

Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS
MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

Tema 1_2. Introducción

Razonamiento Estadístico



Video complementario:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=e1b1a4b3-1803-41fe-b7c5-abdc00f2aa38>

Razonamiento Estadístico

Pensamiento crítico tiene en cuenta las siguientes preguntas:

- ¿Cuál es el contexto?(Incluye una descripción y delimitación espacio-temporal)
- ¿Cuál es el objetivo del estudio? (comprensión del problema en su contexto)
(Dedicarle tiempo-delimitarlo)
- ¿Cuál es la unidad de análisis?
- ¿Cuáles son las fuentes de los datos?
- ¿Qué tipos de variables se tienen?
- ¿Con qué tipo de muestreo han sido obtenidos los datos? (error de muestreo)
- ¿Existen variables (intermedias) que influyan en los resultados y se hayan omitido?
- ¿Las tablas y gráficas resumen adecuadamente los datos?
- ¿Qué tipos de errores y sesgo (“manipulación” voluntaria o involuntaria) que podemos encontrar?
- ¿Las conclusiones se extraen directa y naturalmente de los datos?

Tabla de contenido

- ❑ Tema 1: Introducción a la estadística.
 - Distribución de frecuencias.
 - Tabulación de variables
 - Gráficas básicas
 - Elegir gráficos adecuados
 - Aplicación de las TIC
 - Retos de la estadística en el Big Data

Tabla de contenido

Representación de datos

Distribución de frecuencias

Absolutas n_i

Relativas $\frac{n_i}{N}$

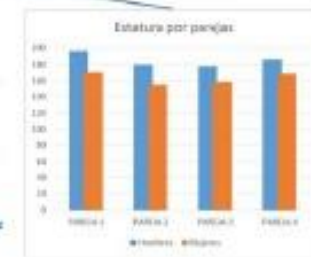
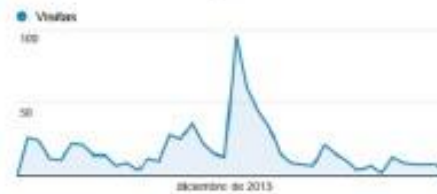
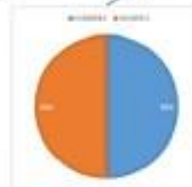
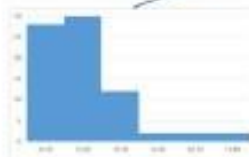
Acumuladas

Porcentajes

Tabulaciones

SUJETO	ALTURA	SEXO
Sujeto 1	171	M
Sujeto 2	154	M
Sujeto 3	159	M
Sujeto 4	196	H
Sujeto 5	169	M
Sujeto 6	180	H
Sujeto 7	178	H
Sujeto 8	187	H

Gráficos



Tablas de frecuencias-Tabulación de variables

Modalidades	Frecuencias (absolutas)	Frecuencias relativas	Frecuencias absolutas acumuladas	Frec. relativas acumuladas
1	n_1	f_1	N_1	F_1
2	n_2	f_2	N_2	F_2
...
k	n_k	f_k	N	1
SUMA	N	1		

X: número de dormitorios en viviendas de una determinada localidad:
3, 1, 2, 1, 2, 1, 2, 4, 1, 3, 5, 2, 2, 5, 4, 4, 4, 5, 1 y 2

x_i	n_i	f_i	N_i	F_i
1	5	5	0,25	0,25
2	6	11	0,30	0,55
3	2	13	0,10	0,65
4	4	17	0,20	0,85
5	3	20	0,15	1
Total	20		1	

Tablas de frecuencias-Tabulación de variables

X: Tipos de reclamaciones en un departamento de atención al cliente

Causas	n_i	N_i	f_i	F_i
Mal funcionamiento	6	6	0,30	0,30
Retrasos	5	11	0,25	0,55
Personal	4	15	0,20	0,75
Incompetencia	3	18	0,15	0,90
Compatibilidad	2	20	0,10	1
Total	20		1	

¿Tienen sentido?

Tablas de frecuencias-Tabulación de variables

X: Calificaciones en una prueba [0,6].

2,87	2,44	3,49	3,83	3,97	4,69	3,35	1,89	3,90	3,55
4,69	3,03	3,00	4,96	3,10	1,84	2,23	3,64	1,96	4,39
3,15	3,61	4,43	2,96	2,04	2,62	3,96	2,41	4,03	4,70
5,33	3,19	3,19	5,03	3,92	1,93	2,74	2,83	3,03	2,64
3,70	1,41	3,87	1,04	2,43	2,87	3,44	0,92	4,22	2,88

Intervalo	f_i
[0,1)	1
[1,2)	6
[2,3)	14
[3,4)	19
[4,5)	8
[5,6]	2
Total	50

► Número de intervalos: k

$k = \sqrt{n}$, Sturges, Scott, Freedman-Diaconis, ...

<https://r-charts.com/distribution/histogram-breaks/>

https://osoramirez.github.io/R_Para_Biologos/distribucion-de-frecuencias.html

Tablas de frecuencias-Tabulación de variables

Para X y Y variables categóricas (**tablas de contingencia**):

n_{ij}	$Y = 1$	$Y = 2$	\dots	$Y = J$	Totals
$X = 1$	n_{11}	n_{12}	\dots	n_{1J}	n_{1+}
$X = 2$	n_{21}	n_{22}	\dots	n_{2J}	n_{2+}
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
$X = I$	n_{I1}	n_{I2}	\dots	n_{IJ}	n_{I+}
Totals	n_{+1}	n_{+2}	\dots	n_{+J}	$n = n_{++}$

Ej. Doce individuos se clasificaron según el sexo (hombre, mujer) y su deseo de ver o no una final de campeonato de fútbol que será televisada:

sexo	futbol		Sum
	si	no	
hombre	6	1	7
mujer	1	4	5
Sum	7	5	12

Tabla de frecuencias absolutas

sexo	futbol	
	si	no
hombre	0.50000000	0.08333333
mujer	0.08333333	0.33333333

Tabla de frecuencias relativas

sexo	futbol	
	si	no
hombre	0.8571429	0.2000000
mujer	0.1428571	0.8000000

Tabla de frecuencias relativas al total por columna

sexo	futbol	
	si	no
hombre	0.8571429	0.1428571
mujer	0.2000000	0.8000000

Tabla de frecuencias relativas al total por fila

Gráficas básicas

(Pennstate2.csv)

Gráfico de tarta (pie)
Gráfico de barras (barplot)

Gráfico de tarta para la variable Tatuajes

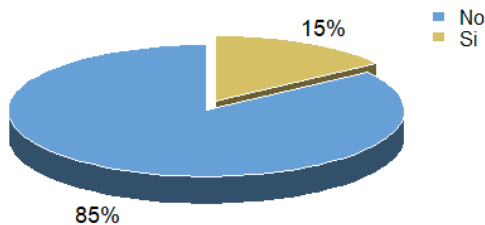
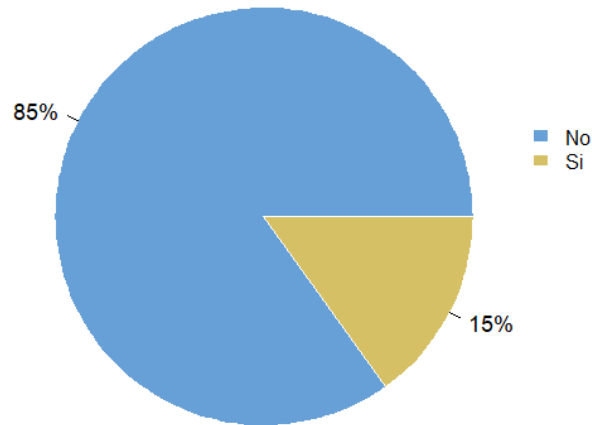
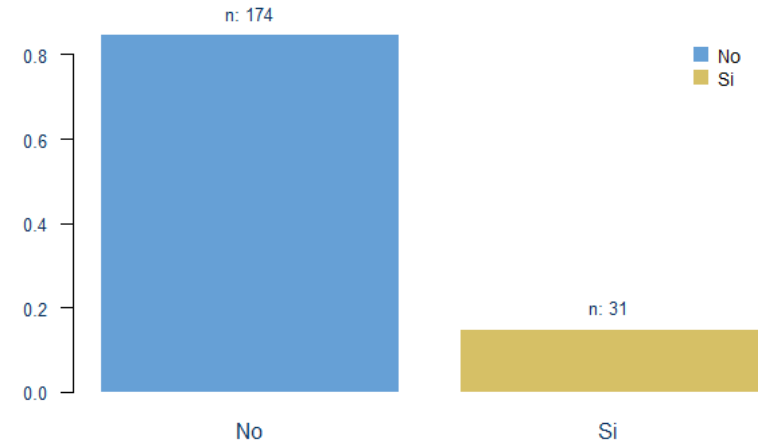


Gráfico de barras para la variable Tatuajes



Compara magnitudes o frecuencias de varias categorías

<https://r-graph-gallery.com/>

Una variable categórica

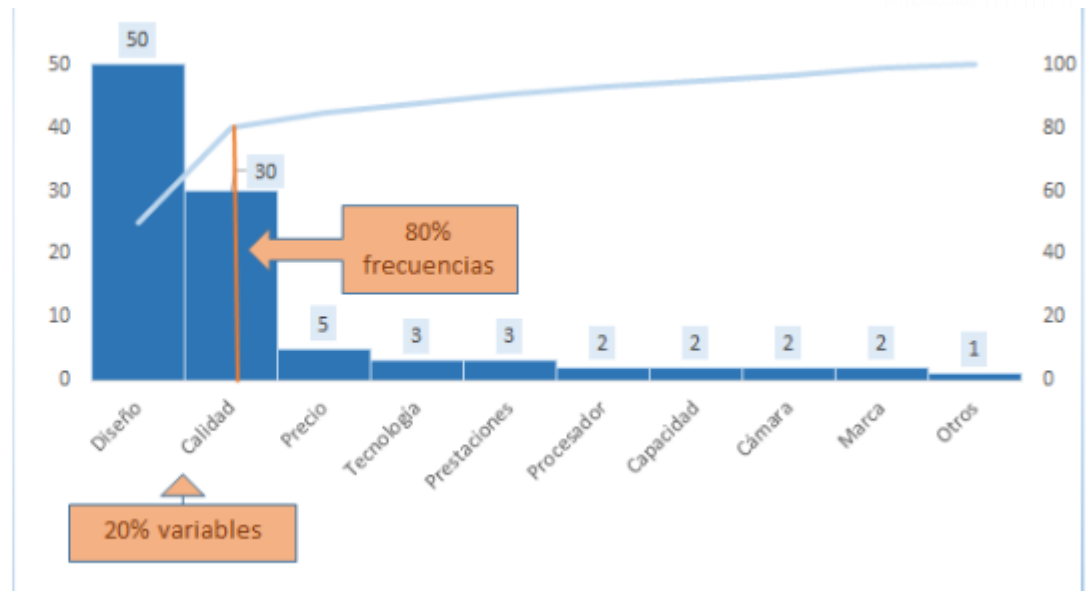
Gráficas básicas

Gráfico de Pareto

Ej. Queremos saber qué mueve a los clientes para comprar un determinado producto, por ejemplo, un teléfono móvil.

Busca establecer prioridades. Focalizar lo importante. Regla 80-20 (20% del esfuerzo genera el 80% resultado)

Motivación	Frec. Abs.	Frec. Ac.	%
Diseño	50	50	50,00%
Calidad	30	80	80,00%
Precio	5	85	85,00%
Tecnología	3	88	88,00%
Prestaciones	3	91	91,00%
Procesador	2	93	93,00%
Capacidad	2	95	95,00%
Cámara	2	97	97,00%
Marca	2	99	99,00%
Otros	1	100	100,00%
Total	100		

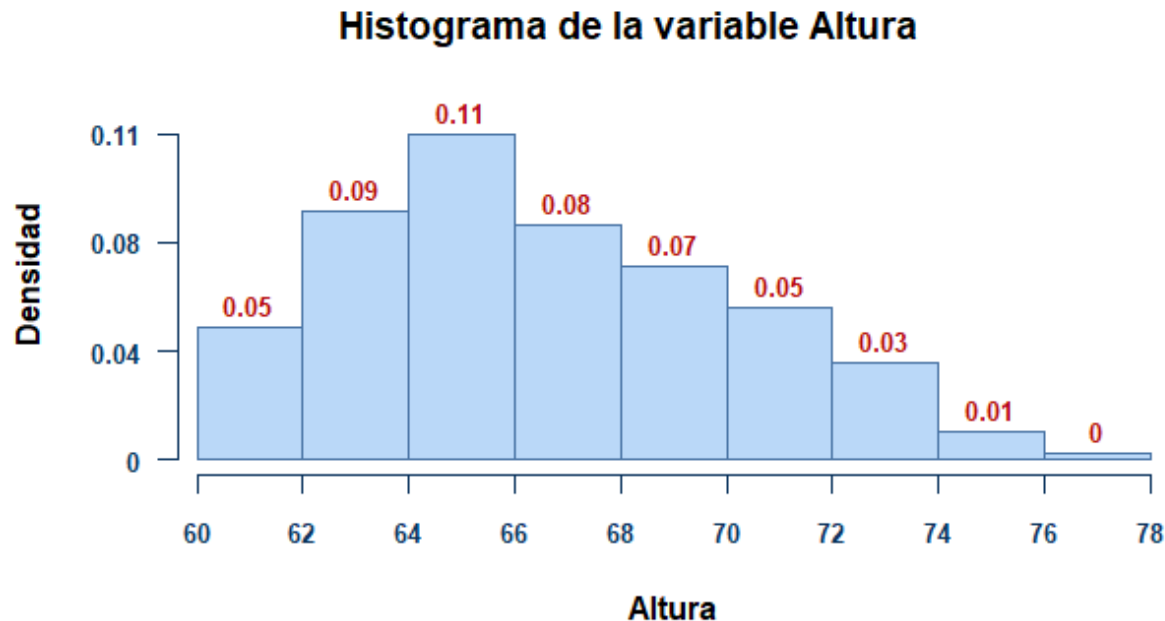


Gráficas básicas

Histograma

El histograma es útil para estudiar la **forma** en que se distribuyen (forma, centralidad y dispersión) los datos cuantitativos.

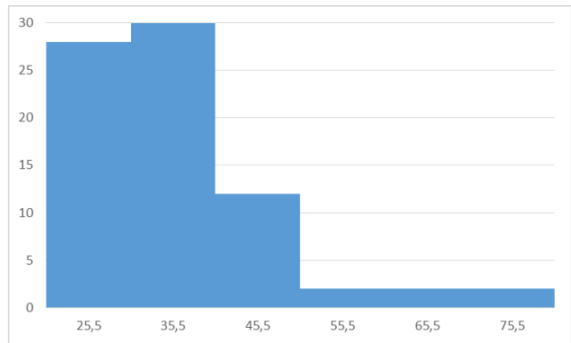
Lo importante de un histograma son las áreas de los rectángulos, no sus alturas.



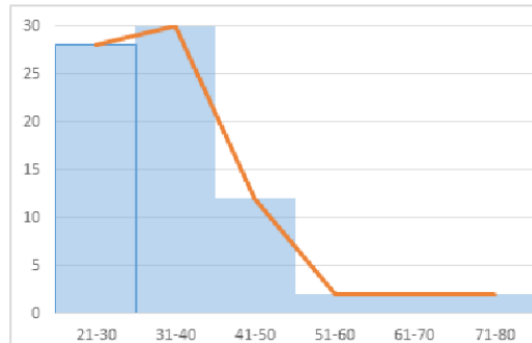
Gráficas básicas

Histograma

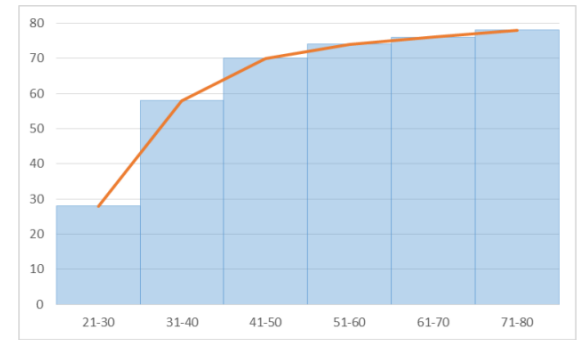
X: Edad de las actrices que han recibido un Oscar (Triola, 2009)



Histograma



Polígono de
frecuencias



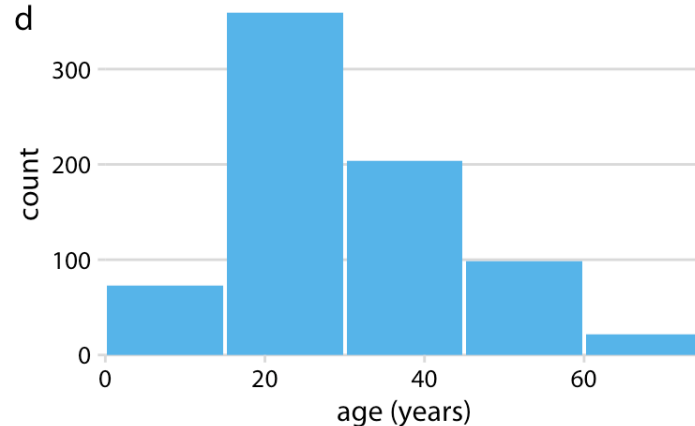
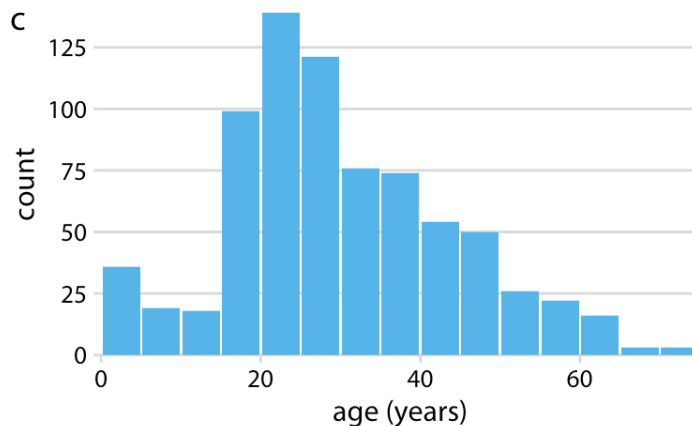
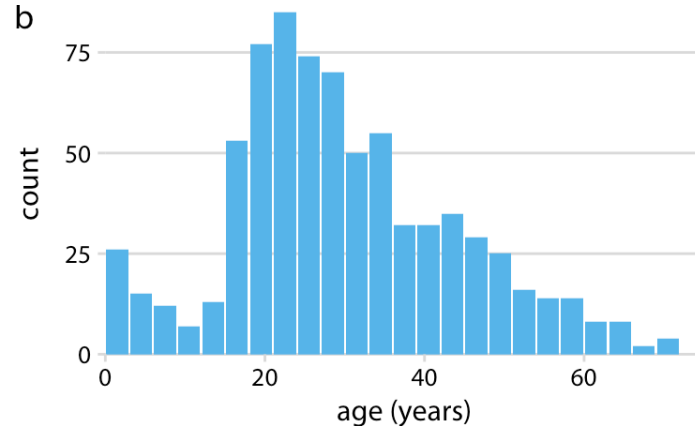
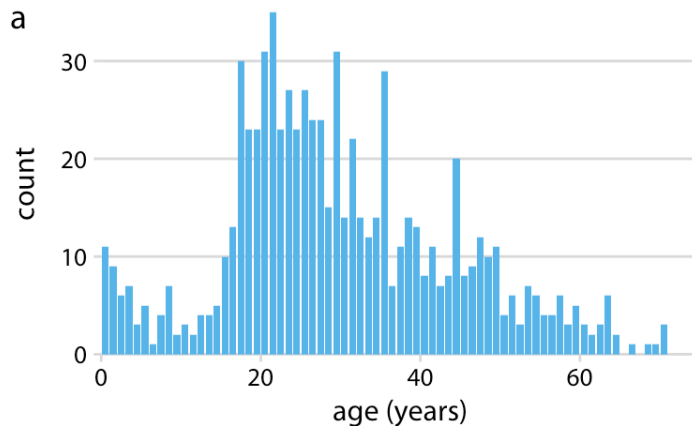
Polígono de
frecuencias
acumuladas-ojiva

Una variable cuantitativa continua

Gráficas básicas

Histograma

La visualización del histograma depende del número de intervalos y de la longitud de cada intervalo definido para la variable en estudio. Lo define el investigador.

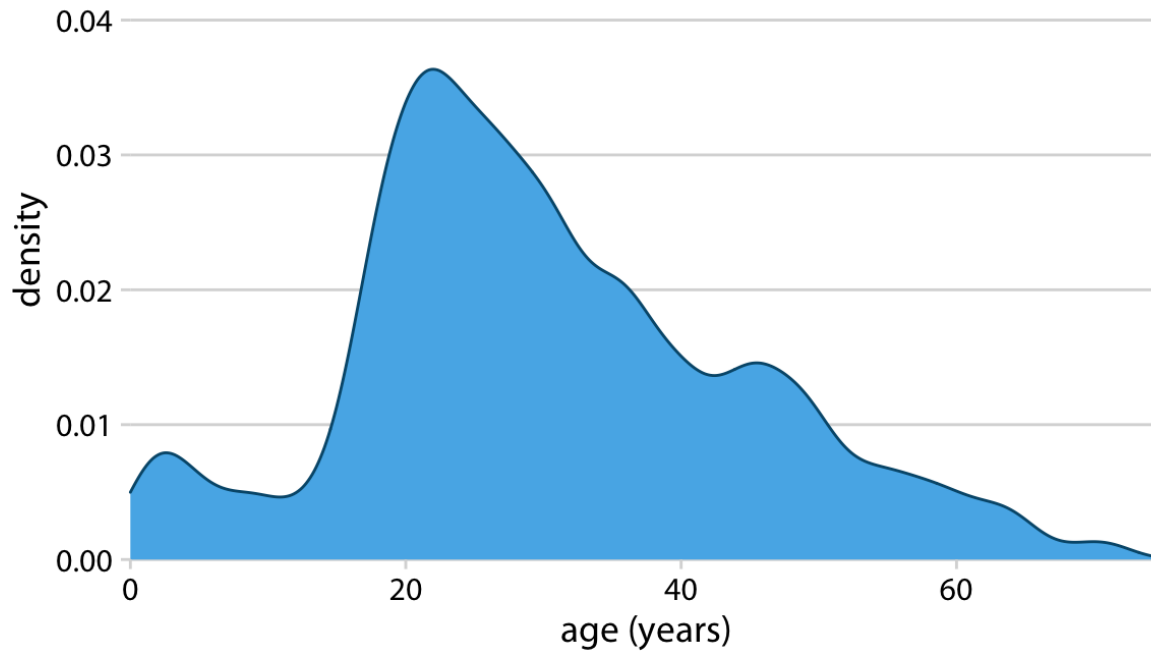


Age distribution of Titanic passengers

Gráficas básicas

Gráficos de densidad

Se basan en Kernel density estimation. Producen una estimación de la distribución (densidad) de la variable.



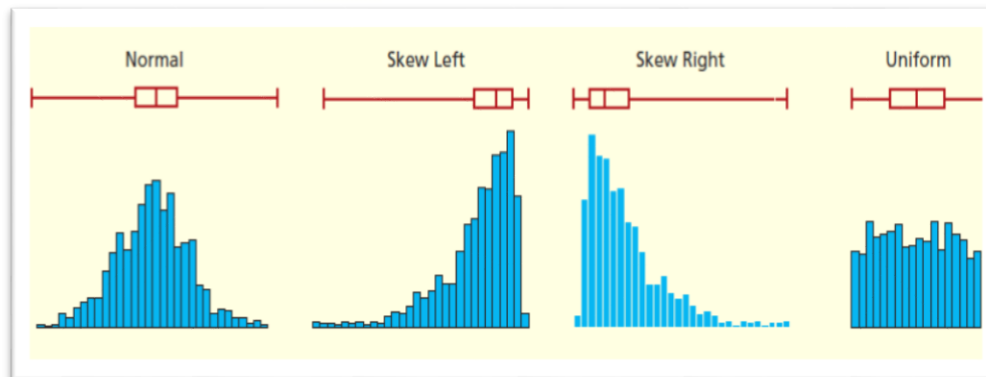
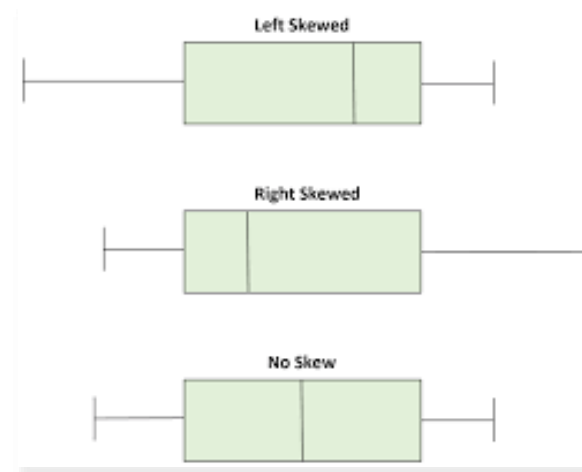
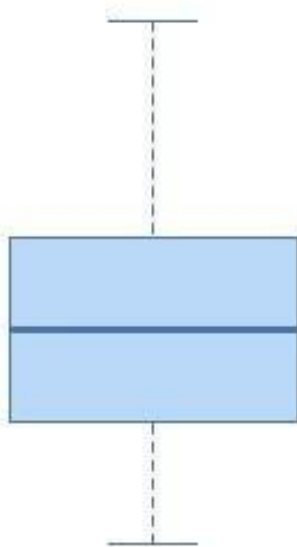
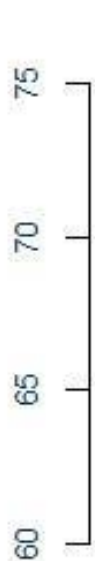
Age distribution of Titanic passengers

<https://clauswilke.com/dataviz/histograms-density-plots.html>

Gráficas básicas

Gráfico de caja y bigotes (Box-Plot)

Boxplot de la variable Altura



Gráficas básicas

Dos variables cualitativas

Diagrama de barra (adosado-agrupado)

Compara distribuciones

Diagrama de barra (anidado-apilado)

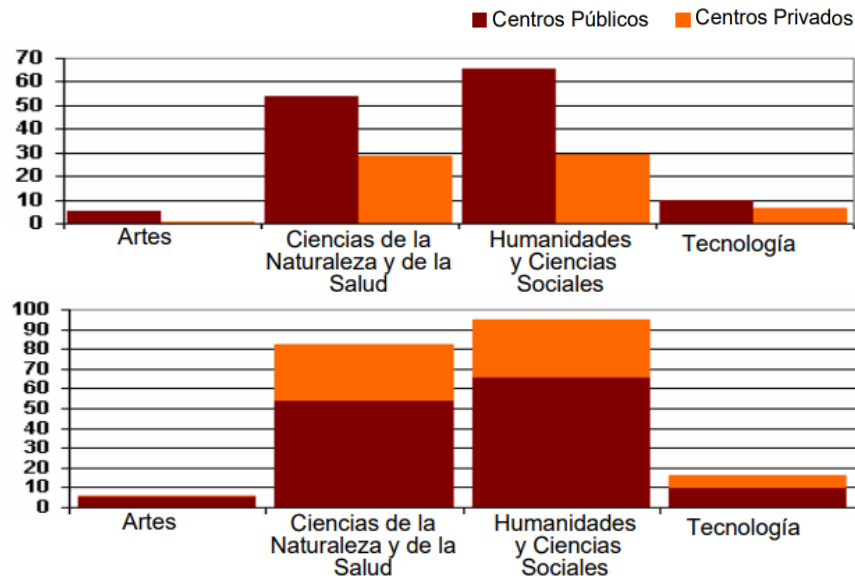
Distribución total,

Participación en cada categoría

Diagrama de barra (bidireccional)

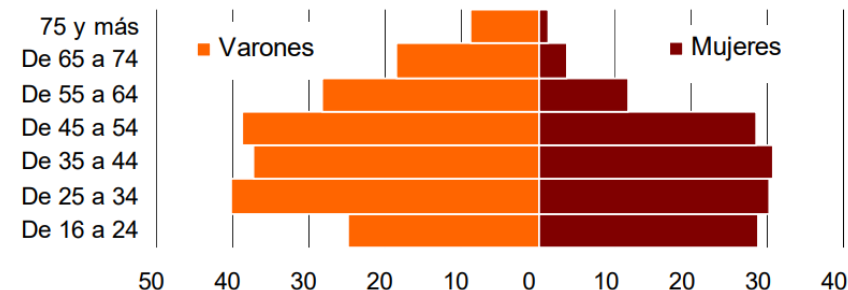
Compara distribuciones

Alumnado que terminó Bachillerato por su opción académica (Miles de alumnos)



Consumo de tabaco según sexo y grupos de edad

Fumadores diarios (porcentajes)



Fuente: Encuesta Nacional de Salud 2006. INE

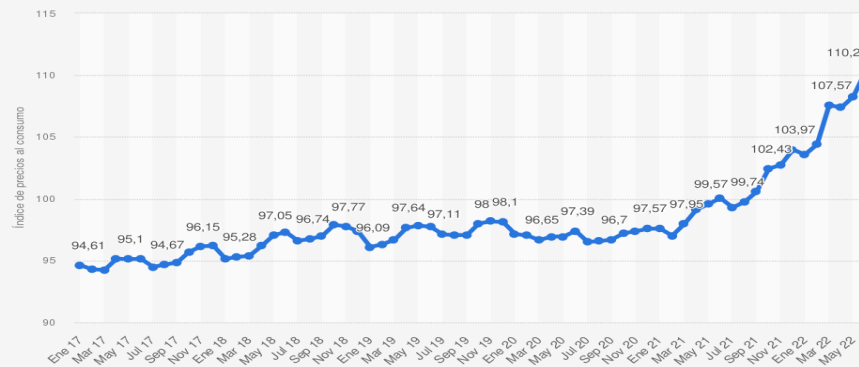
Gráficas básicas

Gráfico de línea

Usada para estudiar tendencias temporales y comportamiento histórico.



Índice de precios de consumo (IPC) en España de enero de 2017 a junio de 2022, por mes



Fuente:
INE (Spain)
© Statista 2022

Información adicional:
enero de 2017 - junio de 2022; año base (2021=100)

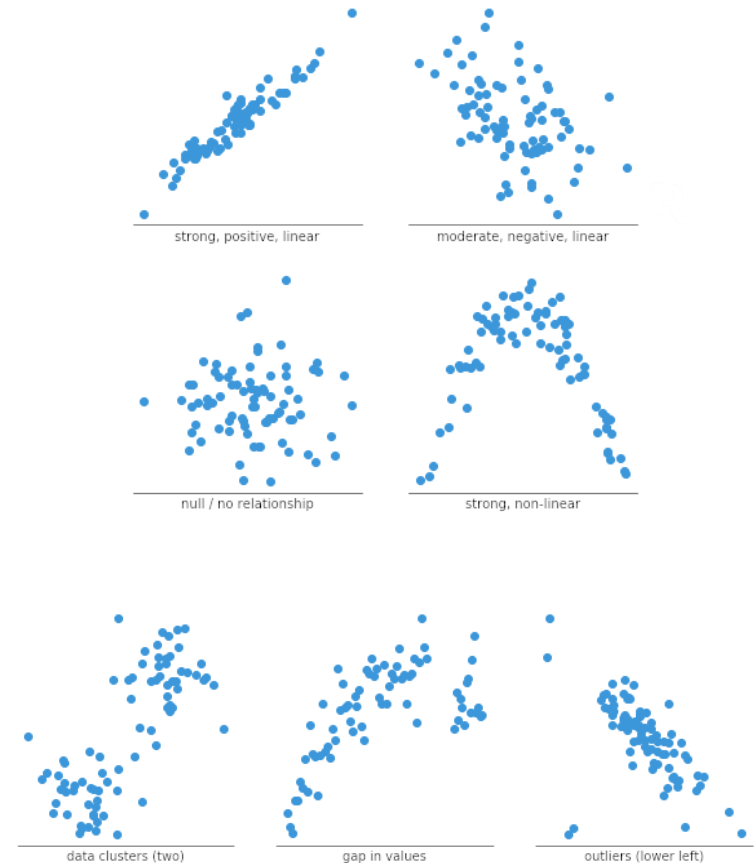
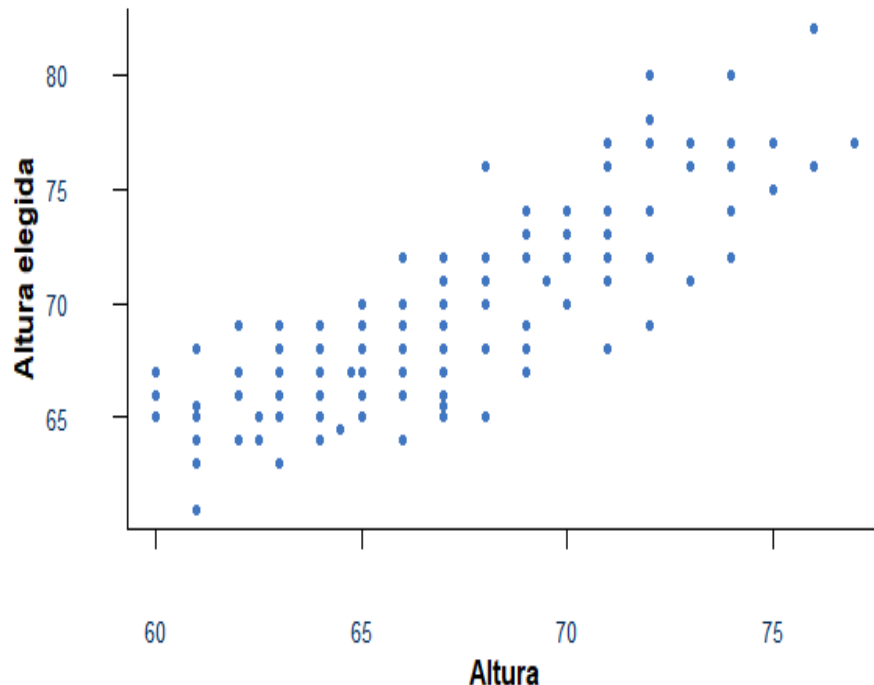
Una variable cuantitativa temporal

Gráficas básicas

Gráfico de dispersión

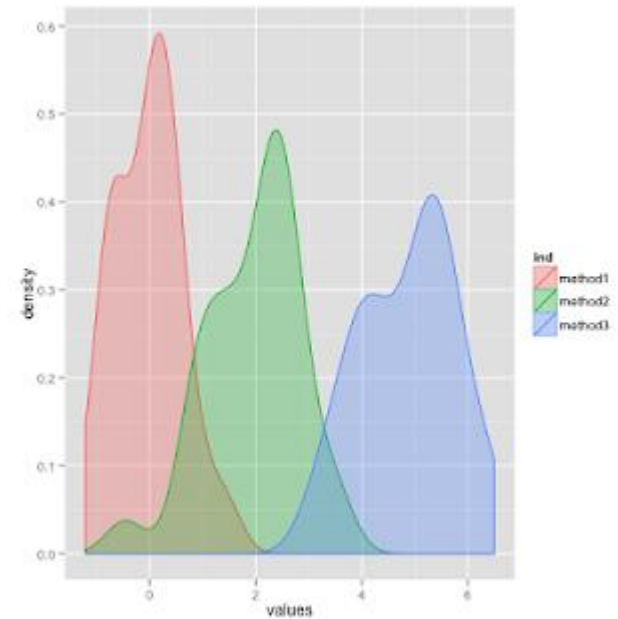
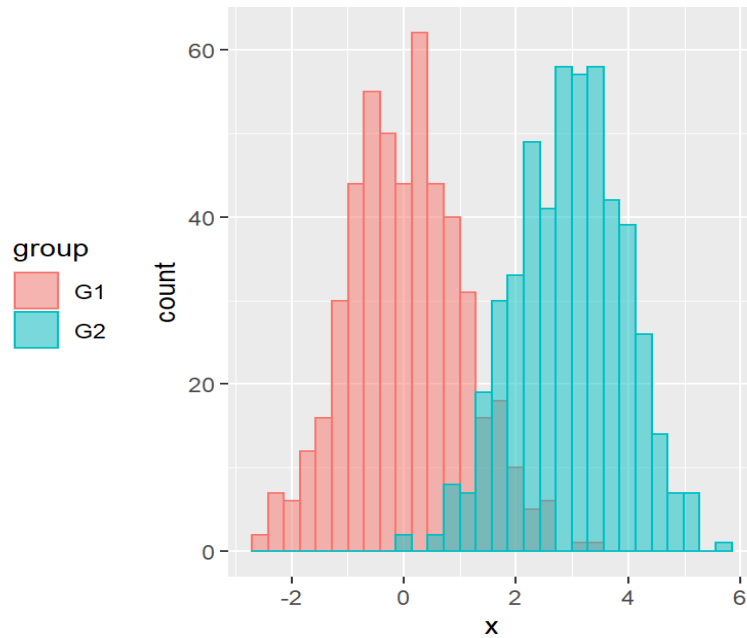
Utiliza puntos para representar los valores de dos variables numéricas diferentes. La posición de cada punto en el eje horizontal y vertical indica los valores de un punto de datos individual. Los gráficos de dispersión se utilizan para observar las relaciones entre las variables.

Dos variables cuantitativas

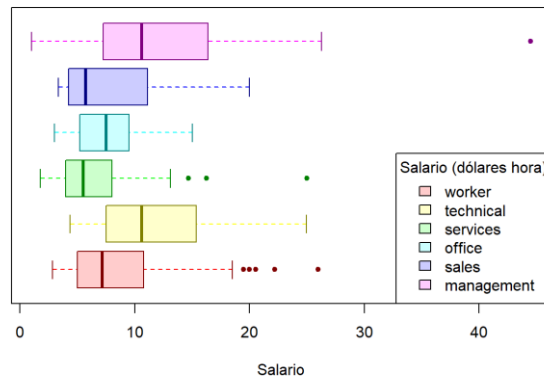


Gráficas básicas

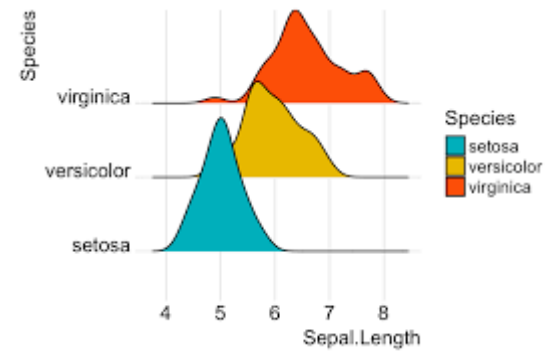
Dos variables: una cuantitativa y una cualitativa



Salario (dólares hora) según ocupación

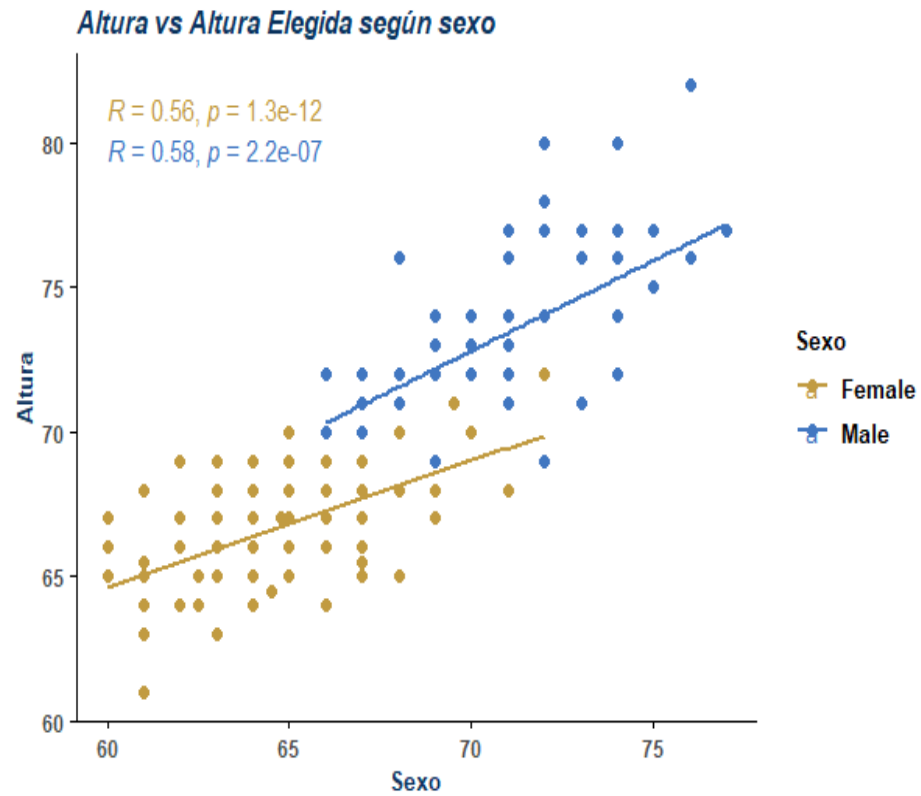


Ridgeline plot



Gráficas básicas

Tres variables: dos cuantitativas y una cualitativa



El arte de elegir el gráfico adecuado

VISUALIZACIÓN DE DATOS MEDIANTE GRÁFICOS

Gráfico de tarta



Gráfico de barras

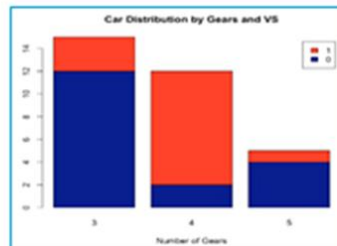


Gráfico de frecuencias

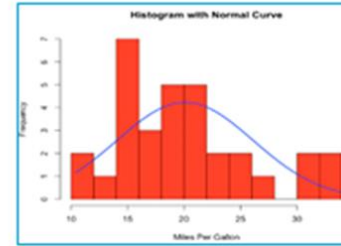


Gráfico de dispersión

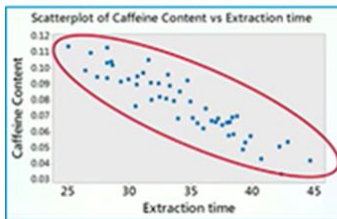


Gráfico de Cajas o Bigote

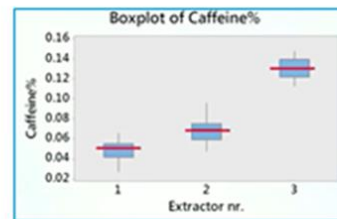
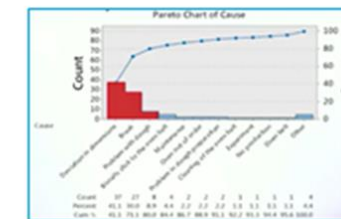
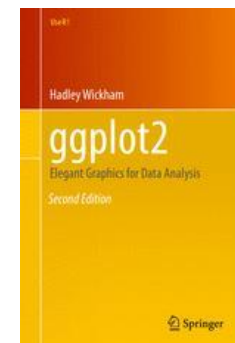
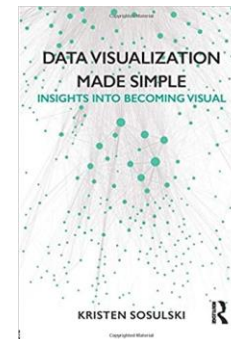
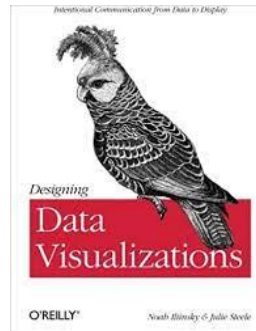
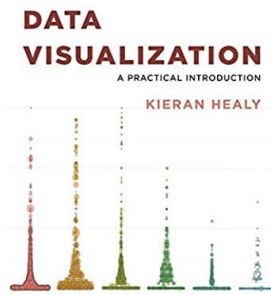


Gráfico de Pareto



Utilizamos diferentes tipos de gráficos según la naturaleza de las variables de estudio, numéricas, categóricas, y según el objetivo del análisis.

Hoy la visualización de data es una “ciencia”.



Retos de la estadística en el Big Data :

1. Excesiva cantidad de información y datos
 2. Complejidad de los datos
- Volumen, Velocidad, Variedad de los datos, Veracidad de los datos, Viabilidad, Visualización de los datos, Valor de los datos
3. Necesidad de infraestructuras potentes de análisis
 4. Políticas de privacidad
 5. Recogida de datos sin previa especificación del problema -Medición de errores
 6. Explicabilidad y transparencia

https://ec.europa.eu/info/sites/default/files/business_economy_euro/banking_and_finance/documents/191113-report-expert-group-regulatory-obstacles-financial-innovation_en.pdf



<https://www.iic.uam.es/innovacion/big-data-infografia-7-v/>

Próxima sesión

□ Tema 3: Medidas que resumen la información.

- Medidas de tendencia central.
- Medidas de tendencia central robustas.
- Medidas de dispersión.
- Medidas de dispersión robustas.
- Medidas de posición y forma.
- Gráficos de caja.
- Datos atípicos y análisis exploratorio de datos.

Learn by **DOING**.



**Estudiar Tema 2. Estadística computacional. Principios básicos.
Ámbitos de aplicación. Técnicas básicas de programación.
Presentación del software R**

Instalación de R

Para instalar estos programas se puede usar lo siguiente:

Instalar R para Windows:

<https://cran.r-project.org/bin/windows/base/>

Instalar R para MacOS:



<https://cran.r-project.org/bin/macosx/>

Instalar Rstudio para Windows o MacOS:

Descarga e instala la última versión que corresponda.

<https://www.rstudio.com/products/rstudio/download/>

Free-Open-Source

OS	Download	Size	SHA-256
Windows 10/11	 RStudio-2022.07.2-576.exe	190.49 MB	b38bf925
macOS 10.15+	 RStudio-2022.07.2-576.dmg	224.49 MB	35028d02

Documento Introducción a R. ESIT

Tema 1. Introducción

1.1. Presentación	3
1.2. ¿Qué es R?	3
1.3. Un poco de historia	4
1.4. ¿Por qué usar R?	5
1.5. El entorno de trabajo RStudio	6
1.6. Instalación de R y RStudio	7
1.7. Formato del código en el texto	13
1.8. La ayuda en línea de R	14
1.9. Referencias bibliográficas	16

Documento Introducción a R. ESIT

Tema 2. Empezando con R: algunos conceptos básicos

2.1. Introducción y objetivos	3
2.2. Entorno de trabajo de RStudio	3
2.3. La consola de R	7
2.4. Variables	8
2.5. Objetos	11
2.6. Directorio de trabajo	15
2.7. Scripts	17
2.8. Creación de Proyectos	20
2.9. Manejo de la biblioteca: paquetes adicionales	22
2.10. Referencias bibliográficas	27

Documento Introducción a R. ESIT

Tema 3. Estructura de datos

3.1. Introducción y objetivos	3
3.2. Estructuras de datos	3
3.3. Vectores	4
3.4. Factores	17
3.5. Matrices	23
3.6. Arrays	37
3.7. Data frames	40
3.8. Listas	63
3.10. Cuaderno de ejercicios	70

Tema 4. Programación básica

4.1. Introducción y objetivos	3
4.2. Operadores en R	3
4.3. Estructuras de control	14
4.4. Funciones	43
4.5. Referencias bibliográficas	58
4.6 Cuaderno de ejercicios	58

Documento Introducción a R. ESIT

Tema 5. Manejo de datos

5.1. Introducción y objetivos	3
5.2. Importando datos desde un archivo	4
5.3. Leer datos desde un paquete	24
5.4. Guardar datos desde R	25
5.5. Manipulación de datos: una introducción al <i>tidyverse</i>	32
5.6. Referencias bibliográficas	43
5.7. Cuaderno de ejercicios	43

Documento Introducción a R. ESIT

Tema 6. Visualización de datos

6.1. Introducción y objetivos	3
6.2. La función plot	4
6.3. La función hist()	34
6.4. La función boxplot()	39
6.5. Gráfico de barras: la función barplot()	42
6.6. Otros gráficos	47
6.7. El paquete ggplot2	47
6.8. Referencias bibliográficas	80
6.9. Cuaderno de ejercicios	80

Documento Introducción a R. ESIT

Tema 7. Introducción a R Markdown

7.1. Introducción y objetivos	3
7.2. Elementos básicos de RMarkdown	4
7.3. Formatos de salida	10
7.4. Bloques de código y código en línea.	14
7.5. Elementos de sintaxis	22
7.6. Escritura de expresiones matemáticas	39
7.7. Referencias bibliográficas	44
7.8. Cuaderno de ejercicios	44



www.unir.net