteniendo el error de cada periodo, dividiendo éste entre el valor real de ese periodo y luego promediando estos errores. El resultado usualmente se multiplica por 100 y se expresa como porcentaje. Si el enfoque del pronóstico no tiene sesgo, el EPM producirá un resultado que está cercano a cero.

$$\text{EPM} = \frac{1}{n} \sum_{i=1}^n \frac{(Y_t - \hat{Y}_t)}{Y_t}$$

Formulas

Desviación Absoluta Media

$$DAM = \frac{1}{n} \sum_{i=1}^n |Y_t - \hat{Y}_t|$$

Error Cuadrático Medio

$$ECM = \frac{1}{n} \sum_{i=1}^n (Y_t - \hat{Y}_t)^2$$

Error Porcentual Absoluto Medio

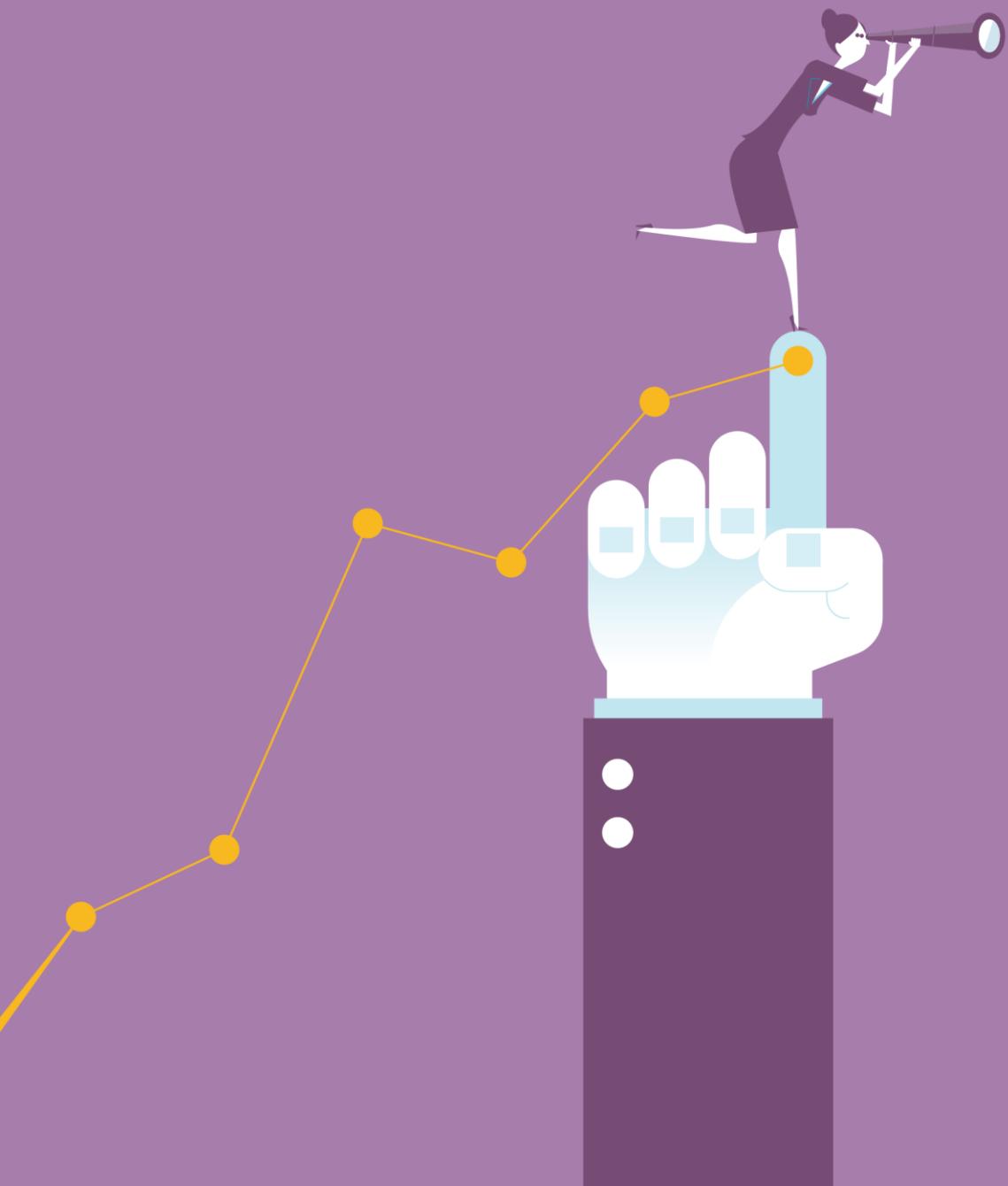
$$EPAM = \frac{1}{n} \sum_{i=1}^n \frac{|Y_t - \hat{Y}_t|}{Y_t}$$

Error Porcentual Medio

$$EPM = \frac{1}{n} \sum_{i=1}^n \frac{(Y_t - \hat{Y}_t)}{Y_t}$$

Tiempo	Valor real Yt	Pronóstico Ygorro	Error del pronóstico	Valor absoluto del error	Error cuadrático	Error absoluto porcentual	Error porcentual
1	58	-	-				
2	54	58	-4	4	16	0.074074074	-0.074074074
3	60	54	6	6	36	0.1	0.1
4	55	60	-5	5	25	0.090909091	-0.090909091
5	62	55	7	7	49	0.112903226	0.112903226
6	62	62	0	0	0	0	0
7	65	62	3	3	9	0.046153846	0.046153846
8	63	65	-2	2	4	0.031746032	-0.031746032
9	70	63	7	7	49	0.1	0.1
		12	34	188	0.555786269	0.162327875	
n=	8						
$DAM = \frac{1}{n} \sum_{i=1}^n Y_t - \hat{Y}_t $							
Desviación absoluta media							
$ECM = \frac{1}{n} \sum_{i=1}^n (Y_t - \hat{Y}_t)^2$							
ECM							
$EPM = \frac{1}{n} \sum_{i=1}^n \frac{(Y_t - \hat{Y}_t)}{Y_t}$							
EPM							
$EPAM = \frac{1}{n} \sum_{i=1}^n \frac{ Y_t - \hat{Y}_t }{Y_t}$							
EPAM							

Tema 9. Estimación de coeficientes por el método de mínimos cuadrados y análisis de correlación.



Análisis causal

Uno de los objetivos principales en la estadística y en la toma de decisiones es realizar análisis causal, es decir:

- Analizar cualitativa y cuantitativamente como ciertos factores afectan a una variable asociada a un fenómeno de interés.

El análisis causal permite:

Caracterizar y cuantificar la relación de comportamiento entre variables, de acuerdo con lo que sugiere la teoría. En economía se suelen hacer análisis de regresión basándose en teoría económica.

El conjunto de técnicas que se utilizarán para construir y evaluar modelos que describen la relación entre variables que permitan formular inferencias basadas en los modelos obtenidos se conocen como **técnicas de regresión**, mientras que al análisis estadístico que resulta de aplicarlas se le denomina **análisis de regresión**.

Mediante el modelo de regresión lineal simple se pretende describir la relación entre dos variables, una llamada variable independiente o predictora y otra llamada variable dependiente, además de realizar inferencias sobre el comportamiento de la variable dependiente. En lo sucesivo se denotará por **X** a la variable independiente y por **Y** a la dependiente.

Tipos de análisis de regresión y modelos econométricos.

Dependiendo del problema a analizar, del interés de investigador y de la naturaleza de los datos, se formulan diferentes tipos de modelos de regresión:

Algunos tipos de modelos son:

- **Mínimos cuadrados ordinarios** – análisis de regresión simple
- **Modelos probit y logit**: integra variables dicotómicas como variable dependiente
- **Modelos no lineales**
- **Modelos de regresión espacial**: integra el concepto de correlación espacial.

El modelo de regresión lineal simple.

Un **modelo de regresión** es un modelo que permite describir cómo influye una variable X sobre otra variable Y.

Y: **Variable dependiente** o respuesta o endógena

X: **Variable independiente** o explicativa o exógena

El objetivo es obtener estimaciones razonables de Y para distintos valores de X a partir de una muestra de n pares de valores.

El modelo de regresión lineal simple

El **modelo de regresión lineal** simple supone que,

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

donde:

y_i representa el valor de la variable dependiente para la observación i.

x_i representa el valor de la variable independiente para la observación i.

u_i representa el error para la observación i que se asume normal,

$$u_i \sim N(0, \sigma)$$

El **objetivo** de este modelo es analizar la relación entre dos variables, X y Y, ambas de carácter cuantitativo. Para definir las relaciones entre las variables X y Y se necesita conocer el valor de los coeficientes β_0 y β_1 del modelo lineal; sin embargo, estos coeficientes son parámetros poblacionales, los cuales son casi siempre desconocidos. Para estimarlos, se extrae una muestra aleatoria de la población de interés y se calculan las estadísticas muestrales que se necesitan.

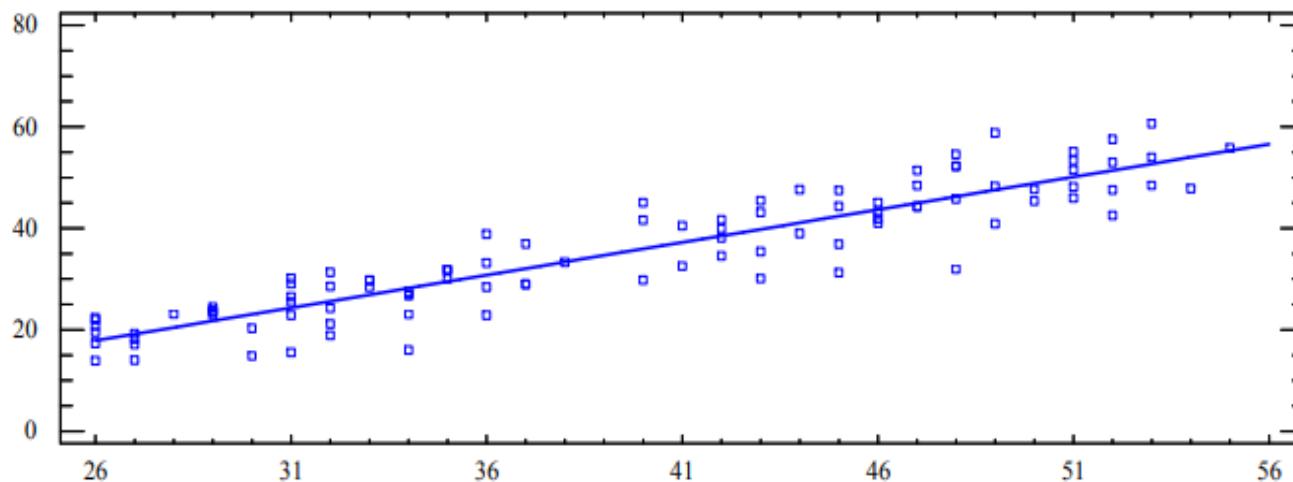
β_0 y β_1 son los coeficientes de regresión:

β_0 : intercepto

β_1 : pendiente

El objetivo es obtener estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

que se ajuste lo mejor posible a los datos.



Hipótesis del modelo de regresión lineal simple

Linealidad: La relación existente entre X e Y es lineal

$$f(x) = \beta_0 + \beta_1 x$$

Homogeneidad: El valor promedio del error es cero, $E[u_i] = 0$

Homocedasticidad: La varianza de los errores es constante

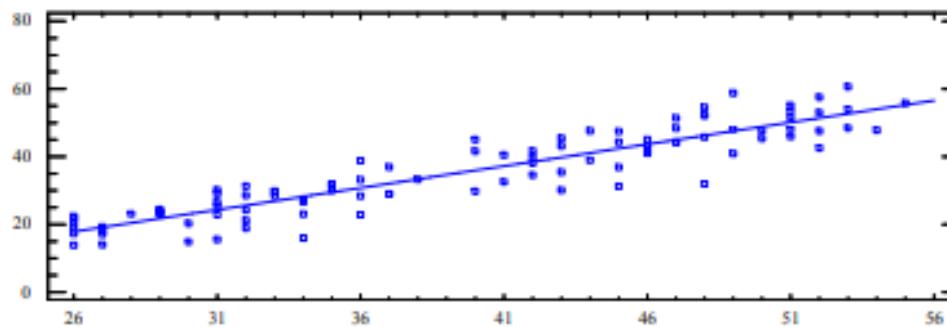
Independencia: Las observaciones son independientes

Normalidad: Los errores siguen una distribución normal

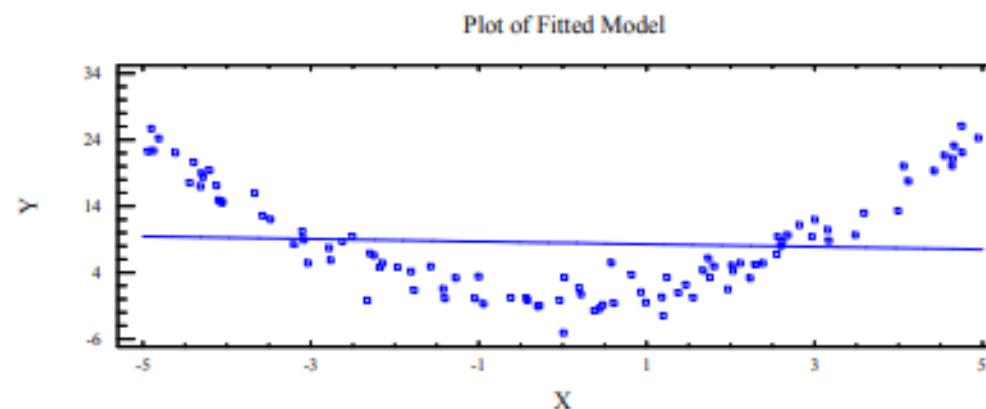
$$u_i \sim N(0, \sigma)$$

Hipótesis del modelo de regresión lineal simple

Linealidad: Los datos deben ser razonablemente rectos:

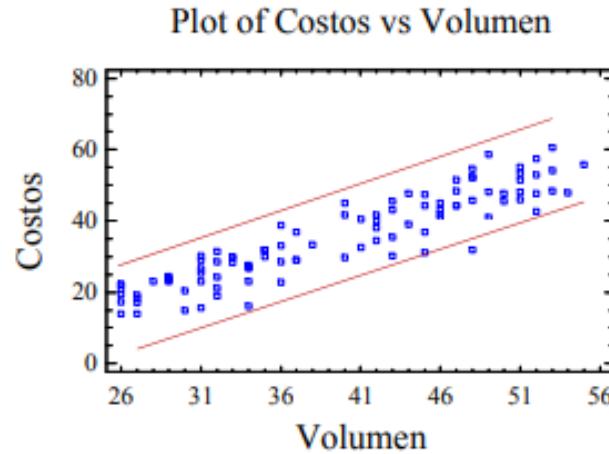


Si no, la recta de regresión no representa la estructura de los datos.

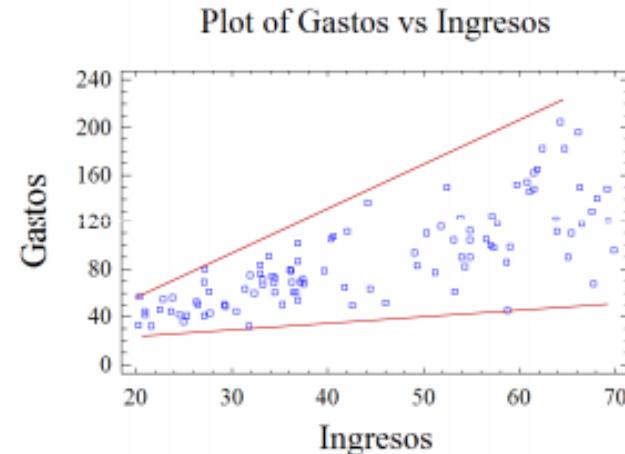


Homocedasticidad

La dispersión de los datos y residuos debe ser constante:



Datos homocedásticos



Datos heterocedásticos

Independencia

Los datos deben ser independientes

Una observación no debe dar información sobre las demás.

En general, las series temporales no cumplen la hipótesis de independencia.

Ejemplo: un modelo de regresión para los determinantes económicos de la felicidad en Latinoamerica

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$$

donde

Y_i = índice de felicidad para cada país i

X_1 = representa la carga fiscal para cada país i

X_2 = representa el producto interno bruto para cada país

X_3 = representa el índice de libertad económica por país

ε_i = representa el término de error

Resultados

```
. regress felicidad PIB impuestos BF
```

Source	SS	df	MS	Number of obs	=	265
Model	82.3683909	3	27.4561303	F(3, 261)	=	67.97
Residual	105.424805	261	.403926455	Prob > F	=	0.0000
				R-squared	=	0.4386
Total	187.793196	264	.711337863	Adj R-squared	=	0.4322
				Root MSE	=	.63555

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
PIB	.0000108	4.58e-06	2.36	0.019	1.78e-06 .0000198
impuestos	-.0264122	.0066139	-3.99	0.000	-.0394356 -.0133887
BF	.0284101	.0042827	6.63	0.000	.019977 .0368431
_cons	6.298143	.5555101	11.34	0.000	5.204291 7.391995

Coeficiente de determinación: R2

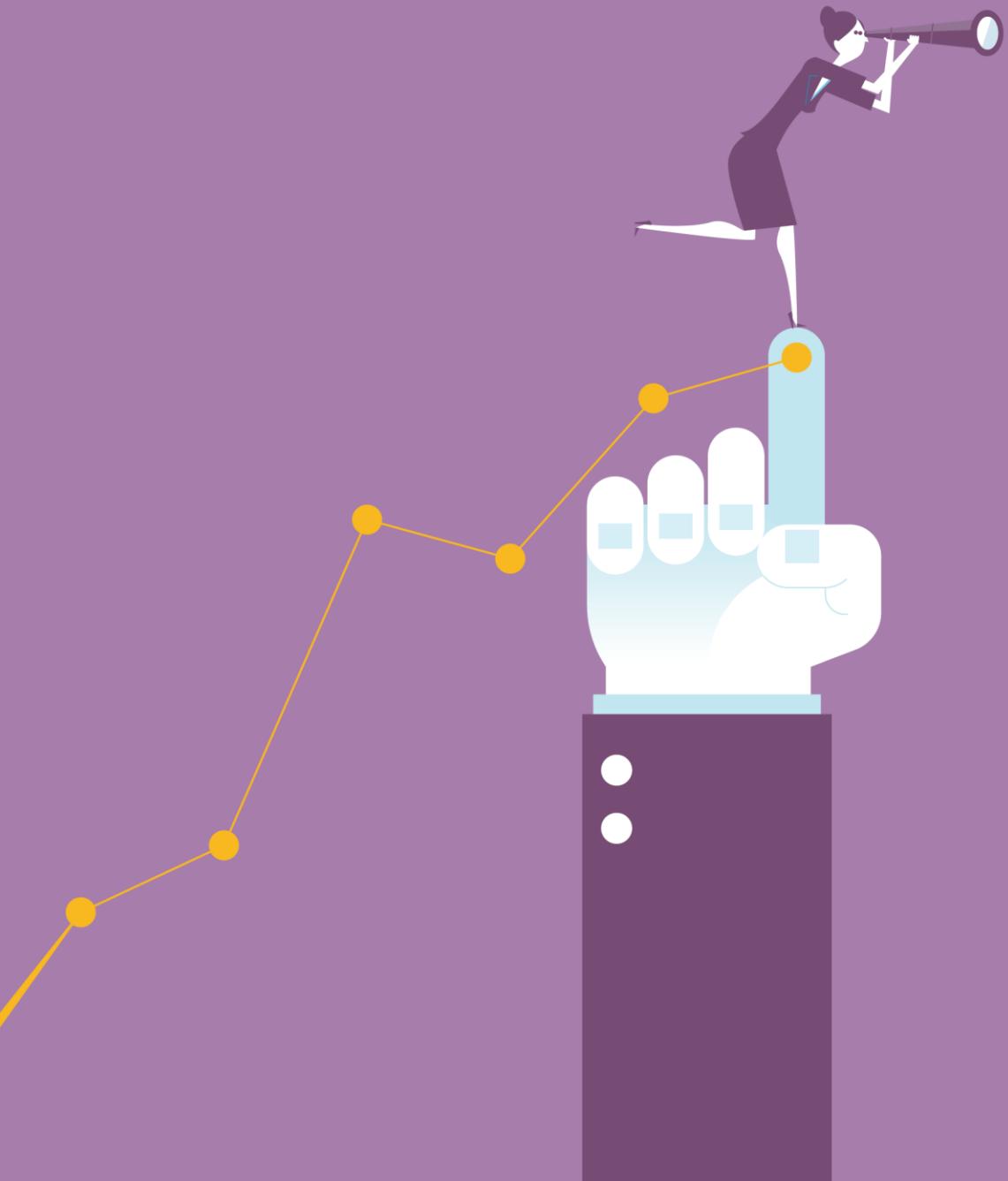
El **coeficiente de determinación**, R-cuadrado se emplea para valorar la bondad de ajuste del modelo. Se define como:

$$R^2 = r_{(x,y)}^2 \in [0, 1]$$

nos indica el porcentaje de la variabilidad muestral de la variable y que es explicada por el modelo, esto es, por su dependencia lineal de x

Los valores próximos a 60% indican que el modelo de regresión proporciona un buen ajuste para los datos.

Tema 10. Inferencia estadística: contraste de hipótesis e intervalos de confianza.



Evaluación del modelo

El **método de mínimos cuadrados produce la mejor línea recta**. Sin embargo, puede no haber relación entre las variables o quizás esta relación sea de otro tipo. Si es así, el modelo será probablemente inapropiado.

Consecuentemente, es importante evaluar cómo el modelo lineal se ajusta a los datos. Si el ajuste es pobre, se debería descartar el modelo y buscar otro.

Para medir **la confiabilidad de la ecuación de estimación**, se utiliza el error estándar de la estimación, el cual mide la variabilidad o dispersión de los valores observados alrededor de la recta de regresión.

El **error estándar de estimación** se denota por σ_{ε} . Si σ_{ε} es grande implica que algunos de los errores son grandes, por lo tanto, el ajuste del modelo es pobre. Si σ_{ε} es pequeño, entonces los valores observados se encuentran cerca de la recta de regresión, lo cual da como resultado un modelo que se ajusta bien a los datos experimentales.

Ecuación para calcular el error estándar

$$s_e = \sqrt{\frac{\sum Y^2 - b_0 \sum Y - b_1 \sum XY}{n - 2}}$$

En donde:

X = valores de la variable independiente

Y = valores de la variable dependiente

b_0 = ordenada al origen

b_1 = pendiente de la ecuación de regresión

n = número de puntos utilizados para ajustar la línea de regresión

Formulas importantes en el análisis de regresión

Calculo de coeficiente beta 1

$$B_1 = \frac{\sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

Calculo de coeficiente beta 0 (intercepto)

$$B_0 = \bar{Y} - B_1 \bar{X}$$

Calculo de R cuadrado

$$R^2 = 1 - \frac{\text{Suma de cuadrados de los residuos}}{\text{Suma de cuadrados total}}$$

Ejemplo: antigüedad vs reparación

$$y_i = \beta_0 + \beta_1 X_1$$

y_i = Gastos de reparación
 X_i = antigüedad del camión

Antigüedad del camión en años (X)	Gastos de reparación durante el último año en miles (Y)	XY	X^2	Y^2
5	7	35	25	49
3	7	21	9	49
3	6	18	9	36
1	4	4	1	16
$\sum X = 12$	$\sum Y = 24$	$\sum XY = 78$	$\sum X^2 = 44$	$\sum Y^2 = 150$

$$\begin{aligned}
 S_{\epsilon} &= \sqrt{\frac{\sum Y^2 - b_0 \sum Y - b_1 \sum XY}{n - 2}} \\
 &= \sqrt{\frac{150 - (3.75)(24) - (0.75)(0.78)}{4 - 2}} \\
 &= \sqrt{\frac{150 - 90 - 58.5}{2}} \\
 &= \sqrt{0.75} \\
 &= 0.8660 \leftarrow \text{Error estándar}
 \end{aligned}$$

. regress Y X						
Source	SS	df	MS	Number of obs	=	4
Model	4.5	1	4.5	F(1, 2)	=	6.00
Residual	1.5	2	.75	Prob > F	=	0.1340
Total	6	3	2	R-squared	=	0.7500
				Adj R-squared	=	0.6250
				Root MSE	=	.86603
Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
X	.75	.3061862	2.45	0.134	-.567413	2.067413
_cons	3.75	1.015505	3.69	0.066	-.6193645	8.119365

no hay evidencia que indique que existe relación entre la edad del camión y los gastos de reparación.

Estadístico de contraste:
No son significativas para nuestro análisis

Interpretación: Cuando S_{ϵ} es pequeño, el ajuste es excelente y el modelo lineal es probablemente una herramienta efectiva para el pronóstico. Si S_{ϵ} es grande, el modelo es pobre y probablemente el modelo debería mejorarse o descartarse.

El **error estándar** S_{ϵ} es útil en la comparación de modelos. Si se obtienen varios modelos, se elegirá el que tenga el menor valor de S_{ϵ} .

Prueba de hipótesis de la pendiente β_1

El proceso para probar la hipótesis acerca de β_1 es el mismo que el proceso para cualquier otro parámetro, a saber:

1. Establecimiento de hipótesis:

$$H_0 : \beta_1 = 0 \text{ en oposición a } H_a : \beta_1 \neq 0$$

La hipótesis nula especifica que no existe relación lineal entre X y Y, lo cual implica que la pendiente es cero.

2. Estadística de prueba:

$$t_{\text{calculada}} = \frac{b_1 - \beta_1}{s_{b_1}}$$

En donde s_{b_1} es error estándar de b_1 y se define como:

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - n(\bar{x})^2}}$$

3. Establecer la región de rechazo α .

4. Regla de decisión: Rechazar H_0 si $|t_{\text{calculada}}|$ es mayor que $t_{\alpha/2}(n - 2)$.

5. Conclusión, en el contexto del problema.

Intervalos de confianza para β_1

Un intervalo de confianza al 100 $(1 - \alpha)\%$ para la pendiente b_1 de la regresión lineal simple.

$$b_1 \pm t^* \frac{s_\varepsilon}{\sqrt{\sum x^2 - n(\bar{X})^2}}$$

con $t^* t_{\alpha/2, (n - 2)}$

Tarea interpretar: felicidad en países asiáticos

```
. regress felicidad PIB impuestos BF
```

Source	SS	df	MS	Number of obs	=	121
Model	45.3130005	3	15.1043335	F(3, 117)	=	54.51
Residual	32.4177759	117	.277075008	Prob > F	=	0.0000
				R-squared	=	0.5829
				Adj R-squared	=	0.5723
Total	77.7307764	120	.64775647	Root MSE	=	.52638

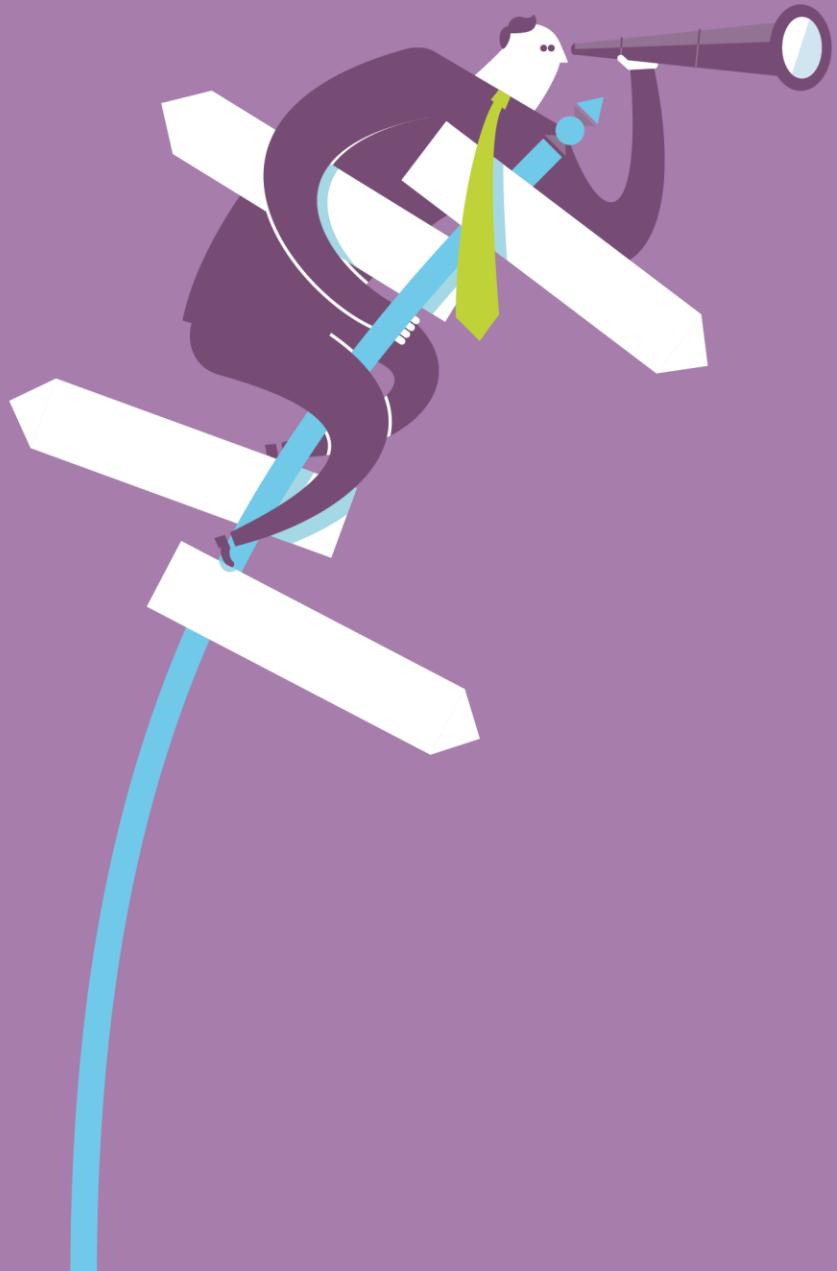
felicidad	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
PIB	.0198184	.0173983	1.14	0.257	-.014638 .0542747
impuestos	-.0207977	.0067393	-3.09	0.003	-.0341445 -.0074509
BF	.0376636	.0031797	11.85	0.000	.0313664 .0439608
_cons	4.30835	.5661711	7.61	0.000	3.187078 5.429622

$$\text{Ecuación: } y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$$

X1 = PIB, X2 = Impuestos, X3 = BF (libertad económica)

Modulo 3: Regresión lineal múltiple.





Tema 11: Predicción y análisis de residuos.

Recapitulando

Mediante el método de mínimos cuadrados se obtiene la mejor línea recta que ajusta a los datos.

La línea de regresión puede usarse para estimar el valor de Y para un valor determinado X.

Sin embargo, si no se cumplen los supuestos. Podemos caer en algo llamado: regresión espuria.

Una relación **espuria** se refiere a la apariencia en que existe una relación de causalidad entre variables cuando en realidad esta no existe.

Existen dos fuentes de incertidumbre asociadas con una predicción puntual generada por la ecuación de regresión adaptada:

1. Incertidumbre debida a la dispersión de los datos respecto a la línea de regresión de la muestra.
2. Incertidumbre debida a la dispersión de la muestra respecto a la línea de regresión de la población.

Es posible elaborar un intervalo de predicción de Y que tome en cuenta estas dos fuentes de incertidumbre.

El error estándar del pronóstico mide la variabilidad de Y prevista sobre la Y real para un valor determinado de X . El error estándar del pronóstico está dado por la expresión:

$$S_{\varepsilon}^2 \left(\frac{1}{n} + \frac{(x_0 + \bar{x})^2}{\sum x^2 - n(\bar{x})^2} \right)$$

El primer término, S_{ε}^2 , mide la dispersión de los datos sobre la línea de regresión de la muestra (primera fuente de incertidumbre).

El segundo término:

$$\sqrt{s_{\varepsilon}^2 + s_{\varepsilon}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x^2 - n(\bar{x})^2} \right)}$$

Mide la dispersión de la línea de regresión de la muestra sobre la línea de regresión de la población (segunda fuente de incertidumbre).

Supuestos del análisis de regresión y análisis de los residuos

El hecho de ajustar un modelo por mínimos cuadrados, construir intervalos de predicción y probar hipótesis, no completa el estudio de regresión.

En la mayoría de los estudios las inferencias pueden ser seriamente engañosas si los supuestos elaborados en la formulación del modelo son extremadamente incompatibles con los datos. Es esencial verificar cuidadosamente los datos para evitar violaciones de los supuestos.

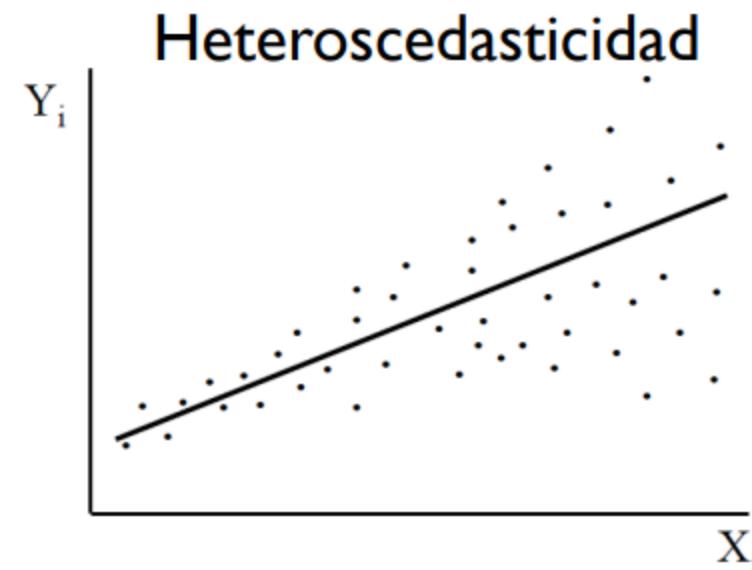
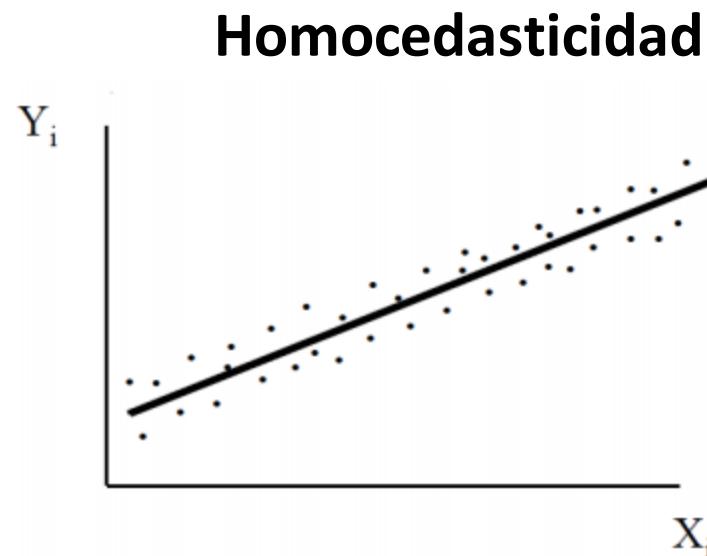
Supuestos del modelo de regresión

Linealidad: el primer principio es que debe existir una relación lineal entre X y Y. En caso contrario, debería de usarse un modelo no lineal.

Independencia: Los datos y los términos de error e son independientes uno del otro. Para poder probar este supuesto, existe el test Durbin – Watson. El valor de este test siempre está entre 0 y 4. Si el DW se ubica entre 1.5 y 2.5 se puede asumir que existe independencia de los datos.

Normalidad: El término de error se distribuye como una función de densidad de probabilidad normal con media cero y varianza constante. Cuando no se cumple el supuesto de normalidad se pierde eficiencia en los coeficientes beta.

Homocedasticidad: Tanto los datos como las varianzas de los errores deben ser constante.



La principal consecuencia de la **heteroscedasticidad** es que **los estimadores pierden eficiencia**.

Un análisis para detectar homocedasticidad es hacer algunos gráficos simples de los datos y ver como se distribuyen, sin embargo, tambien existe la **Prueba Breusch-Pagan** de heteroscedasticidad.

Donde la hipótesis nula es que existe homocedasticidad. Y la hipótesis alternativa es que existe heterocedasticidad.

Para verificar los méritos de un modelo tentativo, se pueden examinar diversas gráficas de residuales y test de comprobación.

¿Cómo identificar el mejor modelo?

Consistencia con la teoría (teoría económica o de la ciencia que se vaya a contrastar)

Coherencia con los datos y los supuestos.

Criterios	Pruebas generales	Pruebas y test
Coherencia con los datos y supuestos	Independencia Homocedasticidad	Durbin – Watson Prueba de White, Breusch - Pagan
Teoría económica	Valores de coeficientes	P - Valor
	Correlación	

Consistencia con la teoría económica. Implica que el modelo estimado debe satisfacer las restricciones impuestas por la teoría económica sobre la especificación inicial y los valores de los coeficientes.

Coherencia con los datos y los supuestos. El modelo debe reproducir adecuadamente el comportamiento de los datos y los supuestos. De manera que el modelo no debe presentar autocorrelación y heteroscedasticidad.

Ejemplo: relación entre millas que recorre un auto y el peso del auto

El conjunto de datos que utilizaremos es un conjunto de datos sobre automóviles antiguos de 1978 en los Estados Unidos.

$$y_i = \beta_0 + \beta_1 weight + \beta_2 price_2 + \beta_3 foreign_3 + \varepsilon_i$$

donde y_i es la variable dependiente que son las millas recorridas versus weight (peso del auto), price (precio) y foreign (si el auto es extranjero)

Resultados de la estimación

Source	SS	df	MS	Number of obs	=	26
Model	382.079636	3	127.359879	F(3, 22)	=	15.25
Residual	183.766518	22	8.35302354	Prob > F	=	0.0000
Total	565.846154	25	22.6338462	R-squared	=	0.6752
				Adj R-squared	=	0.6309
				Root MSE	=	2.8902

mpg	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
weight1	-7.121111	1.604674	-4.44	0.000	-10.449	-3.793222
price	.0002258	.0002654	0.85	0.404	-.0003245	.0007761
foreign	-2.507127	2.056569	-1.22	0.236	-6.772189	1.757935
_cons	42.1662	4.264753	9.89	0.000	33.32164	51.01075

Prueba de Breusch-Pagan

Source	chi2	df	p
Heteroskedasticity	12.71	8	0.1223
Skewness	4.59	3	0.2044
Kurtosis	1.94	1	0.1637
Total	19.24	12	0.0829

Prueba de Breusch-Pagan: Hipótesis de heterocedasticidad: es mayor que el valor 0.05 por lo tanto se rechaza la hipótesis nula de presencia de homocedasticidad.

Skewness = Asimetría



Kurtosis = Observaciones anómalas.