



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 ...

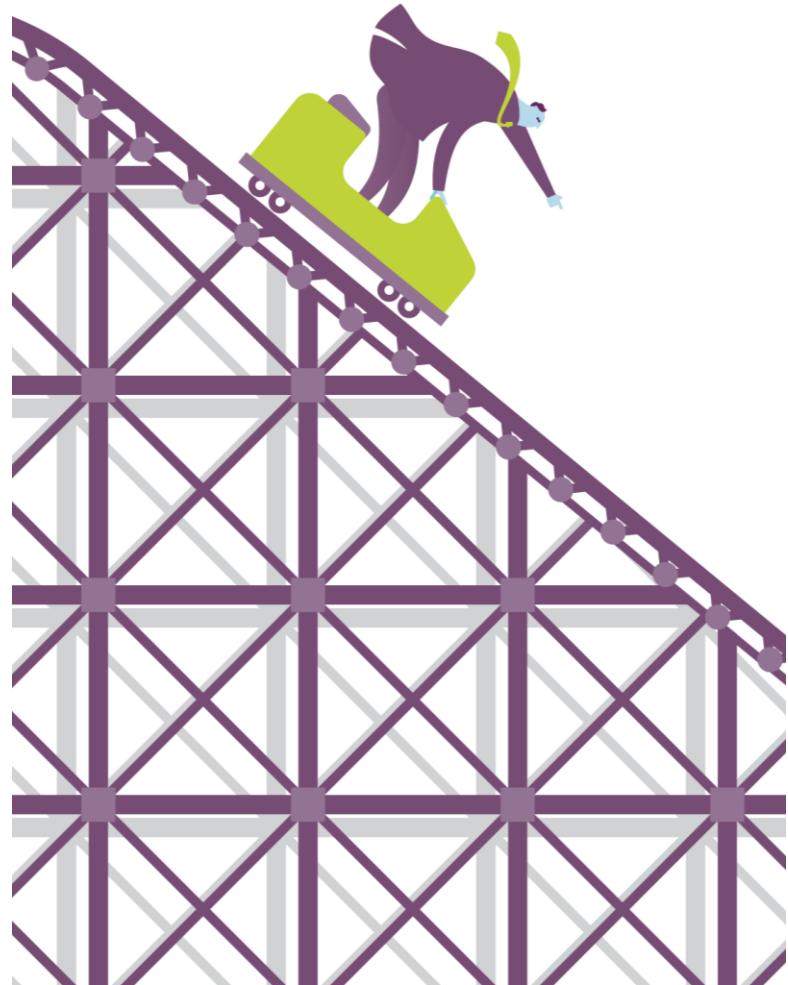
Dos sugerencias para el curso:

- 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

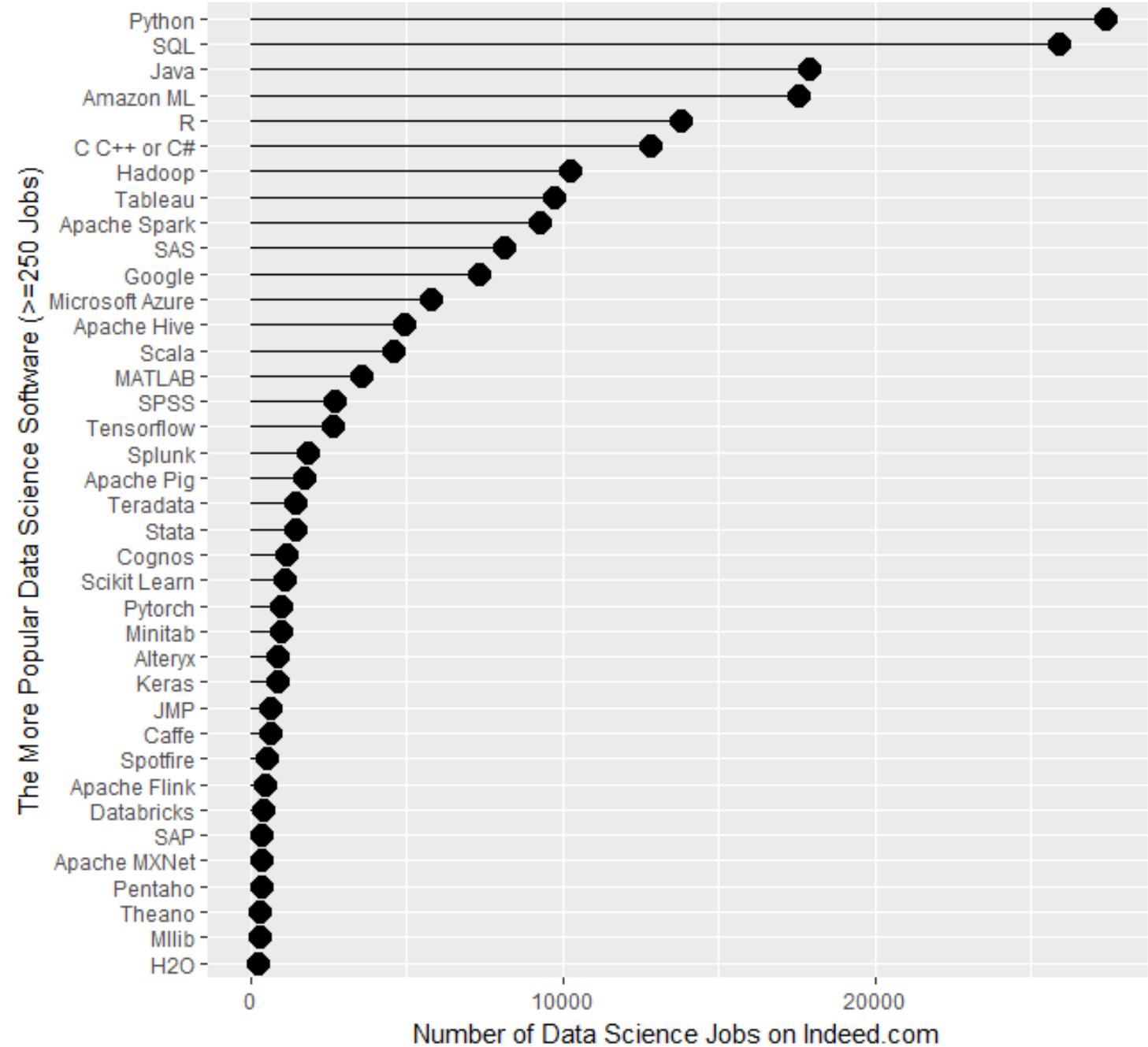
Software estadístico y lenguaje de programación



Hojas de cálculo



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



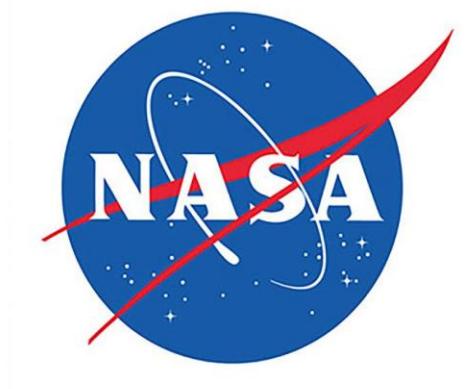
Algunas empresas que utilizan lenguaje de programación estadística...



INTELIGENCIA APLICADA
A DECISIONES



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



Laboratorio Nacional
de Políticas Públicas



Consejo Nacional de Evaluación
de la Política de Desarrollo Social

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



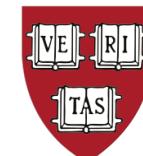
Tecnológico
de Monterrey



Massachusetts
Institute of
Technology



Berkeley
UNIVERSITY OF CALIFORNIA



HARVARD
UNIVERSITY



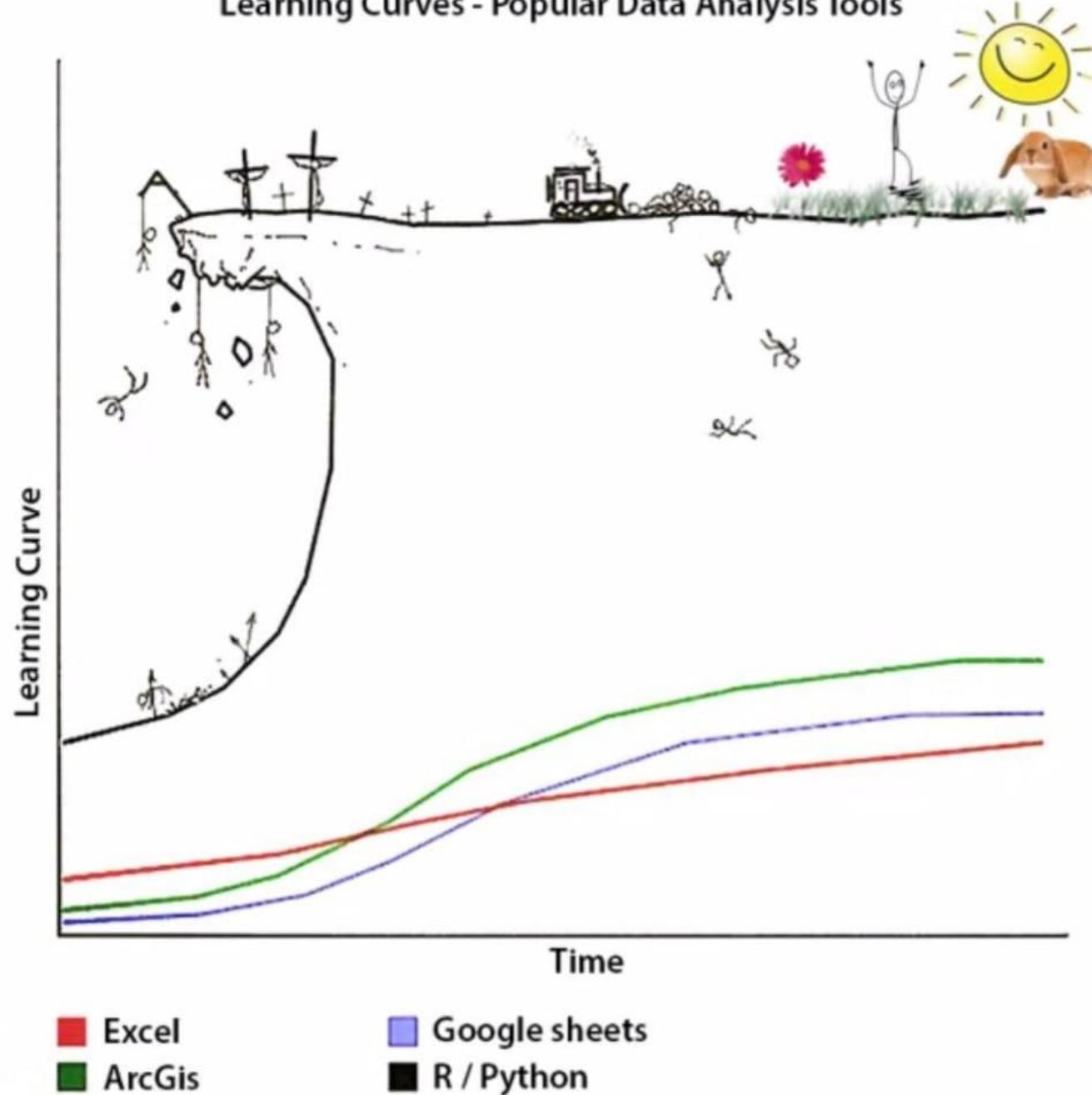
UANL

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?

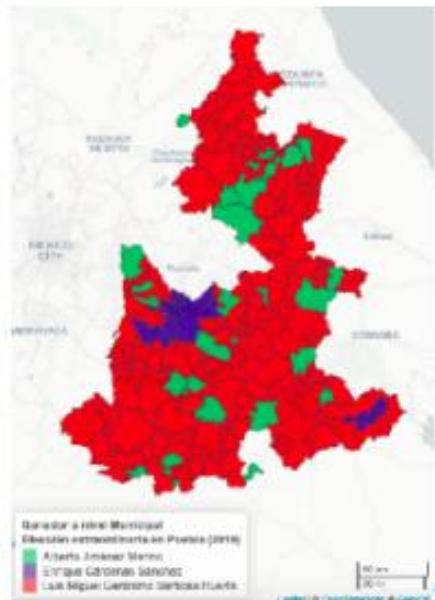
*Esto lo vamos a ver
en clase*



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           )
```



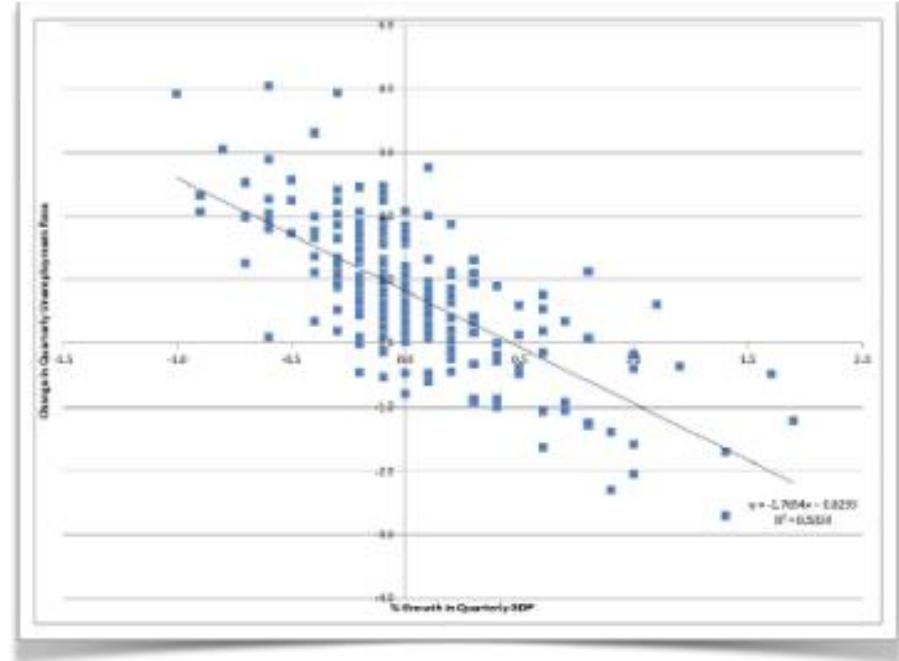
¿Qué se puede hacer con R?

*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*



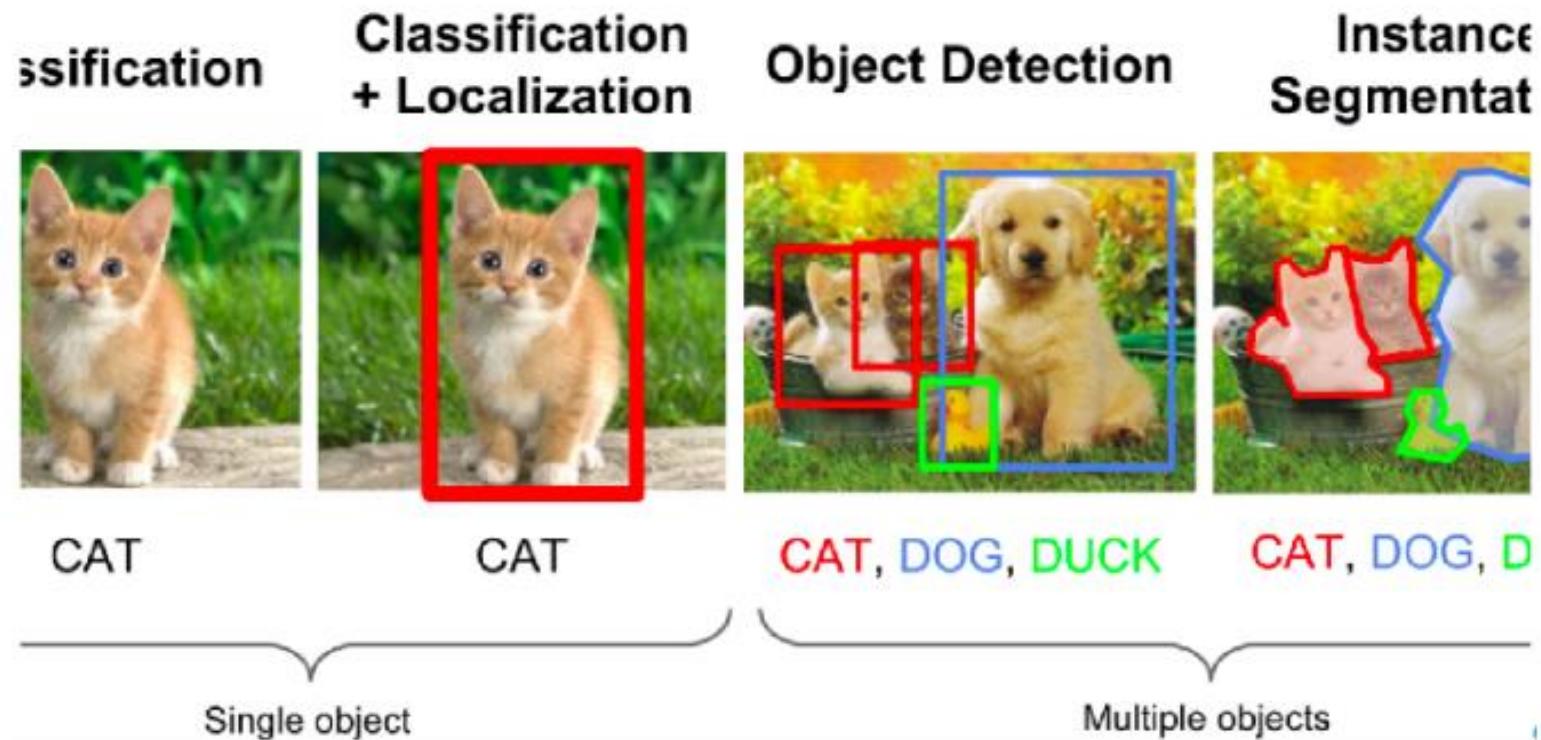
3. Machine Learning y Deep Learning.

```
library(e1071)
```

```
library(tensorflow)
```

```
library(caret)
```

```
library(rpart)
```



¿Qué se puede hacer con R?

*Esto lo vamos a ver
en clase*



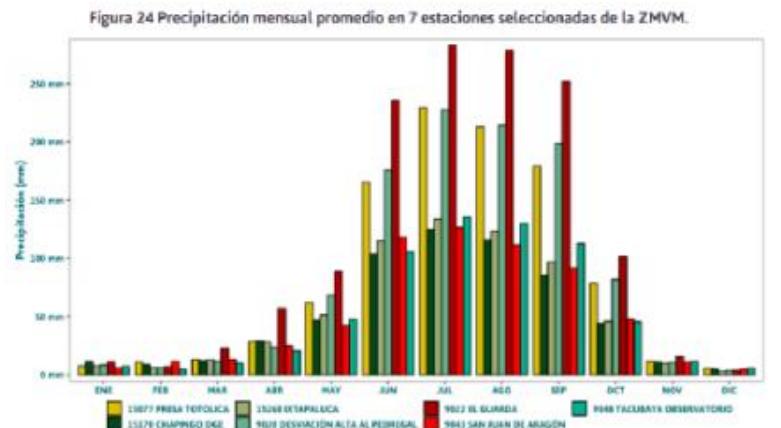
4. Visualización de datos.

```
library(ggplot2)
```

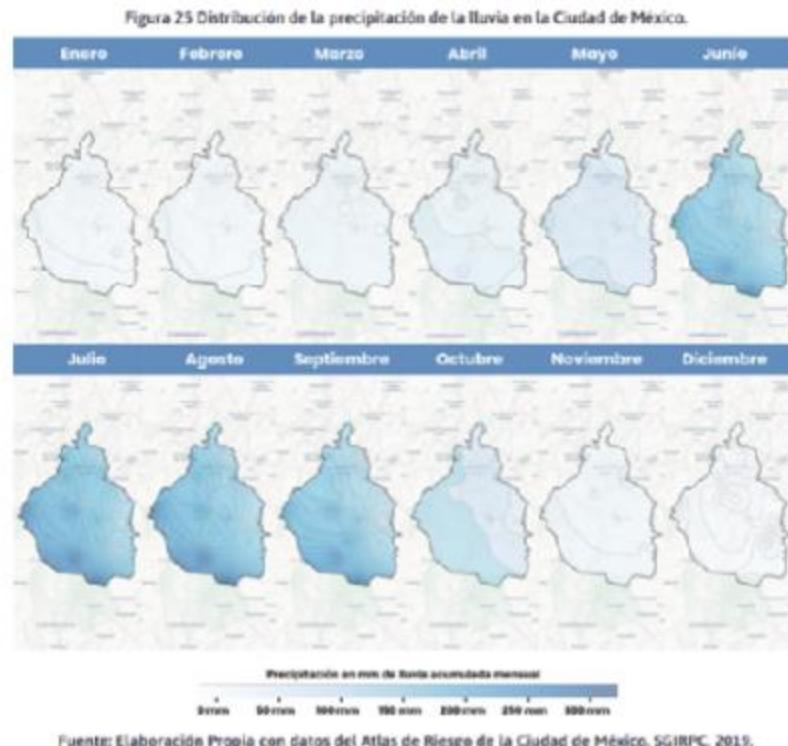
```
library(plotly)
```

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.

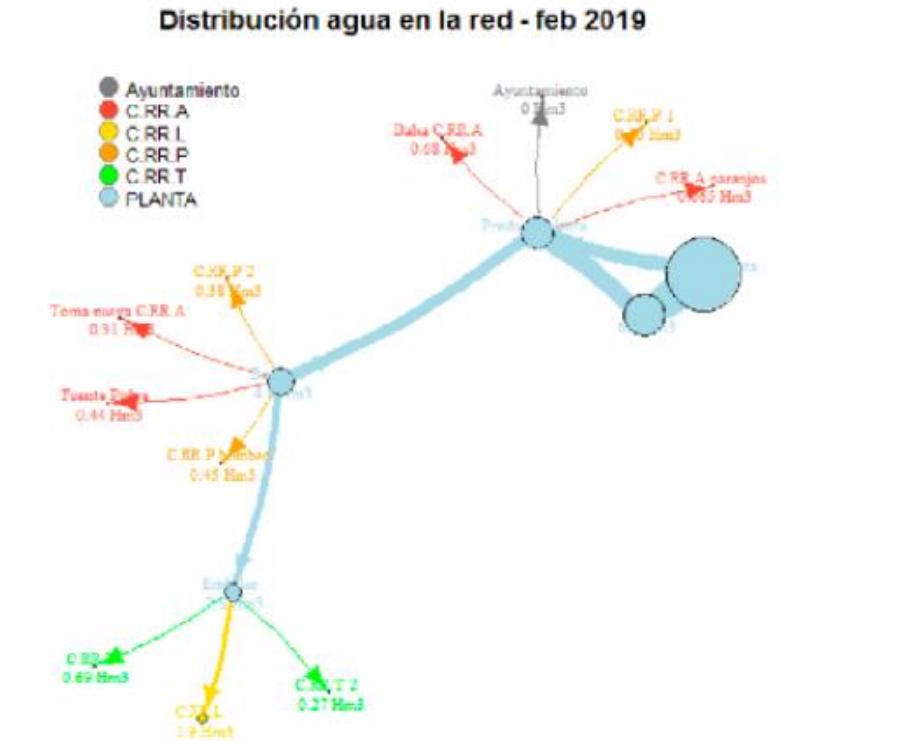
Plan Nacional de Desarrollo, 2019.

*Esto lo vamos a ver
en clase*



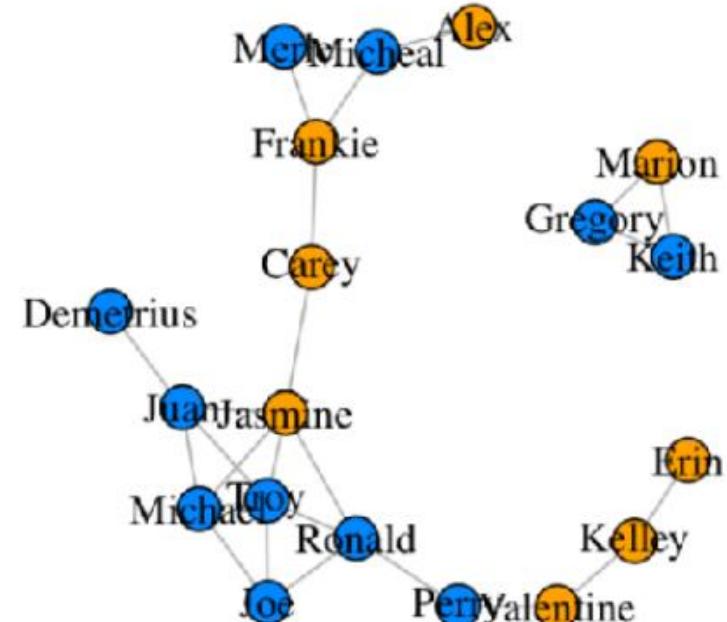
¿Qué se puede hacer con R?

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 JAH	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

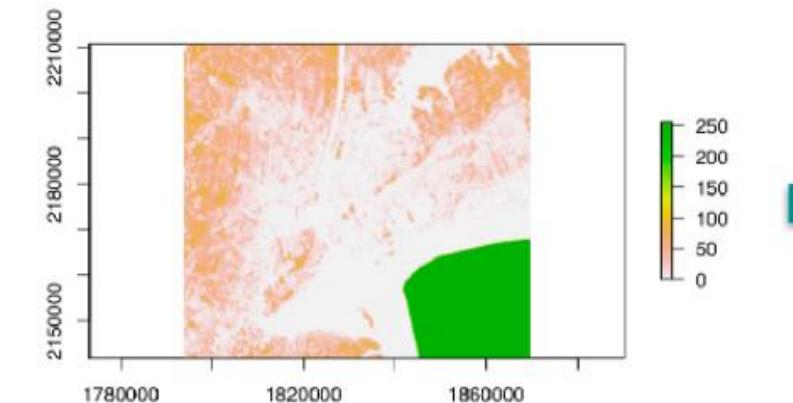
https://www.google.com/flights?lite=0#flt=MEX.NRT.2019-10-21*NRT.MEX.2019-11-05;c:MXN;e:1;sd:1;t:f

¿Qué se puede hacer con R?

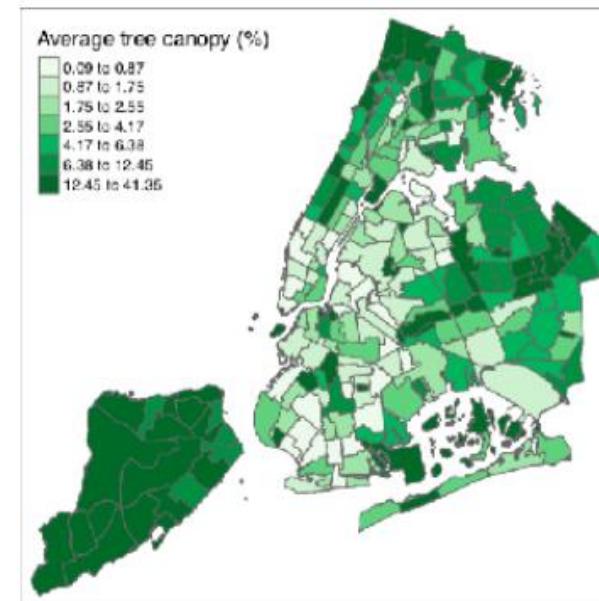
*Esto lo vamos a ver
en clase* 😊

8. Análisis Geoespacial. `library(sf)`

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



Datos Crudos
(Raw Data)



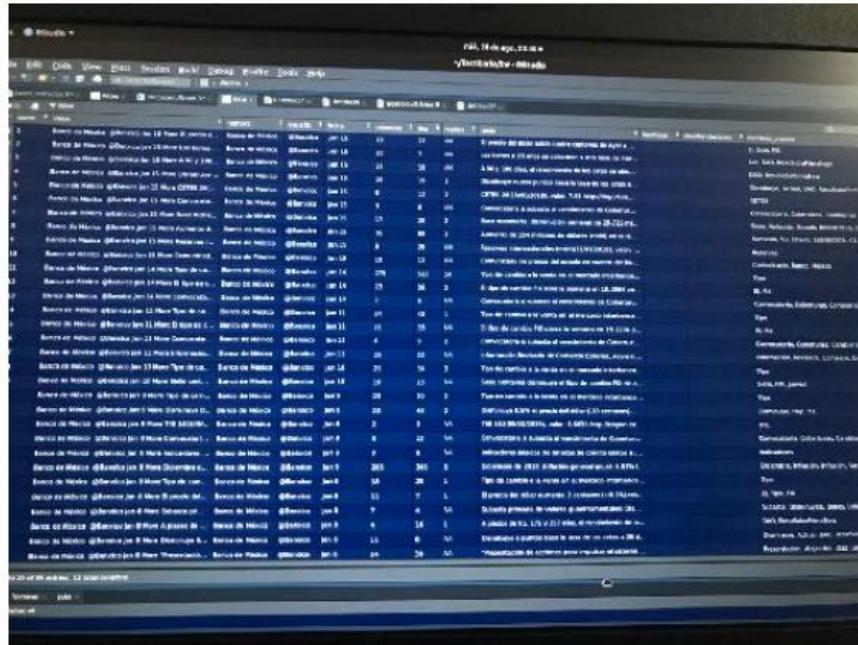
Datos Procesados que permiten llegar a conclusiones

¿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?

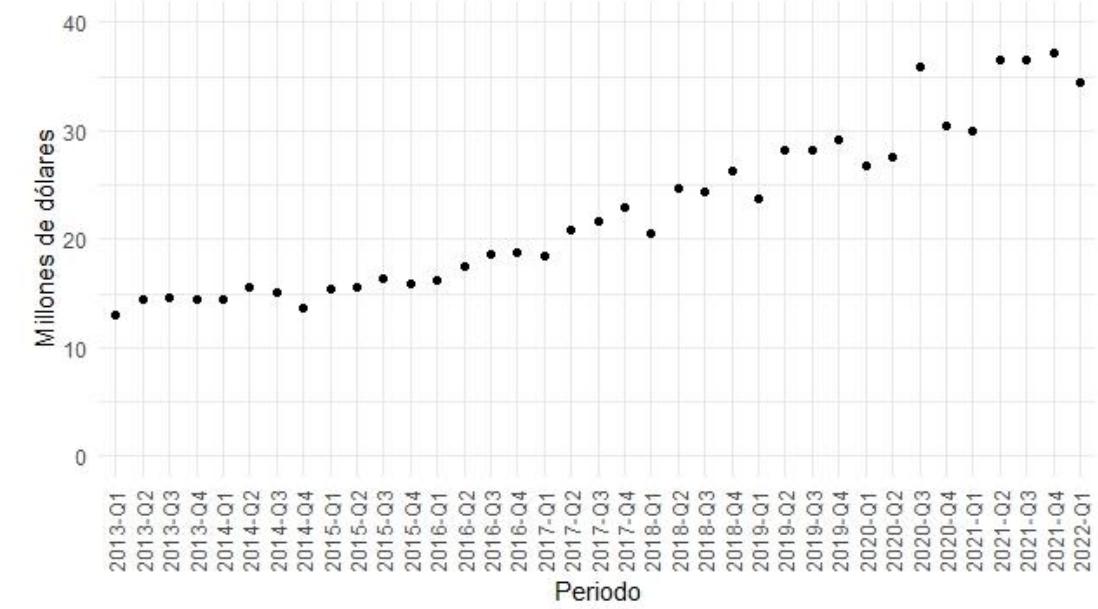
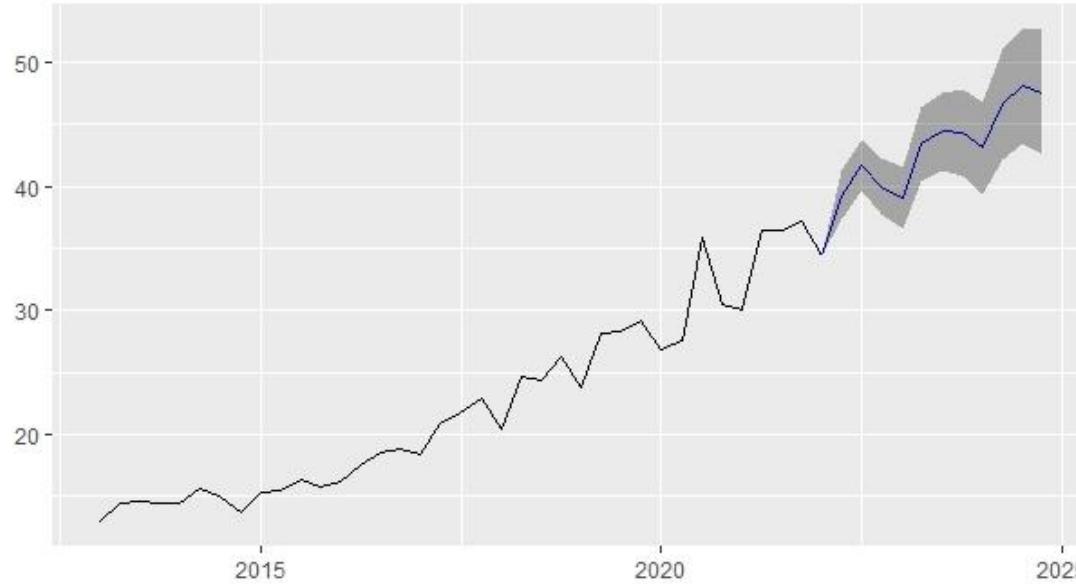
*Esto lo vamos a ver
en clase*



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

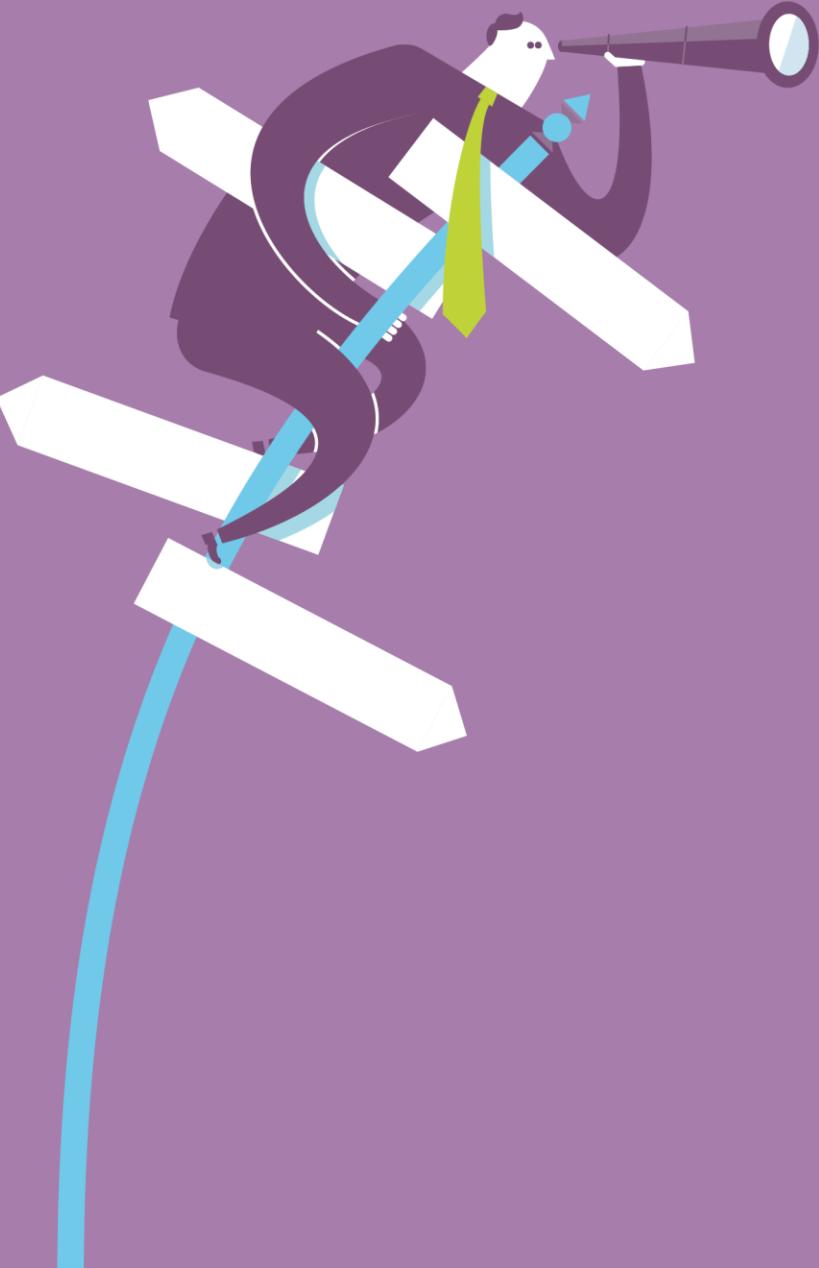
- **Horario:** martes, jueves de 7:30am a 8:55am y viernes de 7:00am a 8:55am
- **Inicio de clases:** 8 de agosto de 2022
- **Asuertos:** 16 de septiembre y 21 de noviembre
- **Primer parcial:** 9 de septiembre
- **Segundo parcial:** 14 de octubre
- **Último día de clases:** 29 de noviembre
- **Exámenes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



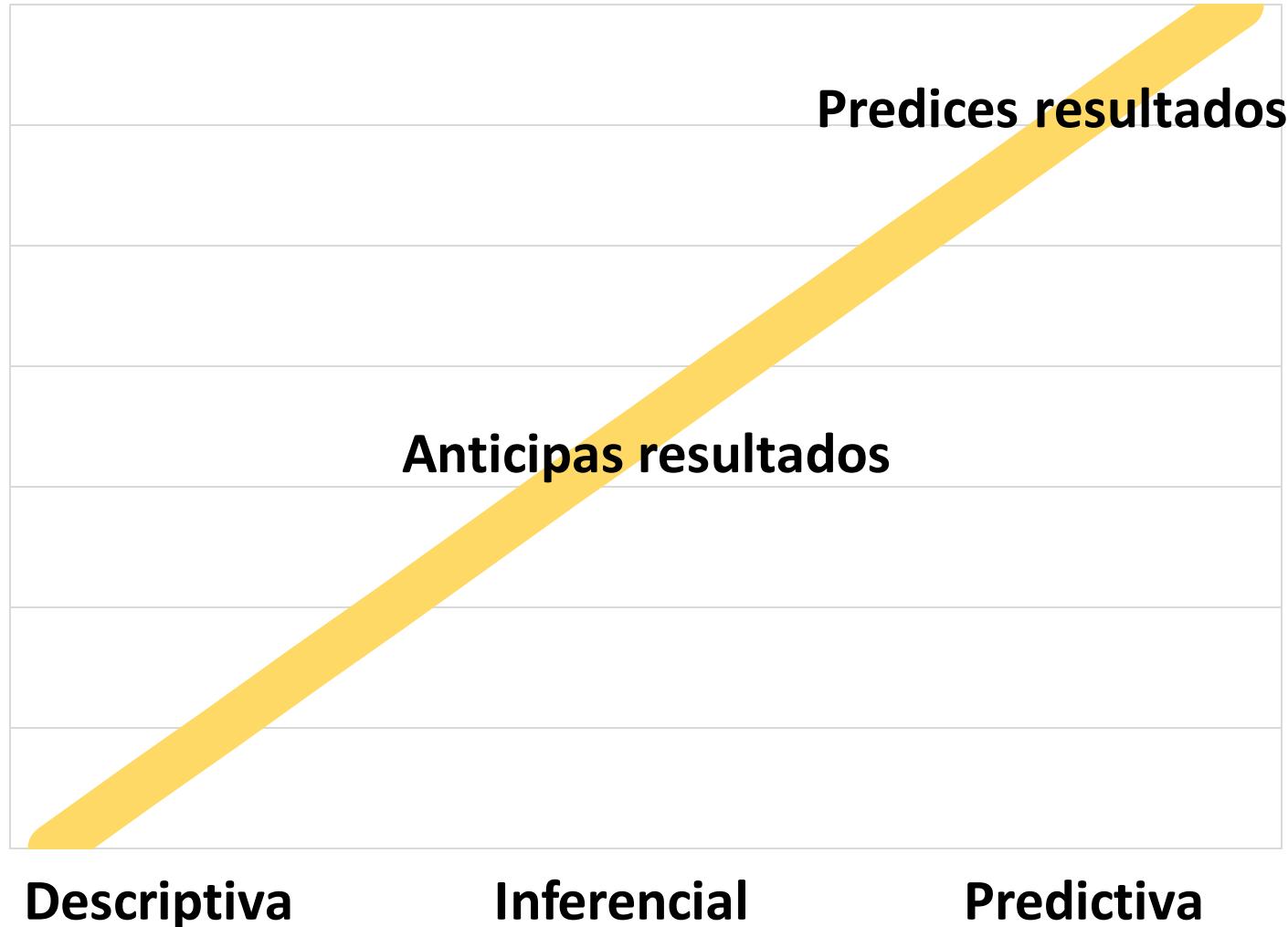
Definición #2: econometría

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

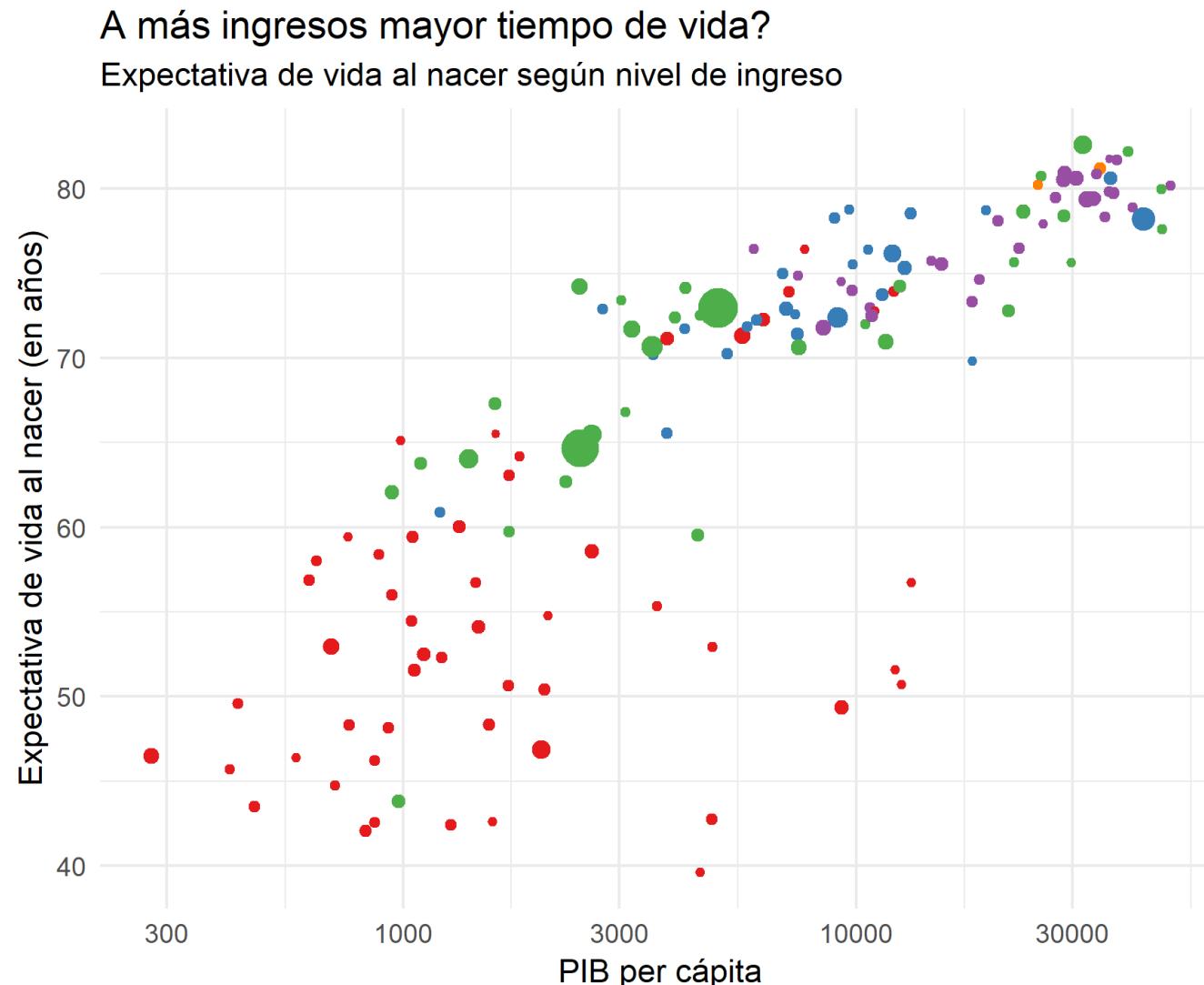
Contraste de hipótesis, probabilidad

Modelos econométricos: regresión, probit, logit

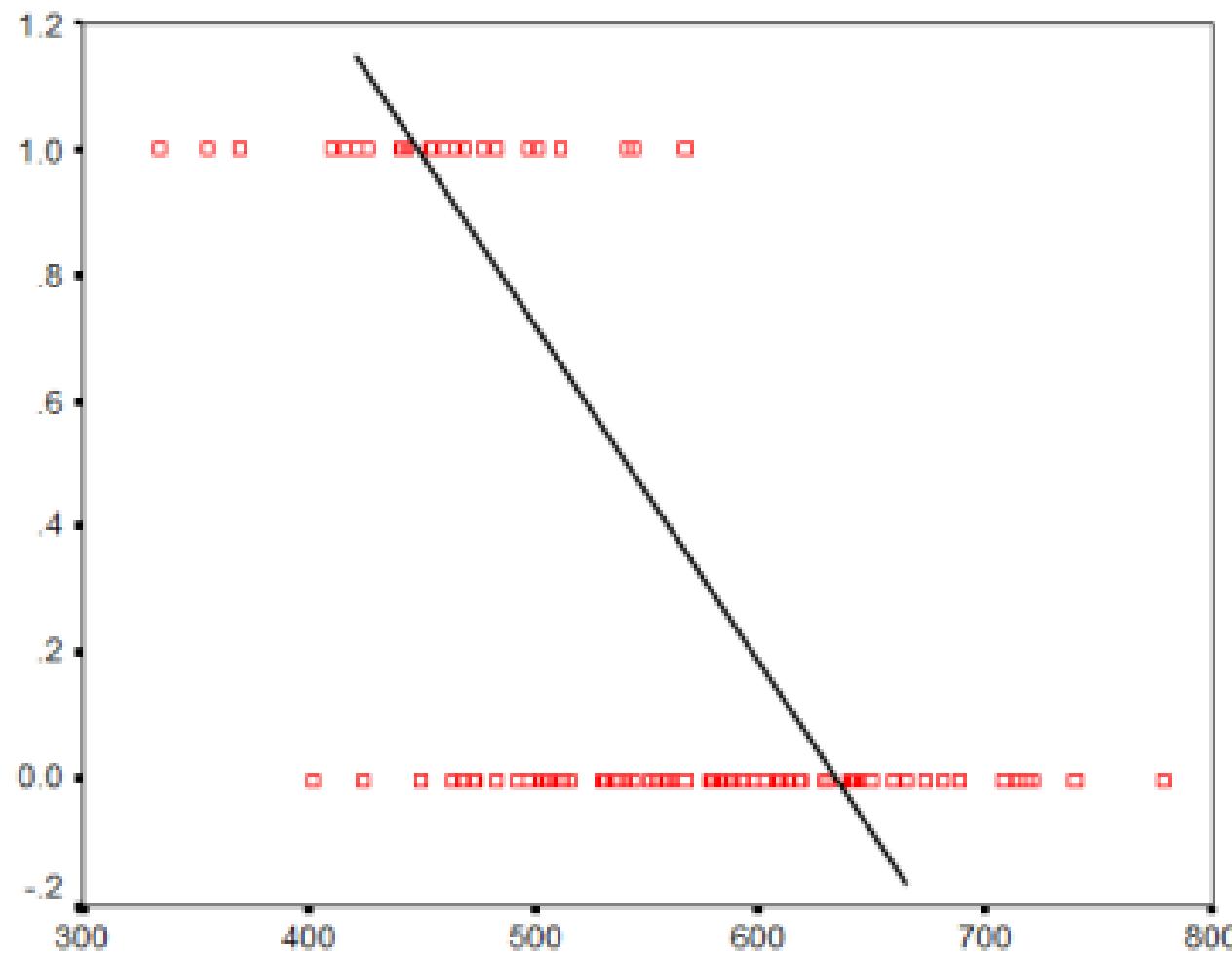
Tipo de variables en la estadística

#	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)

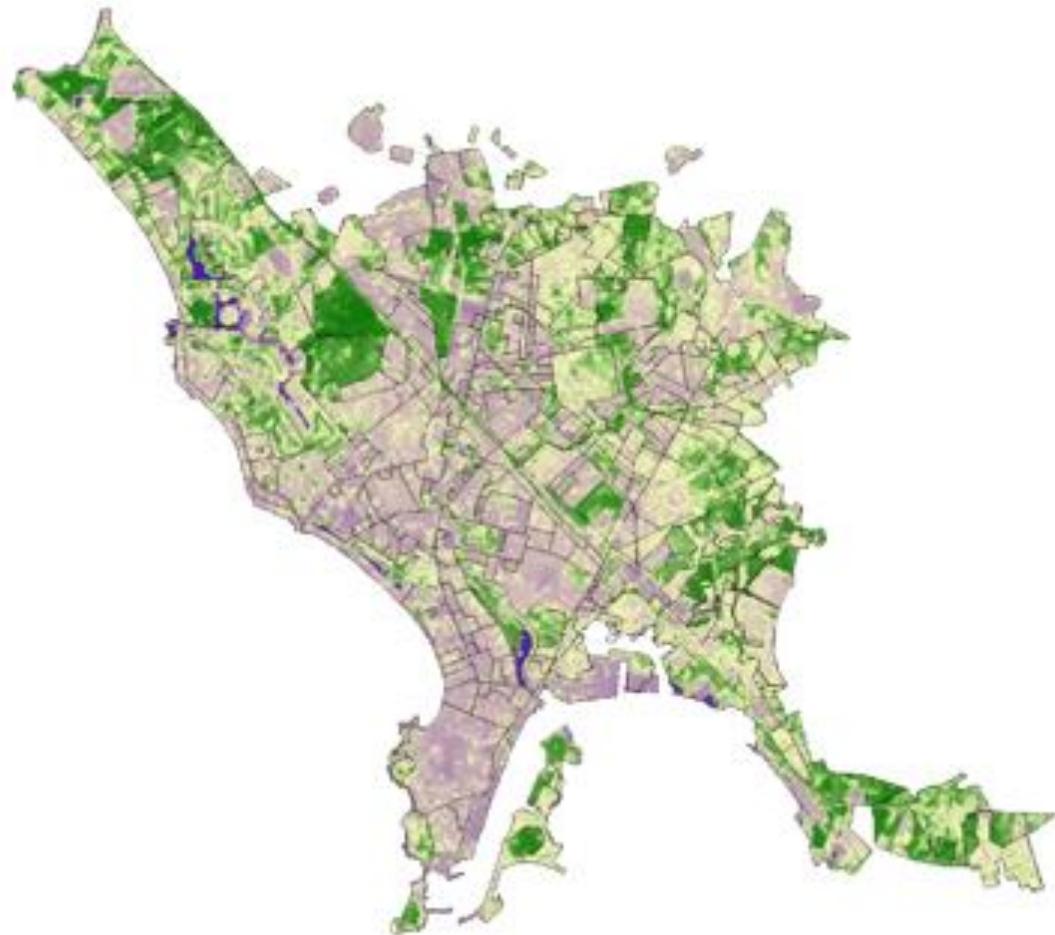


Variables cualitativa con cuantitativa: se contrasta la variable cualitativa de si una persona ha contratado un servicio, 1 = si contrata, y 0 = si no contrata vs variable cuantitativa: ingreso.

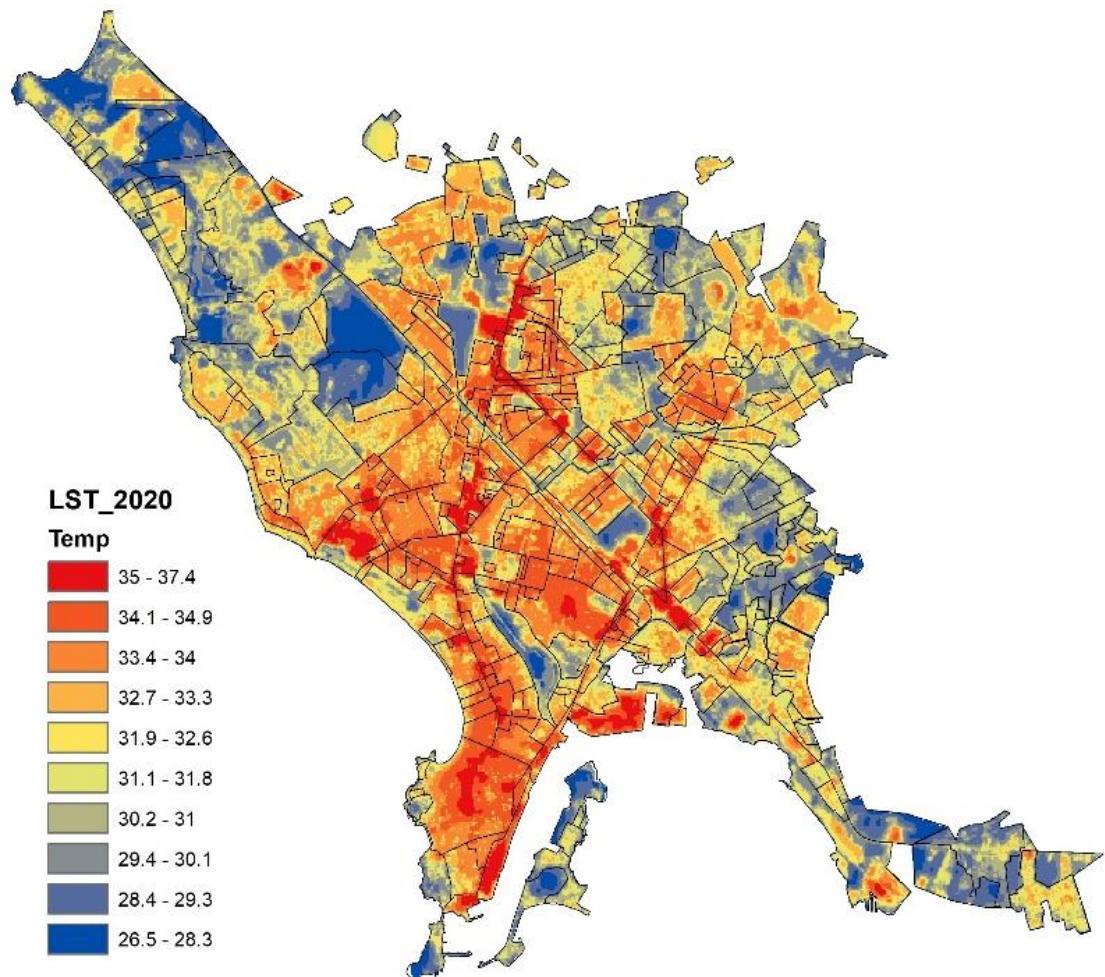


Variable georreferenciada: áreas verdes en Mazatlán y evidencia de isla de calor

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.

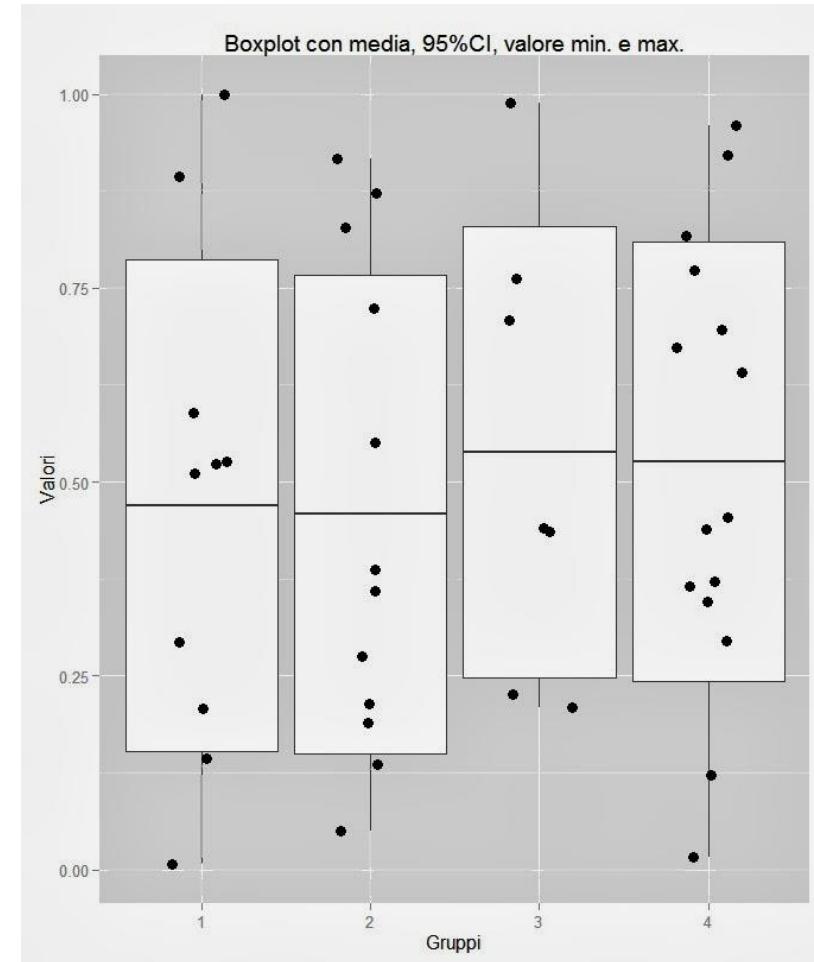
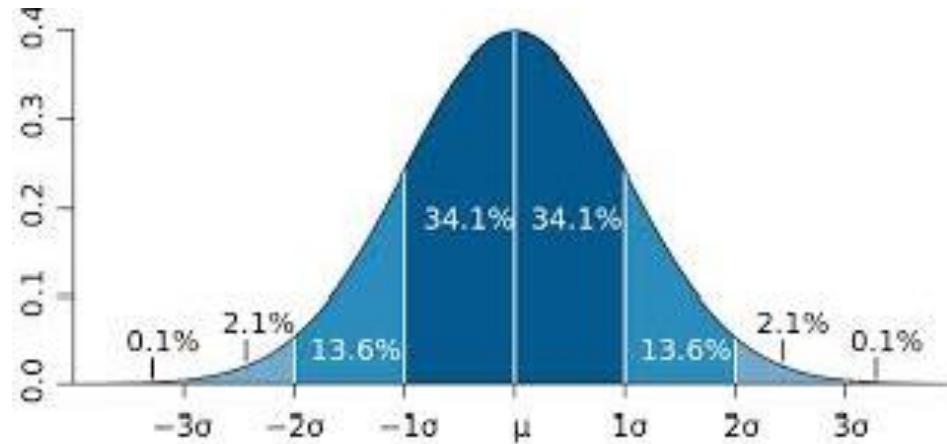
Ejemplo

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

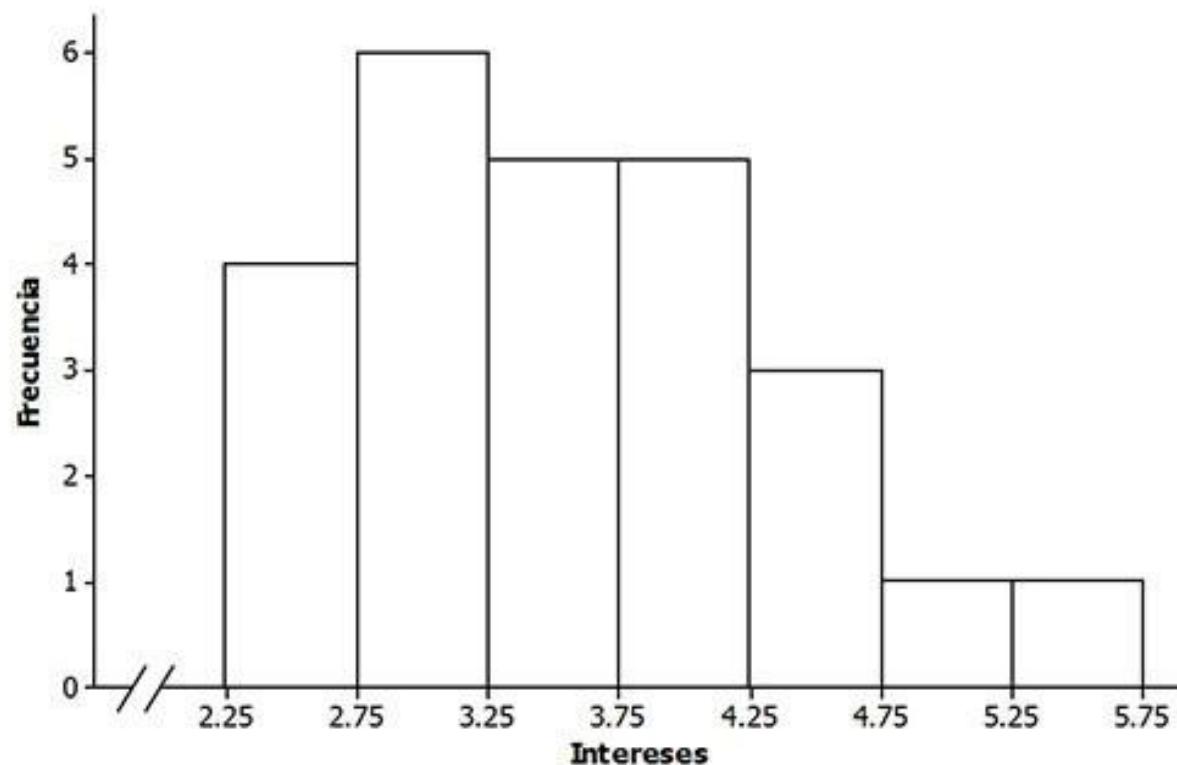
$$T_c \geq \frac{793}{6} \geq 132.1 \geq 133$$

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

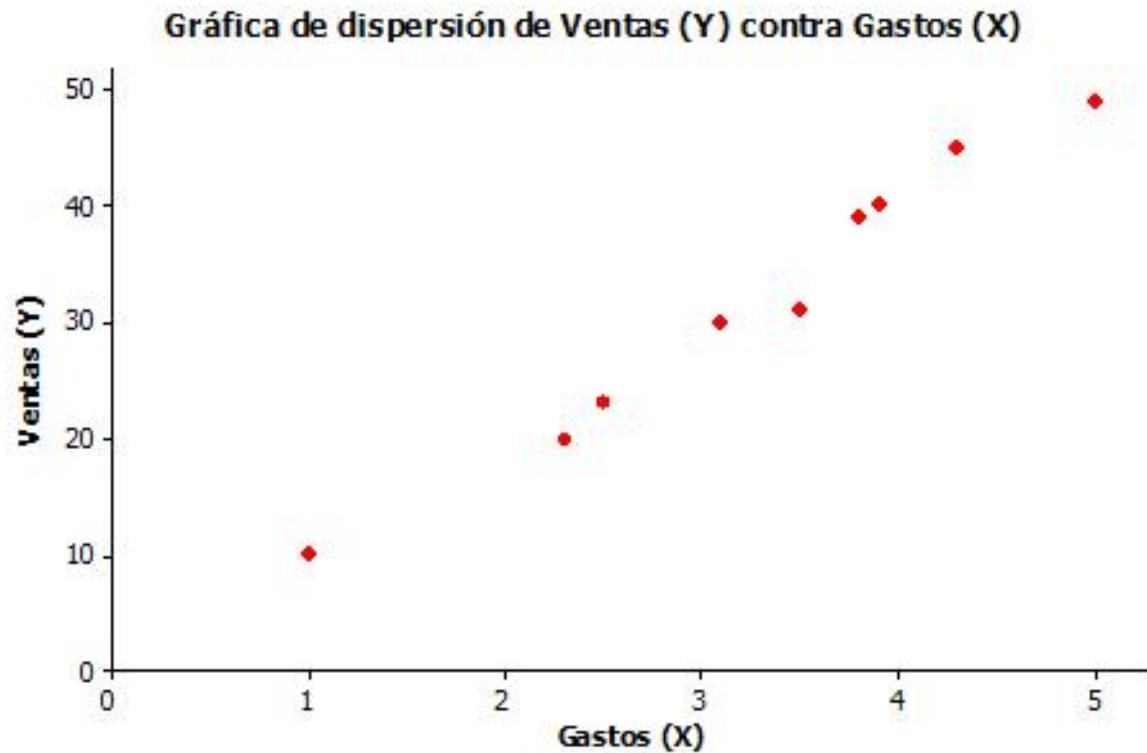
“El histograma condensa los datos, agrupando valores similares en clases. Se puede construir un histograma colocando la variable de interés en el eje horizontal, y la frecuencia, frecuencia relativa o frecuencia porcentual, en el eje vertical” (Hanke y Wichern, 2010).



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:

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

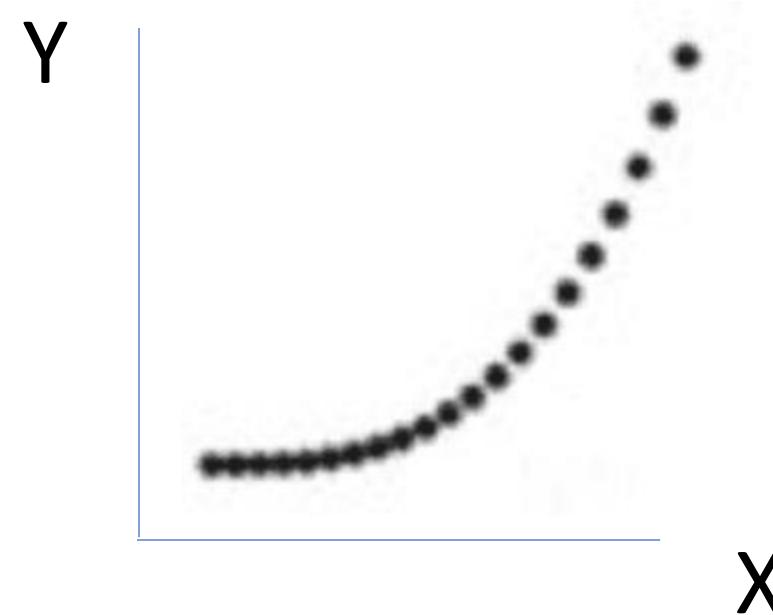
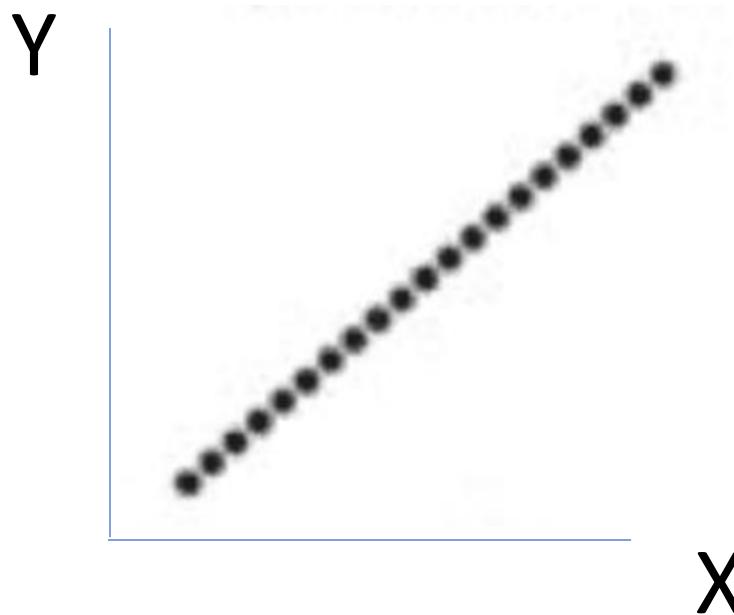
Dependiendo de su valor, nos dirá si hay una relación positiva o negativa. Existe una clasificación para medir su intensidad.

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

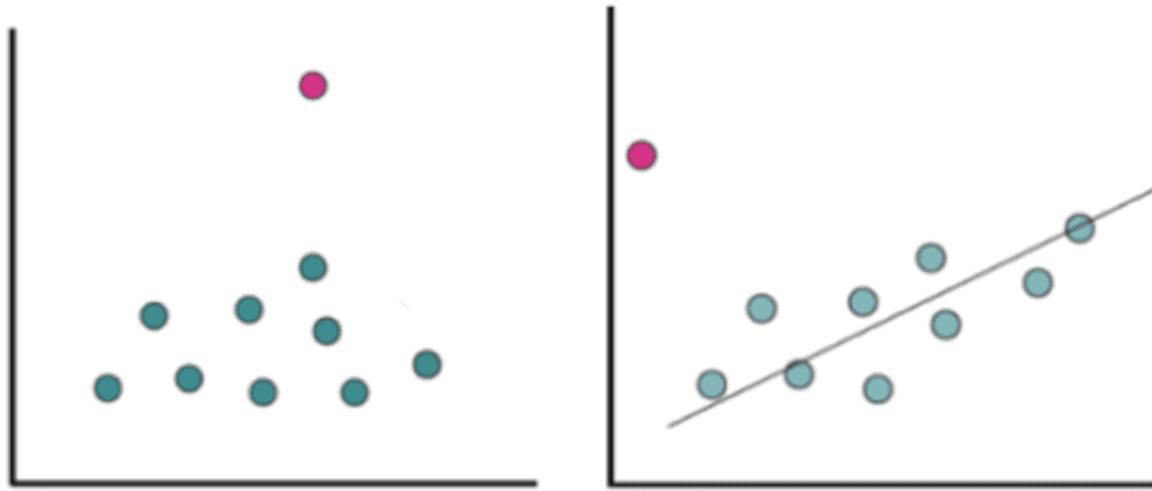
Resultado			Coeficiente de correlación lineal (negativo)
0.00	a	0.09	Nula
-0.10	a	-0.19	Muy débil
-0.20	a	-0.49	Débil
-0.50	a	-0.69	Moderada
-0.70	a	-0.84	Significativa
-0.85	a	-0.95	Fuerte
-0.96	a	-1.00	Perfecta

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

Forma: lineal o no lineal



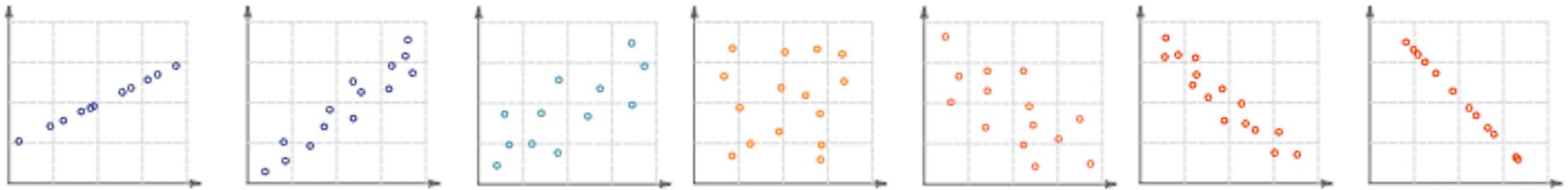
La **Presencia o no de datos atípicos (Outliers)**, puntos que no se ajustan al comportamiento del resto de la nube.



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:

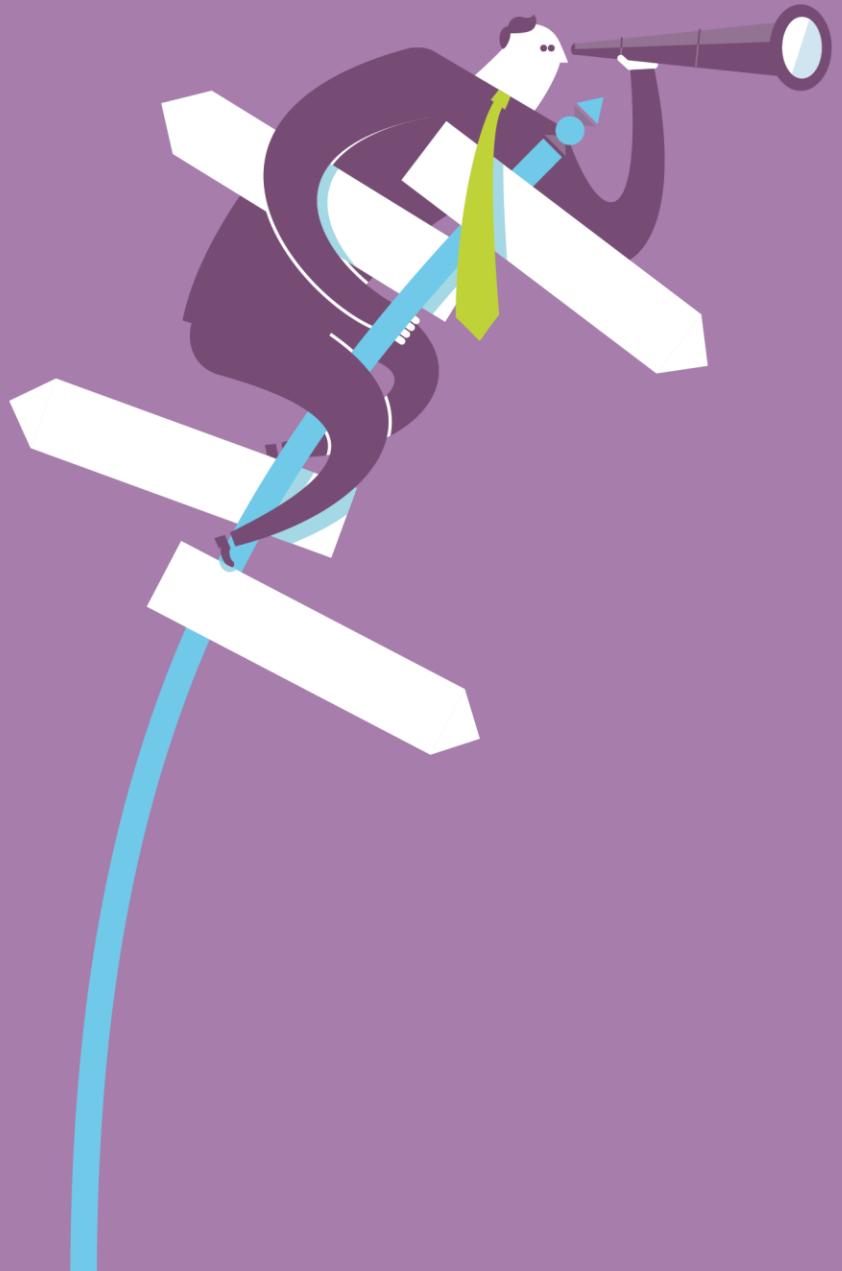
- a. ¿Qué han hecho durante la semana, referente a la materia?**
- b. ¿Que piensan hacer respecto a la materia durante la siguiente semana?**
- c. ¿En qué les podemos ayudar para que logren sus objetivos con el menor sufrimiento posible?**

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 $\Omega = \{e_1, \dots, e_k\}$.

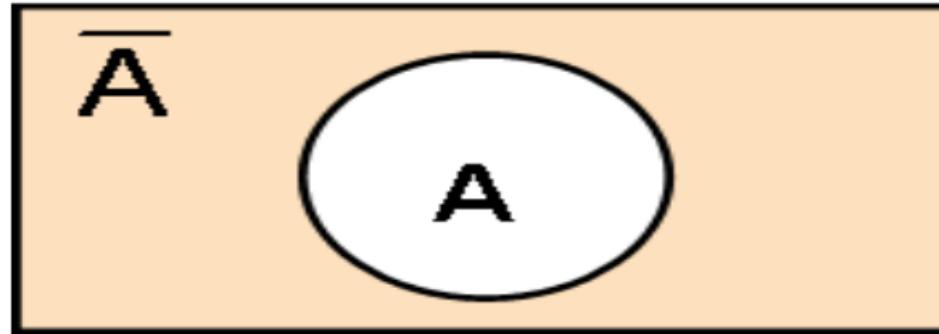
Ejemplo 4. En el caso de lanzar la moneda dos veces, los sucesos elementales son $e_1 = (C, C)$, $e_2 = (C, X)$, $e_3 = (X, C)$ y $e_4 = (X, X)$.

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

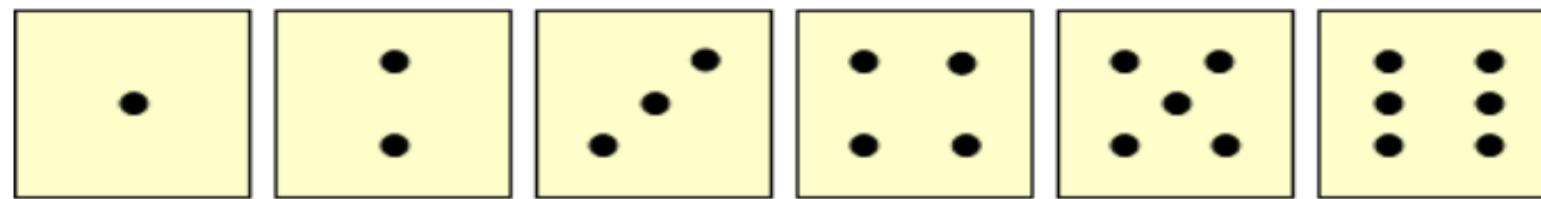
Ejemplo 5. En el caso de lanzar la moneda dos veces, el suceso A = “sale una cara y una cruz” es $A = \{(C, X)\}$.

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,

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

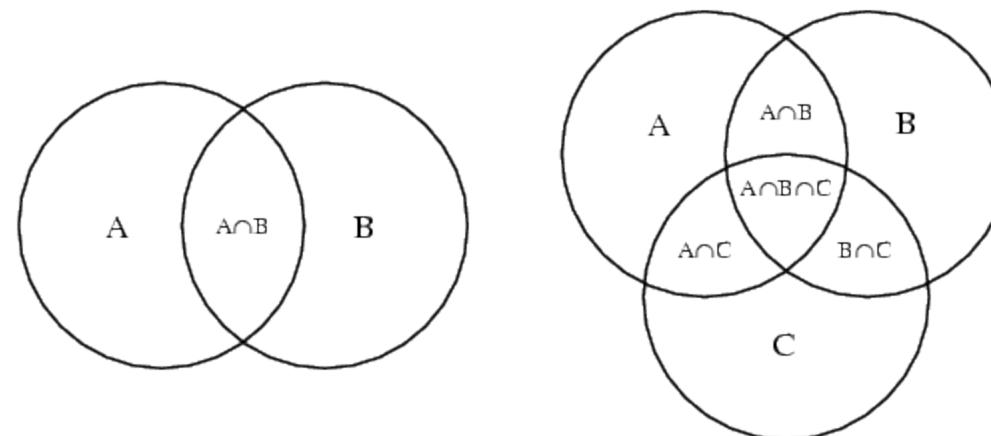
Reglas de la probabilidad

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,

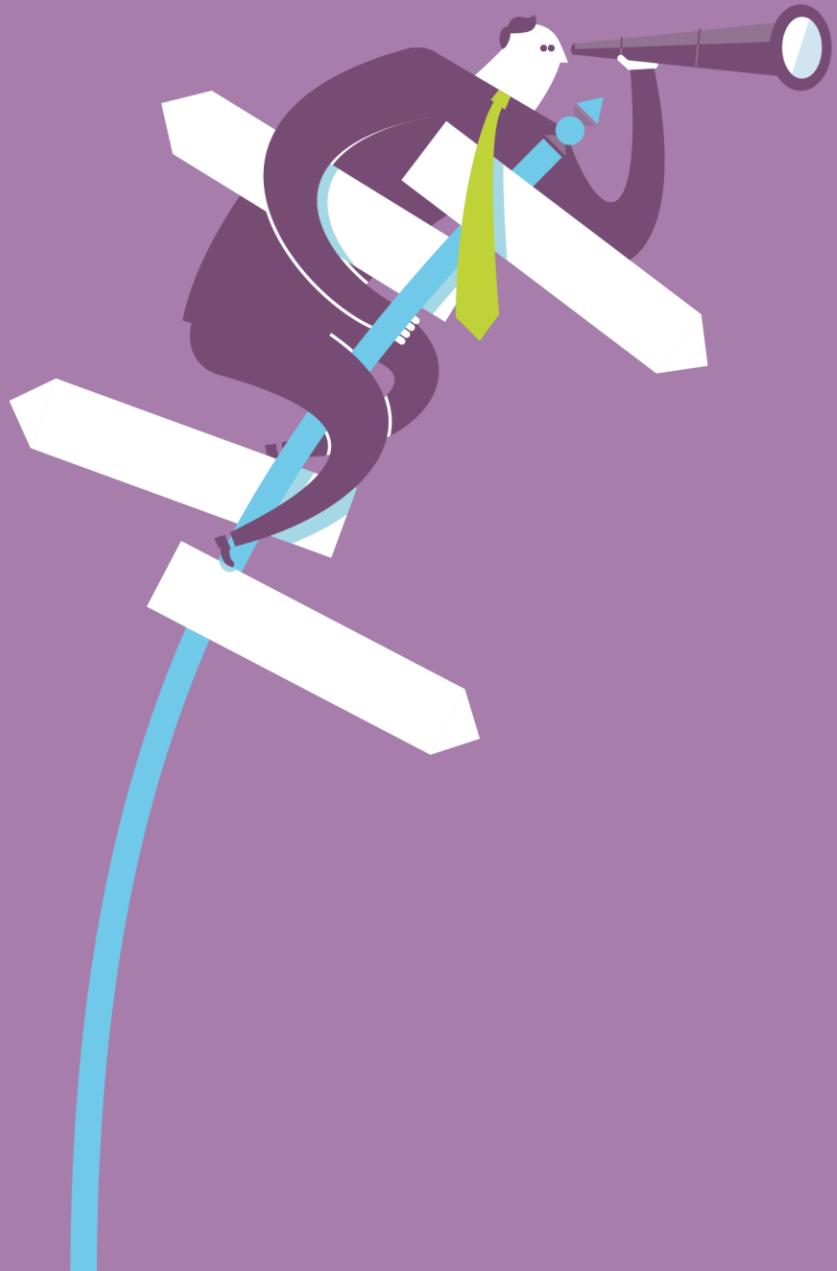
$$P(\text{ambas son blancas}) = \frac{2}{3} \times \frac{7}{11} = \frac{14}{33} = 0.424$$

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.

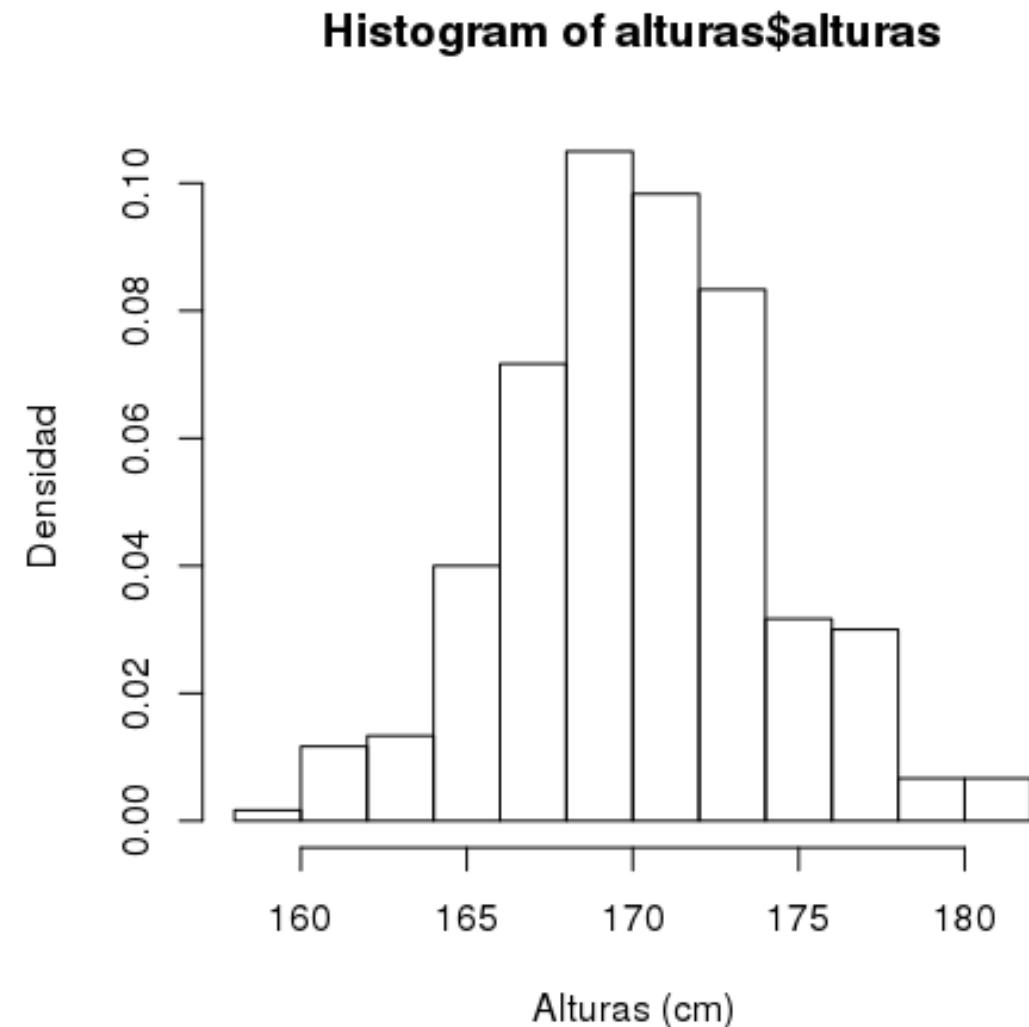
Distribución normal: una pequeña intuición.

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**.

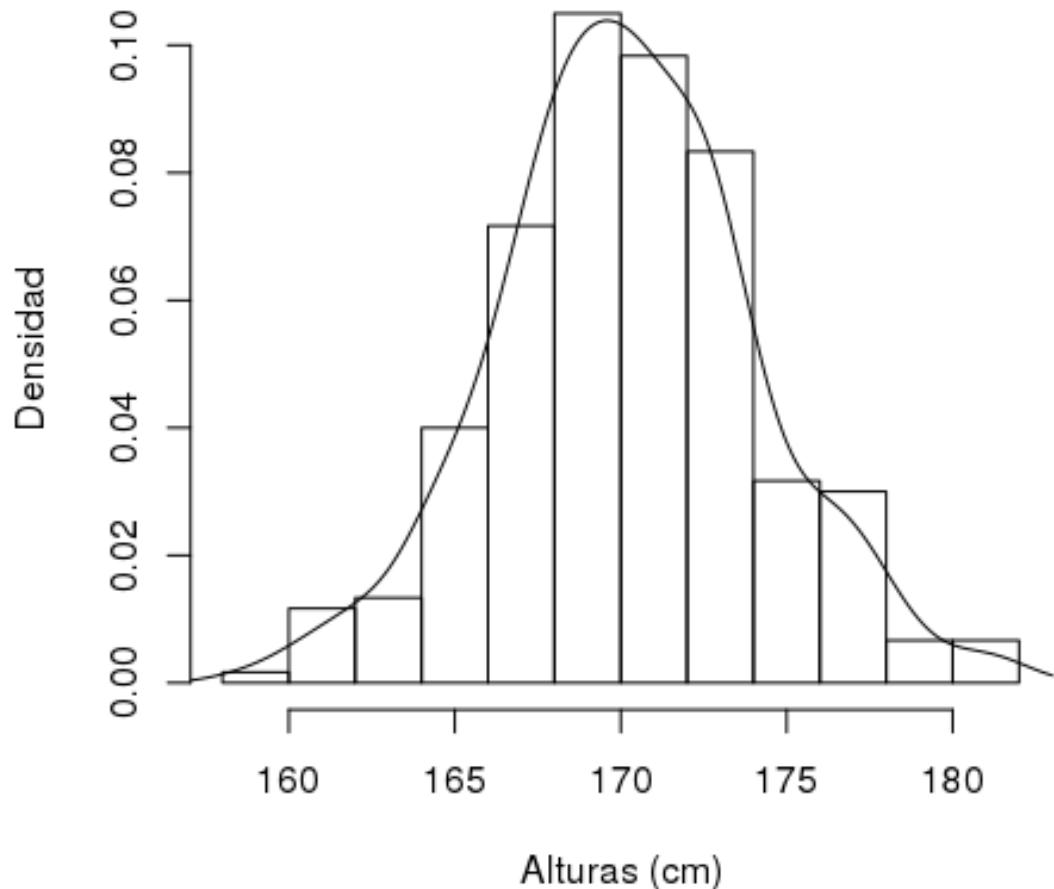
Para entender la distribución normal pensemos el siguiente ejemplo:

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

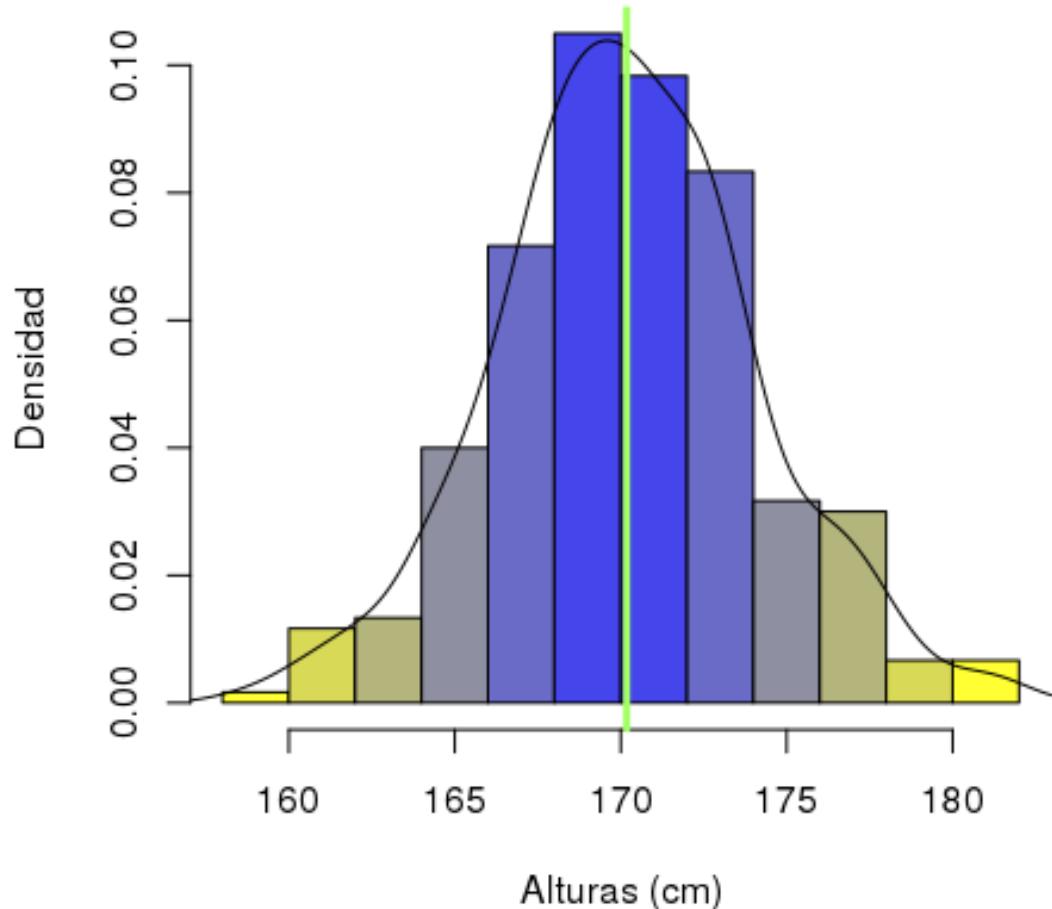


Adicionalmente si dibujamos una curva de densidad en el histograma ...

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.

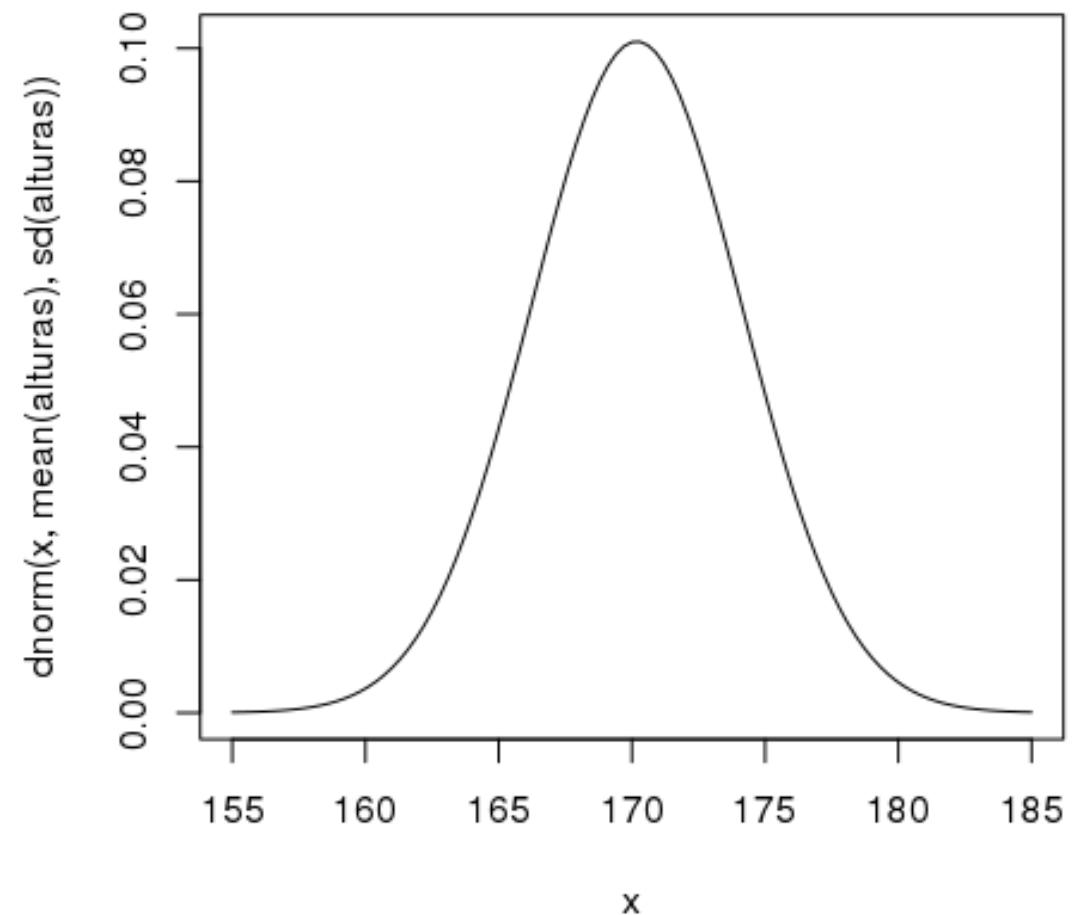
Esta distribución también es simétrica, es decir, si trazamos una línea recta vertical (la media), la parte izquierda será similar a la parte derecha dividiendo el conjunto de valores en 50%|50%.



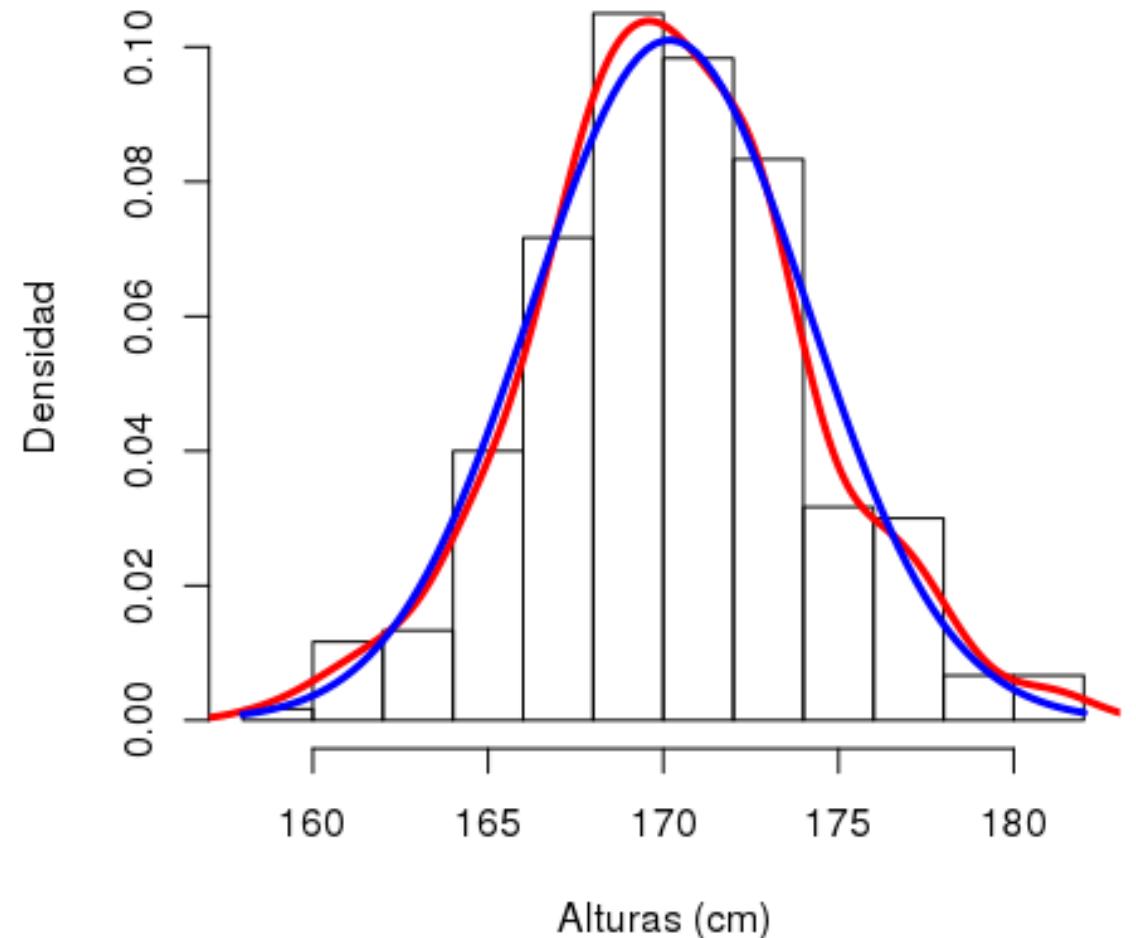
Se dice que la
distribución normal
es simétrica ...

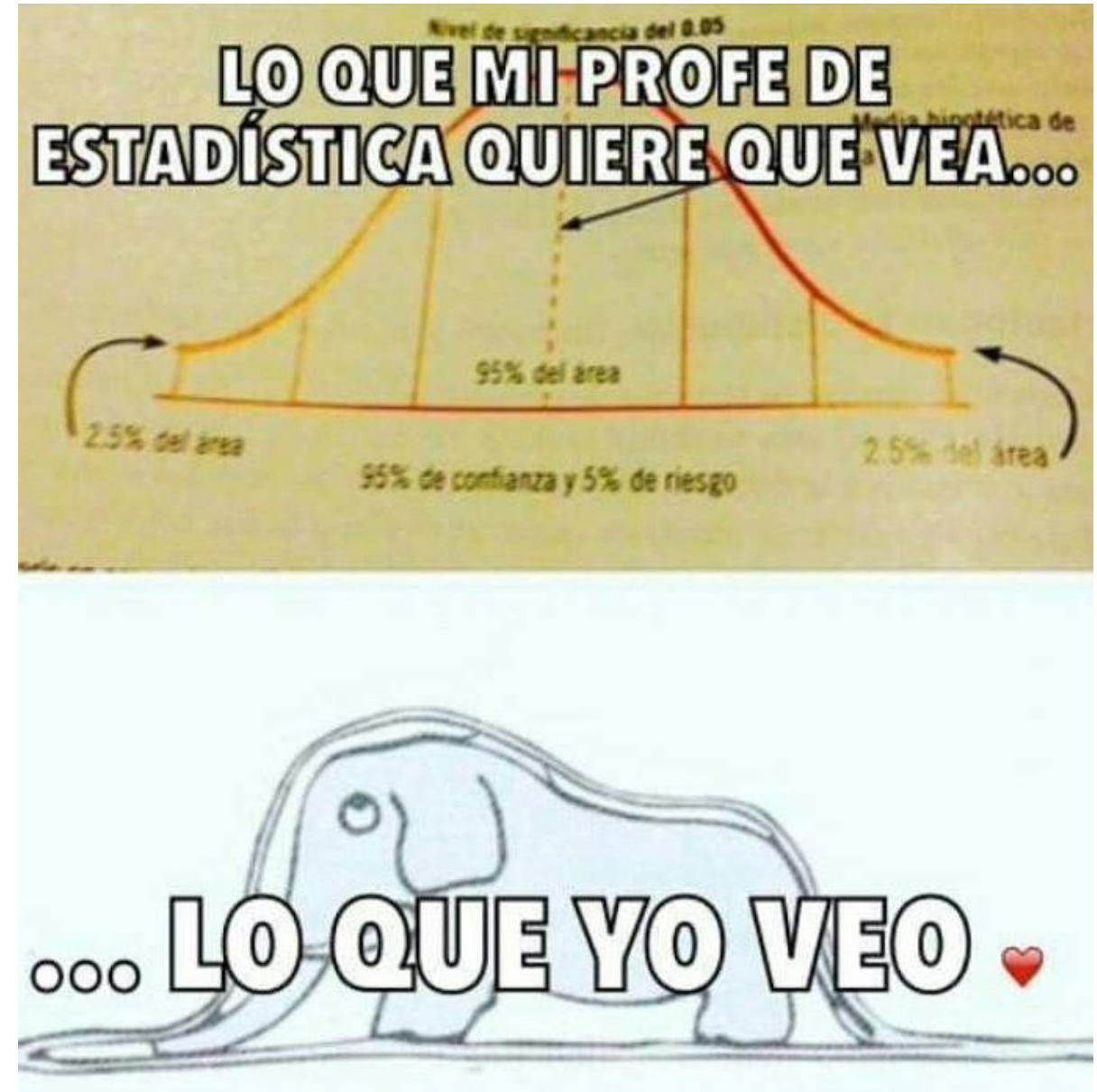
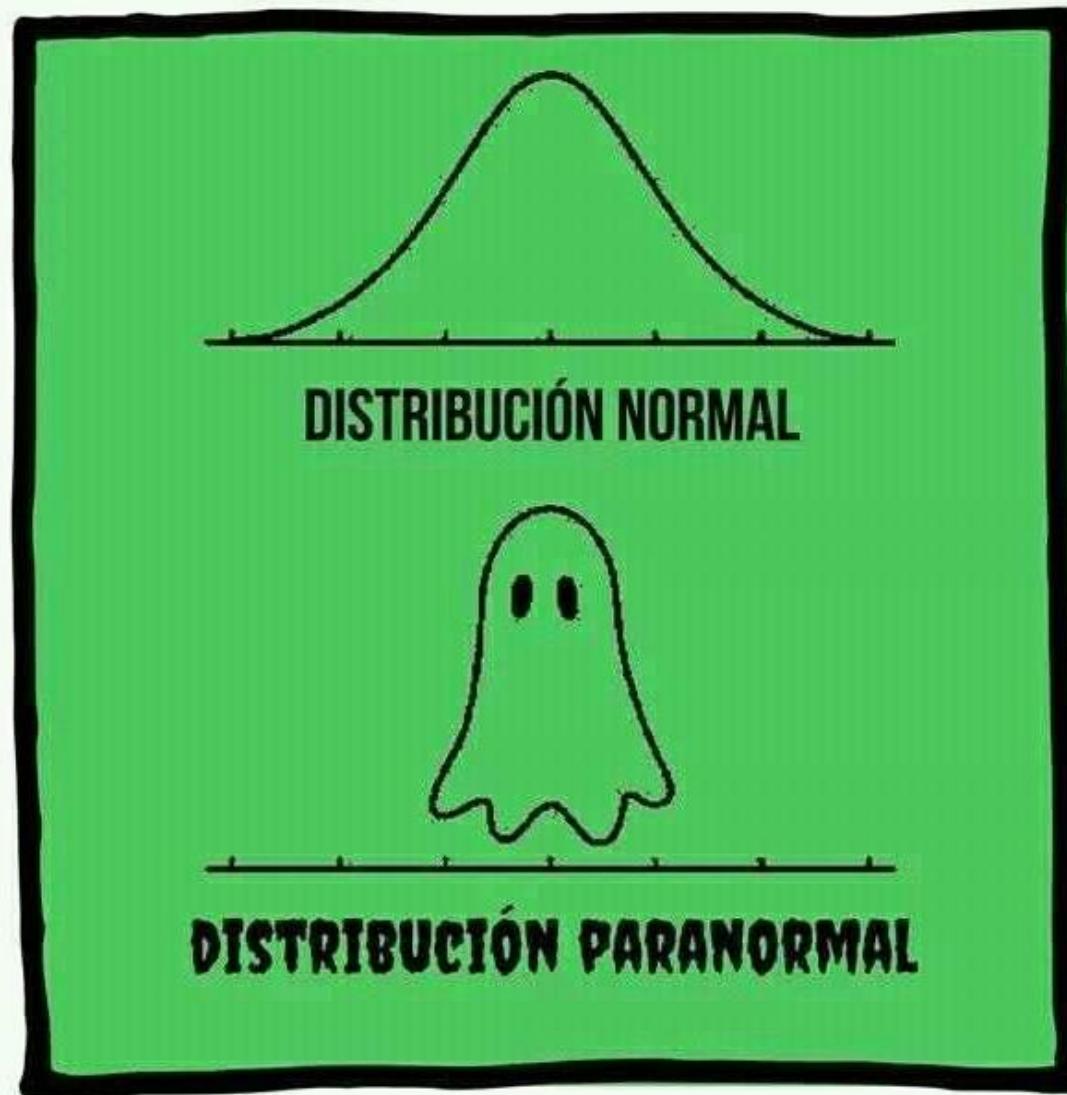
Una ventaja de la distribución normal es que puede explicarse con la media y la desviación estándar, por ejemplo, podemos generar una curva teórica a partir de los valores de la media y la desviación estándar de nuestro conjunto de alturas.

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



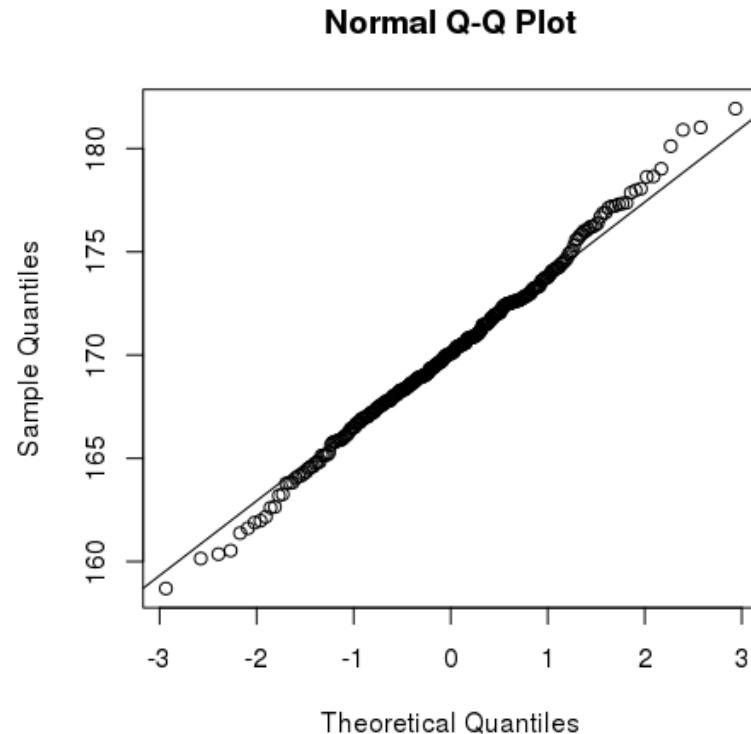
Como se puede ver la curva roja (observada) y la azul (teórica) se parecen mucho. Cuando esto ocurre decimos que la curva de nuestra muestra se ajusta a una distribución normal (posiblemente porque la variable poblacional también sigue la misma distribución).





Tema extra: *Quantile – Quantile Plot*

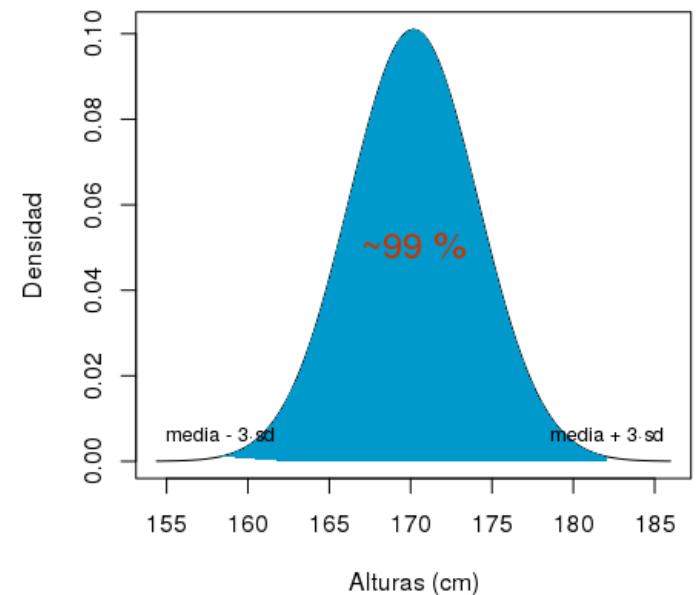
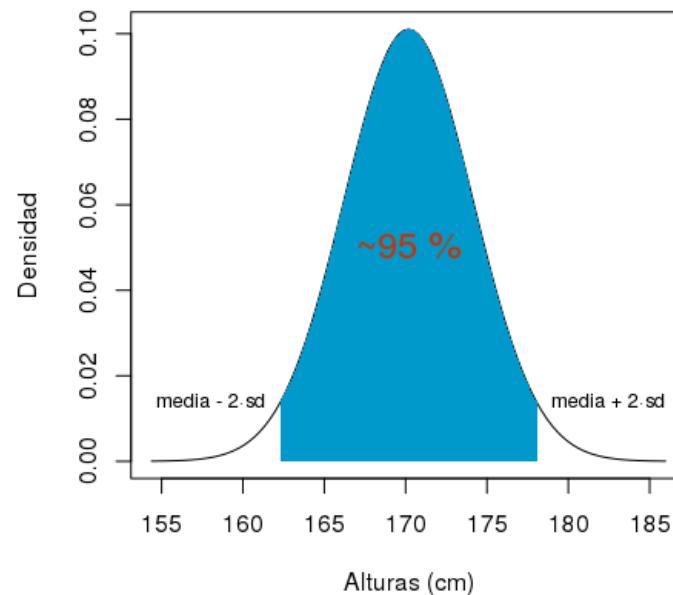
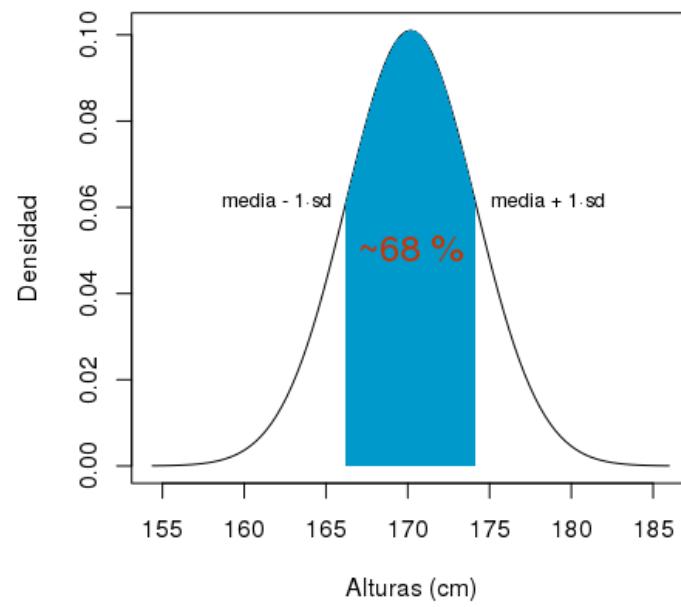
Otro tipo de gráficos que podemos hacer para conocer el ajuste a una distribución normal es el *Quantile-Quantile Plot* más conocido como *q-q plot*.



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.

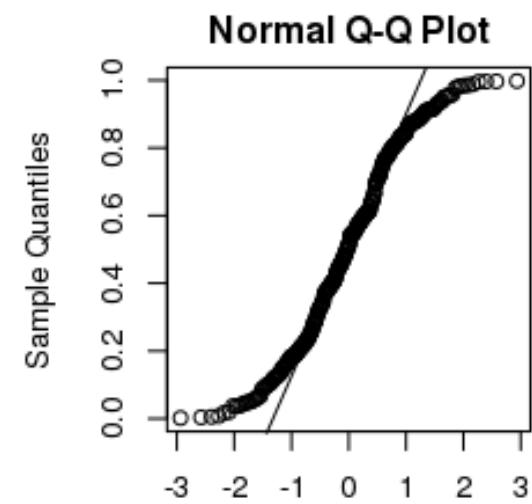
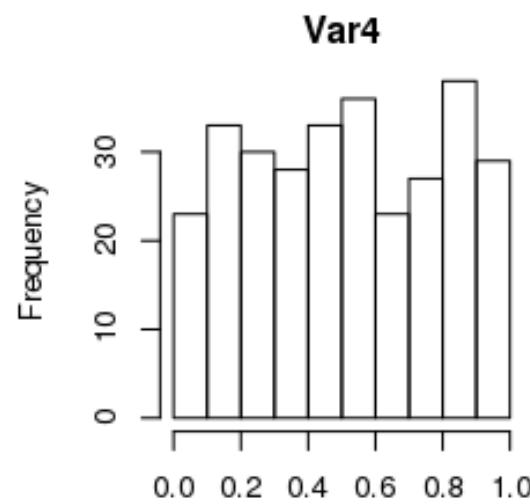
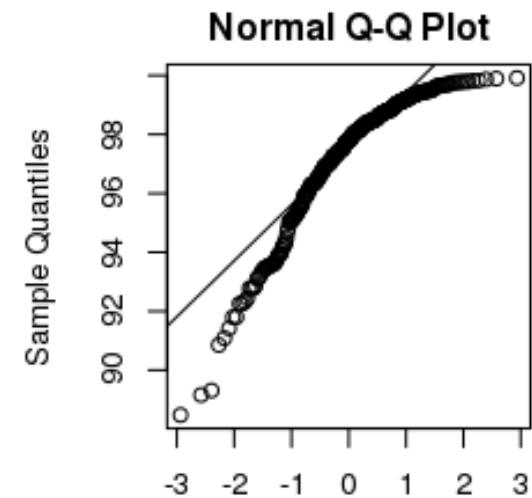
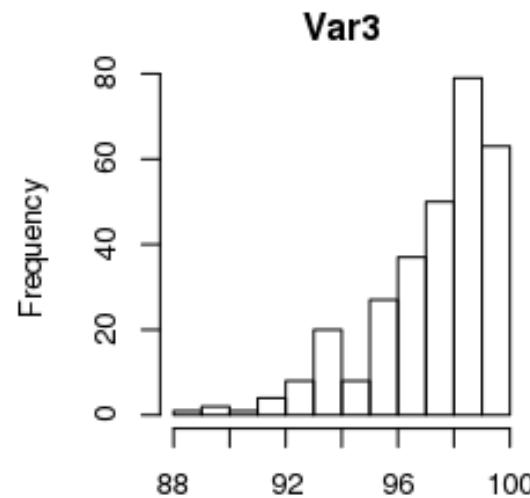
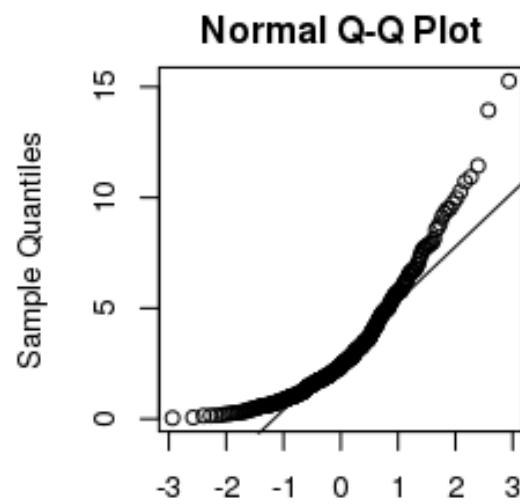
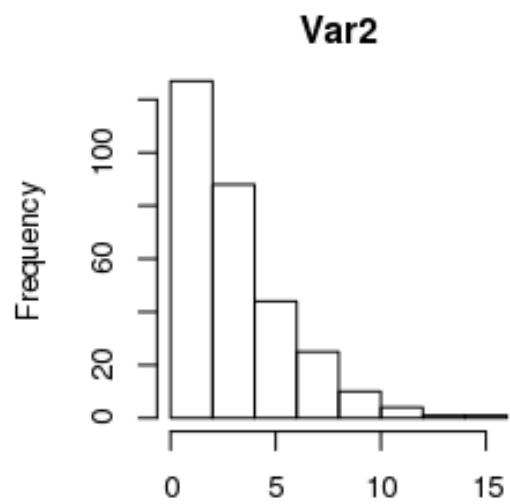
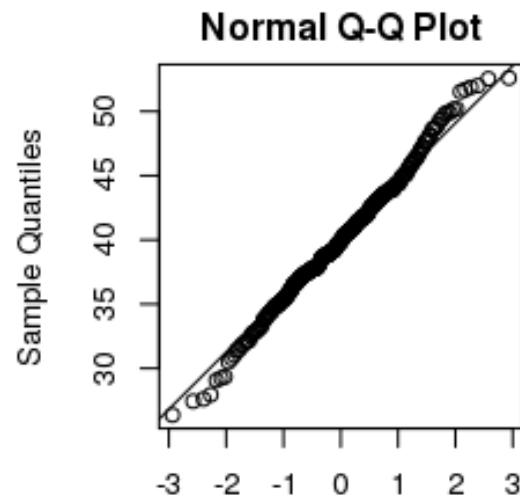
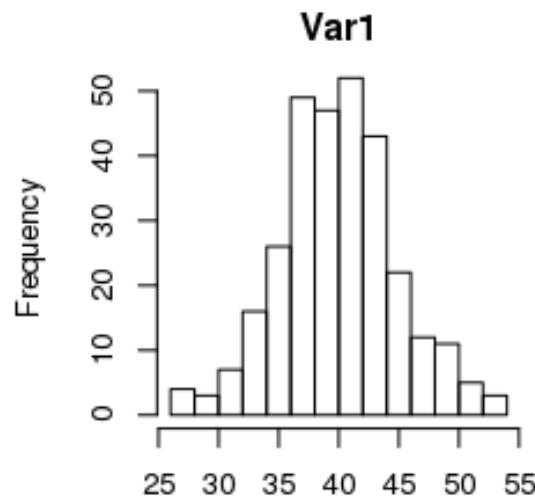
Volviendo a la curva normal, existe una característica muy interesante cuando consideramos la media y la desviación estándar.

- 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.

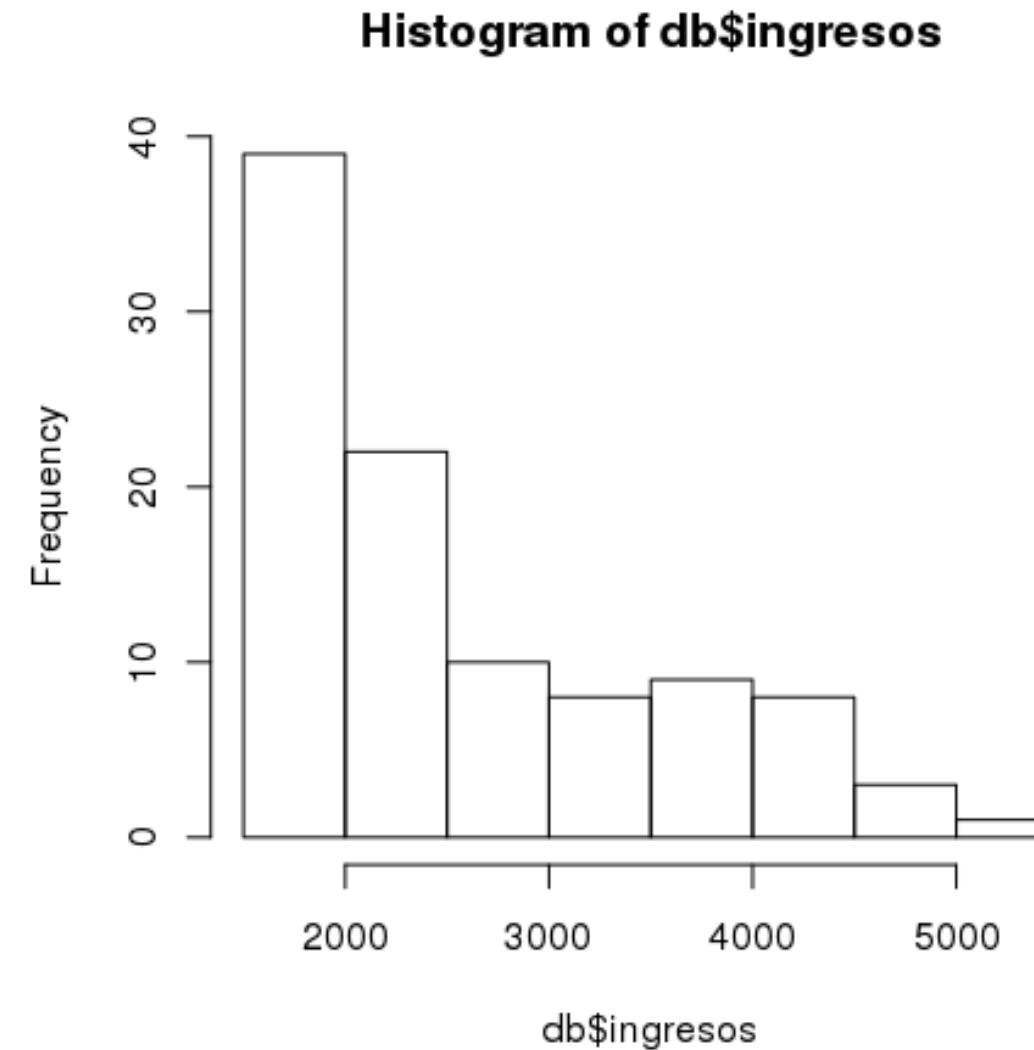
Si tenemos una variable que sigue una distribución normal, podemos usar la media y la desviación estándar para describirla pero qué hacemos si tenemos una que no lo hace, por ejemplo, los ingresos económicos.

Utilizaremos otro tipo de modelos.

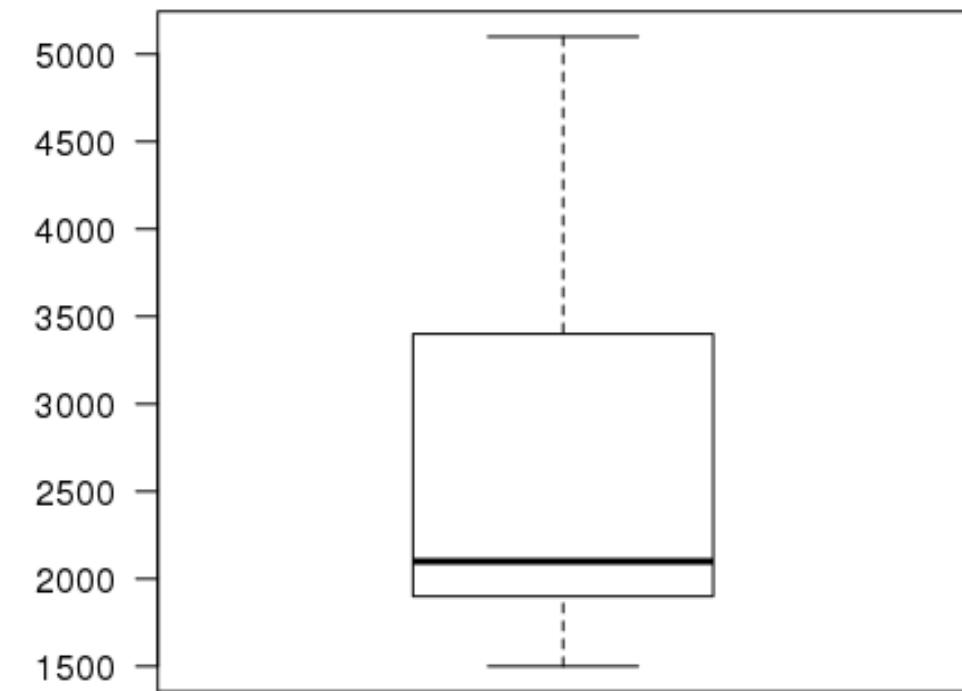
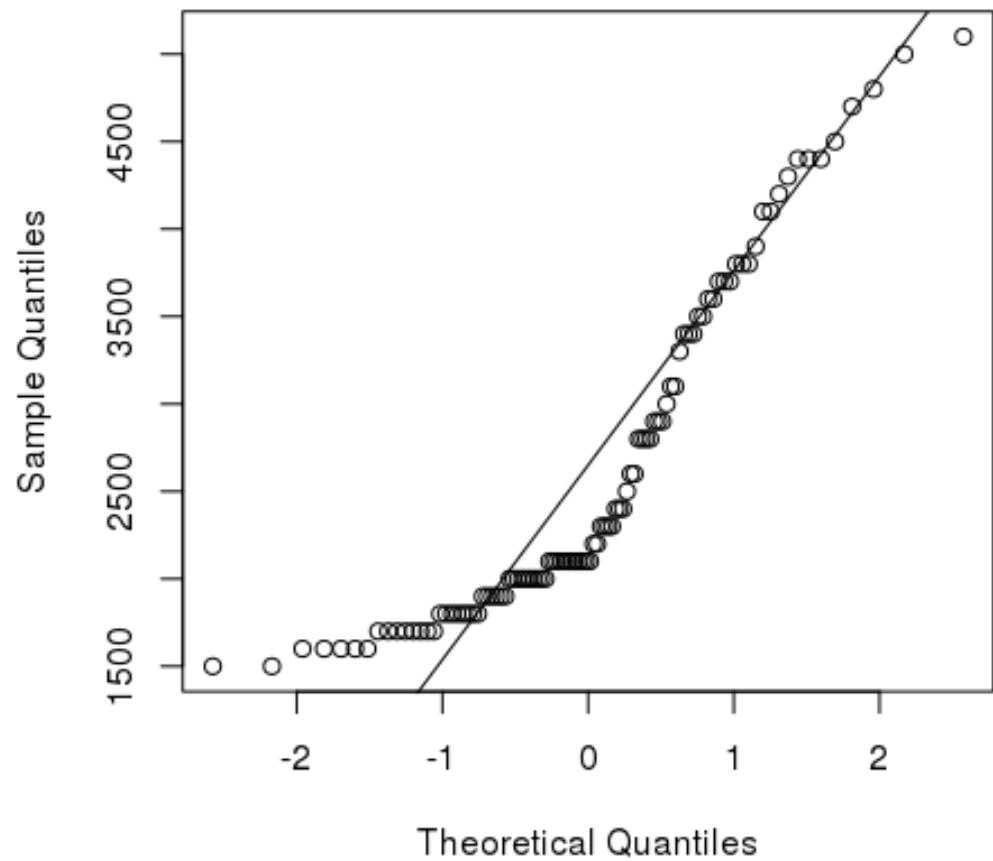
Por ejemplo, veamos los ingresos económicos de personas en una ciudad X.

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:

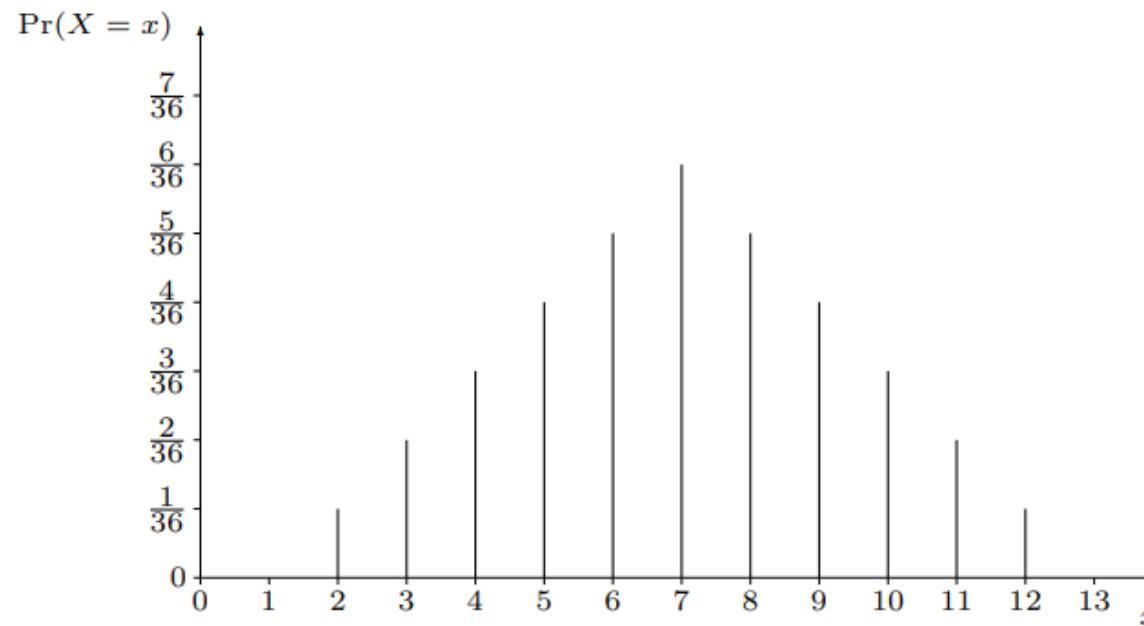
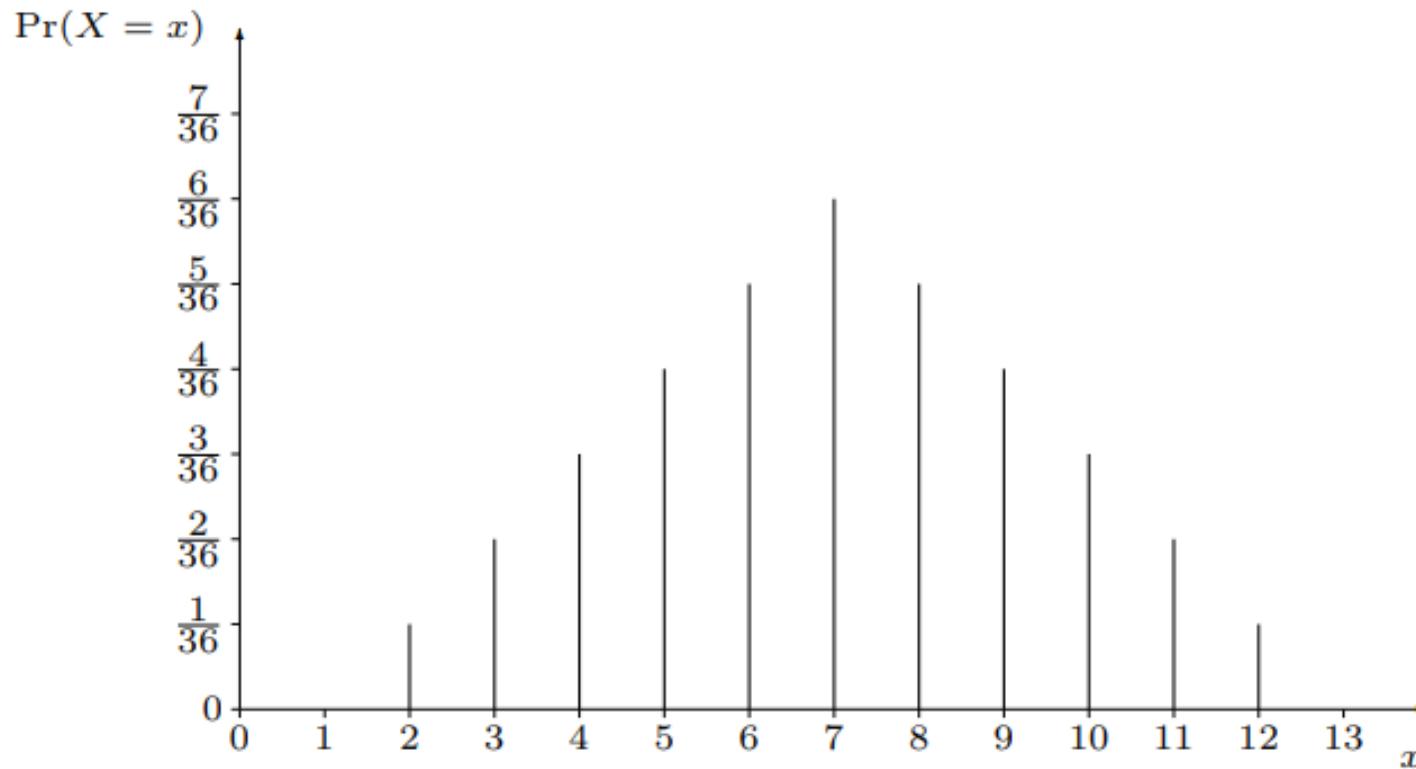


Gráfico de la función de probabilidad de X



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

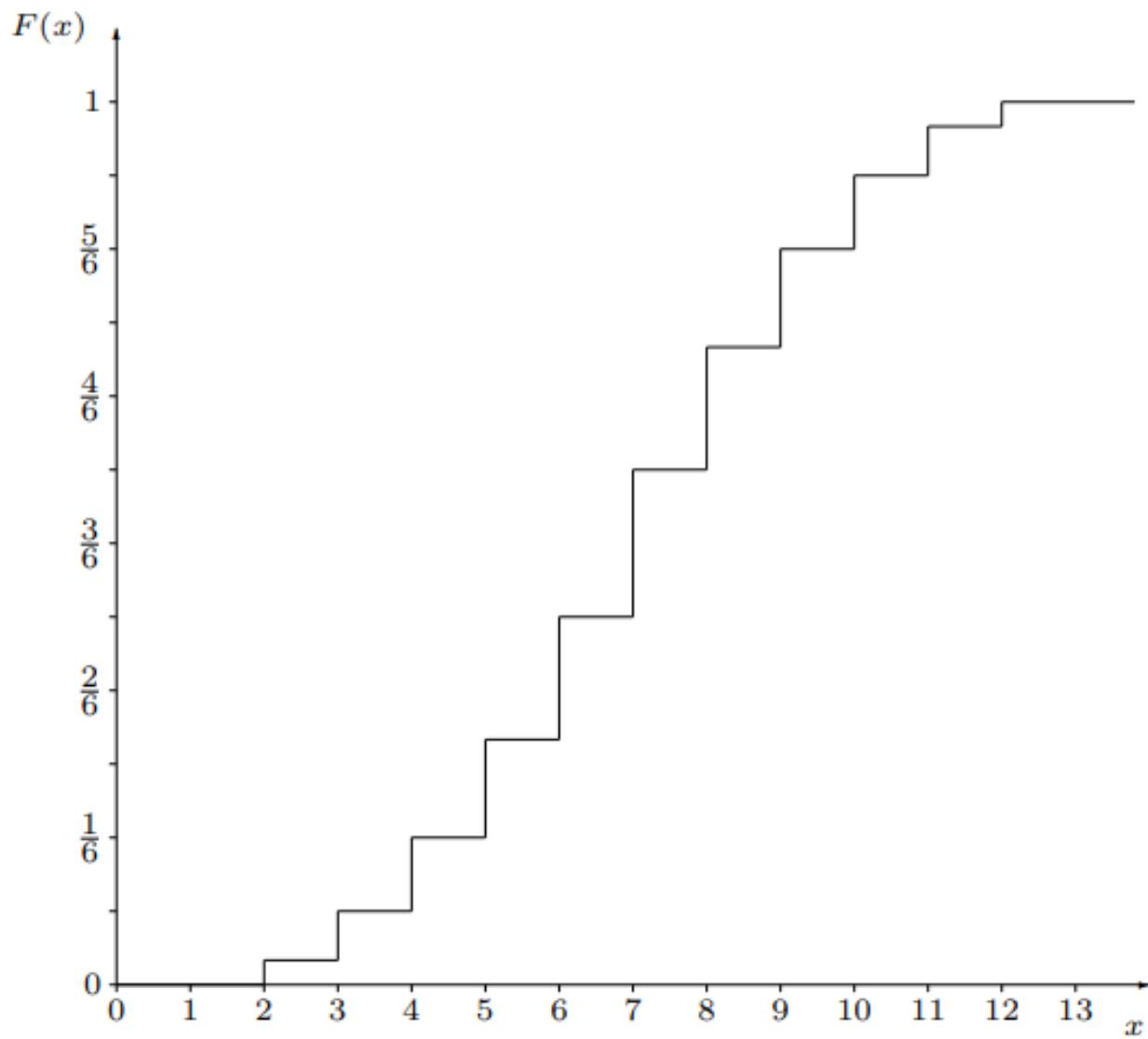
Función de distribución de una v.a. discreta

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

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

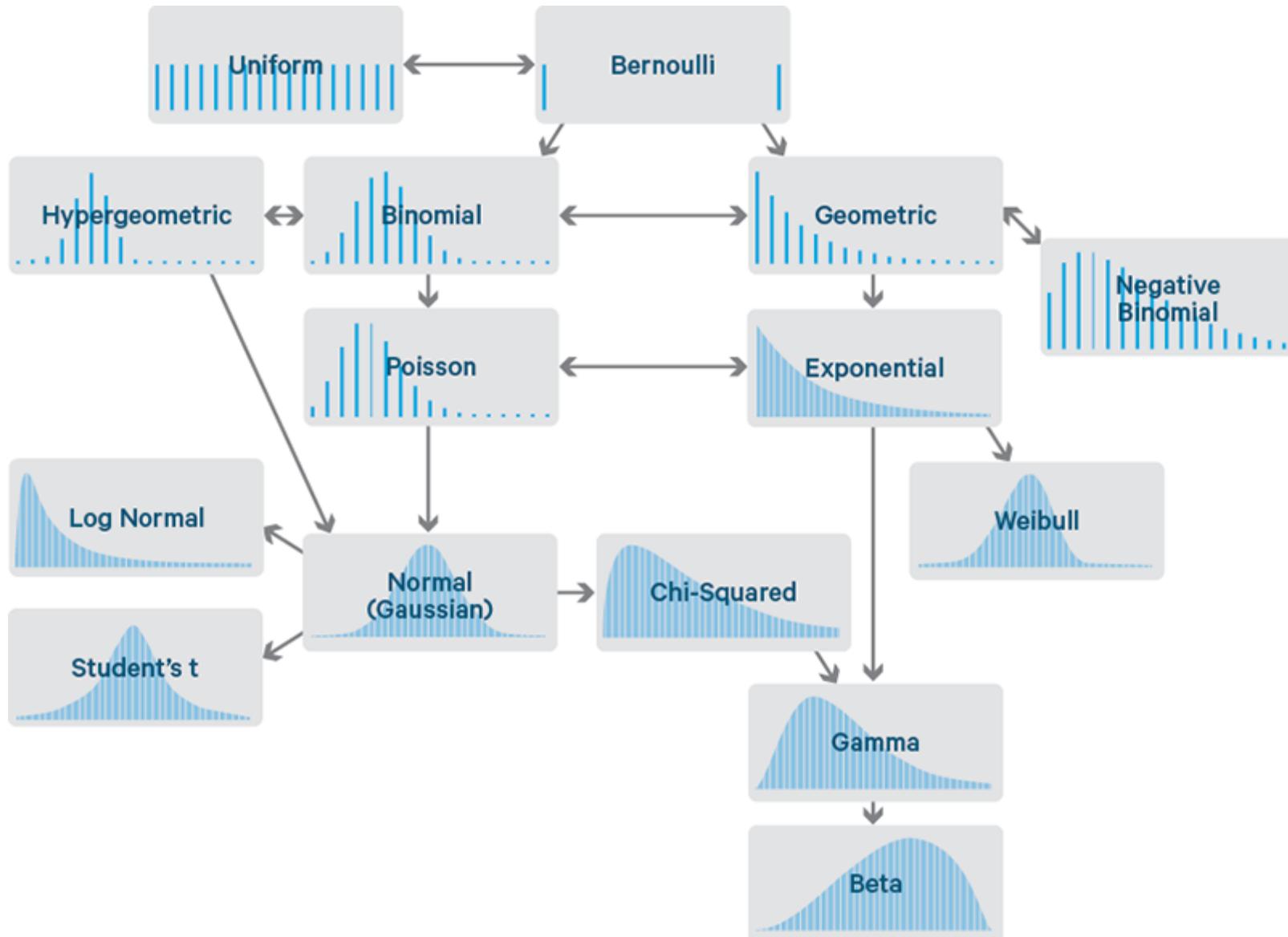
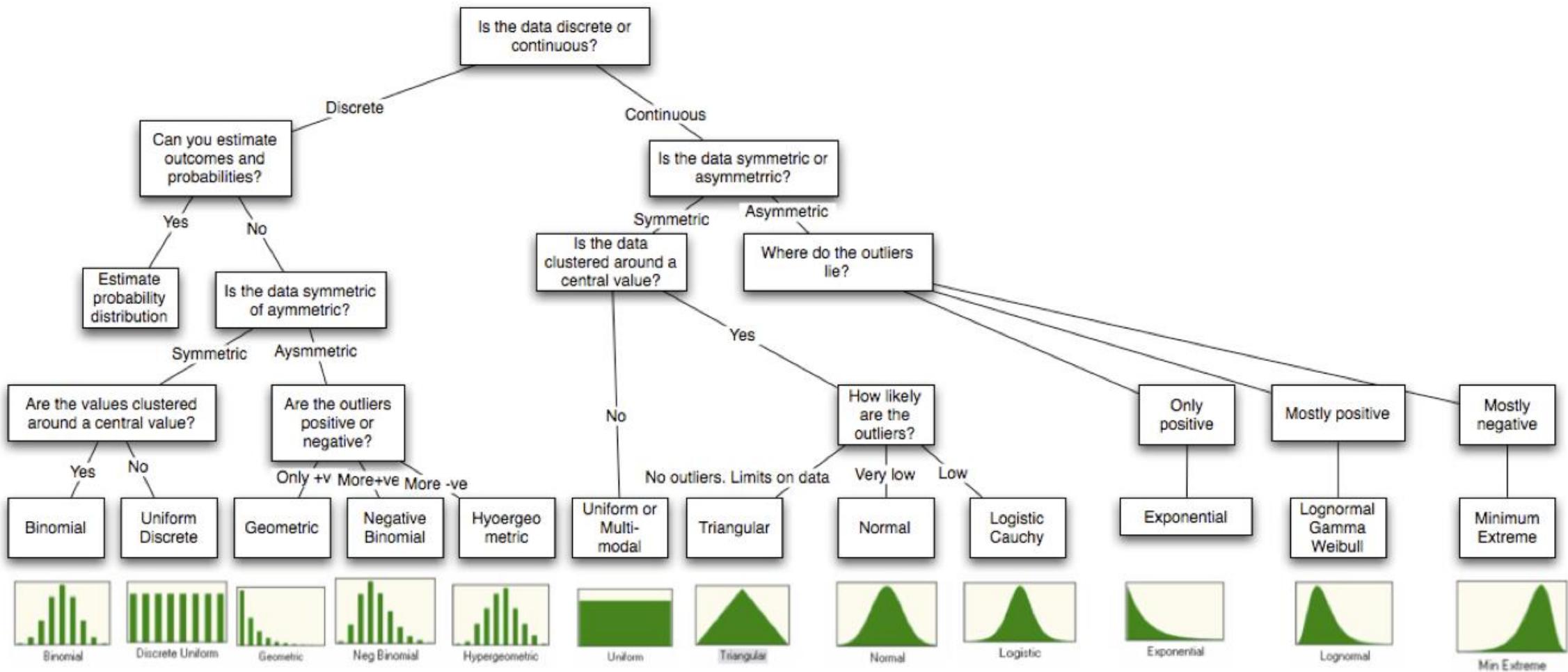


Figure 6A.15: Distributional Choices



¿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?

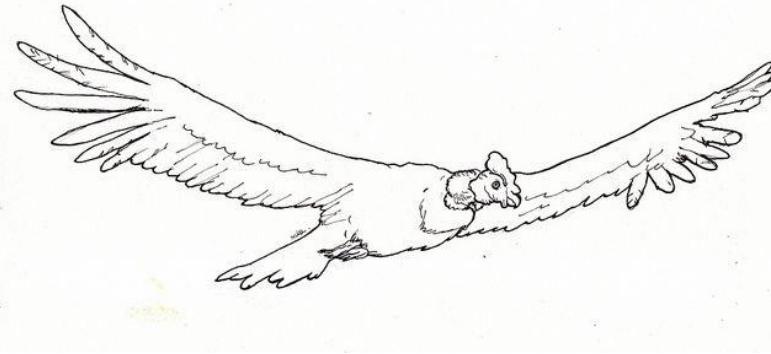
Inicialmente para conocer una característica o variable de la realidad, requerimos información completa sobre la población que analizamos. En ocasiones esto se busca lograr mediante el levantamiento de información censal o registros administrativos (por ej. censos poblacionales, reportes de resultados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

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

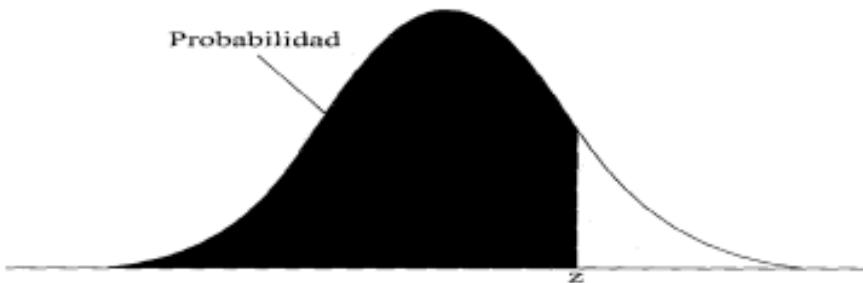
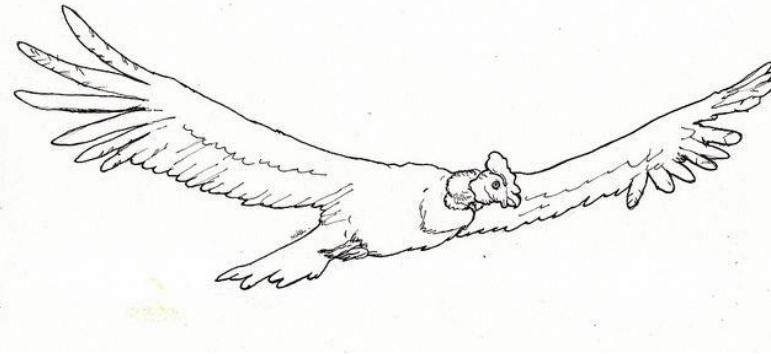


Tabla 3. (continuación) Probabilidad de que una variable normal de media cero y desviación típica uno tome un valor menor que z

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:

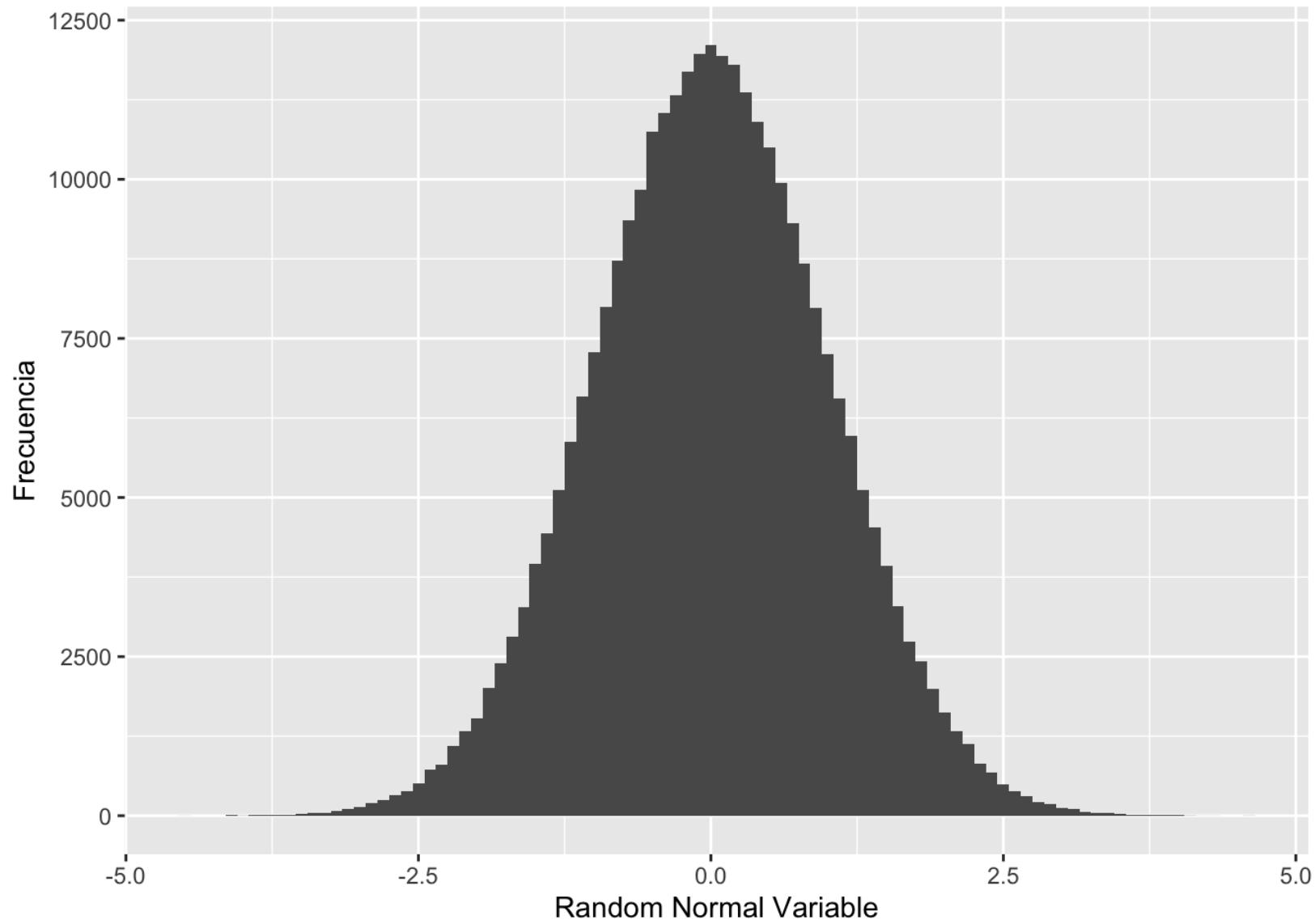
$$Y_i = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-x_i^2}{2}}$$

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)
```