



# 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 limites.

# 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](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* (9ª 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.solveymath.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

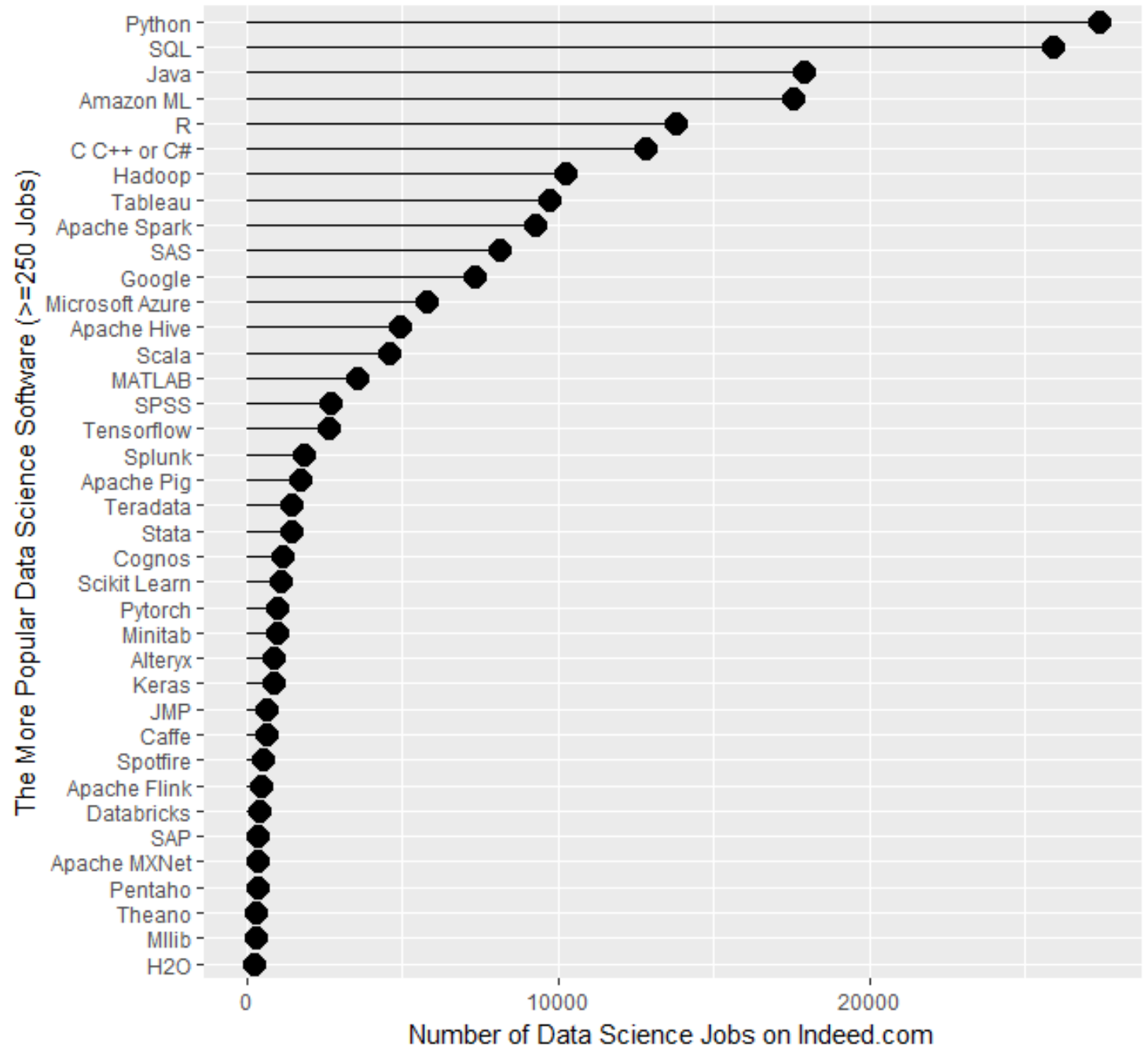
## Software estadístico y lenguaje de programación



## Hojas de cálculo



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



# Algunas empresas que utilizan lenguaje de programación estadística...

Bank of America

Microsoft

Zillow

NETFLIX

bing



JOHN DEERE

FDA



UBER

Culture Amp

cfpb

Consumer Financial  
Protection Bureau



FOURSQUARE



AMAI

INTELIGENCIA APLICADA  
A DECISIONES

KICKSTARTER

Google



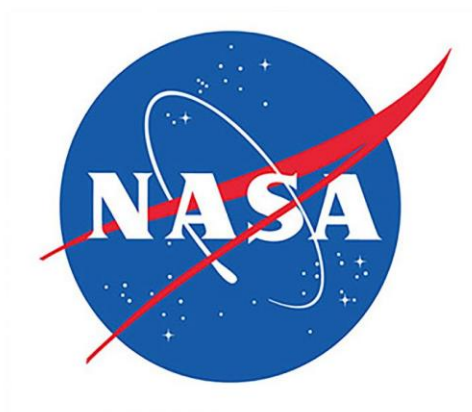
airbnb

The New York Times



The  
Economist

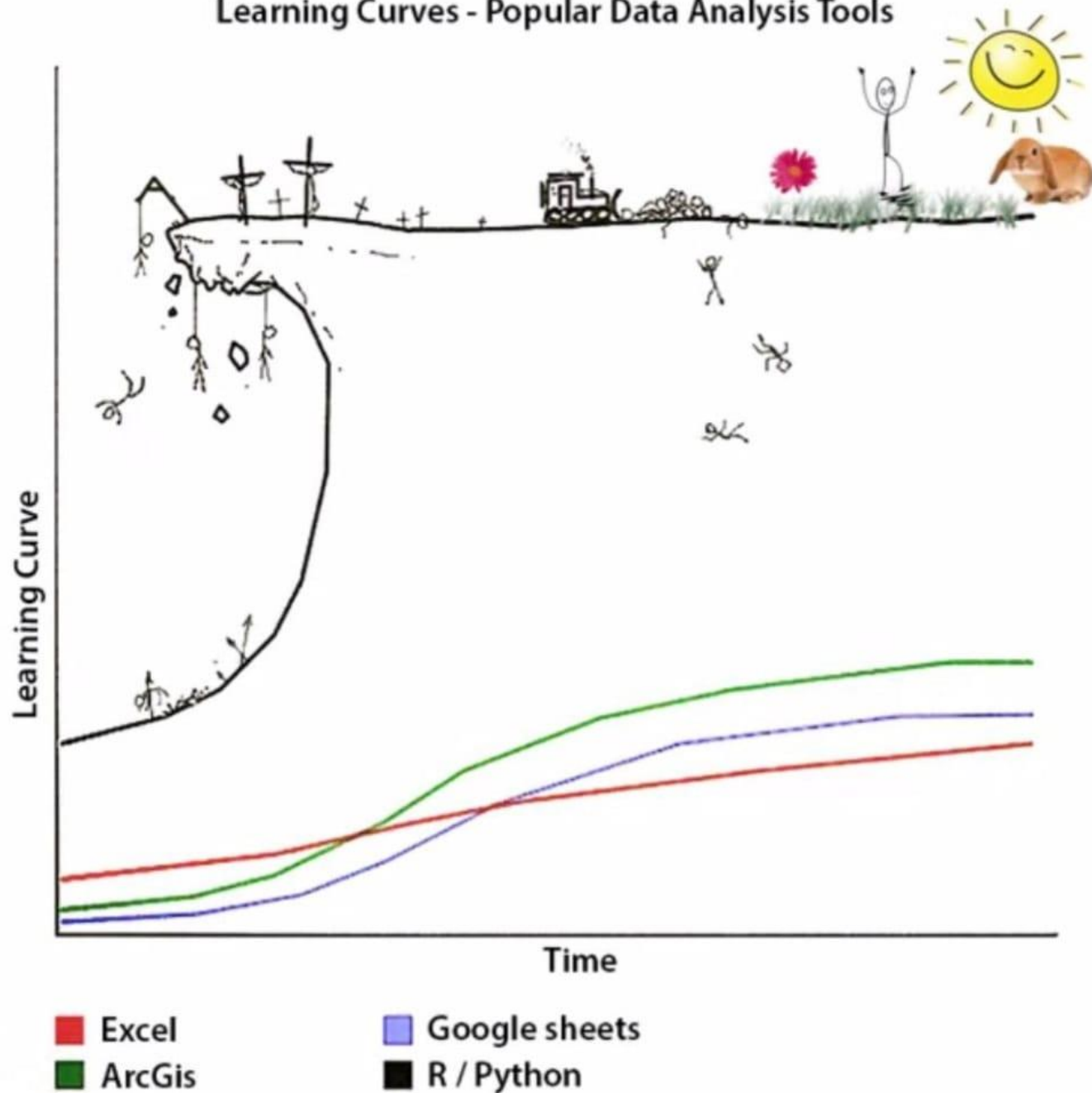
# Algunas organizaciones públicas que utilizan lenguaje de programación estadística...



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



## 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?

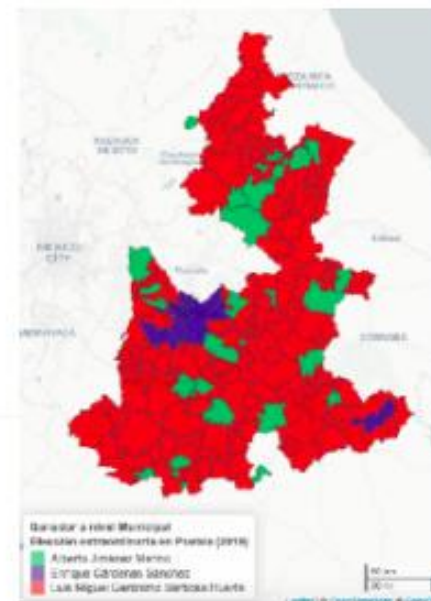
Esto lo vamos a ver  
en clase



## 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?

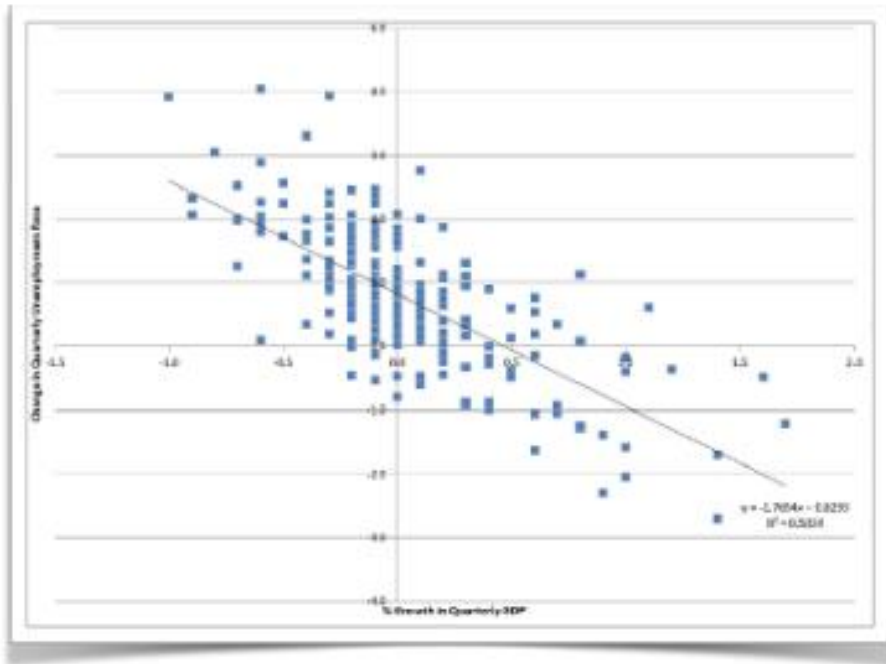
Esto lo vamos a ver  
en clase



## 2. Análisis estadístico y econometría.

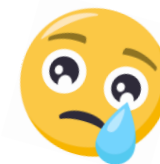
`library(base)`

`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)
```

Classification



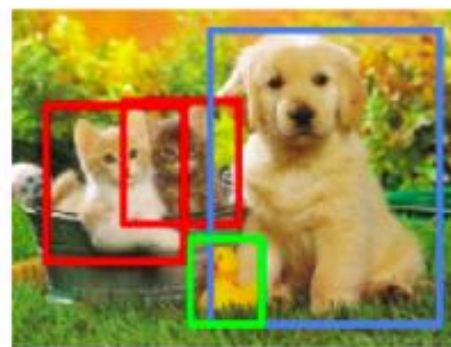
CAT

Classification  
+ Localization



CAT

Object Detection



CAT, DOG, DUCK

Instance  
Segmentation



CAT, DOG, D

Single object

Multiple objects

# ¿Qué se puede hacer con R?

Esto lo vamos a ver  
en clase



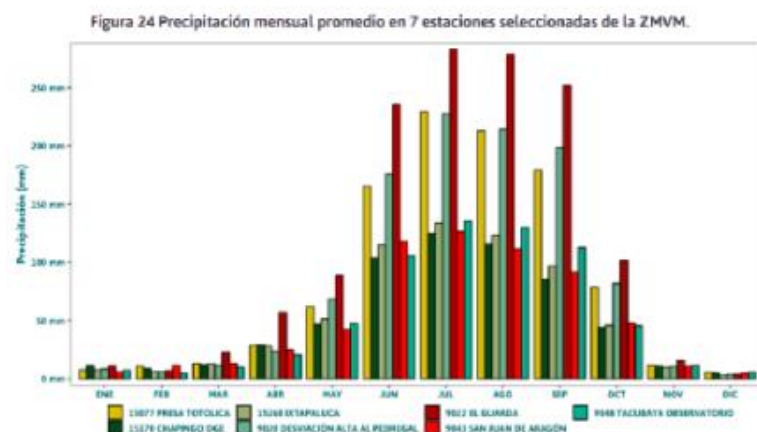
## 4. Visualización de datos.

```
library(ggplot2)
```

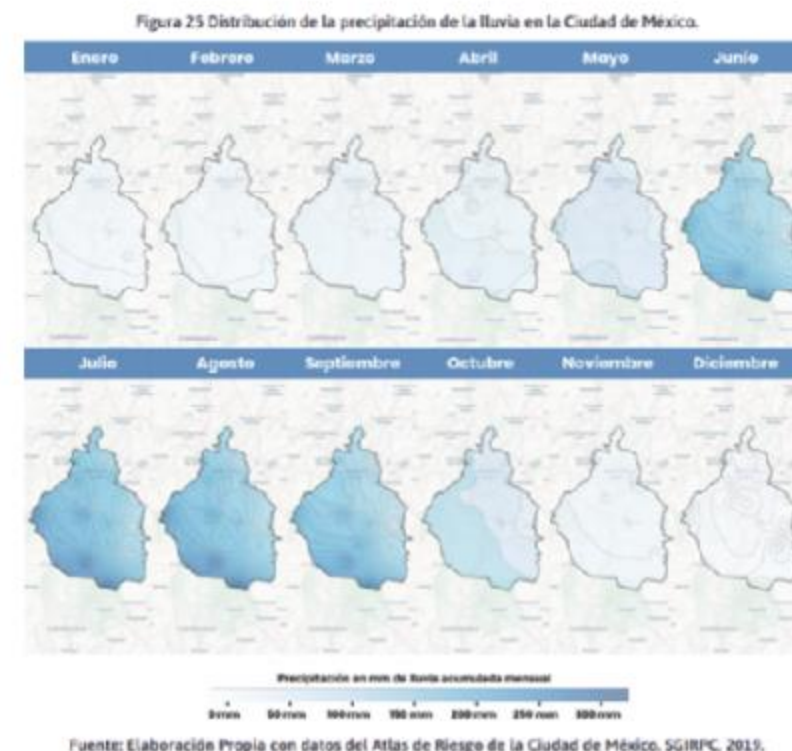
```
library(plotly)
```

```
library(leaflet)
```

```
library(htmlwidgets)
```



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



Fuente: Elaboración Propia con datos del Atlas de Riesgo de la Ciudad de México, SGIRPC, 2019.

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



# ¿Qué se puede hacer con R?

**Esto lo vamos a ver en clase**



***Si nos da el tiempo ...***

## 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.

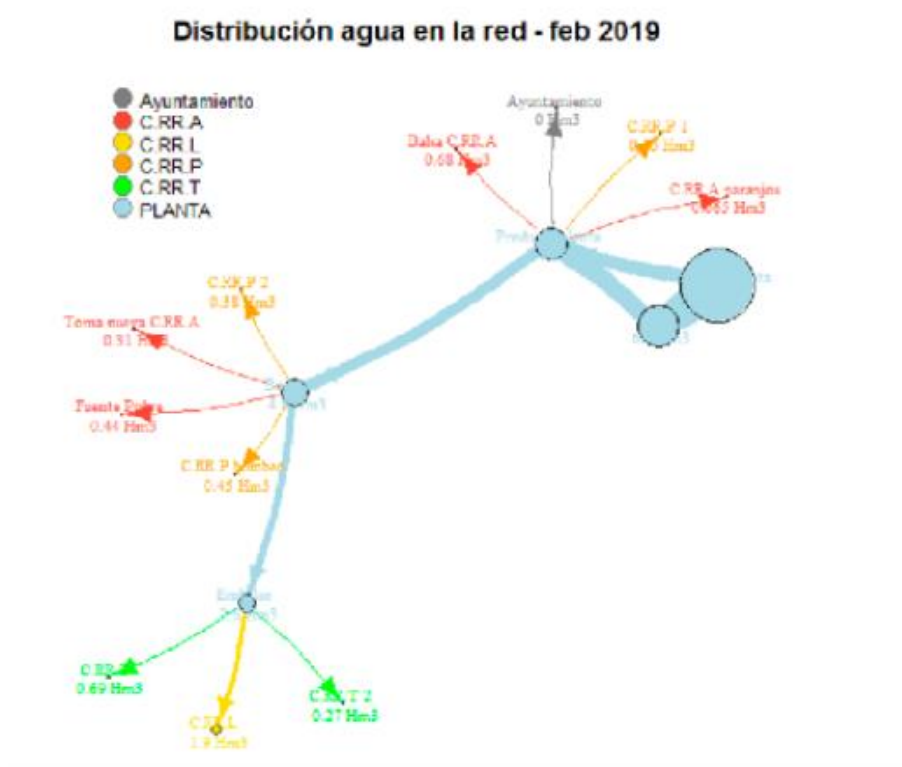
# ¿Qué se puede hacer con R?

**Esto no lo vamos a ver en clase**

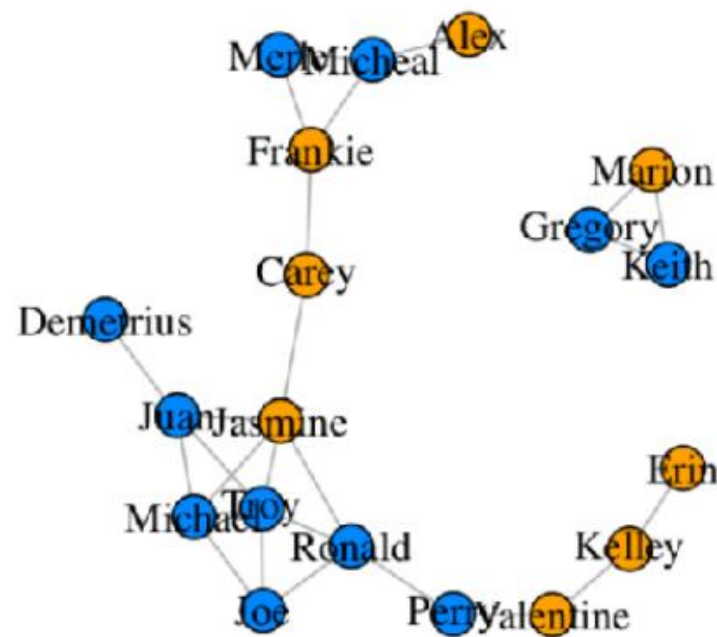


## 6. Análisis de redes.

## library(igraph)



## Red de distribución de Agua



## 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

Best departing flights				
Total price includes taxes + fees for 1 adult. Additional bag fees and other fees may apply.				
	06:05 – 14:35 <sup>1</sup>	18 h 30 m	1 stop	MX\$20,089
United, ANA		MEX–NRT	2 h 35 m SFO	round trip
	07:30 – 15:20 <sup>1</sup>	17 h 50 m	1 stop	MX\$20,089
United, ANA		MEX–NRT	1 h 35 m IAH	round trip
	02:20 – 06:45 <sup>1</sup>	14 h 25 m	Non-stop	MX\$21,889
ANA		MEX–NRT		round trip
	01:30 – 06:20 <sup>1</sup>	14 h 50 m	Non-stop	MX\$31,471
Aeromexico, JAL		MEX–NRT		round trip
Prices are currently typical for your trip.				
Other departing flights				
Prices are not available for: Korean Air. Flights with unavailable prices are at the end of the list.				
	17:30 – 14:20 <sup>2</sup>	30 h 50 m	2 stops	MX\$20,089
United, ANA		MEX–NRT	1 h 10 m ORD	round trip
	17:30 – 14:20 <sup>2</sup>	30 h 50 m	2 stops	MX\$20,089
United, ANA		MEX–NRT	1 h 10 m ORD	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](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)

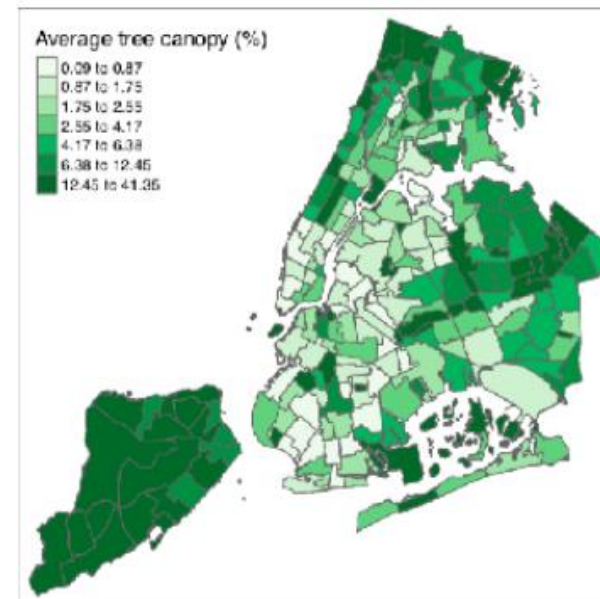
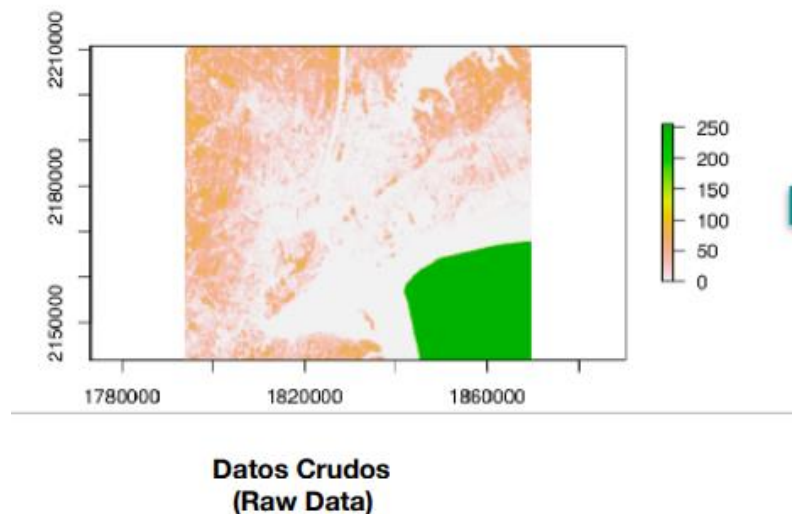
# ¿Qué se puede hacer con R?

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.





# ¿Qué se puede hacer con R?

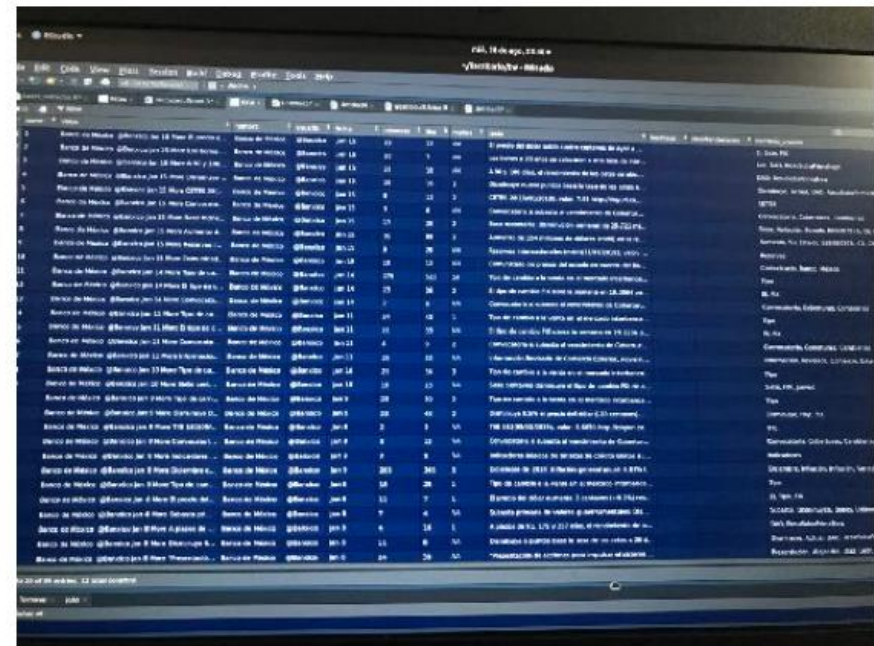
Esto no lo vamos a ver en clase



## 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.).



ID	URL	Fecha	Origen	Tipo	Estado	Detalle
1	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	El primer día de la semana...
2	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	La semana es un día de...
3	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	A la vez, por...
4	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	Revisando...
5	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	El primer día de la semana...
6	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	La semana es un día de...
7	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	A la vez, por...
8	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	Revisando...
9	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	El primer día de la semana...
10	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	La semana es un día de...
11	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	A la vez, por...
12	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	Revisando...
13	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	El primer día de la semana...
14	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	La semana es un día de...
15	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	A la vez, por...
16	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	Revisando...
17	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	El primer día de la semana...
18	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	La semana es un día de...
19	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	A la vez, por...
20	https://twitter.com/...	2018-01-15	Twitter	Tweet	OK	Revisando...

# ¿Qué se puede hacer con R?

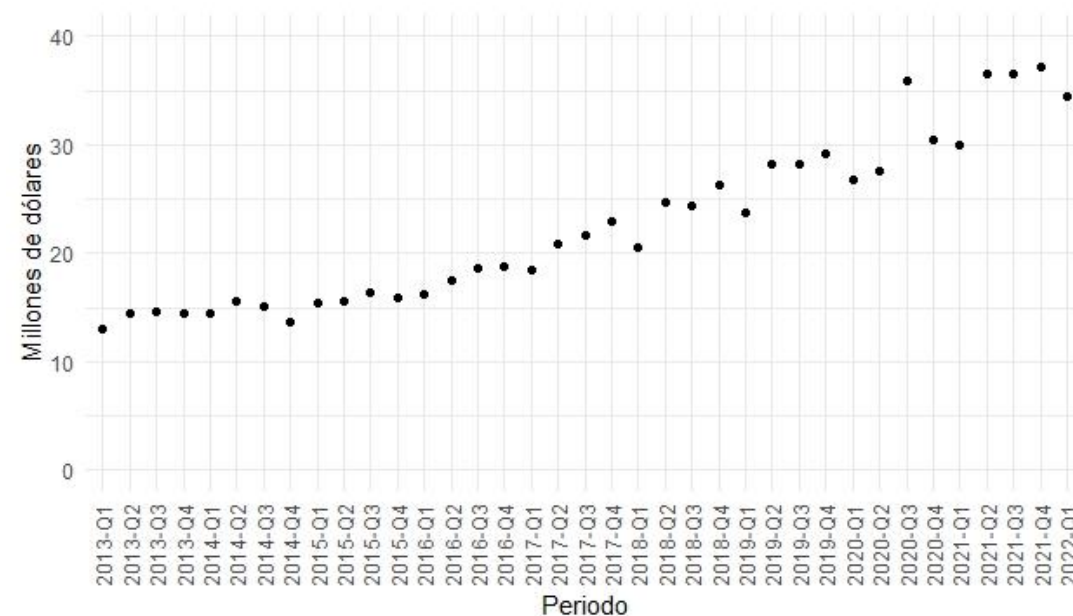
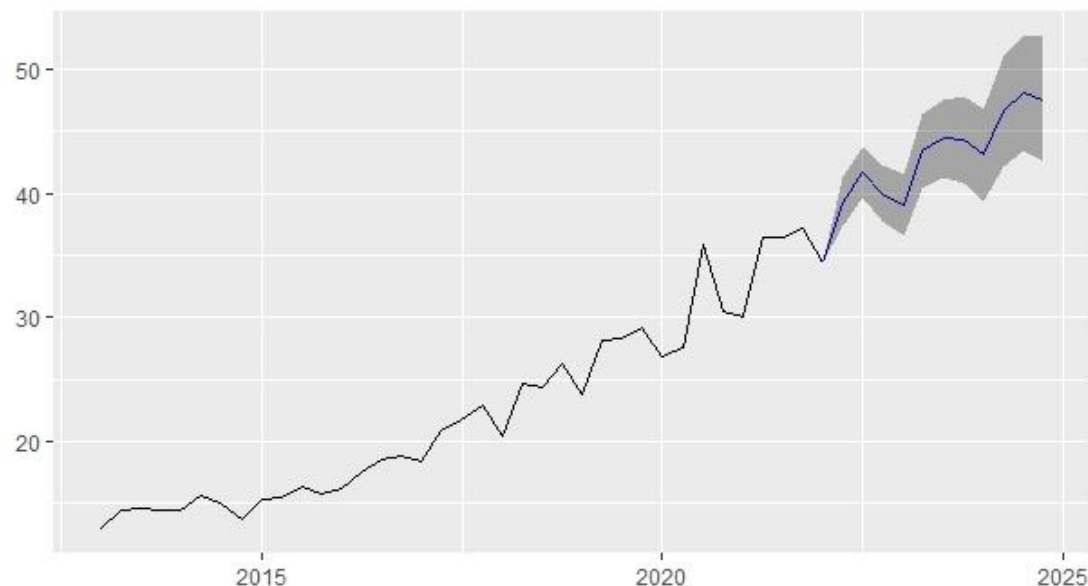
*Esto lo vamos a ver  
en clase*



## 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 ...



R-Ladies Taipei



# Horario y fechas importantes

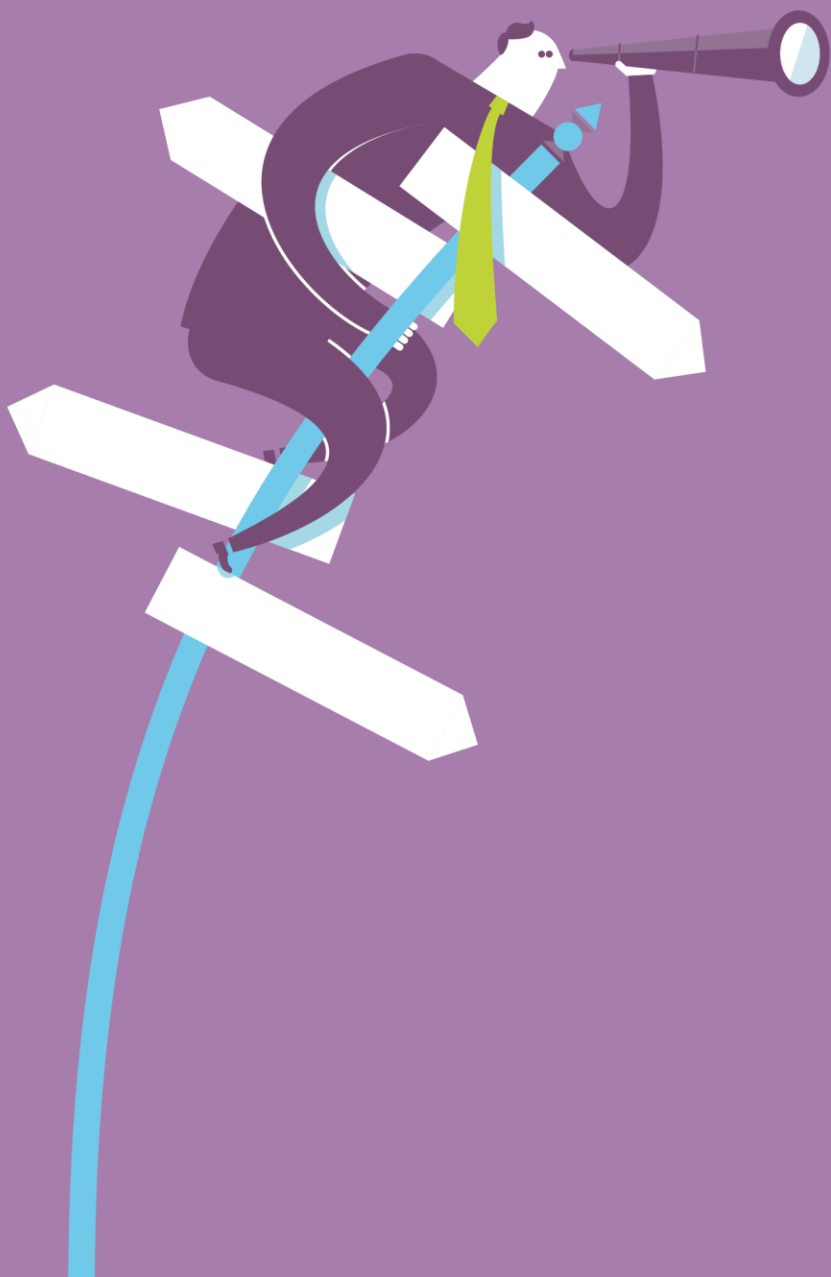
- **Horario:** martes, jueves de 7:30am a 8:55am y viernes de 7:00am a 8:55am
- **Inicio de clases:** 8 de agosto de 2022
- **Asuetos:** 16 de septiembre y 21 de noviembre
- **Primer parcial:** 9 de septiembre
- **Segundo parcial:** 14 de octubre
- **Último día de clases:** 29 de noviembre
- **Exámenes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

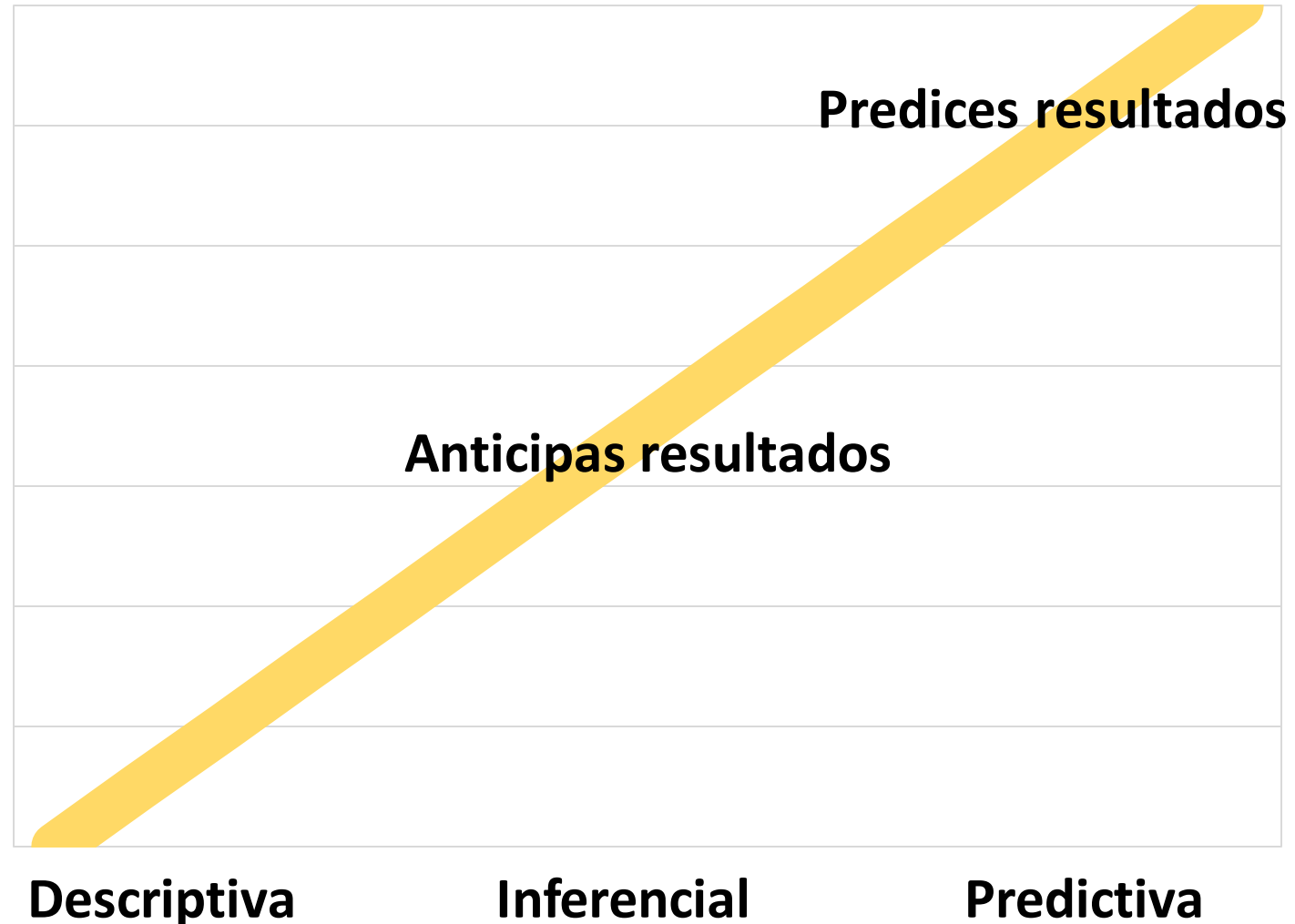
# Evolución en la estadística y analítica de datos

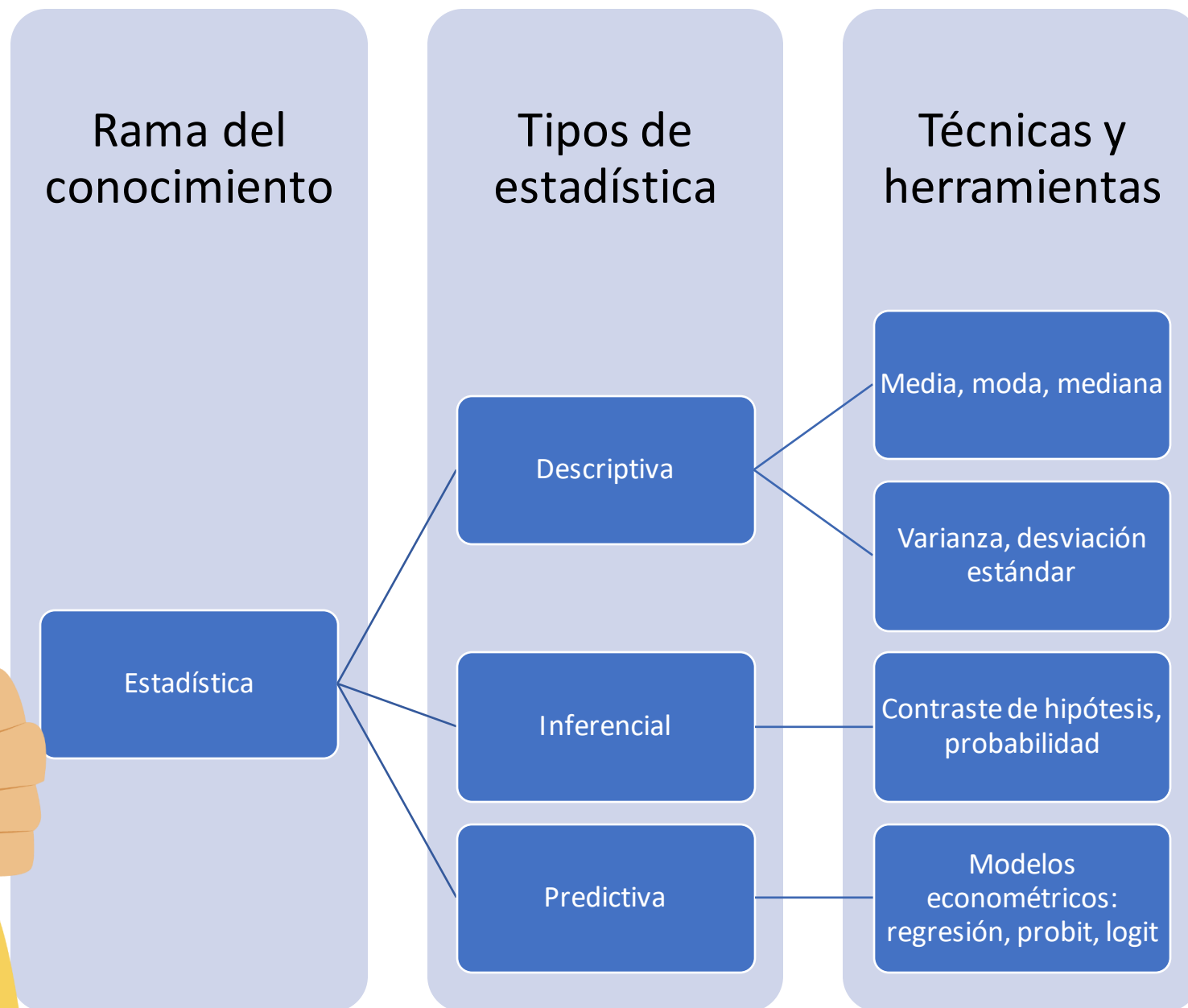
La estadística puede dividirse en dos grandes apartados, descriptiva e inferencial pero con la ciencia de datos y la econometría se puede lograr mejores predicciones...

**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?

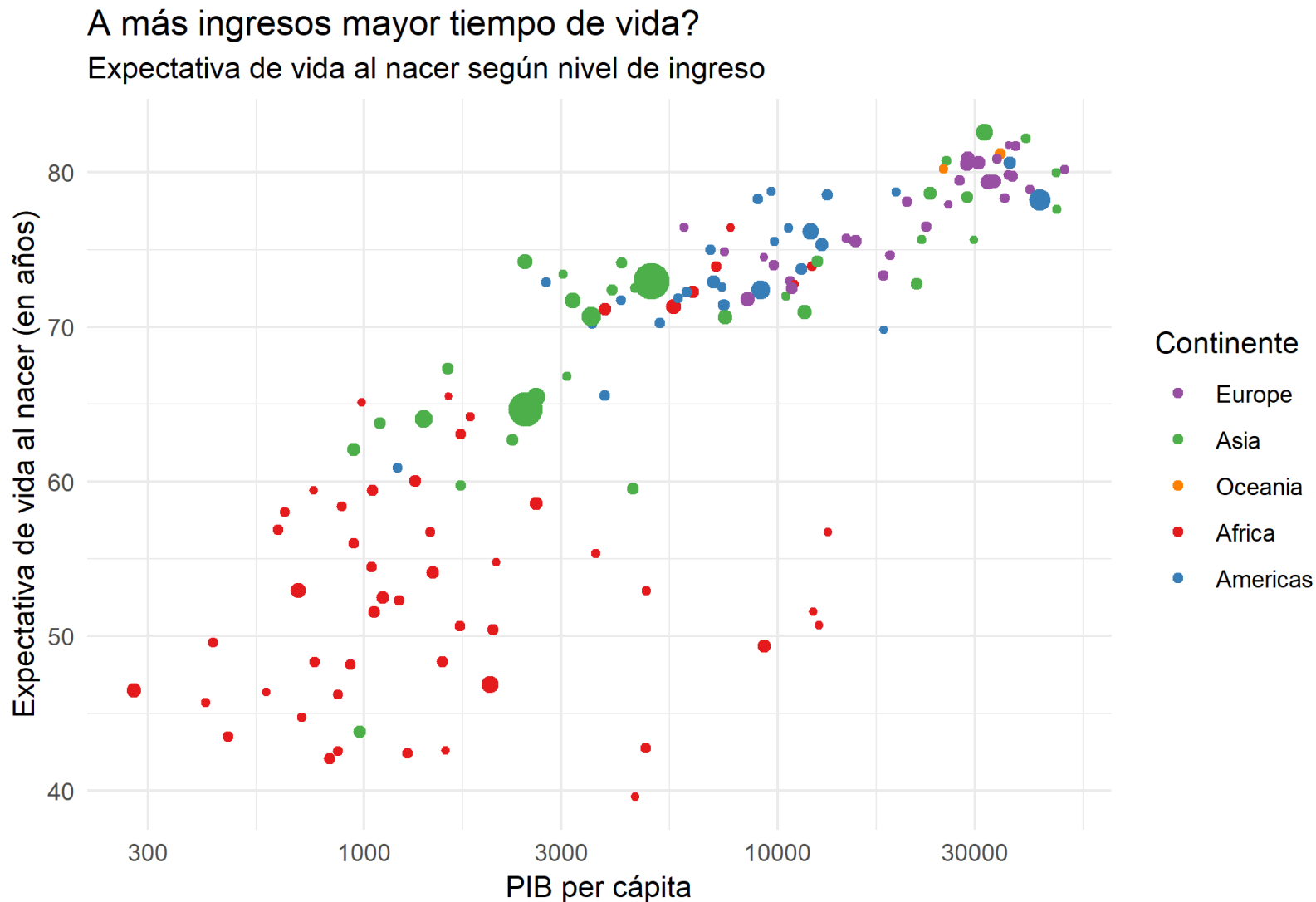




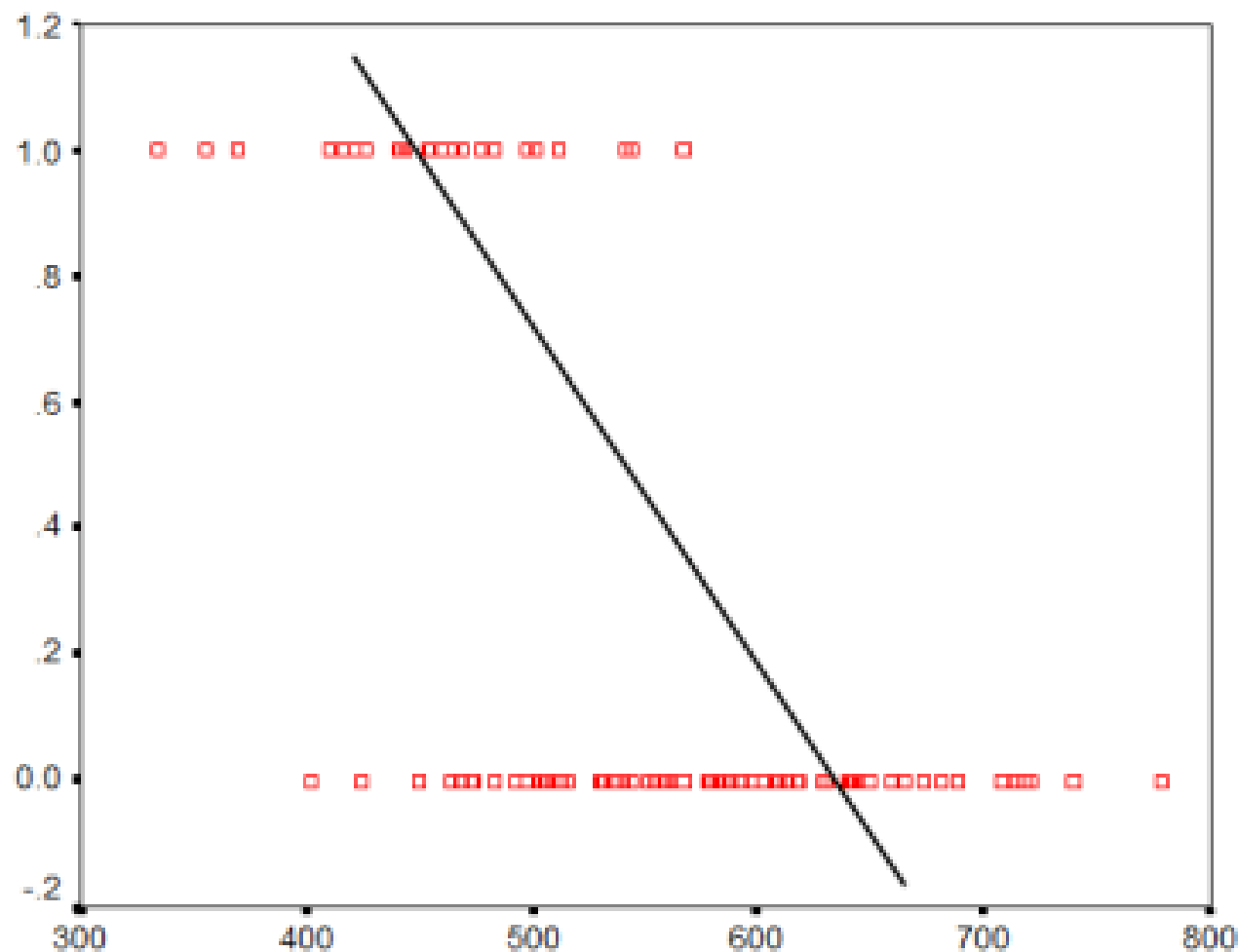
# Tipo de variables en la estadística

#	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)



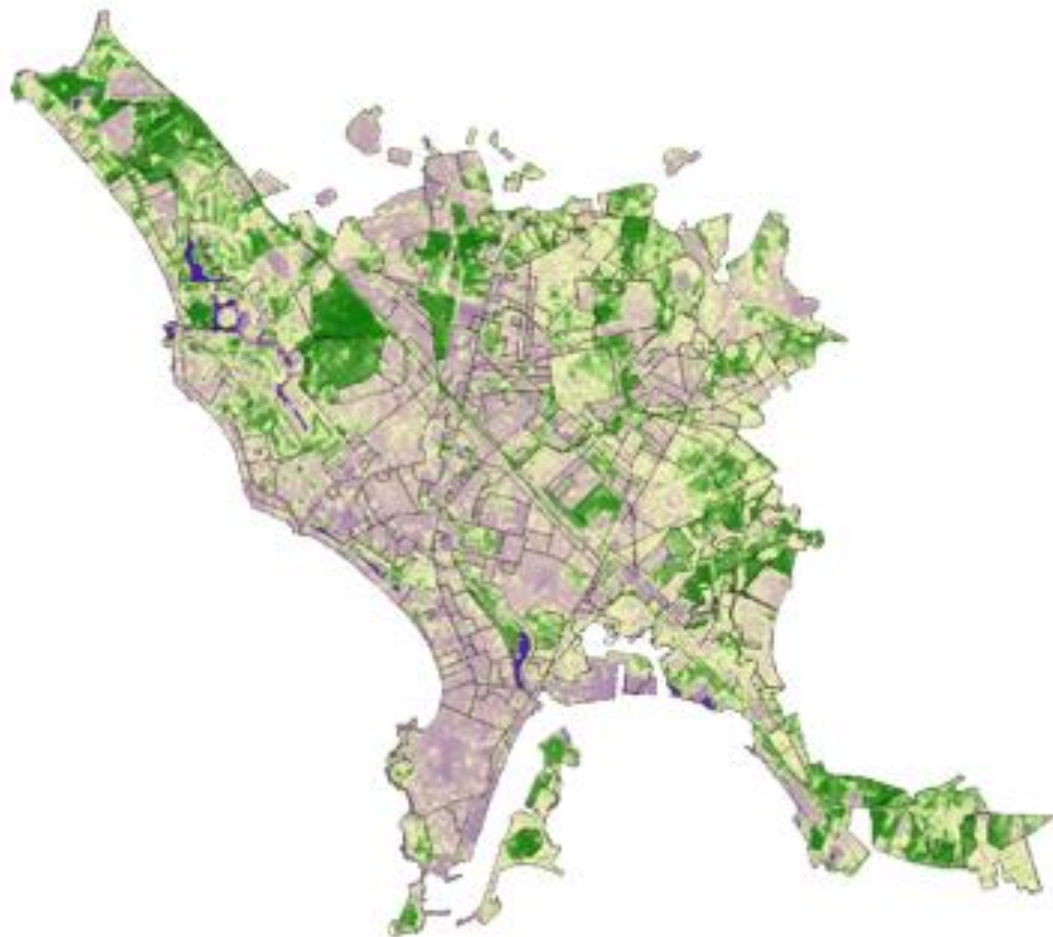
Variables cualitativa con cuantitativa: se contrasta la variable cualitativa de si una persona ha contratado un servicio, 1 = si contrata, y 0 = si no contrata vs variable cuantitativa: ingreso.



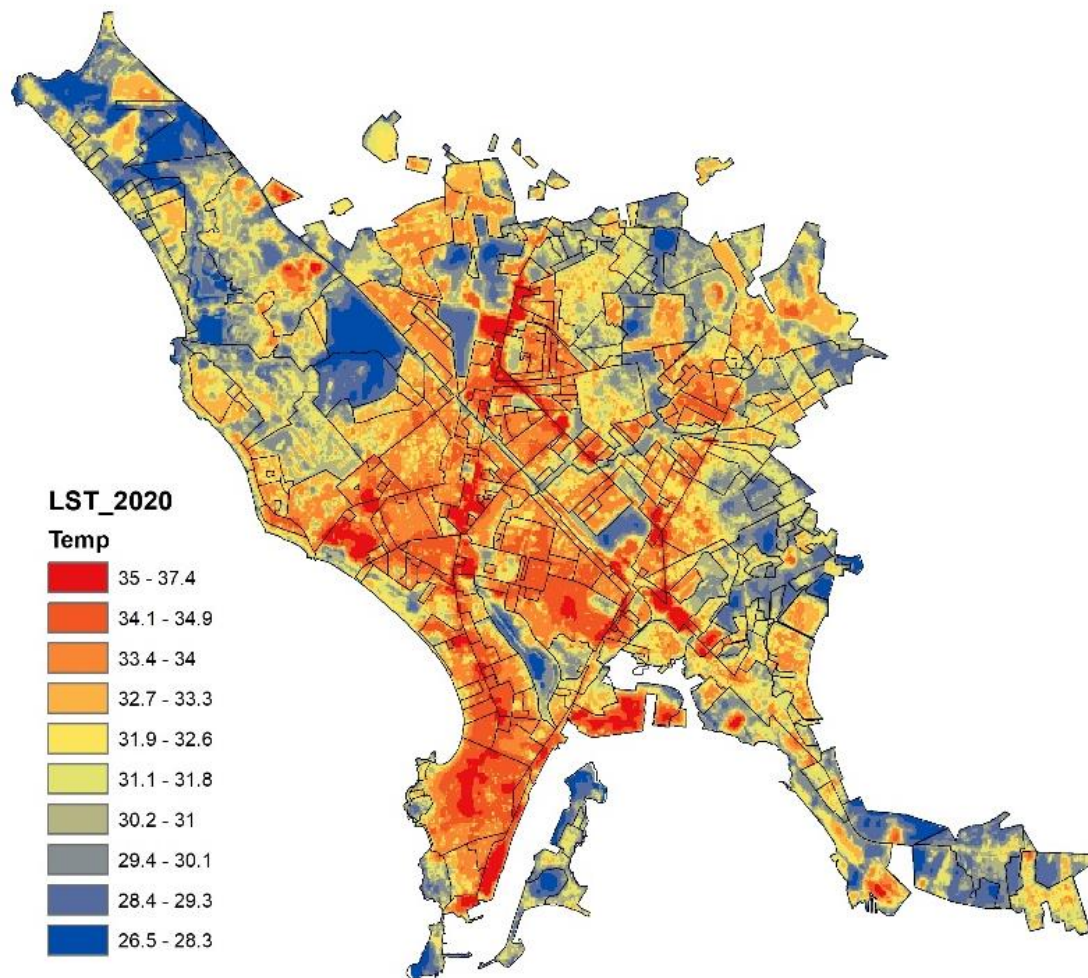


## Variable georreferenciada: áreas verdes en Mazatlán y evidencia de isla de calor

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	$\Sigma$	Este símbolo (llamado sigma) significa "sumatoria". Por lo tanto, si ves este simbolo " $\Sigma x_i$ " solo significa "sumar todos los valores recopilados"..
2	$\Pi$	Este símbolo (pi) significa "multiplicar". Entonces, si ves algo como " $\Pi x_i$ " 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	$\sigma$	Significa la desviación estándar de un conjunto de datos..
2	$\beta_i$	Coeficiente asociado a variable en el análisis de regresión..
3	$\rho$	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,  $\Sigma$ .

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

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

## 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
-----------------	----	----	----	----	----	----	----	----

$$36 + 42 = 78$$

$$78 / 2 = 39$$

$$\text{Mediana} = 39$$



## 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

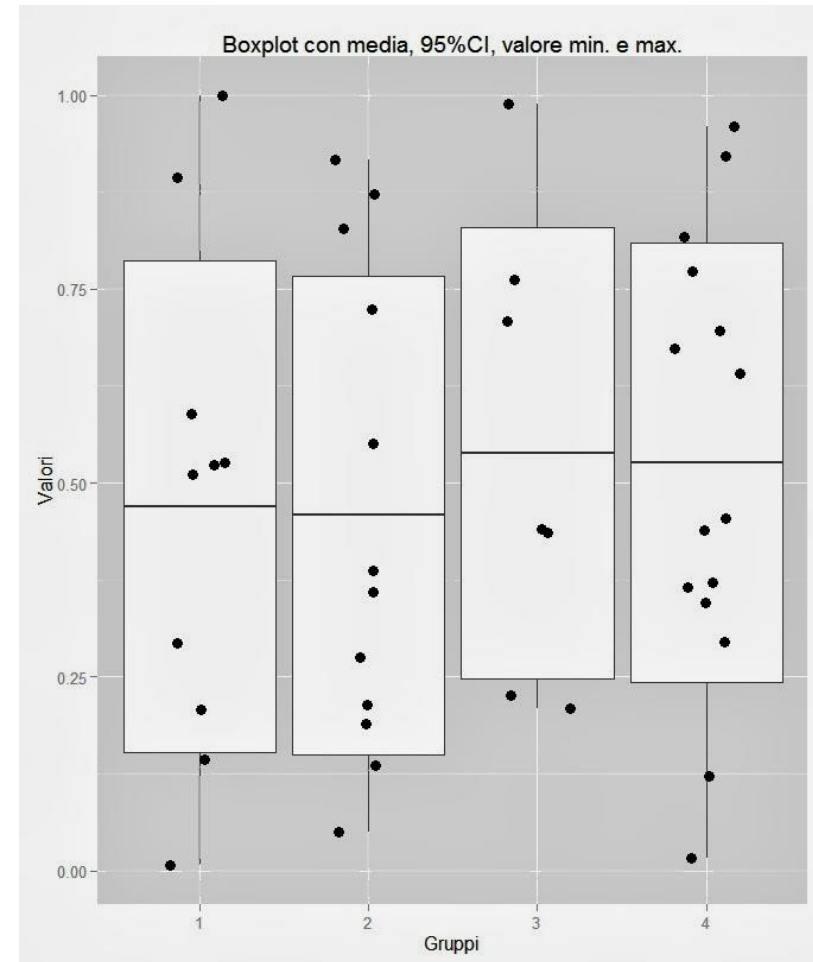
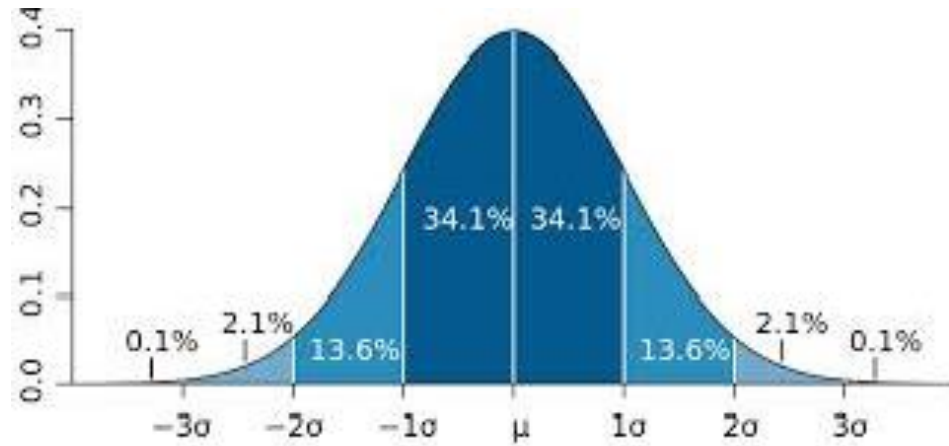
Las calificaciones obtenidas por un alumno en ocho exámenes del curso de Estadística son:

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

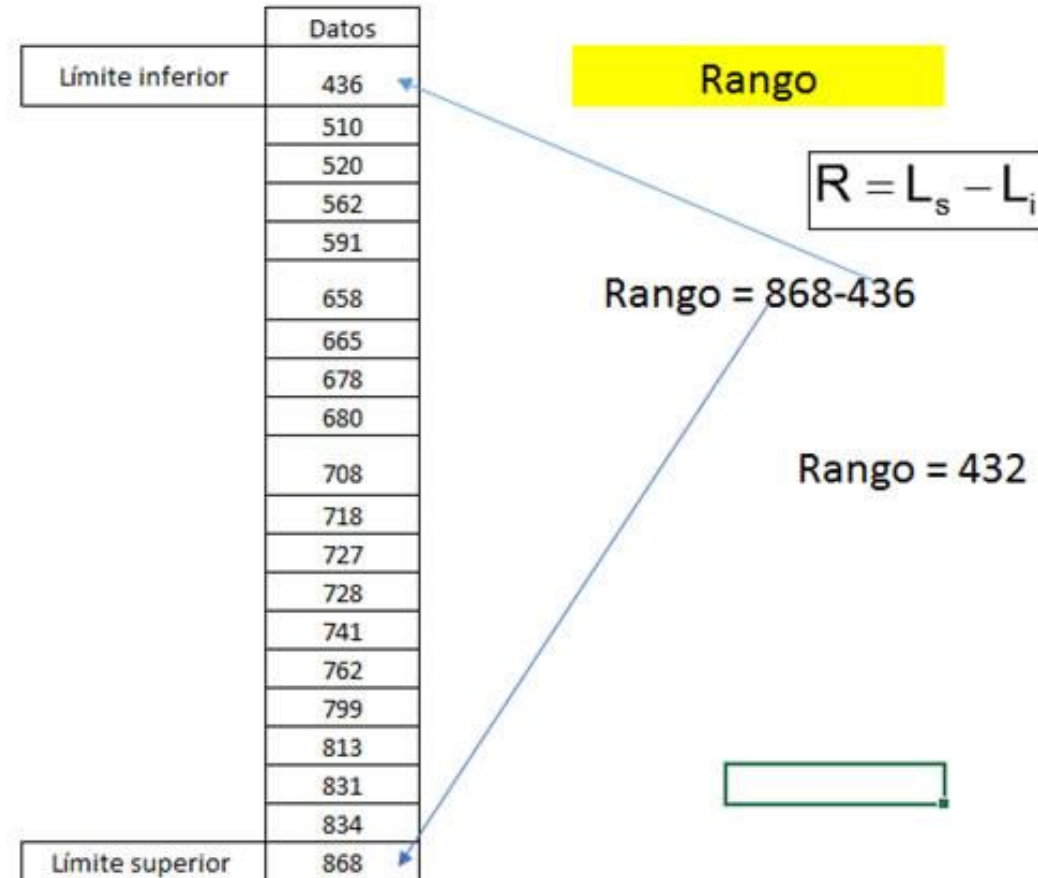


La dispersión se refiere a la separación de los datos en una distribución, es decir, al grado en que las observaciones se separan. Aquí por ejemplo el símbolo  $\mu$  significa la media muestral de los datos que tenemos, y  $\sigma$  significa la desviación estándar.

# 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.



# Varianza y desviación estándar

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 número de observaciones menos uno. Su ecuación es la siguiente:

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

	Datos	Media o promedio	$x - \bar{x}$	$(x - \bar{x})^2$
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
	762	691.45	70.55	4977.3025
	799	691.45	107.55	11567.0025
	813	691.45	121.55	14774.4025
Límite superior	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
<b>Varianza</b>				<b>13940.47105</b>

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

# 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
	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
Límite superior	868	691.45	176.55	31169.9025
		Suma	$\sum (x - \bar{x})^2$	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

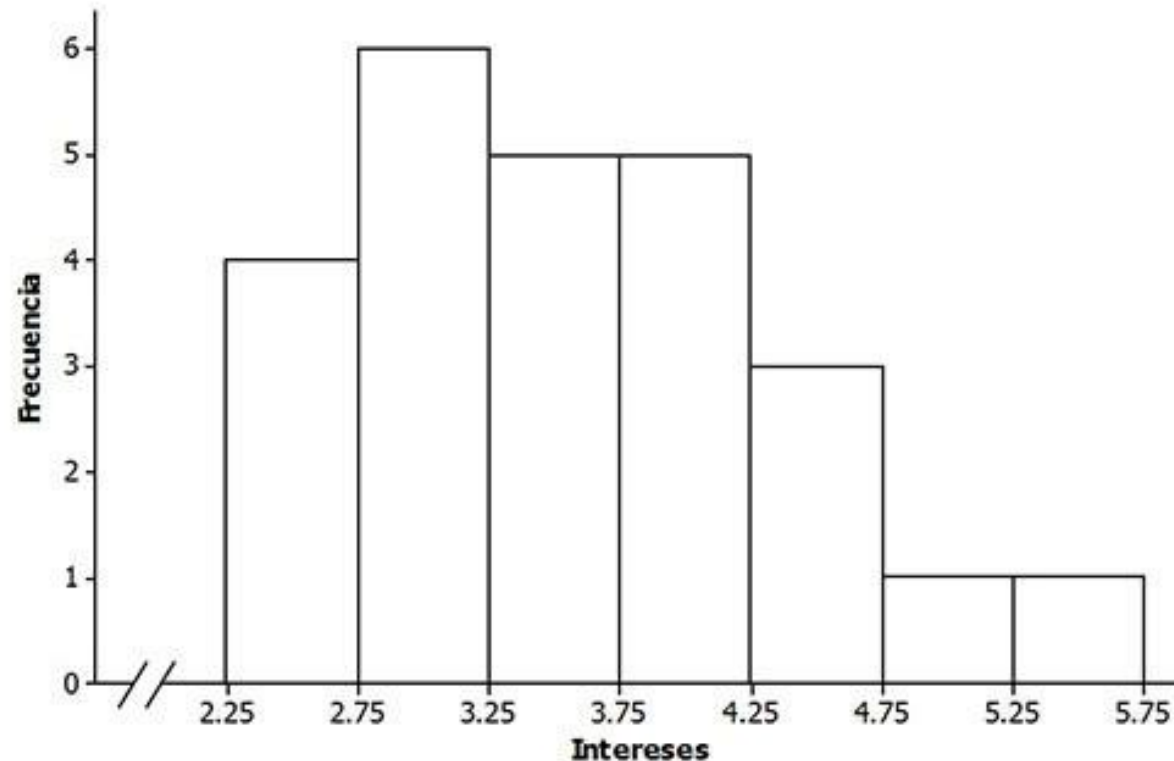
$$T_c \geq \frac{793}{6} \geq 132.1 \geq 133$$

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

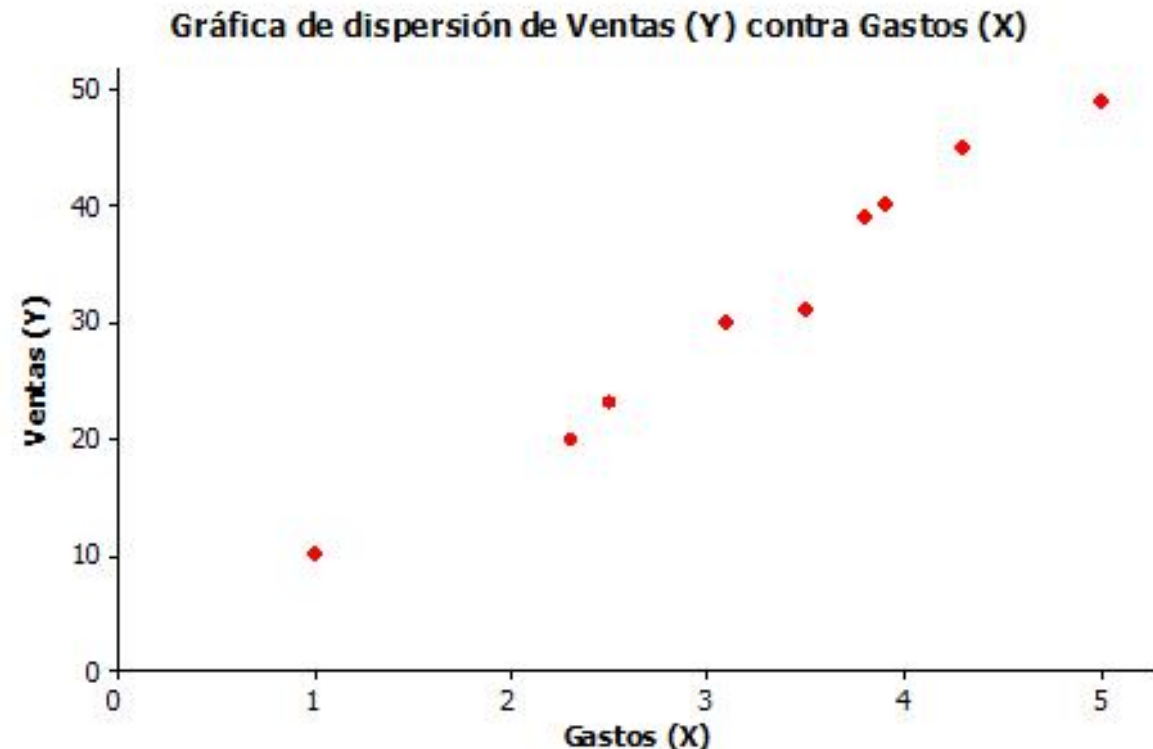
“El histograma condensa los datos, agrupando valores similares en clases. Se puede construir un histograma colocando la variable de interés en el eje horizontal, y la frecuencia, frecuencia relativa o frecuencia porcentual, en el eje vertical” (Hanke y Wichern, 2010).



## 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.**

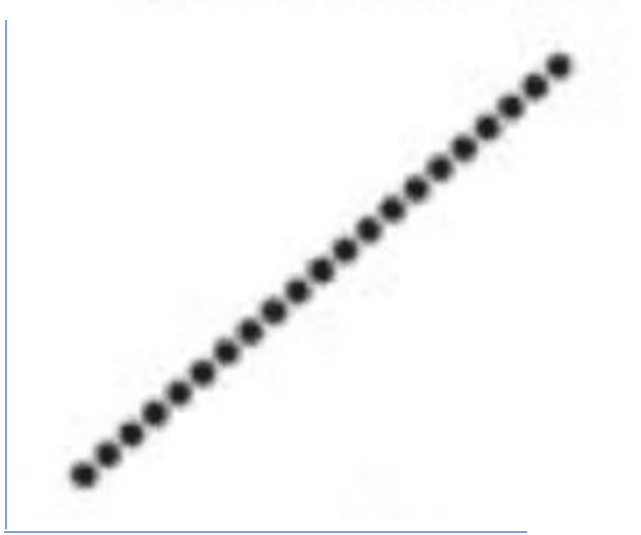
Resultado			Coeficiente de correlación lineal (positivo)
0.00	a	0.09	Nula
0.10	a	0.19	Muy débil
0.20	a	0.49	Débil
0.50	a	0.69	Moderada
0.70	a	0.84	Significativa
0.85	a	0.95	Fuerte
0.96	a	1.00	Perfecta

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.

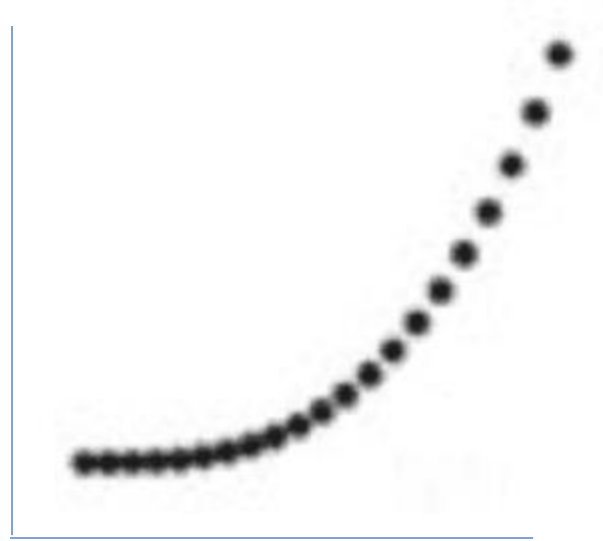
**Forma:** lineal o no lineal

Y



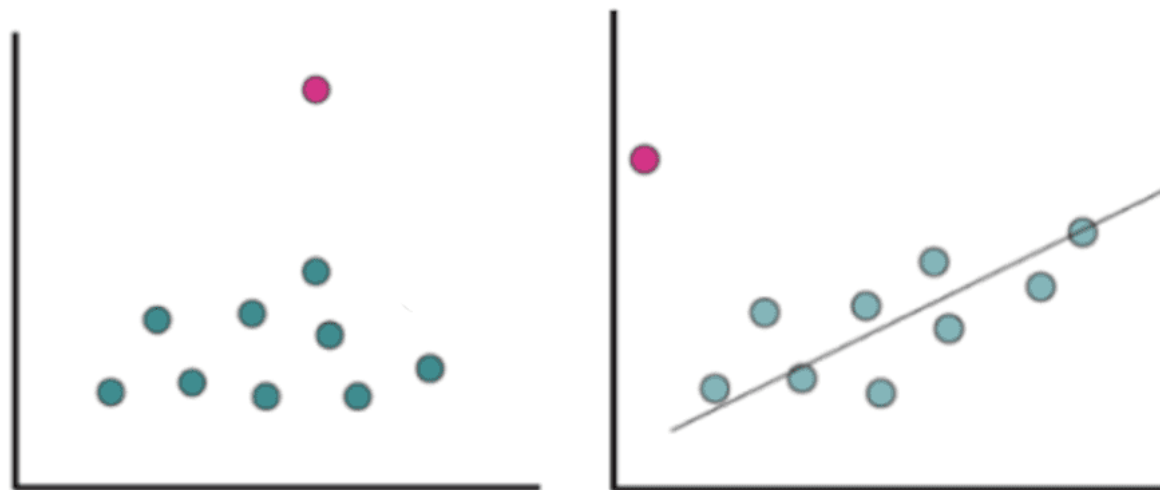
X

Y



X

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

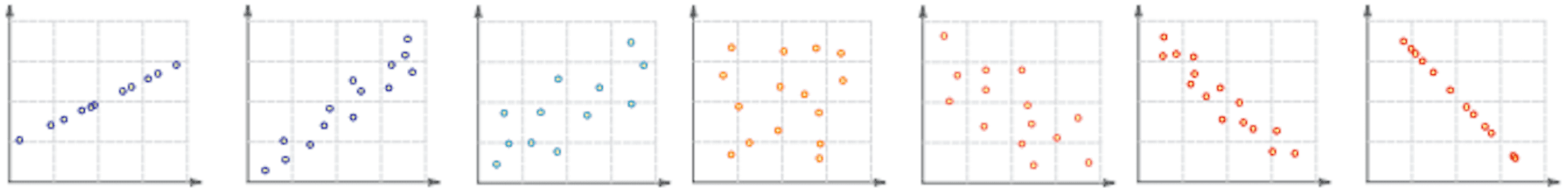


**Los outliers pueden afectar los análisis de correlación y/o regresión (que veremos en temas mas adelante) debido a que la relación entre las dos variables cambia con la presencia de estos valores.**



**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

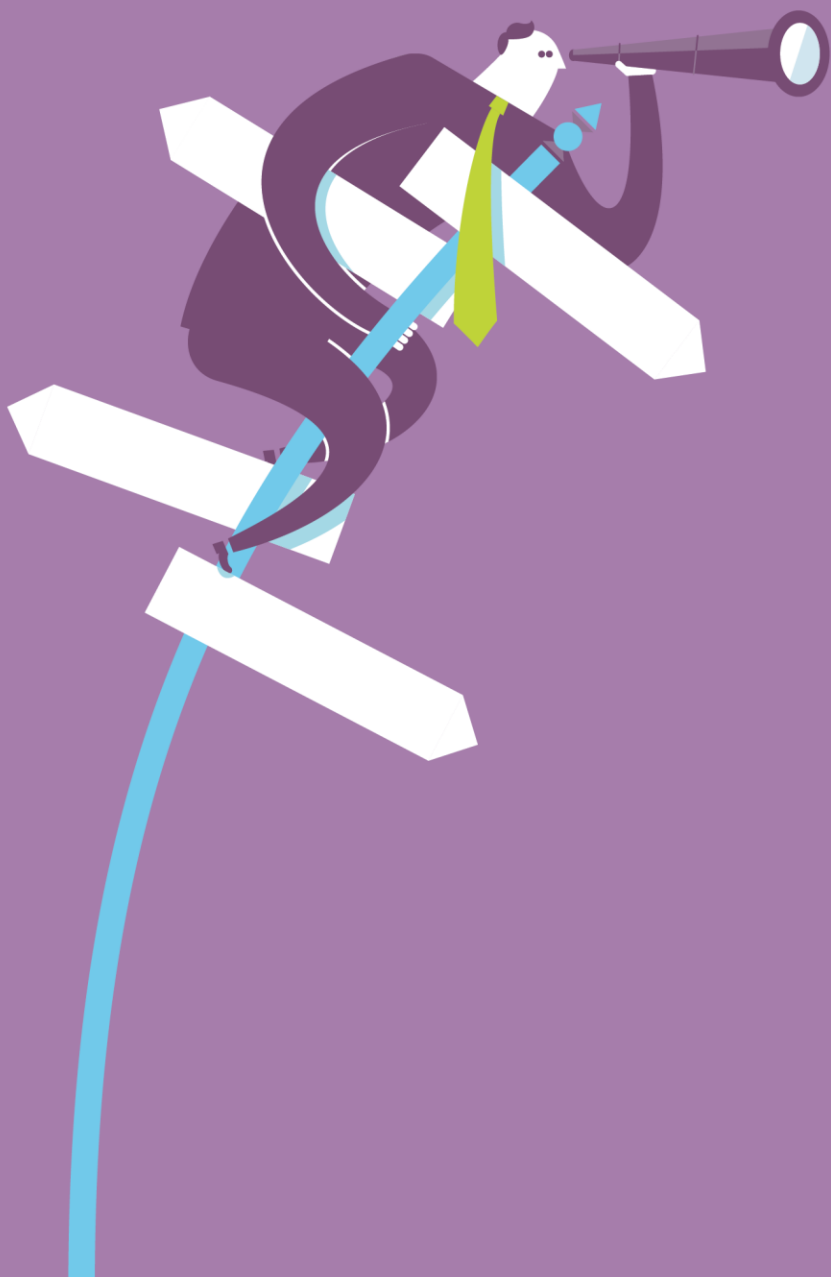
Todas los viernes iniciaremos clase con las siguientes preguntas:

- 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**  $\Omega$  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 contraintuitivo... 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  $\Omega$  (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

$\Omega = \{\mathbf{C}, \mathbf{X}\}$  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  $\Omega = \{e_i, \dots, e_k\}$ .

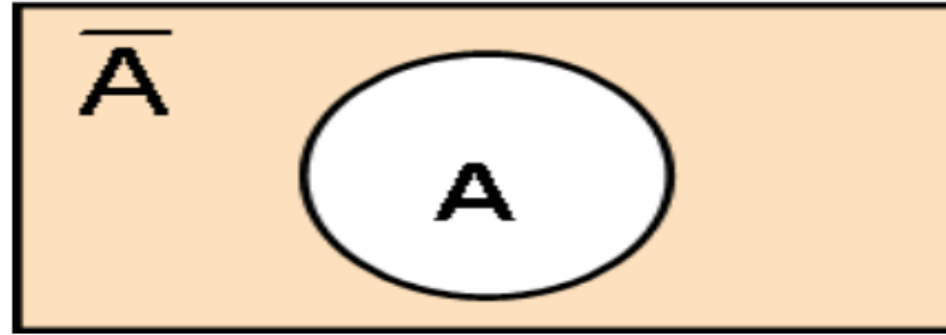
**Ejemplo 4.** En el caso de lanzar la moneda dos veces, los sucesos elementales son  $e_1 = (C, C)$ ,  $e_2 = (C, X)$ ,  $e_3 = (X, C)$  y  $e_4 = (X, X)$ .

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

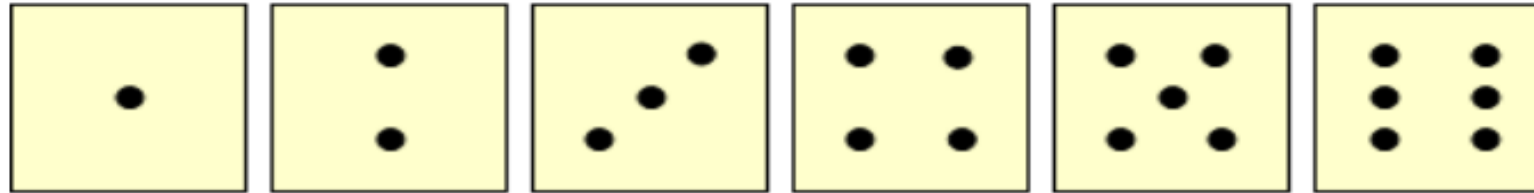
**Ejemplo 5.** En el caso de lanzar la moneda dos veces, el suceso  $A$  = “sale una cara y una cruz” es  $A = \{(C, X)\}$ .

**Suceso seguro:** El espacio muestral completo  $\Omega$ . 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  $\Omega$  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,

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



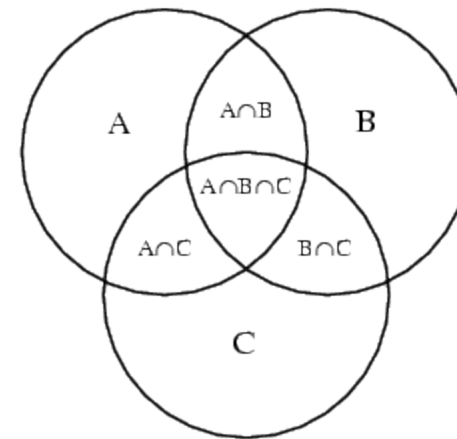
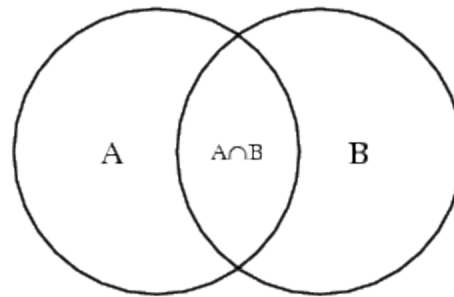
# Reglas de la probabilidad

## 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?

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

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

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

## 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